الجمهورية العربية السورية
جامعة البعث
كلية الهندسة المعلوماتية
قسم هندسة البرمجيات ونظم المعلومات

تحسين عملية اتخاذ القرار من خلال التنقيب في البيانات الضخمة

(البيانات الطبية)

دراسة أعدت لنيل درجة الماجستير في هندسة البرمجيات ونظم المعلومات

إعداد

المهندس قتيبة بركات

إشراف

الأستاذ الدكتور محسن حسين

أستاذ دكتور في قسم هندسة البرمجيات ونظم المعلومات

كلية الهندسة المعلوماتية

جامعة البعث

2019 م - 1441 هـ

كلمة شكر وتقدير

تشاء الأقدار...أن تصل الينابيع إلى الأنهار...وأن ترتطم بالشواطئ أمواج البحار...وأن تغرب الشمس معلنة نهاية النهار... وأنه بعد البدء سينتهي المشوار...

فبينما ترتفع أشرعة مركبي معلنة قرب الرحيل عن محطة الماجستير يهتف قلبي بالشكر إلى جميع الدكاترة في كلية الهندسة المعلوماتية الذين كانوا وما زالوا وسيبقون النجوم التي أستضيء بنورها في درب العلم الطويل الذي لا ينتهي...

وأخص بالشكر والعرفان لمعلمي الرائع وأستاذي الغالي الذي أفتخر بكوني أحد طلابه، قدوتي وتاج رأسي الأستاذ الدكتور **محسن حسين** الذي تكرم فضلاً منه وعزاً لي بالإشراف على هذه الرسالة فكان كما اعتدنا عليه الملهم والسند الذي تعجز عن وصف إبداعه الكلمات...

م. قتيبة بركات

جامعة البعث

كلية الهندسة المعلوماتية

الرقم:

التاريخ:

بيان بإجراء التعديلات العلمية واللغوية

قام الطالب قتيبة حسن بركات بتصحيح الملاحظات التي وردت في متن رسالة الماجستير والتي أشار إليها الدكاترة أعضاء لجنة الحكم خلال مناقشة الرسالة.

عضواً	مشرفاً وعضواً	رئيساً للجنة
د.علي أحمد	د. محسن حسين	د. كمال السلوم

رئيس قسم هندسة البرمجيات ونظم المعلومات

د. كمال السلوم

الجمهورية العربية السورية
جامعة البعث
كلية الهندسة المعلوماتية
قسم هندسة البرمجيات ونظم المعلومات

تصريح

أصرح بأن هذا البحث:

" تحسين عملية اتخاذ القرار من خلال التنقيب في البيانات الضخمة (البيانات الطبية) "

لم يسبق أن قُبل للحصول على أية شهادة، ولا هو مقدم حالياً للحصول على أي شهادة أخرى.

المرشح

قتيبة حسن بركات

DECLARATION

It is declared that this work:

" Improving Decision-Making Process Using Big Data Mining (Medical Data)"

Has not been accepted for any degree, and it is not submitted for any other degree.

Candidate

Kotyba Hasan Barakat

الجمهورية العربية السورية
جامعة البعث
كلية الهندسة المعلوماتية
قسم هندسة البرمجيات ونظم المعلومات

شهادة

نشهد بأن العمل الموصوف في هذه الرسالة هو نتيجة بحث علمي قام به المرشح **قتيبة حسن بركات** تحت إشراف الدكتور **محسن حسين** أستاذ في قسم هندسة البرمجيات ونظم المعلومات كلية الهندسة المعلوماتية جامعة البعث مشرفاً علمياً وأي رجوع إلى بحث آخر في هذا الموضوع موثق في هذا النص.

المرشح
قتيبة حسن بركات

الدكتور المشرف العلمي
أ.د. محسن حسين

Certificate

It is hereby certified that the work described in this thesis is the result of the candidate's own investigation, under the supervision of **Dr. Mohsen Hussien** Professor in Software Engineering and Information System Department, Al-Baath University. Any references to other research work has been acknowledged in this text.

Candidate
Kotyba Hasan Barakat

Supervisor
PhD. Mohsen Hussien

المساهمة العلمية

AL Baath University
Journal of Al Baath University

جامعة البعث
مجلة جامعة البعث

قبول نشر بحث
طالب دراسات عليا

الرقم: ١٩٠٢
التاريخ ٢ / ٥ / 2019

طالب الدراسات العليا: قتيبة بركات
الدكتور المشرف: **محسن حسين**
كلية: **الهندسة المعلوماتية** – جامعة: **البعث**

نود إعلامكم بقبول بحثكم الموسوم:

تطوير خوارزمية التنقيب عن المحفزات في سلاسل DNA

للنشر في مجلة جامعة البعث بالمجلد **41** لعام **2019**
بعد أن تم تحكيمه من قبل مختصين.

نشكر لكم هذه المساهمة الطيبة ونتطلع إلى استمرار تواصلكم مع مجلتنا ومدها بما لديكم من جديد والاطلاع على الأبحاث المنشورة في المجلة على موقع المجلة والرابط المدونين في أسفل الصفحة.
موقع الجامعة www.albaath-univ.edu.sy والرابط magazine @ albaath-univ.edu.sy

رئيس تحرير
مجلة جامعة البعث

Journal of AL Baath University – Syria – Homs
Est. 1983
P.O. Box: 77 – Fax: ++963 31 2138071
Tel: ++963 31 9910 (1180 1132)
magazine @ albaath-univ.edu.sy

مجلة جامعة البعث – سورية – حمص
تأسست عام 1983
ص. ب.: 77 فاكس: ++963 31 2138071
هاتف:(1132 1180) 9910 31 963++
www.albaath-univ.edu.sy

الفهرس

مقدمة عامة 1
1- تمهيد 2
2- مشكلة البحث 3
3- أهداف البحث 4
4- ترتيب البحث 5
الفصل الأول: البيانات الضخمة 7
1.1- مقدمة 8
2.1- لمحة تاريخية 8
3.1- مفهوم البيانات الضخمة 9
4.1- أبعاد البيانات الضخمة 9
5.1- استخدام البيانات الضخمة 10
6.1- مثال عن البيانات الضخمة 11
7.1- خلاصة 13
الفصل الثاني: دور البيانات الضخمة في صناعة القرارات 14
1.2- مقدمة 15
2.2- لمحة عن علم التنقيب في البيانات 15
3.2- أسباب ظهور مفهوم التنقيب في البيانات 16
4.2- كيفية التنقيب في البيانات 16
5.2- مستودع البيانات 18
6.2- أهمية عملية اتخاذ القرارات 19
7.2- خطوات عملية اتخاذ القرارات 20
8.2- دور نماذج المعرفة في صناعة القرارات 22
9.2- خلاصة 23
الفصل الثالث: المعلوماتية الحيوية 24
1.3- مقدمة 25
2.3- لمحة تاريخية 25
3.3- تعريف المعلوماتية الحيوية 26
4.3- مجالات المعلوماتية الحيوية 27
5.3- أسباب تطور المعلوماتية الحيوية 27
6.3- تطبيقات المعلوماتية الحيوية 28

7.3- أساسيات المعلوماتية الحيوية 29
8.3- خلاصة 29
الفصل الرابع: سلاسل DNA 30
1.4- مقدمة 31
2.4- لمحة تاريخية 31
3.4- تكوين سلاسل DNA 34
4.4- عملية إنتاج البروتين 35
5.4- معامل النسخ (المحفز) 37
6.4- خلاصة 38
الفصل الخامس: دراسة خوارزميات تحليل سلاسل DNA 39
1.5- مقدمة 40
2.5- خوارزمية **MCES** 40
3.5- خوارزمية **MDWB** 44
4.5- خلاصة 45
الفصل السادس: الخوارزمية المقترحة 46
الفصل السابع: النتائج والأعمال المستقبلية 58
مراجع 62

الملخص

تعتمد فعالية القرارات على كفاءة المعلومات المعتمدة عليها، وتعظم أهمية المعلومات كلما ازدادت دقة وسائل التنقيب في البيانات التي تحوي هذه المعلومات بين ثناياها الضخمة.

يعتبر معامل النسخ (Transcription Factor) TF العنصر الوظيفي الأكثر أهمية في سلاسل الحمض النووي الريبي منقوص الأكسجين (DeoxyriboNucleic Acid) DNA، حيث تدعى مواقع توضع هذا المعامل ضمن تلك السلاسل بمواقع معامل النسخ الملزمة (Transcription Factor Binding Sites) TFBS، ويتم تحديد هذه المواقع باستخدام المحفزات Motifs والتي تعرف بأنها السلاسل الجزئية الأكثر تكراراً في سلاسل DNA.

يعتبر تحديد المحفزات من أكثر التحديات التي يواجهها الباحثون في علم الأحياء، حيث تنوعت وسائلهم وتكاثفت جهودهم من أجل إيجاد الطريقة الأفضل لتحديدها نظراً لدورها الأساسي في المجال الطبي.

ويعظم التحدي وتزداد الصعوبة عندما نعلم بأن سلاسل DNA من أكثر البيانات الضخمة المتنامية في كبرها، وعليه فقد وجب على الباحث الراغب بهذا التحدي أن يبتكر خوارزمية بحث عن المعامل المطلوب ضمن هذه البيانات الضخمة مبحراً في محيطات زخارف سلاسلها البنيوية ليصل إلى الهدف المنشود محققاً مقاييس الدقة والسرعة.

قمنا في هذا البحث بتطوير خوارزمية تقوم على استخدام مفاهيم الرتل والأشجار من أجل تمثيل المحفزات ضمن سلاسل DNA، كما تمت الاستفادة من ميزات الأعداد الأولية للتعبير عن توضع نيكليوتيد معين ضمن كل سلسلة جزئية من سلاسل DNA، ثمّ مقارنة نتائج تطبيق هذه الخوارزمية مع الخوارزميات السابقة المقترحة ضمن هذا المجال.

الكلمات المفتاح : عملية اتخاذ القرار – التنقيب في البيانات – البيانات الضخمة – سلاسل الحمض النووي الريبي منقوص الأكسجين – المحفزات – مواقع معامل النسخ الملزمة – معامل النسخ – التنقيب في النصوص.

قائمة الجداول

الجدول	الصفحة
1.1 إحصائية بطلاب جامعة البعث	11
1.4 الفرق بين أنواع البكتيريا المستخدمة في تجربة غريفث	32
1.7 مقارنة زمن التنفيذ بين الخوارزمية المقترحة و خوارزميتي MCES و MDWB	59
2.7 مقارنة النسبة المئوية للدقة بين الخوارزمية المقترحة و خوارزمتي MCES و MDWB	60

قائمة الأشكال

الشكل	الصفحة
1.2 خطوات عملية صنع القرارات	21
1.4 ارتباط النيكليوتيدات	34
2.4 مراحل تشكيل البروتينات	36
3.4 تموضع معامل النسخ ضمن سلاسل DNA	37
4.4 ظهور المحفز ضمن سلاسل DNA بأشكال متعددة	38
1.6 المخطط التدفقي للخوارزمية المقترحة	49
1.7 مخطط بياني يظهر زمن تنفيذ الخوارزمية المقترحة مقارنة بخوارزميتي MCES و MDWB	60
2.7 مخطط بياني يظهر الدقة بين الخوارزمية المقترحة و خوارزميتي MCES و MDWB	61

قائمة المختصرات

HUGO	HUman Genome Organisation
NCBI	National Center of BIotechnology
BSCS	Biological Sciences Curriculum Study
TF	Transcription Factor
TFBS	Transcription Factor Binding Sites
DNA	DeoxyriboNucleic Acid
RNA	RiboNucleic Acid
PWM	Position Weight Matrix

قائمة المصطلحات

Big Data	البيانات الضخمة
Apache Hadoop	برنامج أو منصة برمجية مفتوحة المصدر مكتوبة بلغة الجافا لتخزين ومعالجة البيانات الضخمة بشكل موزع
Utah Data Center	منشأة تخزين بيانات تابعة لوكالة الأمن القومي الأمريكية مصممة لتخزين بيانات بحجم إكسابايت أو أكبر، وتهدف إلى دعم مبادرة الأمن الوطني الشامل
Structured Data	البيانـات المنظمـة
UnStructured Data	البيانـات غيـر المنظمـة
Semi-Structured Data	البيانات نصف المنظمة
Data Mining	التنقيب في البيانات
Decision Making	صناعة القرارات
Bioinformatics	المعلوماتية الحيوية
HUGO (HUman Genome Organisation)	منظمة الجينوم البشري
Artificial Neural Networks	الشبكات العصبية الاصطناعية
NCBI (National Center of Biotechnology)	المركز الوطني للمعلومات التقنية الحيوية
BSCS (Biological Sciences Curriculum Study)	مركز دراسة منهجية العلوم البيولوجية
TF (Transcription Factor)	معامل النسخ
TFBS (Transcription Factor Binding Sites)	مواقع معامل النسخ الملزمة
Motifs	المحفزات
DNA (DeoxyriboNucleic Acid)	الحمض النووي الريبي منقوص الأكسجين
Nucleotides	نيوكليوتيدات
Adenine	أدينين
Thymine	ثيمين
Cytosine	سيتوزين
Guanine	غوانين
RNA (RiboNucleic Acid)	الحمض النووي الريبي
genetic code	الشفرة الجينية
PWM(Position Weight Matrix)	مصفوفة أوزان المواقع

مقدمة عامة

مقدمة عامة

1- تمهيد:

إن التعامل مع كميات كبيرة من البيانات ليست بالأمر السهل، وخاصة في عصرنا هذا الذي زاد فيه حجم البيانات والمعلومات الممثلة ضمن منتجاته الرقمية إلى حد كبير بزغ فيه مصطلح البيانات الضخمة Big Data الذي يعبر عن كمية لا متناهية الكبر من البيانات المنظمة وغير المنظمة.

إن استخدام التنقيب في البيانات يوفر للمؤسسات في جميع المجالات القدرة على استكشاف المعلومات وبناء التنبؤات المستقبلية وتحديد السلوك والاتجاهات مما يسمح باتخاذ القرارات الصحيحة وفي الوقت المناسب.

ومن هنا فقد دعت الحاجة لتطوير أفضل الطرق اللازمة لمعالجة هذه البيانات متجاوزين الصعوبات الناجمة عن كبرها وذلك بغية الحصول على المعلومات المفيدة المتناثرة بين ثناياها واللازمة للوصول إلى المعرفة والتي يتم استخدامها عادة في بناء القرارات الأمثلية.

ففي المجال الطبي تستخدم تقنيات التنقيب في البيانات لاكتشاف وتوصيف الأمراض الأكثر شيوعاً في مناطق محددة أو أزمنة معينة أو ظروف خاصة، وذلك بهدف وضع الحلول المناسبة واتخاذ سبل الوقاية اللازمة للحد من انتشار الأمراض، بالإضافة لإعداد الأبحاث المتخصصة في دراسة الأدوية والعلاجات الطبية وسبل تطويرها وتحديثها ورفع كفاءتها وفاعليتها وصلاحيتها وقدرتها على العلاج.

ففي الهندسة الوراثية والتي تتعامل بشكل مباشر مع المادة الوراثية للكائن الحي، وإذا ما علمنا بأن المعلومات الوراثية المخزنة داخل كل خلية تبلغ قرابة المترين أي أن كل المعلومات الوراثية المخزنة داخل الكائن الحي تعادل ضعفي قطر نظامنا الشمسي، فإننا نرى الجهد الكبير الذي سنبذله لتحليل هذه المعلومات لاستخلاص المعرفة التي نحتاجها لعلاج الأمراض واكتشاف الأدوية، وتعتبر المحفزات من أهم المعلومات التي نحصل عليها بمعالجة هذه المادة الوراثية.

يعتبر تحديد المحفزات من أكثر التحديات التي يواجهها الباحثون في علم الأحياء، حيث تنوعت وسائلهم وتكاثفت جهودهم من أجل إيجاد الطريقة الأفضل لتحديدها نظراً لدورها الأساسي في المجال الطبي.

قمنا في هذا البحث بتطوير خوارزمية تقوم على استخدام مفاهيم الرتل والأشجار والأعداد الأولية، ثمّ مقارنة نتائج تطبيق هذه الخوارزمية بالخوارزميات السابقة المقترحة ضمن هذا المجال.

2- مشكلة البحث:

تزايد العمل بمصطلح «البيانات الضخمة» وذلك نظراً للنمو المتزايد للبيانات ، فقد زاد حجم البيانات بشكل أمسى فيه من الصعب معالجتها الآن باستخدام برنامج واحد أو جهاز مستقل، أو باستخدام تطبيقات معالجة البيانات التقليدية .

ولذلك فقد قامت الشركات البرمجية بتطوير برامج مساعدة وأدوات جديدة يمكن من خلاله المساعدة في تحليل تلك البيانات الضخمة.

بالإضافة إلى الحجم الهائل من البيانات التي يتم إنتاجها وتخزينها وإتاحتها تحت مظلة «البيانات الضخمة»، تتسم طرق معالجة تلك البيانات بخصائص أخرى تختلف عن البيانات التقليدية، وذلك نظراً لاختلاف خصائص البيانات الضخمة التي تتميز بما يلي[1]:

١ -الحجم: يقدر الخبراء أنه بحلول العام 2020م ستحتوي شبكة الإنترنت على ما يقرب من 40,000 زيتابايت (ألف مليار مليار بايت) من البيانات الجاهزة للتحليل واستخلاص المعلومات.

٢ -السرعة: لمعالجة مجموعة صغيرة من البيانات المخزنة في قواعد البيانات، أو ملف «إكسل»، كانت الشركات تقوم بتحليل كل مجموعة بيانات على حدة وبشكل متسلسل إلى أن يتم الانتهاء منها جميعاً. ولكن مع تضخم حجم البيانات، أصبحت الحاجة مُلِحَّة لإيجاد نظم خاصة تضمن سرعة تحليل البيانات الضخمة وقت وصولها وأدت تلك الحاجة إلى ابتكار تقنيات خاصة لمعالجة تلك البيانات مثل برامج «Apache Hadoop» .

٣ -تنوع الملفات: مع ازدياد أعداد مستخدمي الإنترنت والهواتف النقالة وشبكات التواصل الاجتماعي المختلفة، تغيرت طريقة تخزين البيانات من وجودها في قواعد بيانات تقليدية إلى بيانات مخزنة عشوائياً وبامتدادات متنوعة (مثل الصور ومقاطع الصوت والفيديو والرسائل القصيرة).

وتقدم البيانات الضخمة ميزة تنافسية للشركات إذا تم تحليلها والاستفادة منها لفهم عملائها وطرق تفكيرهم ورغباتهم، ومن ثم اتخاذ القرارات بصورة أكثر فعالية.

ولكن بسبب زيادة حجم تلك البيانات في كل عام وبشكل مطَرد، يظل البحث عن أفضل السبل لتحليلها واستثمارها قيد الدراسة والبحث.....

3- أهداف البحث:

بعد أن بينا أهمية المعلومات ودورها في بناء القرارات الفعالة موضحين الصعوبات المتمثلة في كيفية استخراج هذه المعلومات من محيطات البيانات الضخمة الموجودة في عصرنا هذا، فقد دعتنا الحاجة إلى ابتكار أحدث الطرق اللازمة لإيجاد نماذج المعرفة المطلوبة رغم صعوبة استخلاصها...

فعلى سبيل المثال لا الحصر يولد القطاع الصحي حالياً كميات هائلة من البيانات المعقدة بشكل مستمر، حيث تتضمن هذه البيانات (سلاسل الحمض النووي منزوع الأكسجين DNA، سجلات المرضى الإلكترونية، كيفية تشخيص الأمراض، الدراسات الطبية.....).

تعد سلاسل DNA من أكثر البيانات الضخمة المتنامية في كبرها، ويعتبر العنصر الوظيفي الأكثر أهمية في سلاسل DNA معامل النسخ (Transcription Factor TF)، حيث تدعى مواقع تموضع هذا المعامل ضمن هذه السلاسل بمواقع معامل النسخ الملزمة (TFBS).
ويعتبر تحديد هذه المواقع من أكثر التحديات التي يواجهها الباحثون في علم الأحياء، حيث تنوعت وسائلهم وتكاثفت جهودهم لتحديد الطريقة الأفضل لتحديد هذا المعامل وذلك لأهميته في المجال الطبي.

حيث سنستعرض في هذا البحث أبرز الخوارزميات التي تهدف إلى تحديد معامل النسخ مطورين لخوارزمية تقوم بتحديده بفعالية أكبر قياساً إلى أحدث الخوارزميات المتعلقة بهذا المجال، حيث سعينا في هذا البحث إلى تطوير خوارزمية تحليل للبيانات الضخمة تعتمد بشكل عام على التمثيل الرياضي لأبجدية البيانات وقمنا بتخصيص تطبيقها لنحدد وفقها معامل النسخ الموجود ضمن سلاسل DNA.

وبالتالي فقد هدفنا في هذا البحث إلى ما يلي:

١ -إجراء دراسة مرجعية عن عملية اتخاذ القرارات باستخدام التنقيب في البيانات الضخمة.
٢ -دراسة خوارزميات تحليل سلاسل DNA.
٣ -تطوير خوارزمية تقوم بتحديد المعاملات الهامة ضمن البيانات النصية وذلك بالاعتماد على المفاهيم الرياضية.
٤ -تطبيق الخوارزمية بشكل عملي على سلاسل DNA لتحديد مواقع معامل النسخ TF.

4- ترتيب البحث:

يحتوي البحث على دراسة مرجعية لأهم المصطلحات والمفاهيم اللازمة لدراسة وتحليل منهجية اتخاذ القرارات عن طريق التنقيب في البيانات الضخمة، مع تطبيق هذه المنهجية على سلاسل DNA.

حيث يتألف هذا البحث من ثمانية فصول:

- **الفصل الأول: البيانات الضخمة**

يتضمن سبب نشوء هذا المصطلح وما يحمله من تعقيدات وصعوبات وأهمية، وما يحمله من دور أساسي في اتخاذ القرارات الفعالة.

- **الفصل الثاني: دور البيانات الضخمة في صناعة القرارات**

نستعرض في هذا الفصل كيفية اتخاذ القرارات بالاعتماد على نماذج المعرفة الناجمة عن التنقيب في البيانات بشكل عام، مع توضيح الصعوبات الموجودة ضمن التنقيب في البيانات الضخمة.

- **الفصل الثالث: المعلوماتية الحيوية**

قمنا في هذا الفصل بشرح هذا العلم الناجم عن اتحاد علم الأحياء مع علم الحاسوب، مع توضيح مفهوم ثورة المعلومات الجينية المؤثرة في هذا العلم، وإلقاء الضوء على أهميته المتمثلة في تشخيص الأمراض ومعالجتها.

- **الفصل الرابع: سلاسل DNA**

يتناول هذا الفصل توضيح هيكلية سلاسل DNA مع استعراض أهم المعاملات المخزنة ضمن هذه السلاسل.

- **الفصل الخامس: دراسة خوارزميات تحليل سلاسل DNA**

تم في هذا الفصل استعراض أحدث الخوارزميات التي قامت بتحليل سلاسل DNA لاستخراج المعلومات الهامة والمخزنة ضمن ثناياها.

- **الفصل السادس: الخوارزمية المقترحة**

ونبين فيه الخوارزمية التي قمنا بتطويرها مع إيضاح آلية عملها.

- **الفصل السابع: النتائج والأعمال المستقبلية**

نناقش فيه ايجابيات الخوارزمية المقترحة وسلبياتها، إضافة إلى توضيح مدى فعاليتها قياساً إلى الخوارزميات التي سبقتها.

الفصل الأول

البيانات الضخمة

Big Data

1.1- مقدمة:

"البصمة الرقمية" عبارة قصيرة تعبر بجوهرها عن كميات لا متناهية الكبر من البيانات الناجمة عن الأثر الذي يتولد عند قيام الفرد بأي شكل من أشكال التفاعل على شبكة الإنترنت، وقد استوعبت الشركات العالمية والدول المتقدمة أهمية الاستفادة من تلك البيانات، حيث قامت بوضع خطط مستقبلية وبناء مراكز بيانات متخصصة للاستفادة من تلك البيانات، مثل مشروع وكالة الأمن القومي (National Security Agency (NSA لتطوير قاعدة بيانات وطنية أطلق عليه مشروع مركز بيانات يوتاه (Utah Data Center)، حيث يهدف إلى تحليل بيانات مستخدمي شبكات الإنترنت والاتصالات في العالم لفهم سلوكياتهم ونشاطاتهم.

وتشكل هذه البيانات حجر الأساس في بزوغ مصطلح البيانات الضخمة، فما هي البيانات الضخمة؟ نستعرض في هذا الفصل طيات هذا المصطلح المتنامي أهمية في الفضاء الإلكتروني. [1]

2.1- لمحة تاريخية:

أدى التقدم السريع الذي شهده العالم مؤخرا ويشهده حاليا في مجال تكنولوجيا المعلومات والاتصالات إلى ثورة معلوماتية هائلة نجم عنها قواعد بيانات كبيرة، حيث تشير الدراسات التي قامت بها شركة إنتل (Intel) إلى أن حجم البيانات التي أنتجها البشر منذ بداية التاريخ حتى عام 2003 يبلغ حوالي خمسة إكسابايت (10^{18} بايت)، إلا أن هذا الحجم تضاعف حوالي خمسمائة مرة خلال عام 2012 ليصبح 2.7 زيتابايت (10^{21} بايت)، وتوقعت الدراسات آنذاك أن يتضاعف هذا الرقم إلى حوالي ثلاث مرات بحلول عام 2015، وبذلك فقد أصبحت البيانات الضخمة بمصادرها المختلفة واقعا يعيشه العالم بأسره، واعتمد قاموس أكسفورد الانجليزي مصطلح البيانات الضخمة (Big Data) وأضافه للقاموس مع مصطلحات حديثة أخرى .[2]

وانطلاقاً مما سبق، فقد واجه النمو المتسارع في إنتاج البيانات (من حيث الحجم والمصدر والسرعة والتنوع) تحديات كبيرة في كيفية التعامل مع هذه البيانات وطرق الاستفادة منها، فقد باتت البيانات الضخمة حدث الساعة من قبل معظم الجهات والمؤسسات الوطنية وا لإقليمية والدولية، ومجالاً واسعاً للبحث في إدارة قواعد البيانات، والمنهجيات وا لإجراءات الممكن تبنيها في سبيل استغلال البيانات الضخمة في جميع مجالات الحياة.[3]

3.1- مفهوم البيانات الضخمة[4]:

منذ وقت ليس ببعيد كانت تنحصر البيانات ضمن فئة قواعد البيانات المنظمة في مجلدات وملفات وجداول وغيرها، والتي أصبحت لا تشكل الآن أكثر من 10% من إجمالي البيانات في العالم، وذلك لدخول مصادر معلومات جديدة وغير منظمة مثل رسائل البريد الإلكتروني، مقاطع الفيديو، منشورات الفيسبوك، التغريدات، رسائل الدردشة على الواتس آب، وغيرها من المصادر الأخرى، مما أدى إلى تكوين قواعد البيانات الضخمة.

لا يوجد هناك تعريف واضح للبيانات الضخمة، لأن مفهوم حجم البيانات غير محدد بمكان أو زمان في ظل ما نشهده حاليا من تسارع في تطور تكنولوجيا المعلومات والاتصالات فالبيانات الضخمة في الوقت الحالي قد لا تكون ضخمة في الوقت القادم، وما هو ضخم من بيانات بالنسبة لشخص أو مؤسسة معينة قد لا يعد ضخماً بالنسبة لشخص آخر أو مؤسسة أخرى.

أطلق معهد ماكنزي العالمي (Mackenzie) في عام 2011 تعريفا للبيانات الضخمة بأنها " مجموعة من البيانات بحجم يفوق قدرة قواعد البيانات التقليدية من تخزين وإدارة وتحليل تلك البيانات "، أطلق مصطلح البيانات الضخمة في مجال تقنية المعلومات على مجموعة من حزم البيانات الضخمة جدا والمعقدة والتي يصعب التعامل معها بواسطة نظم إدارة قواعد البيانات التقليدية .

من جانب آخر، تصف الأمم المتحدة البيانات الضخمة بأنها " مصادر البيانات ذات الأحجام الضخمة والسرعات العالية والتنوع في البيانات، والتي تتطلب أدوات وأساليب جديد ة لحفظها وإدارتها ومعالجتها بطريقة فعالة ".

4.1- أبعاد البيانات الضخمة[5]:

إضافة إلى المفهوم المتعارف عليه في البيانات الضخمة وهو كبر حجم البيانات، عرفت شركة غارتنر (Gartner) للأبحاث أبعاداً أخرى تميز البيانات الضخمة، وذلك كما يلي:

١ -**حجم البيانات:** والتحديات المتعلقة بعمليات الجمع والتخزين والتحليل لقواعد البيانات .

٢ -**تعدد وتنوع البيانات:** وذلك من حيث مصادرها المختلفة التي تم ذكرها سابقا، إضافة إلى تنوع تركيبها وهيكليتها التي تنقسم إلى عدة أنواع:

أ - البيانات المنظمة (Structured Data): وهي البيانات المخزنة ضمن قواعد بيانات منظمة، حيث يميزها إمكانية البحث فيها وتحليلها، كما يمكن إدارتها باستخدام الوسائل التقنية التقليدية، ويمثل هذا النوع من البيانات نحو 10% فقط من إجمالي "البيانات الضخمة".

ب -البيانات غير المنظمة (UnStructured Data): هي كل ما لا يمكن تصنيفه بسهولة كالصور والرسوم البيانية، ومقاطع الفيديو، وصفحات الويب، وملفات PDF، والعروض

التقديمية، ورسائل البريد الإلكتروني، والتغريدات، ومنشورات الفيسبوك، ورسائل الدردشة، وغيرها. ورغم أن هذه الأنواع من الملفات لها هيكل داخلي يخصها، تعتبر "غير منظمة" لأن بياناتها لا تتسق تماما كقواعد البيانات.

ت – البيانات نصف المنظمة (Semi-Structured Data): وهي بيانات شبه منظمة ويمكن تصنيفها بين النوعين السابقين، فهي خليط بين الاثنين، مثل ملفات XML.

٣ –**سرعة تواتر حدوث البيانات:** وسرعة الوصول إليها وتحليلها، فهناك معلومات مثل التغريدات يتم الحصول عليها وحصرها خلال فترات قياسية قصيرة جداً.

٤ –**مدى دقة هذه البيانات:** بحيث تمكن من الاستخدام الفاعل لها من قبل أصحاب القرار .

5.1– استخدام البيانات الضخمة[5]:

مما لا شك فيه أن إتاحة البيانات الضخمة بأنواعها المختلفة للاستخدام سيؤدي إلى فتح آفاق واسعة للبحث والتطوير على مختلف الأصعدة، تقول شركة IBM أن البيانات الضخمة تعطيك فرصة لاكتشاف رؤى مهمة في البيانات، وتضيف شركة Oracle بأن البيانات الضخمة تتيح للشركات أن تفهم زبائنها بعمق أكثر. إن الفائدة المتوقعة من استخدام البيانات الضخمة ستكون في مختلف الجوانب والمجالات، فعلى سبيل المثال لا الحصر نذكر ما يلي:

١ –دعم البحث العلمي وذلك في مختلف المجالات، من خلال توفيرها لقواعد بيانات شاملة على مستويات تفصيلية للمجتمعات المراد دراستها، وضمن النطاق الزمني المحدد لمرجعية البيانات. مما ينعكس إيجابا على دقة وكفاءة نتائج الأبحاث والدراسات.

٢ –تشجيع الإبداع والابتكار، من خلال استغلال قواعد البيانات الضخمة لإجراء دراسات تفصيلية عليها لتطوير وابتكار خدمات جديدة تساهم في تحسين مستوى أداء المؤسسة أو الشركة .

٣ –الاستخدام في المجالات الاقتصادية المختلفة كاستغلال قواعد البيانات الضخمة في تحفيز الاستثمار وخلق فرص العمل، إضافة إلى تشجيع التنافسية في مؤسسات الأعمال مما يزيد من كفاءة وجودة الخدمات والسلع المنتجة.

٤ –إتاحة البيانات الضخمة لتحسين مستوى الخدمات المختلفة، ففي مجال الرعاية الصحية تستخدم في مجال البحث حول أساليب توفير الرعاية الصحية المثلى، وغيرها من القضايا .

٥ –المساهمة في التحول من الاقتصاد المعرفي إلى الاقتصاد الإبداعي، وسينتقل الأفراد والهيئات من إنتاج البيانات إلى الاستفادة من البيانات بما يصب في مصلحة الجميع.

6.1- مثال عن البيانات الضخمة:

لنلق نظرة على إحصائية الطلاب المسجلين في جامعة البعث للعام الدراسي 2019/2018 وفق الجدول 1.1

الجدول 1.1 إحصائية بطلاب جامعة البعث

الكلية		السنة الأولى		السنة الثانية	السنة الثالثة	السنة الرابعة	السنة الخامسة	السنة السادسة	المجموع
		مستجد	قديم						
الهندسة المعلوماتية		464	110	371	379	256	322	0	1902
السنة التحضيرية		1061	0	0	0	0	0	0	1061
الطب البشري		0	8	679	370	400	319	282	2058
طب الأسنان		0	0	271	87	61	0	0	419
الصيدلة		0	0	661	365	395	376	0	1797
الهندسة المعمارية		451	32	302	306	338	360	0	1789
الهندسة المدنية	العامة	879	86	594	591	536	416	0	3102
	البيئية	161	29	121	122	95	106	0	634
	الموارد المائية	146	29	142	119	100	116	0	652
الهندسة الميكانيكية والكهربائية	الطاقة الكهربائية	223	45	132	155	228	225	0	1008
	القوى الميكانيكية	216	44	161	145	115	114	0	795
	التصميم والإنتاج	177	92	269	259	153	83	0	1033
	الالكترونيات والاتصالات	350	35	270	281	229	216	0	1381
	التحكم الآلي والحواسيب	140	66	300	242	160	140	0	1048
	الميكاترونيك	198	43	71	145	104	107	0	668
	المعادن	159	72	98	107	108	45	0	589
الهندسة الكيميائية والبترولية	البترولية	436	140	356	307	233	265	0	1737
	الكيميائية	252	75	256	214	202	175	0	1174
	الغذائية	237	66	177	134	115	126	0	855
	الغزل والنسيج	130	46	103	108	72	74	0	533
الزراعة		219	62	391	476	349	306	0	1803
العلوم الصحية		510	100	397	305	285	0	0	1597

الكلية		السنة الأولى		السنة الثانية	السنة الثالثة	السنة الرابعة	السنة الخامسة	السنة السادسة	المجموع
		مستجد	قديم						
العلوم	الرياضيات	310	450	819	617	492	0	0	2688
	الفيزياء	197	94	326	240	159	0	0	1016
	الكيمياء	398	138	393	294	165	0	0	1388
	الإحصاء الرياضي	54	101	143	56	174	0	0	528
	علم الحياة	148	152	173	147	208	0	0	828
	الجيولوجيا	96	83	132	59	94	0	0	464
العلوم الثانية	الرياضيات	5	5	8	5	11	0	0	34
	الكيمياء	2	7	15	5	14	0	0	43
التربية	الإرشاد النفسي	750	730	800	359	282	310	0	3231
	معلم صف	950	880	1150	901	1100	0	0	4981
	مناهج وطرائق التدريس	700	450	500	340	300	200	0	2490
	رياض الأطفال	250	294	312	211	287	0	0	1354
التربية الثانية		51	55	260	162	160	0	0	688
الحقوق		798	400	881	680	630	0	0	3389
الاقتصاد		164	23	265	126	53	0	0	631
السياحة		73	114	163	79	92	0	0	521
الآداب والعلوم الإنسانية	اللغة العربية	579	1207	1807	1614	2123	0	0	7330
	اللغة الإنكليزية	79	1118	714	1059	1359	0	0	4329
	اللغة الفرنسية	209	503	716	588	1901	0	0	3917
	اللغة الفارسية	12	29	31	55	101	0	0	228
	التاريخ	149	2087	1233	619	1223	0	0	5311
التربية الموسيقية		57	125	160	85	73	0	0	500
المجموع النهائي		12440	10225	17123	13518	15535	4401	282	73524

بالتمعن في الجدول السابق نلاحظ أنه على الرغم من العدد الكبير للطلاب المسجلين فيمكن تخزين البيانات الشخصية للطلاب وفق قواعد البيانات التقليدية وبطريقة مهيكلة، ولكن ماذا عن باقي المعلومات الضرورية التي يتم تسجيلها للطالب خلال دراسته ونذكر منها على سبيل المثال لا الحصر:

١ – نوع القبول.

٢ – تاريخ تسجيله لأول مرة.

٣ – حصوله على مصدقة التأجيل.

٤ – قيامه بإيقاف التسجيل مع تحديد الفصل.

٥ – درجاته في الامتحان العملي لكل المواد.

٦ – درجاته في الامتحان النظري لجميع المواد.

٧ – أحقيته بدخول دورة معينة.

٨ – عدد الفصول التي بقيت له قبل أن يستنفذ.

٩ – الرسم الواجب دفعه لقاء التسجيل.

وغيرها الكثير من البيانات الواجب ربطها مع بيانات كل طالب والتي بعضها يحسب استنادا إلى قوانين وزارة التعليم العالي، وبعضها الآخر متغير كل سنة بل وكل فصل!

وعليه فإننا نقف أمام نوع من أنواع البيانات الضخمة الواجب معالجتها بغية الوصول إلى خطة إستراتيجية لازمة لتطوير منهجية التعليم.

7.1– خلاصة:

استعرضنا في هذا الفصل مفهوم البيانات الضخمة والذي يعد نفط العصر الواجب تكريره ومعالجته بغية الاستفادة الكلية من عصرنا الرقمي.

الفصل الثاني

دور البيانات الضخمة في صناعة القرارات

The Role of Big Data in Decision Making

1.2- مقدمة:

"المنتخب الألماني يحقق كأس العالم عام 2014 بفضل علم البيانات الضخمة" عبارة رددها الكثير من المحللين الرياضيين والمتابعين لمنتخب ألمانيا الذين ارتأوا بأن اختيار مدرب المنتخب للاعب ماريو غوتزه لينزل كبديل في الدقائق الأخيرة من المباراة النهائية ليحرز هدف اللقب فور دخوله، لم يكن نا جماً عن قراءة فنية فقط بل عن دراسة تحليلية للبيانات الطبية والمهارية لكل لاعب من لاعبي المنتخب...
ففعالية أي قرار يعتمد بشكل أساسي على مدى أهمية وكفاءة البيانات التي يرتكز عليها هذا القرار.
سنلقي الضوء في هذا الفصل على مفهوم البيانات الضخمة، موضحين الدور الذي تلعبه هذه البيانات في فعالية صناعة القرارات، وذلك باستخدام علم التنقيب في البيانات للحصول على نماذج المعرفة اللازمة لصنع القرارات الفعالة.

2.2- لمحة عن علم التنقيب في البيانات[8][7][6]:

"من يملك المعلومات يملك العالم"، عبارة أكدتها حادثة لقاء الرئيس الأمريكي باراك أوباما عام 2013 مع المديرين التنفيذيين لكبرى الشركات أمثال فيسبوك، غوغل، ياهو و تويتر....
حيث ناقش معهم موضوع "التنقيب في البيانات"، وذلك بغية الحصول على المعلومات التي يريدها كسلاح فعال يجابه به من يريد...
تعرف عملية التنقيب في البيانات بأنها عملية دمج الطرق التقليدية لتحليل البيانات مع خوارزميات معقدة من أجل استخلاص معلومات دقيقة و مفيدة من ثنايا كم هائل من البيانات.
فمع وجود كميات كبيرة من البيانات المخزنة في قواعد البيانات ازدادت الحاجة إلى تطوير أدوات تمتاز بالكفاءة لتحليل البيانات و استخراج المعلومات و المعارف منها. ومن هنا ظهر ما يسمى بالتنقيب في البيانات كتقنية تهدف إلى استخراج المعرفة من كميات هائلة من البيانات، و هي تقنية حديثة فرضت نفسها بقوة في عصر المعلوماتية، لأن استخدامها يوفر للشركات و المنظمات في جميع المجالات القدرة على استكشاف و التركيز على أهم المعلومات في قواعد البيانات، كما تركز تقنيات التنقيب على بناء التنبؤات المستقبلية و استكشاف السلوك و الاتجاهات مما يسمح باتخاذ القرارات الصحيحة في الوقت المناسب.

3.2- أسباب ظهور مفهوم التنقيب في البيانات[9] :

ظهر مفهوم التنقيب في البيانات في أواخر الثمانينيات من القرن الماضي وأثبت وجوده كأحد الحلول الناجحة لتحليل كميات ضخمة من البيانات، وذلك بتحويلها من مجرد معطيات متراكمة وغير مفهومة (بيانات) إلى معلومات قيمة يمكن استغلالها والاستفادة منها بعد ذلك ، ونستعرض فيما يلي أهم أسباب ظهور هذا المفهوم:

١ -تطور تكنولوجيا المعلومات وخاصة تكنولوجيا تخزين البيانات و معالجتها.
٢ -انخفاض تكاليف الاتصالات الإلكترونية مما سهل آلية الوصول إلى قواعد البيانات.
٣ -ظهور أساليب تحليل جديدة كالشبكات العصبونية وشجرة القرار.
٤ -تطور صناعة البرمجيات.
٥ -تطور أساليب وتقنيات تخزين البيانات.
٦ -تزايد الحاجة إلى نتائج تحليلية سريعة من قبل الشركات بما يحقق لها المواكبة المستمرة والفعالة في مجال عملها.

4.2- كيفية التنقيب في البيانات [10] :

تنقسم نماذج التنقيب في البيانات إلى نوعين أساسيين:

أولاً : النماذج التنبؤية

تهدف إلى التنبؤ بقيمة بعض الخصائص. مثل التنبؤ باحتمال شراء زبون لسلعة معينة.

ثانياً : النماذج الوصفية

وتنقسم هذه النماذج إلى صنفين:

١ -نماذج العنقدة التي تسمح بتجميع الأفراد، والأحداث، أو المنتجات في عناقيد.
٢ -نماذج الارتباط التي تسمح بتحديد العلاقات فيما بينهم.

وهناك عدة أدوات للتنقيب في البيانات، نذكر أهمها:

- **التلخيص:**

يشير التلخيص إلى أساليب تفتيت كتل البيانات الكبيرة إلى مقاييس موجزة، توفر وصفا عاما للمتغيرات وعلاقاتها. ومن الأمثلة لأساليب التلخيص نذكر: المتوسطات، والمجاميع، والإحصائيات الوصفية التي تتضمن مقاييس النزعة المركزية مثل المتوسط الحسابي و الوسيط والمنوال، ومقاييس التشتت مثل الانحراف المعياري. وعلى الرغم من أن مقاييس التلخيص تعطي صورة كبيرة عن بعض التفاصيل ذات العلاقة فإنها غالبا ما تهمل تفاصيل أخرى ذات أهمية كبيرة تتعلق بسلوك المستهلك خصوصا.

- **التصنيف:**

يتمثل التصنيف في تفسير أو التنبؤ بخاصية فرد ما من خلال خصائص أخرى، ويمكن انجاز التصنيف بالاعتماد على الأساليب الإحصائية القديمة مثل الانحدار والتحليل التمييزي، أو بالاعتماد على أساليب حديثة نسبيا مثل قوى الارتباط والاستنتاج المستند إلى الحالة والشبكات العصبونية.

- **التنبؤ:**

يشبه التنبؤ التصنيف أو التقدير، ما عدا أن البيانات تصنف على أساس التنبؤ بسلوكها المستقبلي أو تقدير قيمتها المستقبلية. حيث إن المتغير التابع المتنبأ به هو متغير كمي . ومن الأدوات التقليدية المستخدمة في التنبؤ نذكر على سبيل المثال: الانحدارات بأنواعها و التحليل التمييزي. أما الأساليب الجديدة فتشتمل على قواعد الارتباط وشجرة القرار والشبكات العصبونية والخوارزميات الوراثية.

- **العنقدة أو التجزئة:**

يتمثا التجميع العنقودي أو التجزئة إلى قطاعات في البحث عن مجموعات متجانسة في مجتمع من الأفراد.و يشير التجميع العنقودي أو التجزئة إلى قطاعات إلى عملية تشكيل مجموعات أو قطاعات مؤلفة من أفراد أو أصحاب أسر، وذلك بالاستناد إلى معلومات متضمنة في مجاميع من المتغيرات التي تصفهم. و الغرض من التجميع العنقودي المساعدة على تطوير برامج تسويقية مصممة على مقاسات الزبائن أنفسهم، و التي بالإمكان استخدامها لاستهداف أعضاء لكل قطاع من هذه القطاعات على أمل ترغيبهم في تكرار الشراء أو التحول إلى زبائن دائمين.

وتتم أساليب التجميع العنقودي غالبا بمساعدة أساليب التحليل العنقودي الإحصائية و الأساليب المستندة إلى شجرة القرار، والشبكات العصبونية والخوارزميات الوراثية.

- **تحليل الارتباط :**

يتمثل الارتباط في البحث عن علاقات أو ارتباطات موجودة بين عدة خصائص. و يشير تحليل الارتباط إلى مجموعة من الأساليب التي تستخدم لربط أنماط الشراء عبر القطاعات المتقاطعة أو عبر الوقت. فمثلا يقوم أسلوب تحليل سلة السوق (نوع من أنواع الارتباط) باستخدام المعلومات الكامنة في السلع التي اشتراها المستهلكون فعليا للتنبؤ بالسلع المحتمل شراؤهم إياها إذا ما تم تقديم عروض خاصة لهم أو إذا تم تعريفهم بهذه السلع.

- **الكشف عن التغيرات أو الانحرافات:**

يرتكز على استكشاف التغيرات المهمة جدا في البيانات من خلال قياسات سابقة أو قيم معيارية.

5.2- مستودع البيانات[11] :

هو عبارة عن" مجموعة من بيانات دائمة تاريخية متكاملة تساعد في اتخاذ القرارات الإدارية "، للمساعدة في الوصول إلى البيانات لأغراض التحليلات الزمنية واكتشاف المعرفة واتخاذ القرارات، فهي مصممة خصيصاً لاستخراج واستخلاص البيانات ومعالجتها وتقديمها وتمثيلها في صورة مناسبة لهذا الغرض، وتتضمن كميات ضخمة من البيانات القادمة من مصادر مختلفة، أو من عدة قواعد مختلفة من أنظمة وأماكن مختلفة، وتتميز تلك النوعية من قواعد البيانات بتطابق بنيتها الداخلية مع ما يحتاجه المستخدم من مؤشرات ومحاور التحليل.

خصائص وصفات مستودعات البيانات:

١ -تستخدم النموذج المتعدد الأبعاد.

٢ -تدعم السلاسل الزمنية وتحليل التوجهات اللذين يحتاجان لبيانات تاريخية لا تستطيع قواعد البيانات العادية أن توفرها .

٣ -تحديث البيانات وهو يتم كل فترة بواسطة أجزاء منه تختص بهذا الأمر .

٤ -استرجاع البيانات وتحليلها وهو أساس عملها.

٥ -دعم معمارية العميل/الخادم وتعددية المستخدمين .

٦ -الاحتفاظ بكمية ضخمة من البيانات قد تصل إلى عدة تيرابايتات .

٧ -محددة لموضوع على سبيل المثال في منشأة طبية ما تكون موضوعاتها تتعلق بالأطباء، المرضى، الممرضين، الأدوية...

٨ -متكاملة من خلال وجود علاقة بين البيانات .

٩ -غير قابلة للتعديل أن تلك البيانات عند تحميلها في المستودعات تستخدم فقط للتحليل والدراسة والعرض.

١٠ -مرتبطة بعامل الوقت وهي أهم نقطة في تلك البيانات .

خطوات بناء مستودعات البيانات:

١ -إنشاء مساحة للبيانات وهي قاعدة بيانات ذات سعة تخزين عالية جداً تقوم بتخزين كافة البيانات القادمة من أنظمة التشغيل المختلفة لكي يتم تنقية وتعديل البيانات فيها قبل تحميلها في مستودع البيانات، ويراعى فيها أن يكون تصميم هذه القاعدة متوافقاً مع تصميم مستودع البيانات.

٢ -بناء مستودع البيانات وتصمم المستودعات دائما بحيث تسمح بوجود علاقات ذات أبعاد مختلفة.

٣ – تجزئة مستودع البيانات إلى مجموعة من متاجر البيانات بحيث يكون هناك بيانات خاصة بالإدارة المالية، وأخرى خاصة بإدارة الموارد البشرية أو يكون التقسيم مبنياً على فروع المنظمة.

٤ –دمج وتنقية ونقل البيانات وفي هذه المرحلة يتم جلب البيانات من مصادرها المختلفة، ونقوم بتحويلها من صورة إلى أخرى إذا تطلب الأمر ذلك، وفي أحيان كثيرة تدمج بعض البيانات مع بعضها الآخر، أو نقوم بتعريف بيانات جديدة لم تكن موجودة من قبل، بالإضافة إلى تنقية البيانات غير الصحيحة وحذف غير المهم منها.

٥ –تحميل البيانات في مستودع البيانات وفي هذه المرحلة يتم اختبار البيانات.

٦ –تحليل البيانات وإنشاء تطبيقات نظم دعم اتخاذ القرار وفي هذه المرحلة تنفذ التطبيقات الخاصة بعرض مستودع البيانات وتحليلها، حيث تقوم هذه التطبيقات بعرض البيانات بعدة أبعاد وتقوم باستخدام خوارزميات معقدة لتحليل البيانات.

وتجدر الإشارة إلى وجود بعض القضايا التي يجب مراعاتها في عملية البناء، ونذكر منها:

١ –استخلاص البيانات من عدة مصادر قد تكون غير متجانسة.

٢ –مراقبة وضبط حجم مستودع البيانات أثناء وبعد تحميله بالبيانات.

٣ –تهيئة البيانات لضمان انسجامها داخل مستودع البيانات.

٤ –تحديث البيانات كل فترة من الزمن.

٥ –تحديد الوقت اللازم للبناء وما الجدوى المنتظرة.

٦ –تنظيف البيانات لضمان جودتها، ويتم من خلال قاعدة البيانات التي أخذت منها البيانات.

ويعد مستودع البيانات حجر الأساس في صناعة نماذج المعرفة، وقبل أن نستعرض دور مستودع البيانات في صناعة القرارات علينا أن نشرح عملية اتخاذ القرارات وذلك في الفقرات القادمة.

6.2– أهمية عملية اتخاذ القرارات[12][11] :

يعرف **القرار** بأنه عملية عقلية يقوم بها المرء لاختيار طريقة القيام بفعل معين أو قول معين من بين عدة خيارات ممكنة، فيما تعرف عملية اتخاذ القرار بأنها عملية إدراكية معقدة تهدف إلى صناعة قرارات سليمة وحلول واقعية تقلل نسبة الأخطاء الناجمة عن هذا القرار...

تعتبر عملية اتخاذ القرارات من أهم الأعمال التي يقوم بها المدير، بل هي صلب عمل المدير، فهي وظيفة أساسية يمارسها المديرون في كل وقت من الأوقات.

فالعمل الإداري ما هو إلا سلسلة متصلة من القرارات، ويرتبط نجاح المنظمة واستمرارها وتفوقها بمدى كفاءة القرارات التي يتم اتخاذها في مستوياتها المختلفة، لذلك يجب أن تتوافر القدرة على اتخاذ القرارات في جميع من يشغلون المناصب الإدارية.

علماً أن اتخاذ القرارات لا يشمل المفهوم الإداري فقط، وإنما يتعداه ليشمل السلوك اليومي للإنسان في جميع الأعمال التي يفعلها والكلمات التي يتلفظ بها.

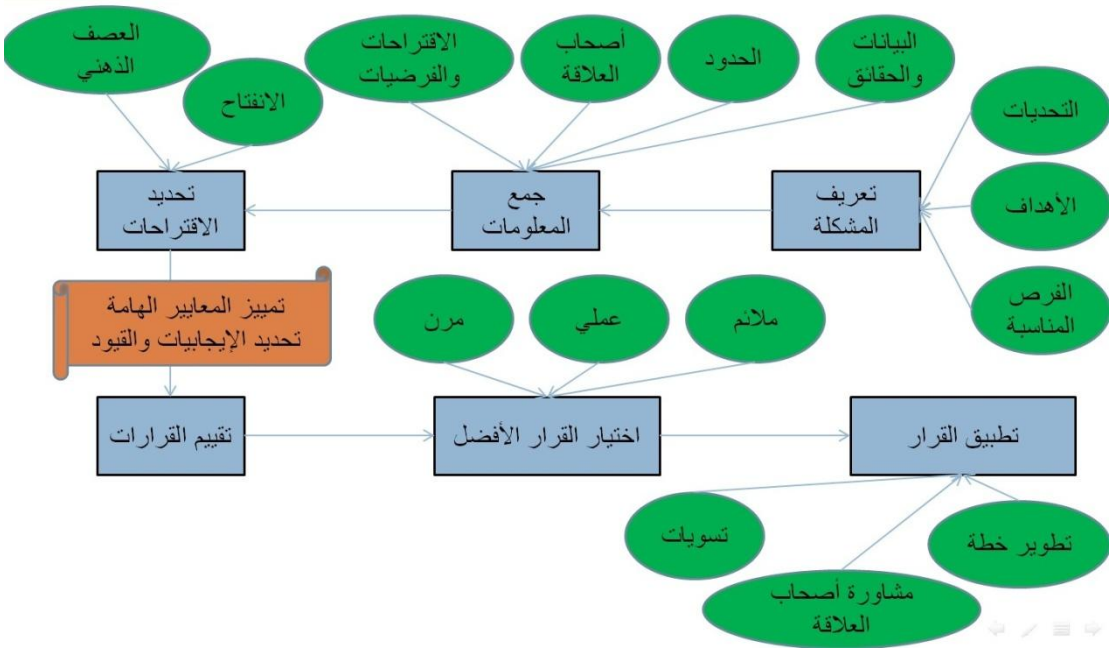
7.2- خطوات عملية اتخاذ القرارات[13][12] :

لا بد بداية من التمييز بين مفهومي عملية صنع القرار واتخاذ القرار ، فهما يفسران أو يستخدمان على أنهما عملية واحدة يمكن أن تؤدي الغرض نفسه، والواقع خلاف ذلك، فعملية صنع القرار هي عملية واسعة تتضمن أكثر من مرحلة، أما اتخاذ القرار فإنه يمثل آخر مرحلة في عملية صنع القرار، وبالتالي فهو مرحلة من عملية صنع القرار وليست مرادفة لها.

ليس هناك طريقة مثالية لصنع القرارات بسبب حالة عدم التأكد التي تكتنف عملية صنع القرارات، وليست هناك معادلة لتوضيح كيفية اتخاذ القرارات الناجحة،

ولذلك فإن اتخاذ القرارات لا يقوم فقط على المنطق، وإنما في كثير من الأحيان يبنى على الحكم الشخصي والمبادرة من قبل متخذ القرار ، الأمر الذي يضعه موضع النجاح النسبي أو الفشل النسبي، إلا في الحالات النادرة التي يصنف بها هذا القرار بالصفة المطلقة نجاحاً كانت أم فشلاً.

ورغم تنوع الأسس المعتمدة في تصنيف القرارات، فإنها تتوحد في مراحل صنعها، حيث تمر عملية اتخاذ القرارات بالخطوات الموضحة في الشكل 1.2



الشكل 1.2 خطوات عملية صنع القرارات

١ -تعريف المشكلة:

في هذه الخطوة يتم تحديد الأهداف المنشودة والمعوقات المقيدة لنا كما يتم تحديد الإيجابيات المتاحة والمرتكزة على التجارب السابقة، وذلك بالدراسة العميقة والتشخيص الفعال للمشكلة التي نواجهها ونبتغي حلها.

٢ -جمع المعلومات:

وذلك بتحليل المشكلة تحليلاً منهجياً عن طريق تحديد الجهات المتفاعلة مع هذا القرار، وحدود تنفيذ القرار، ومدى تأثيره وذلك استنادا إلى البيانات والحقائق المتوفرة، مع تقديم عدة اقتراحات قابلة للتنفيذ وتحقق الهدف المنشود رغم اختلافها.

٣ -تحديد الاقتراحات:

في هذه المرحلة يتم دراسة الاقتراحات التي تم تحديدها في الخطوة السابقة، لتحديد الاقتراحات الأكثر قابلية للتنفيذ مع مراعاة إيجابيات وسلبيات كل منها وفي ظل وجود عدة قيود تعوق تنفيذها.

٤ -تقييم الاقتراحات:

يتم في هذه المرحلة توصيف الحلول التي تم اقتراحها في الخطوة السابقة، وذلك بربط كل حل بما يلي:

١) قابلية تطبيقه.
٢) الآثار الناتجة عن تنفيذه.
٣) مدى تحقيقه للهدف المنشود.
٤) الإيجابيات المنتظرة جراء تطبيقه.
٥) تكاليف تنفيذه.

٥ -اختيار القرار الأفضل:

بما أنه لا يوجد قرار مثالي وعليه يتم تحديد القرار الأفضل والناتج عن دراسة كل مواصفات القرارات المقترحة، ويجب أن يتصف القرار المتخذ بما يلي:

١) ملائم: يحقق الهدف المنشود.
٢) عملي: قابل للتطبيق.
٣) مرن: قابل للتوسع والتطوير مع مرور الوقت واختلاف المعطيات بشكل كمي لا نوعي.

٦ -تطبيق القرار:

حيث يتم تنفيذ القرار، مع معالجة كل المشاكل غير المتوقعة والناجمة عن تطبيقه.

8.2- دور نماذج المعرفة في صناعة القرارات[15][14][13]:

تكتسب المعلومات أهميتها من كفاءة الدور الذي تمثله في تزويد الإنسان بما يحتاج إليه من معارف يستمد منها تقديراته وتصوراته....
وعبر مراحل تاريخية متتالية تزايدت أهمية المعلومات بشكل كبير وذلك بما تحدثه من أثار عميقة في توسيع المعرفة الإنسانية وتنمية وعي الفرد وإدراكه لما يحيط به من ظواهر ومتغيرات مختلفة.

واليوم في ظل عالمنا المعاصر غدت المعلومات صناعة العصر التي تمكن من يمتلكها من امتلاك زمام التطور، فقد أصبحت أداة فعالة يعتمد عليها في إدارة تشكيل الحاضر ورسم صورة المستقبل، ولا شك أن ذلك يمثل الشيء الكثير بالنسبة لراسم السياسة وصانع القرار.

وإذا كانت المعلومات على تلك الدرجة من الأهمية والأثر الفاعل في إيصال المعرفة وتسهيل الإلمام بمكونات الواقع وتفاعلاته وتأمين مقدرة اكتشاف الحاضر ودقة التنبؤ بالمستقبل وتدعيم عوامل النمو العلمية والفنية والمادية فإن القيام بعملية صنع القرار في أي من المجالات دونما الارتكاز على المعلومات يفقد متخذ القرار الاستفادة من عامل جوهري وربما حاسم لضمان تحقيق الهدف الذي يتطلع إليه بقراراته المتخذة بل يقود ذلك في حالات مختلفة إلى التعرض لتقديرات خاطئة والوقوع في اتخاذ قرارات خاطئة.

إن دور المعلومات بالنسبة لصانع القرار وإن كان يتخذ أبعاداً ومفاهيم شاملة، إلا أنه يتباين في مستوياته وآثاره ارتباطاً بتباين مستويات التطور والواقع الذي يؤدي مفعوله فيه، وفي كل الأحوال فإن الأثر الفعلي لذلك الدور يتحدد عملياً بمدى إنتاج وتبادل المعلومات واستخدامها كمرجعية شرطية لازمة لعملية اتخاذ القرار.

حيث يجب أن تتوفر لصانع القرار بصورة عامة تغطية واضحة ودقيقة لما يلي :

١ -إيضاح طبيعة الموضوع أو المشكلة المطروحة وما يرتبط بذلك من خلفيات ومسببات ودوافع.
٢ -التحليل الدقيق لمكونات الموضوع وما يتداخل معه من تأثيرات وتفاعلات متبادلة.
٣ -إيضاح متطلبات ودواعي اتخاذ القرار.
٤ -تقديم الخلاصات والتصورات وتحديد البدائل المتعلقة باتخاذ القرار.
٥ -تحديد الإمكانيات المتوفرة والمطلوبة واللازمة لتنفيذ أي من البدائل المعروضة لاتخاذ القرار.
٦ -إيضاح حدود اختصاصات ودور الجهات الأخرى فيما يتعلق بموضوع القرار.
٧ -تحديد الآثار المحتملة لاتخاذ وتنفيذ القرار.

9.2- خلاصة:

استعرضنا في هذا الفصل مفهومي التنقيب في البيانات واتخاذ القرارات، قبل أن نوضح دور التنقيب في البيانات بصناعة القرارات الفعالة.

الفصل الثالث

المعلوماتية الحيوية

Bioinformatics

1.3- مقدمة:

أدت التطورات التقنية الحديثة إلى تفجر كميات هائلة من البيانات في مختلف المجالات ومنها مجالات العلوم الحيوية إلى حد تعذر معه تحليل تلك البيانات واستخراج النتائج منها بالعقل البشري المجرد.
و واكب ذلك تطور كبير في مجال المعلوماتية، مما أسهم في اندماج معظم هذه التقنيات لتحليل تلك البيانات الضخمة، بغية التوصل إلى حلول عملية كثيرة أدت إلى ثورة علمية في العديد من القطاعات وفك غموض العديد من الأسرار الدفينة.
وتطلب ذلك التطور المعرفي انبثاق علم جديد هو علم المعلوماتية الحيوية، بوصفه علماً يدمج الحاسوب مع الرياضيات والجينوم والأحياء، ويتعامل مع كافة البيانات الحيوية العائدة للكائنات الحية.
سنستعرض في هذا الفصل نشأة هذا العلم وأهم الخصائص والأسس التي يستند إليها.

2.3- لمحة تاريخية [16]:

في منتصف القرن الماضي تم فك الشيفرة الوراثية وذلك بالتعرف على هيكلية سلاسل الحمض النووي الريبي منقوص الأكسجين DNA وعلاقته بسلاسل الحمض النووي الريبي RNA.

ومع تنامي دور علم المعلوماتية آنذاك قامت العالمة مارجريت داي هوف عام 1964م بعمل خريطة للتابعات البروتينية لتوضيح تركيب وتتابع مجموعة من البروتينات الفيروسية وذلك بمساعدة برنامج لبحث مناطق التشابه والاختلاف بين التتابعات البروتينية والجينية المختلفة، الأمر الذي يعتبر بداية استخدام المعلوماتية الحيوية.

و تلا ذلك معالجة بعض الأمراض و حل العديد من المشكلات الصحية التي يواجهها الإنسان وذلك باستخدام مفهوم الهندسة الوراثية والاستعانة بالتركيب الوراثي والبيولوجي لبكتيريا الإشريكية القولونية والتي تسكن في الأمعاء الغليظة في الإنسان.

وبعد النجاح الذي حققه علم المعلوماتية الحيوية ظهرت مجالات جديدة مثل علم الجينات بغية دراسة جينوم الكائنات الحية المختلفة وذلك باستخدام التقنيات المعلوماتية.

وفي نهاية القرن الماضي تأسست منظمة الجينوم البشري (HUman Genome Organisation) HUGO، والتي تهدف إلى دراسة وتحليل المواد الوراثية المكونة للكائنات الحية.

وفي مطلع القرن الحالي تم التعرف على الكثير من التراكيب الجينية الموجودة ضمن سلاسل DNA.

ولا يزال العمل مستمرا لكشف كامل المعلومات الكامنة والشفرات المتوارية ضمن ثنايا المواد الوراثية المشكلة للكائنات الحية.....

3.3- تعريف المعلوماتية الحيوية [17]:

عرف المركز الوطني للمعلومات التقنية الحيوية (National Center of Biotechnology) NCBI المعلوماتية الحيوية بأنها "مجال العلم الذي يدمج فيه كل من البيولوجي وعلوم الكمبيوتر وتكنولوجيا المعلومات في نسق واحد".

وقد حدد مركز دراسة منهجية العلوم البيولوجية (Biological Sciences Curriculum Study) BSCS مفهوم المعلوماتية بأنها "دراسة تجميع وترتيب وتحليل معلومات وبيانات DNA والبروتين باستخدام الكمبيوتر والتكنيكات الإحصائية".

فهي العلم الذي يدرس ويدمج بين التكنولوجيات المستخدمة في الحصول على البيانات البيولوجية واستخدام الحاسوب لتخزين وترتيب وتحليل تلك البيانات وما يرتبط بذلك من تطبيقات مختلفة ، وهو مجال علمي حديث ومتنامٍ نتج من التداخل بين علم البيولوجي وعلم الحاسب وتكنولوجيا المعلومات لتدعيم تخزين وتنظيم واستعادة البيانات البيولوجية .

قبل ظهور علم المعلوماتية الحيوية كان هناك مصطلحان لوصف مكان إجراء التجارب البيولوجية هما :

١ -In vivo: أي إجراء التجارب على الكائن الحي بشكل مباشر.

٢ -In Vitro: أي إجراء التجارب داخل المعمل.

وبعد استخدام علم المعلوماتية الحيوية تم إدخال مصطلح (In silico) أي إجراء التجارب نظرياً على الحاسب قبل إجرائها على النظم الحيوية أو في المعامل.

4.3- مجالات المعلوماتية الحيوية [17] [18]:

نستعرض فيما يلي أهم المجلات التي تعتمد على مفهوم المعلوماتية الحيوية:

١ -**علم الوراثة:** هو العلم الذي يدرس المورثات (الجينات) وما ينتج عنه من تنوع الكائنات الحية، وكانت مبادئ توريث الصفات مستخدمة منذ تاريخ بعيد لتحسين المحصول الزراعي وتحسين النسل الحيواني عن طريق تزويج حيوانات من سلالة ذات صفات جيدة ، حيث تخزن المعلومات الوراثية بشكل عام ضمن سلاسل DNA في الصبغيات المشكلة لنواة الخلية.

وبزيادة الأبحاث في هذا العلم تزايدت الصعوبات وازدادت الحاجة إلى التقنيات التي توفرها المعلوماتية من سرعة تحليل ودقة تخزين تهدف إلى دراسة تغيير هذه المورثات ضمن ظروف مناخية وصحية متعددة.

٢ - **علم البروتيوم:** وهي الدراسة الشاملة لجميع أصناف وأنواع البروتينات، خاصة فيما يتعلق ببنية البروتين ووظائفه، حيث تشكل البروتينات الجزء الحيوي الأساسي من الكائنات الحية، كما تعتبر المكون الرئيسي للمسارات الاستقلابية الرئيسية في أي خلية.

٣ -**بيولوجيا النظم:** وهو الأكثر تعقيدا حيث يدرس دور تفاعلات سلاسل DNA مع البروتين وتأثيرها على وظيفة الخلايا والأنسجة والأعضاء ككل، فهو يهدف إلى المقاربة الرياضية للبيولوجيا حيث يتم عمل نماذج رياضية لمختلف الظواهر البيولوجية ومحاولة استعمالها في المحاكاة، وعادة ما تكون النماذج عبارة عن تفاعلات كيميائية مرتبط بعضها ببعض (شبكة تفاعلات).

5.3- أسباب تطور المعلوماتية الحيوية [19] :

تزايدت أهمية المعلوماتية الحيوية وتزايد عدد الباحثين في هذا المجال ويعود ذلك إلى عدة أسباب منها:

١ -زيادة المعلومات البيولوجية (والمخزنة ضمن مشروع الجينوم البشري).

٢ -الوصول إلى البيانات واستخدامها واستغلالها واستخلاص المفيد منها فعملية البحث في البيانات المتكررة الكبيرة التي يصل بعضها إلى آلاف الصفحات تفوق القدرة البشرية فضلا عن أنها مملة وتستغرق الوقت الطويل وتحتاج إلى عمليات رياضية معقدة.

٣ –النمو المتفجر لنظم المعلومات والاتصالات.

٤ –سهولة التخزين والبحث والنسخ حيث يمكن تقصي المعلومات السابقة في غضون ثوانٍ بالإضافة إلى سهولة تحديثها وإنشاء عدة نسخ منها.

6.3– تطبيقات المعلوماتية الحيوية [20] :

1– العثور على الجينات:

عند وجود سلسلة من سلاسل DNA ونرغب في معرفة مواقع الجينات على هذه السلسلة، وأيضاً في حال الحاجة إلى التنبؤ بتركيب الجين أي تحديد الأجزاء التي ترمز إلى البروتينات والتي تسمى exons وتحديد الأجزاء التي لا ترمز إلى البروتينات والتي تسمى introns، ومن أشهر البرامج في هذا المجال برنامج GRAIL الذي يعمل باستخدام الشبكات العصبية الاصطناعية Artificial Neural Networks.

2– محاذاة السلاسل:

لمعرفة ما إذا كانت السلسلة التي تم الحصول عليها حديثا مشابهة لسلسلة أو مجموعة سلاسل أخرى نعرفها مسبقاً فإذا وجد التشابه في السلاسل دل على وجود وظيفة مشتركة أو متشابهة.
ومن أشهر البرامج المستخدمة لمحاذاة السلاسل الثنائية باستخدام البرمجة الديناميكية NEEDLE.

3– تحديد مواقع معامل النسخ الملزمة (TFBS (Transcription Factor Bindnig Sites:

يعتبر معامل النسخ (Transcription Factor) TF العنصر الوظيفي الأكثر أهمية في سلاسل DNA، حيث تدعى مواقع توضع هذا المعامل ضمن تلك السلاسل بمواقع معامل النسخ الملزمة (TFBS (Transcription Factor Binding Sites، ويتم تحديد هذه المواقع باستخدام المحفزات Motifs والتي تعرف بأنها السلاسل الجزئية الأكثر تكراراً في سلاسل DNA.

٤ –مجال تصنيع الأدوية:

من خلال معرفة ارتباط البروتينات وتحديد التراكيب ثلاثية الأبعاد للبروتينات 3D-Structures وغيرها من المعلومات التي سهلت طرق تطوير الأدوية بكفاءة عالية وبأقل التأثيرات الجانبية.

٥ –مجال الطب:

وذلك من خلال التعرف السريع على الأمراض الوراثية وتطوير عقاقير متنوعة من خلال الحاسوب، وذلك من خلال تحديد تمثيل الأمراض افتراضيا ومعرفة تركيب المادة الوراثية المؤثرة فيها.

٦ –إيجاد العلاقات التطورية:

وذلك لفهم العلاقات الوراثية والتطورية بين الأحياء إذ تساعد المعلوماتية الحيوية في إيجاد الأبعاد الزمنية لتطور الأحياء بواسطة الرسوم وغيرها من الوسائل .

٧ –التنبؤ بتراكيب العناصر:

وتشمل التنبؤ بتراكيب DNA و RNA والبروتينات، حيث تتضمن هذه المهمة تحديد مواقع العناصر المتكررة، والمناطق غير المساهمة في بعض العمليات البيولوجية.

٨ –استرجاع البيانات:

توفر المعلوماتية الحيوية كماً هائلاً من البيانات المخزنة ضمن قواعد بيانات منظمة تسهل عملية استرجاع البيانات التي يتم استخدامها في مجال علم الأحياء.

7.3– أساسيات المعلوماتية الحيوية [20] :

تعتمد كفاءة تنفيذ مهام المعلوماتية الحيوية على ثلاثة محاور رئيسية:

أولاً: تنظيم البيانات بطريقة تسمح وتمكن الباحثين من الوصول إليها مع إمكانية التحديث والإضافة عليها، وتعد تهيئة البيانات من المهام الأساسية لفعالية تحليل البيانات.

ثانياً: تطوير الأدوات والمصادر التي تساعد في تحليل البيانات ، وخاصة حين يتعلق الموضوع بالتطابق البيولوجي بين البروتينات.

ثالثاً: استخدام الأدوات اللازمة لتفسير النتائج الخاصة بالجوانب البيولوجية وبأسلوب واضح ومفهوم .

8.3– خلاصة:

استعرضنا في هذا الفصل أساسيات علم المعلوماتية الحيوية ومجالاته وأهدافه وسبل تطبيقها.

الفصل الرابع

DNA سلاسل

DNA Sequences

1.4- مقدمة:

الحمض النووي الريبي منقوص الأكسجين (DNA DeoxyriboNucleic Acid) هو جزيء ضخم يوجد داخل كل خلية من خلايا الكائنات الحية ويحتوي على المعلومات الوراثية التي تسمح بعمل وتكاثر وتطور هذه الكائنات.

حيث يتكون هذا الجزيء من سلسلتين بيولوجيتين تلتف إحداهما على الأخرى على شكل لولب مزدوج، و تسمى سلسلة DNA الواحدة بالسلسة العديدة النيكليوتيدات لأنها مكونة من وحدات أساسية تسمى النيكليوتيدات والتي تعتبر الوحدة الأساسية في سلاسل DNA، حيث تعتبر هذه النيكليوتيدات بمثابة الحروف الأساسية التي تكتب بها الجينات أو المورثات التي تنقل أوصاف الطفل من الأم والأب [21].

2.4- لمحة تاريخية [22] [23] [24]:

حازت أسباب التشابهات الوراثية بين الأجيال المتعاقبة من نفس السلالة على اهتمام العلماء منذ الأزل، حتى تم التوصل إلى الحامض النووي المعروف باسم DNA والذي اكتشفه العلماء بعد العديد من الأبحاث العلمية، ونستعرض فيما يلي أهم المحاولات الرامية لاكتشاف مكان حفظ كل المعلومات البيولوجية الخاصة بالكائن الحي.

بدأت قصة اكتشاف الحمض النووي DNA في القرن التاسع عشر، حيث تم التعرف على الجزيء المعروف الآن باسم DNA لأول مرة عام 1869م من قبل العالم فريدريك ميسشر.

حيث بحث ميسشر عن المكونات الرئيسية لخلايا الدم البيضاء، وهي جزء من الجهاز المناعي لأجسامنا، فأجرى تجارب باستخدام محاليل الملح لفهم المزيد من مكونات خلايا الدم البيضاء.

وقد لاحظ أنه عندما أضاف حمضاً إلى محلول يضم خلايا فإن هناك مادة يتم فصلها عن المحلول، حيث تذوب هذه المادة عند إضافة القلويات لها، وعند فحص هذه المادة وجد أن لها خصائص غير متوقعة تختلف عن تلك الموجودة في البروتينات الأخرى.

وصف ميسشر هذه المادة الغامضة بالنكليوين "nuclein" لأنه يعتقد أنها جاءت من النواة، ومن دون علمه، اكتشف ميسشر الأساس الجزيئي لكل أنواع الحمض النووي، ثم بدأ في البحث عن طرق لاستخراجها بشكلها النقي.

كان ميسشر مقتنعًا بأهمية النكليوين واقترب كثيراً من الكشف عن دورها، وذلك على الرغم من الأدوات والأساليب البسيطة المتاحة له والتي لم تسانده في كشف ما يريد.

وبعد سنوات طويلة تفشى مرض التهاب الرئة في لندن عام 1928م، وكان العالم فريدريك غريفث من الباحثين عن كيفية تسبب بكتيريا الالتهاب الرئوي لذلك المرض محاولاً اكتشاف علاج مناسب لذلك.

حيث درس غريفث تأثير سلالتين من البكتيريا S و R المسببة للالتهاب الرئوي على الفئران،

ويوضح الجدول التالي الفرق بين هاتين السلالتين:

الجدول 1.4 الفرق بين أنواع البكتيريا المستخدمة في تجربة غريفث[23]

بكتيريا S	بكتيريا R
مميتة	غير مميتة
ذات حواف ملساء	ذات حواف خشنة
لديها محفظة من عديد التسكر تقيها من الأجهزة الدفاعية	لا يوجد لديها محفظة

فقد حقن مجموعة من الفئران بهذه البكتيريا فلاحظ ما يلي:

١ عند حقن الفئران ببكتيريا من السلالة S ⟸ ماتت الفئران.
٢ عند حقن الفئران ببكتيريا من السلالة R ⟸ لم تمت الفئران.
٣ عند حقن الفئران ببكتيريا من السلالة S والتي تم قتلها بالحرارة ⟸ لم تمت الفئران.
٤ عند حقن الفئران بسلالة البكتيريا S والتي تم قتلها بالحرارة ممزوجة ببكتيريا من السلالة R ⟸ ماتت الفئران!!!

فإن كانت نتيجة التجربة الثالثة متوقعة وذلك لكون الحرارة قد سببت تلف المحفظة العديدة التسكر والتي تحمي هذه البكتيريا من الجهاز المناعي للفئران مما أزال تأثير هذه البكتيريا المميتة، فإن المفاجأة غير المتوقعة كانت في التجربة الرابعة فما كان للبكتيريا من السلالة R أن تقتل الفئران ولا تستطيع البكتيريا من السلالة S العديمة التأثير بفعل الحرارة من أن تقتل الفئران، فكيف ماتت الفئران؟ وهل عادت البكتيريا من السلالة S إلى الحياة واستعادت تأثيرها القاتل؟؟

وبعد الدراسة تبين للعالم غريفث أن مادة وراثية من السلالة S (تبين فيما بعد أنها سلاسل DNA) انتقلت إلى داخل البكتيريا من السلالة R فحولتها إلى بكتيريا ذات تأثير قاتل، وقد أطلق على هذه الظاهرة اسم التحول البكتيري، وقد أطلق على هذه المادة الوراثية اسم مادة التحول البيكتيري.

وكان من المنطقي أن تكون أول خطوة بعد اكتشاف مادة التحول البكتيري محاول العلماء عزل هذه المادة لتحليها ومعرفة بنيتها.

ففي عام 1944م قرر العالم أوزوالد آفري وزملاؤه العالم كولين ماكلويد والعالم ماكلين ماكرتي محاولة التعرف على مادة التحول البكتيري وتحليلها.

واستنتجوا أن سلاسل DNA قد انتقلت من البكتيريا من السلالة S ورغم قتلها حرارياً فقد انتقلت إلى البكتيريا من السلالة R، وقد امتصتها هذه البكتيريا بطريقة غير معروفة واندمجت مع سلاسل DNA الخاصة بها فاكتسبت خصائصها وهذه الخصائص ستنتقل إلى الأبناء (أي أن كل البكتيريا التي ستتكاثر من هذه البكتيريا ستكون مميتة أيضاً).

وهنا تم الاعتراض على هذه النتيجة من قبل الباحثين في علم الأحياء آنذاك، واستند هذا الاعتراض إلى أن مادة DNA لم تكن نقية تماما لأنها تحمل كمية من البروتين (المختلط به والمجاور له ضمن الخلية) وعليه يحتمل أن تكون تلك الكمية من البروتين هي التي تسببت في إحداث التحول البكتيري.

ولتأكيد فرضية افري قام زملاؤه بتعديل التجربة الرابعة من تجارب غريفث، حيث تم عزل مادة التحول البكتيري النشطة المنتقلة (DNA + بروتينات) وذلك عن طريق انزيمات محللة لكل مادة على حدة.

وتتالت المحاولات وتراكمت الأبحاث والفرضيات حتى تم التوصل إلى توصيف سلاسل DNA بشكلها الدقيق ومكوناتها البيولوجية من قبل العالمين جيمس واطسون و فرنسيس كريك عام 1965م فحازا على جائزة نوبل للسلام لكشفهما إحدى أعظم الاكتشافات في التاريخ والتي عرفت البشرية بكيفية نقل المعلومات الوراثية من جيل إلى جيل.

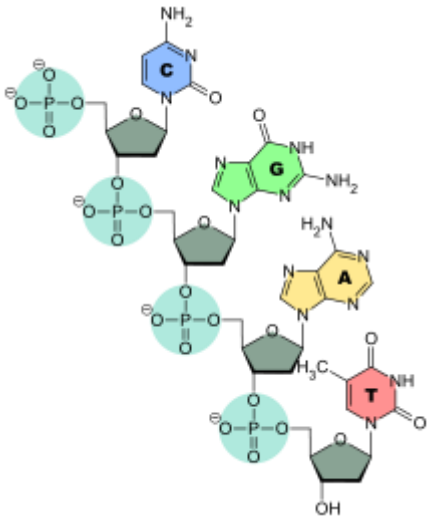
3.4- تكوين سلاسل DNA [25] [26] [27]:

تتكون سلاسل DNA من جزيء طويل يحتوي على الشيفرة الوراثية الفريدة لكل شخص، والتي تحمل تعليمات بناء البروتينات الضرورية لعمل الأجسام، وتنتقل تعليمات الحمض النووي DNA من الوالدين إلى الأبناء، حيث يحصل الابن على نصف الحمض النووي لديه من الأب، ويحصل على النصف الثاني منه من الأم.

وتعرف سلاسل DNA بأنها جزيء رقيق طويل يتكون من النيكليوتيدات (Nucleotides)، حيث توجد أربعة أنواع مختلفة من النيكليوتيدات التي تشكل الأبجدية الوراثية :

١ -أدينين (Adenine) A.
٢ -ثيمين (Thymine) T.
٣ -سيتوزين (Cytosine) C.
٤ -غوانين (Guanine) G.

وترتبط هذه النيكليوتيدات بعمود فقري يتكون من فوسفات، سكر خماسي الكربون مع قواعد نتروجينية، وذلك كما يوضح الشكل 1.4



الشكل 1.4 ارتباط النيكليوتيدات

و على الرغم من وجود أربعة حروف مختلفة فقط، فإن سلاسل DNA تتكون من ملايين الحروف مما يسمح بحدوث مليارات من التركيبات المختلفة.

فإذا تم حل أو تفكيك جميع جزيئات الحمض النووي في الجسم و قم نا بمدها بشكل مستقيم، فإنها ستمتد إلى الشمس و تعود عدة مرات.

يحتوي جسم الإنسان على 210 نوع مختلف من الخلايا، حيث توجد (خلايا الدم) و (خلايا العظام) و (الخلايا التي تصنع عضلاتنا).....

حيث تقوم كل خلية بعمل وظيفي مختلف، وتحصل الخلايا على التعليمات المطلوبة للقيام بوظائفها من سلاسل DNA، التي نستطيع تشبيهها ببرنامج موجود على جهاز الحاسوب (الخلية) يقوم بإعطاء التعليمات لهذا الحاسوب واللازمة للقيام بوظائفه .

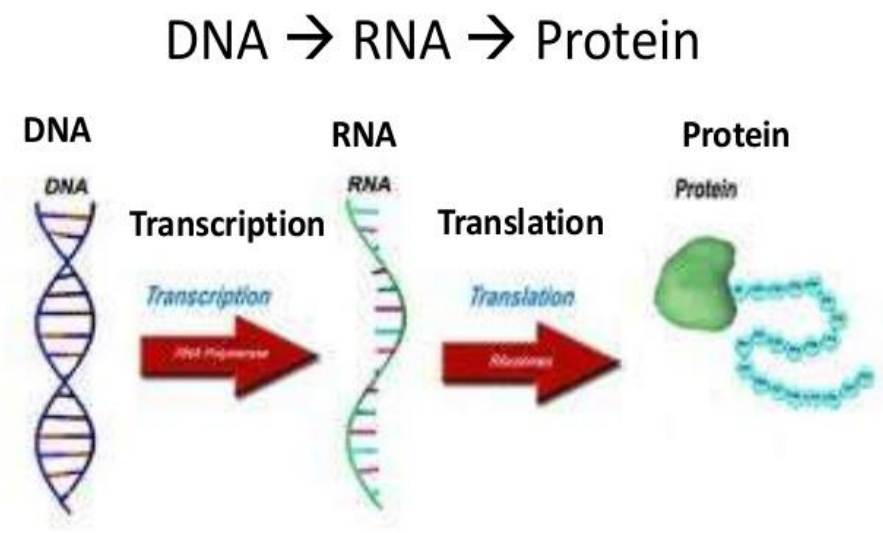
وتجدر الإشارة إلى أن حوالي 99,9% من سلاسل DNA العائدة لكل شخص على هذا الكوكب متشابهة تماماً، و فقط 0.1 % الذي يختلف من كل شخص لآخر و هذا ما يجعلنا كل واحد منا يختلف عن الآخر!!!

4.4- عملية إنتاج البروتين [28]:

تعرف البروتينات بأنها جزيئات كيميائية معقدة تعد من أهم الجزيئات التي تحويها الكائنات الحية ومن أكثرها تنوعا، فأنواعها تعد بالآلاف وتختلف باختلاف وظائفها، وهي بمنزلة جنود مسخرة لخدمة الخلية الحية وتقوم بدور أساسي في جميع التفاعلات الحيوية التي تجري في أجسامنا .

تتكون البروتينات من وحدات تسمى "الحموض الأمينية" وعدد الحموض الأمينية الموجودة في الطبيعة كثيرة جدا، ولكن 20 منها فقط تدخل في تركيب البروتينات، ويمكن تشبيه العشرين بمثابة الحروف في الأبجدية، عددها محدود ولكن يمكن بواسطتها تكوين ما نشاء من جمل وعبارات وكلمات، على أن يرتبط بعضها ببعض وفق قواعد اللغة .

يتم تشكيل البروتينات والسلاسل الببتيدية من المعلومات الوراثية الموجودة في سلاسل DNA وذلك كما هو موضح في الشكل 2.4



الشكل 2.4 مراحل تشكيل البروتينات

حيث تتألف عملية تشكيل البروتينات من العمليتين التاليتين:

الأولى : عملية النسخ Transcription:

هي عملية النسخ الأنزيمية التي يقوم بها أنزيم RNA polymerase لتحويل سلسلة DNA (لمورثة ما) إلى سلسلة الحمض النووي الريبي (RiboNucleic acid) RNA الموافقة المتممة وبهذه العملية يتم نقل المعلومات الوراثية من سلاسل DNA إلى سلاسل RNA، حيث توضح هذه العملية تدفق المعلومات الوراثية داخل الكائنات الحية، ويعرف هذا التدفق أو المسار بالعقيدة المركزية.

الثانية : عملية الترجمة Transcription:

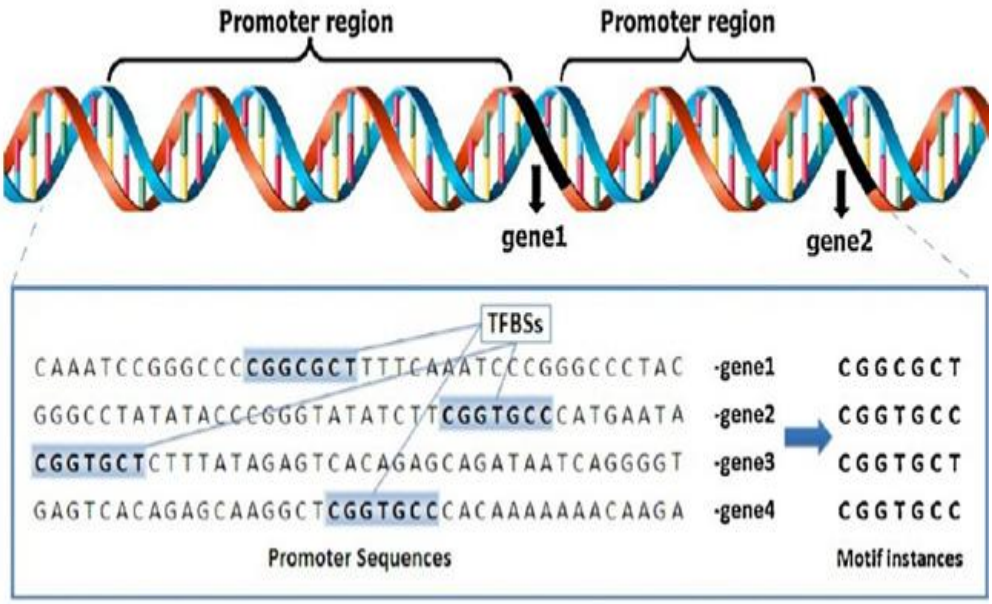
في هذه المرحلة تتم عملية الاصطناع الحيوي للبروتين (والتي تشكل جزءا من عملية التعبير الجيني)، ففي هذه المرحلة يتم فك تشفير المعلومات الواردة عن طريق سلسلة RNA المرسال لإنتاج السلسلة الببتيدية المطلوبة حسب قواعد الشفرة الجينية genetic code (و هي التي تحدد الأحماض الأمينية العشرين التي تدخل في تركيب البروتينات وذلك مقابل كل ثلاثية نيكليوتيدية موجودة في سلسلة DNA أو سلسلة RNA).

5.4 – معامل النسخ (المحفز) [29] [30]:

تعتمد عملية النسخ الوراثي بشكل أساسي على وجود المحفزات Motifs (مجموعة جزئية من سلاسل DNA)، حيث يعرف المحفز بمعامل النسخ (Transcriptin Factor) TF، و تدعى مواقع تموضع هذا المعامل ضمن هذه السلاسل بمواقع معامل النسخ الملزمة TFBS (Transcription Factor Binding Sites).

ويعتبر تحديد هذا المعامل من أهم المعاملات اللازمة لتحديد أساسيات تحليل الشيفرات الوراثية وتحديد الأمراض والمورثات مما يساهم في علاجها.

تتكرر المحفزات في سلاسل DNA وذلك كما يوضح الشكل 3.4

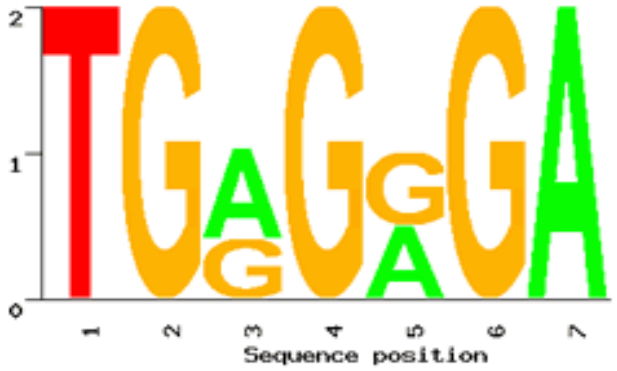


الشكل 3.4 تموضع معامل النسخ ضمن سلاسل DNA

تعد المحفزات السلاسل الجزئية الأكثر تكرارا ضمن سلاسل DNA، ويكون لهذه المحفزات طول L بمقدار اختلاف d حيث:

$$0<=d< L$$

حيث يمكن لمحفز واحد أن يظهر بعدة أشكال مختلفة بمقدار d كما يوضح الشكل 4.4



الشكل 4.4 ظهور المحفز ضمن سلاسل DNA بأشكال متعددة

6.4- خلاصة:

ألقينا الضوء في هذا الفصل على سلاسل DNA مع أهم المعاملات التي تحتويها، وسنستعرض في الفصول القادمة طرق تحليلها...

الفصل الخامس

دراسة خوارزميات تحليل سلاسل DNA

1.5- مقدمة:

تنوعت الطرق والخوارزميات الهادفة إلى تحديد المحفزات الموجودة ضمن سلاسل DNA، وتعتبر الخوارزميات التالية من أشهر هذه الخوارزميات:

١ - 1990 Expectation maximization [31].
٢ - 1992 WordUP [32].
٣ - 1993 Gibbs sampler [33].
٤ - 1995 MEME [34].
٥ - 1999 Consensus [35].
٦ - 2000 SMILE [36].
٧ - 2002 MITRA [37].
٨ - 2007 PhyloScan [38].
٩ - 2015 MCES [39].
١٠ -2018 MDWB [40].

سوف نقوم في هذا الفصل بدراسة وتحليل أحدث خوارزميتين وهما MCES, MDWB وذلك لكونهما تعتمدان على الخوارزميات السابقة وتنتهجان المفهوم نفسه.

2.5- خوارزمية MCES:

في عام 2015 قام الباحثان كيانغ يو وهونغوي هو بتطوير خوارزمية

(Mining and Combining Emerging Substrings) MCES حيث تتكون خوارزمية

MCES من الخطوات الآتية:

الخطوة الأساسية : الخوارزمية الأساسية MCES Step

نحتاج بداية إلى نوعين من مجموعات البيانات:

١ -مجموعة الاختبار Test Set: تحوي مجموعات السلاسل التي تحوي المحفزات.
٢ -مجموعة التحكم Control Set: تحوي مجموعات السلاسل التي لا تحوي المحفزات (وتستخدم كأساس للمقارنة).

فإذا كانت مجموعة الاختبار Dt و مجموعة التحكم Dc صغيرتين يتم تحديد المحفزات على آلة واحدة وذلك بحساب تردد السلاسل الجزئية المشكلة لهما، ومن ثم تحديد السلاسل التي يتجاوز تكرارها تردد العتبة threshold frequency Pf ومن ثم تحديد مصفوفة أوزان المواقع PWM(Position Weight Matrix) لتحديد تشابه السلاسل.

أما في حال كانت هاتان المجموعتان كبيرتين فيتم الانتقال إلى الخطوات الفرعية والتي يتم تنفيذها على عدة آلات على التوازي وذلك بتطبيق تقنية MapReduce مع احتمال إعادة تكرارها حتى الوصول إلى المطلوب، حيث نستخدم هنا المعامل z' الذي يعبر عن إمكانيات الجهاز الذي نعمل عليه فنقارنه بحجم البيانات المراد معالجتها لتحديد فيما إذا كان الجهاز كافي للتنفيذ أم أننا بحاجة لتطبيق تقنية MapReduce على عدة أجهزة وذلك كما يلي:

Input: a test set D_t and a control set D_c of DNA sequences
Output: the set of motifs M
1: set mining parameters
2: **if** $\|D_t\| + \|D_c\| \leq z'$ **then**
3: perform mining step in a single machine
4: **else**
5: perform mining step distributedly using MapReduce
6: perform combining step using Algorithm 1 and get M
7: **if** $M = \Phi$ **then**
8: $\rho_f \leftarrow \rho_f/2$
9: perform mining step and combining step again
10: **return** M

وفي حال عدم وجود محفزات ناتجة يتم تغيير تردد العتبة Pf وذلك بقسمته على اثنان (علما أن Pf يتم حسابه من خلال بعض القوانين الرياضية المعتمدة على علم الاحتمالات) ومن ثم تكرار الخطوات الفرعية، ونبين فيما يلي الخطوات الفرعية لهذه الخوارزمية:

الخطوة الأولى : مرحلة التمثيل Map Step

حيث نقوم بتقسيم مجموعة الاختبار Dt ومجموعة التحكم Dc إلى مجموعات جزئية Di وبطول موحد z، حيث نتعامل في هذه المرحلة مع كل مجموعة جزئية Di وذلك للحصول على متحول خاص بهذه

السلسلة Ti والذي يعبر عن الثلاثية <Q,a,b>،مع العلم بأن Q سلسلة جزئية من Di و a,b أعداد وذلك وفق ما يلي:

Input: a data block D_i
Output: the set of 3-tuples T_i
1: $T_i \leftarrow \Phi$
2: **for** each substring φ **do**
3: get its occurrence count $count(\varphi, D_i)$
4: $freq(\varphi, D_i) \leftarrow count(\varphi, D_i)/|D_i|$
5: **if** $l_{\min} \leq |\varphi| \leq l_{\max}$ and $freq(\varphi, D_i) > \lambda\rho_f$ **then**
6: **if** D_i is from D_t **then**
7: add $< \varphi, count(\varphi, D_i), 0 >$ to T_i
8: **else**
9: add $< \varphi, 0, count(\varphi, D_i) >$ to T_i
10: **return** T_i

نلاحظ بأن Ti يحدد لكل سلسلة جزئية Q من المجموعة Di تكرارها ضمن هذه المجموعة، ويتم إسناد هذه القيمة إلى a إذا كانت Di مجموعة جزئية من Dt، أو إسنادها إلى b إذا كانت Di مجموعة جزئية من Dc.

الخطوة الثانية : مرحلة التخفيض Reduce Step

حيث نقوم في هذه المرحلة بمعالجة كل المتحولات Ti التي تم تشكيلها في الخطوة الأولى وذلك للحصول على مجموعة من السلاسل الجزئية Ses والتي تحقق مجموعة من الشروط كما يلي:

Input: all sets of 3-tuples $T_i(1 \leq i \leq n_{block})//n_{block}$ is the total number of D_i
Output: the set of emerging substrings S_{es}
1: $S_{es} \leftarrow \Phi$
2: **for** $i \leftarrow 2$ to n_{block} **do**
3: merge T_i to T_1
4: **for** each 3-tuple $< \varphi, \alpha, \beta >$ in T_1 **do**
5: $freq(\varphi, D_t) \leftarrow \alpha/|D_t|$
6: $freq(\varphi, D_c) \leftarrow \beta/|D_c|$
7: $growth(\varphi, D_t, D_c) \leftarrow freq(\varphi, D_t)/freq(\varphi, D_c)$
8: **if** $freq(\varphi, D_t) > \rho_f$ and $growth(\varphi, D_t, D_c) > \rho_g$ **then**
9: add φ to S_{es}
10: **return** S_{es}

حيث تعتمد هذه الخطوة على سلسلة من القوانين الرياضية التي تحدد تكرار السلاسل الجزئية التي يتم ترشيحها لتكون محفزات.

الخطوة الثالثة : دمج السلاسل المنبثقة Combine Emerging Substrings

حيث نقوم بتحديد المحفزات وذلك انطلاقا من السلاسل الجزئية Ses التي تم توليدها في المرحلة الثانية، وذلك بحساب مصفوفة أوزان المواقع PWM لتحديد السلاسل المتشابهة وفق ما يلي:

Input: the set of emerging substrings S_{es}
Output: the set of motifs M
1: $S_{ds} \leftarrow \Phi$ // the set of dispersive substrings

2: $S_{pwm} \leftarrow \Phi$ // the set of PWMs
3: $M \leftarrow \Phi$ // the set of motifs
4: **for** $i \leftarrow l_{\min}$ to $l_{\max}$ **do**
5: cluster the emerging substrings in S_{es} of length i
6: **for** each obtained cluster of emerging substrings C_{es} **do**
7: **if** $|C_{es}| > 5$ **then**
8: align the substrings in C_{es} and get a PWM w
9: add w to S_{pwm}
10: **else**
11: add the substrings in C_{es} to S_{ds}
12: cluster the substrings in S_{ds} and add obtained PWMs to S_{pwm}
13: cluster the PWMs in S_{pwm}
14: **for** each obtained cluster of PWMs C_{pwm} **do**
15: align the PWMs in C_{pwm} and get a combined PWM w
16: fetch the segment of w with high information content to obtain a motif m
17: add m to M

18: **return** M

3.5- خوارزمية MDWB:

في عام 2018 قام الباحثان محمد ديفان مسعود و مانجيولا بتطوير خوارزمية MCES وذلك بابتكار خوارزمية (Motif Discover Word-based Algorithm) MDWB، والتي تعتمد بشكل أساس على القوانين الرياضية المستخدمة في خوارزمية MCES مع اختزال بعض الخطوات، حيث تتألف خوارزمية MDWB من الخطوتين الآتيتين:

الخطوة الأولى

ويتم فيها تحديد السلاسل الجزئية الموجودة ضمن مجموعة الاختبار Dt ومجموعة التحكم Dc على أن تحقق مجموعة من الشروط كما يلي:

Input: test (Dt) and control (Dc) set dataset.

Output: Emerging Substring EMr

EMr← Ω

for i←Dt& Dc

merge i and EMr

freq (Ω, Dt) ← α/| Dt|

freq (Ω, Dc) ← β/| Dc|

growth (Ω, Dc, Dt) ← α/| Dt| / β/| Dc|

if Freq (Ω, Dt)>pf and growth (Ω, Dc, Dt)>pg then

add Ω to EMr

return EMr

الخطوة الثانية

ويتم فيها تحديد المحفزات كما يلي:

Algorithm 2: Word-Based Algorithm

Input: Emerging Substring of DNA sequences
Output: the set motif M
Each string set parameter
if ||Dt|| + ||Dc||<= z' then
mining step combine into each string
else
if M = Ω then
pf ←pf/2
return M

وتعد خوارزمية MDWB أحدث الخوارزميات في مجال اكتشاف المحفزات الموجودة ضمن سلاسل DNA.

4.5- خلاصة:

استعرضنا في هذا الفصل أحدث خوارزميتين في مجال الكشف عن المحفزات،وبعد تحليل ودراسة هاتين الخوارزميتين، فإننا نجد أنهما يتصفان بما يلي:

١ -أن التعقيد الزمني الناجم عن العمليات الرياضية المعقدة التي يتم تنفيذها يبلغ $O(||Dt||+||Dc||+n^4)$، n : عدد سلاسل DNA، ||Dt|| : عدد سلاسل مجموعة الاختبار، ||Dc|| : عدد سلاسل مجموعة التحكم

٢ -تحديد القيم المتعلقة بكل سلسلة جزئية يتم في مراحل مختلفة وبالتي يوجد أكثر من وصول إلى كل نيكليوتيد موجود ضمن سلاسل DNA.

٣ -الاعتماد على علم الاحتمالات، وبالتالي وجود احتمال للخطأ.

٤ -معالجة مجموعة التحكم Dc والتي لا تحوي محفزات.

٥ -تزداد الصعوبة والتعقيد بزيادة d.

وعليه فقد قمنا بتطوير خوارزمية تقوم بتلافي المشاكل السابقة كما سنوضح في الفصل القادم.

الفصل السادس

الخوارزمية المقترحة

تقوم الخوارزمية التي قمنا بتطويرها على المفاهيم الآتية:

١) الرتل.

٢) الأشجار: ففي هذه الخوارزمية سنقوم بإنشاء شجرة من الطول L، تتألف عقدها من رموز الأبجدية الوراثية { A, G, C ,T }، حيث تمثل هذه الشجرة التمثيل البياني للمحفزات الموجودة ضمن مجموعة من سلاسل DNA.

خطوات الخوارزمية

١ –نقوم بتهيئة شجرة المحفزات M (بداية نبني الجذر فقط).

٢ –نقوم ببناء رتل من الطول L+d (طول المحفز : L ، مقدار الاختلاف : d).

٣ –نقوم بتهيئة مجموعة أعداد أولية مقترنة بظهور النيكليوتيد ضمن إحدى خانات الرتل، ونؤكد ضرورة أن يكون العدد أولياً وذلك لكون حاصل جداء الأعداد الأولية عدداً مميزاً لا يمكن الحصول عليه إلا من جداء هذه الأعداد تحديداً (عدد عناصر هذه المجموعة 4*L).

٤ –نقوم بتحديد مجموعة مؤلفة من d عدد أولي، ويكون كل عدد من هذه المجموعة أصغر من كل الأعداد الأولية السابقة، ونقوم بتخزين هذه الأعداد في الخانات الأخيرة من الرتل.

٥ –عند المرور على كل نيكليوتيد نقوم بتخزين العدد الأولي المقابل للوصول إلى هذا النيكليوتيد وذلك ضمن الرتل.

٦ –عند امتلاء الرتل نقوم بحساب جداء الأعداد المخزنة ضمن الرتل وليكن W.

٧ –نبني الشجرة M من خلال سلسلة DNA الأولى وذلك من خلال تمرير العدد W على أبناء كل عقدة (ابتداء من الجذر) وذلك بحساب القاسم المشترك الأكبر gcd بين العدد W و وزن العقدة وفي حال كان gcd>1 يتم المرور بهذه العقدة إلى أبنائها مع تغيير قيمة W بحذف gcd أي W = W/gcd.

٨ –نكرر الخطوة السابقة حتى الوصول إلى نهاية الشجرة وفي حال الوصول إلى إحدى الأوراق نزيد وزنها بمقدار واحد.

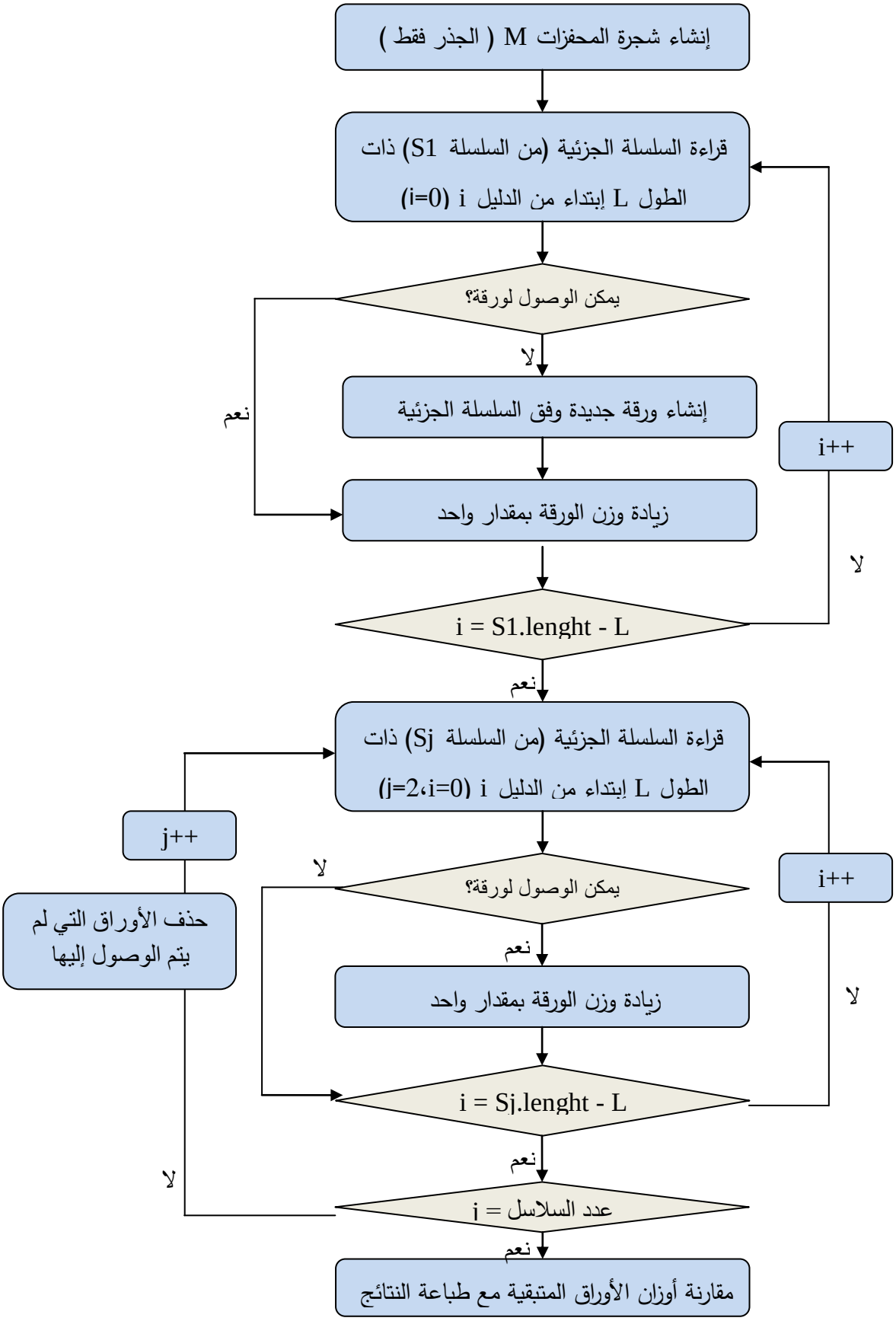
٩ –في حال كان عدد أبناء العقدة أقل من 4 (عدد حروف الأبجدية الوراثية) ننشئ إبناً لهذه العقدة في حال إمكانية المرور منها.

١٠ –يتم بناء الشجرة M وفق السلسلة الأولى مع تحديد تردد الأوراق.

١١ –نقوم بتمرير كل السلاسل الجزئية من الطول L والموجودة ضمن سلاسل DNA المتبقية على الشجرة M مع حذف الأوراق (مع آبائها التي لا تملك إلا إبناً وحيداً) التي لم يتم الوصول إليها في هذه السلسلة.

ملاحظة: في حال كان d=0 فيمكن الاستغناء عن الرتل ومجموعة الأعداد الأولية حيث تتم المقارنة بين النيكليوتيد المقروء و النيكليوتيد المخزن ضمن العقدة عن طريق المطابقة وليس عن طريق القاسم المشترك الأكبر.

ونبين في الشكل 1.6 المخطط التدفقي لهذه الخوارزمية.



الشكل 1.6 المخطط التدفقي للخوارزمية المقترحة

وسنستعرض فيما يلي بعض الأمثلة التي توضح خطوات تنفيذ الخوارزمية

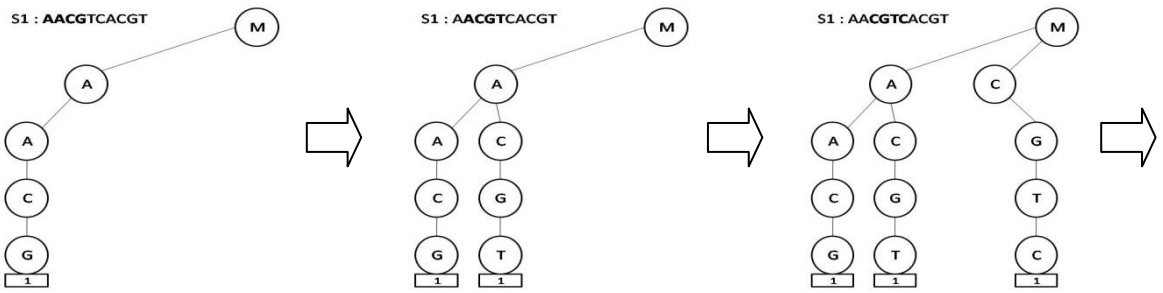
مثال 1:

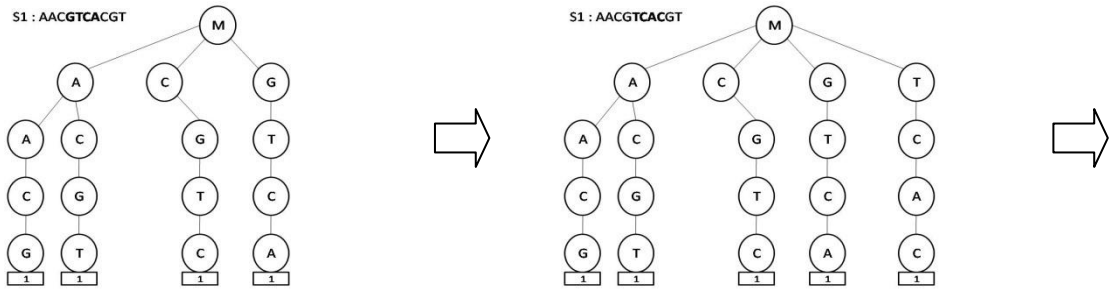
S1 = AACGTCACGT ; **L=4, d=0**

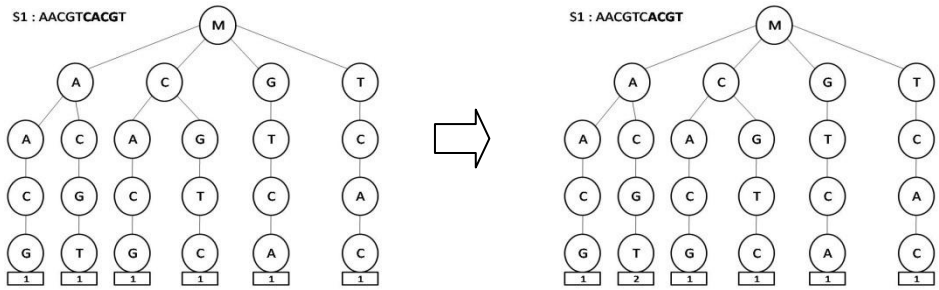
S2 = GCACGTCGCA

S3 = TCAGTCACGT

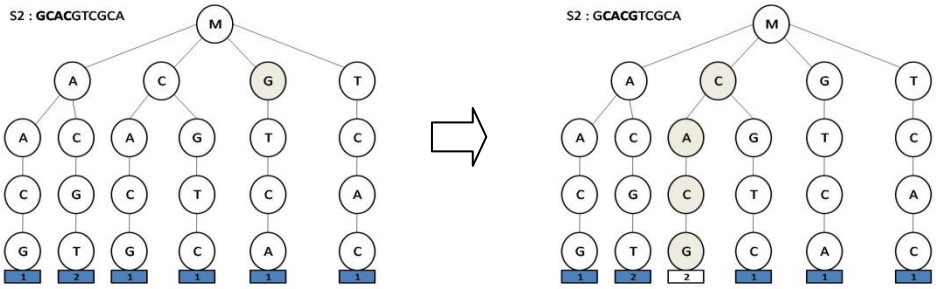
بداية نقوم بإنشاء شجرة المحفزات M من السلسلة S1 بقراءة كل سلسلة جزئية من الطول L كما يلي:



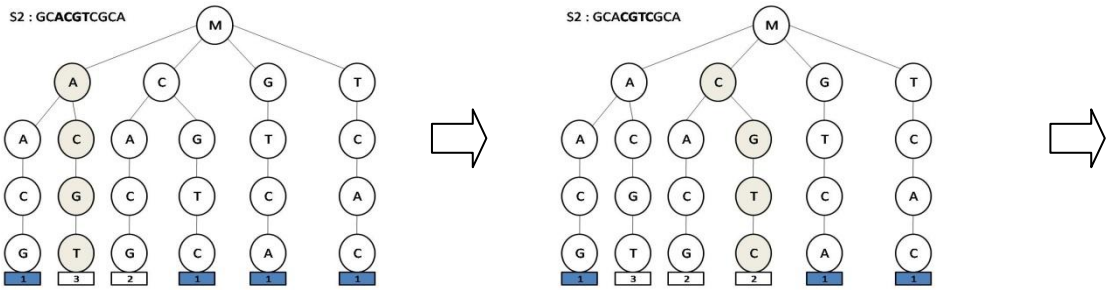


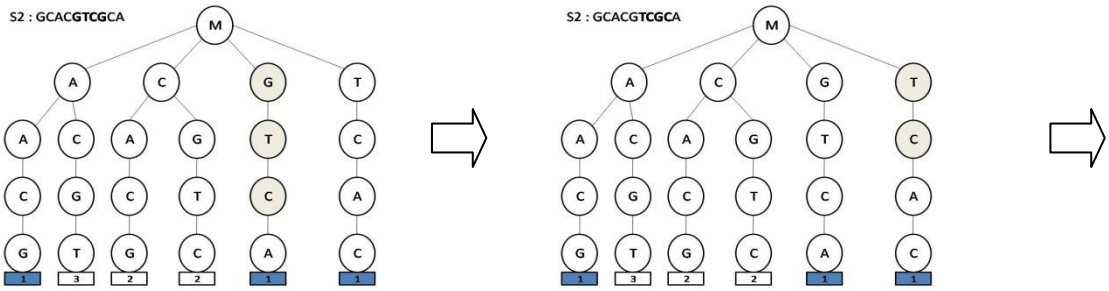


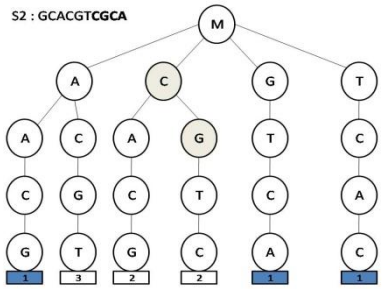
نقوم الآن بقراءة السلسلة S2 وفق ما يلي:



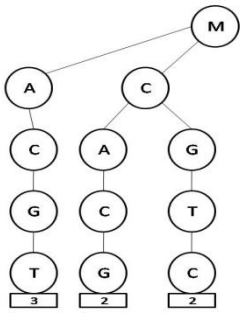
نلاحظ أن السلسلة الجزئية الأولى GCAC لا يمكنها المرور إلا بالعقدة G ولم تصل إلى الأوراق وبالتالي لم تحدث أي تأثير، في حين أن السلسلة الجزئية الثانية CACG مرت من الجذر إلى الورقة G عبر العقد المطابقة لتسلسلها، نكمل قراءة بقية السلاسل كما يلي:



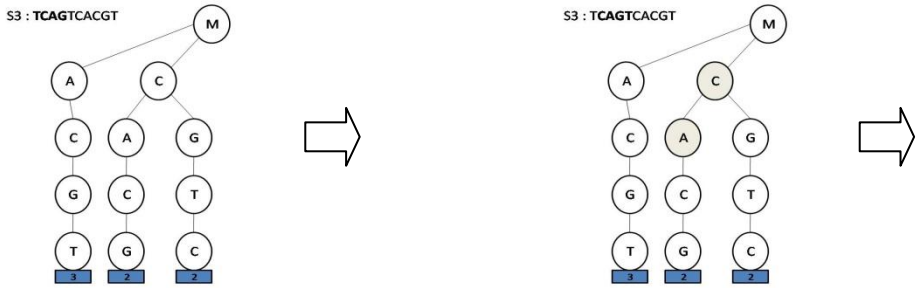


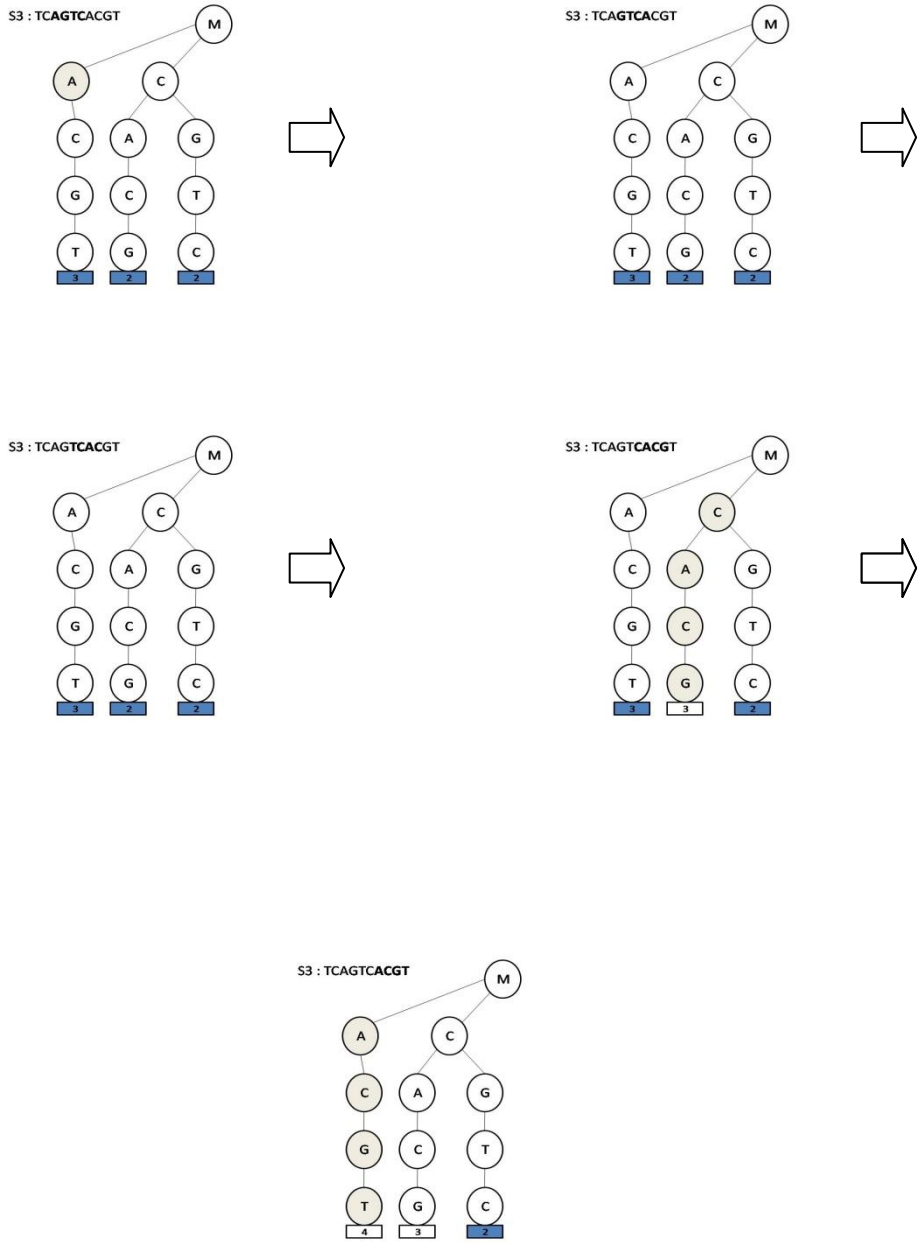


نلاحظ أن ثلاث أوراق فقط تم الوصول وفق السلسلة S2 وتم تعديل قيم أوزانها، وبالتالي نقوم بحذف الأوراق التي لم يتم الوصل إليها مع آبائها التي لا تملك إلا إبناً وحيداً لتصبح شجرة المحفزات كما يلي:



نقوم الآن بتمرير السلاسل الجزئية للسلسلة S3 على شجرة المحفزات فنحصل على ما يلي:





نلاحظ بعد حذف الورقة التي لم يتم الوصول إليها أن ورقتين فقط تم الوصول إليهما وفق السلسلة S3، وبالتالي إن السلسلتين الجزئيتين المتكررتين في كل السلاسل هما CACG و ACGT وبما أن وزن ACGT أعلى فهو المحفز.

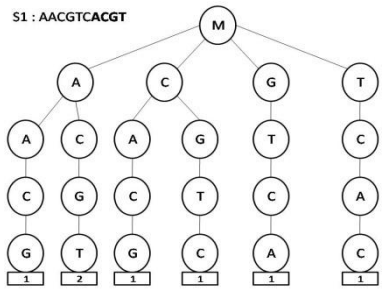
مثال 2:

S1 = AACGTCACGT ; **L=4, d=1**

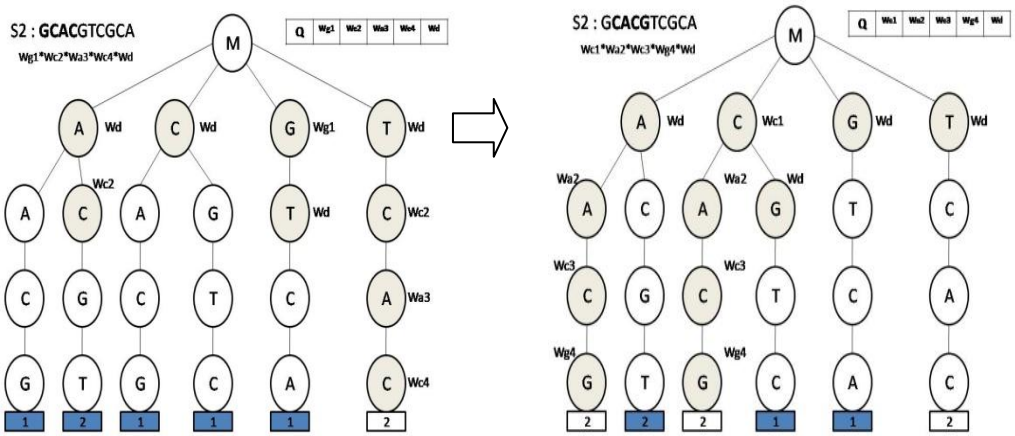
S2 = GCACGTCGCA

S3 = TCAGTCACGT

بداية نقوم بإنشاء شجرة المحفزات M من السلسلة S1 كما سبق في المثال الأول فنحصل على الشجرة التالية:

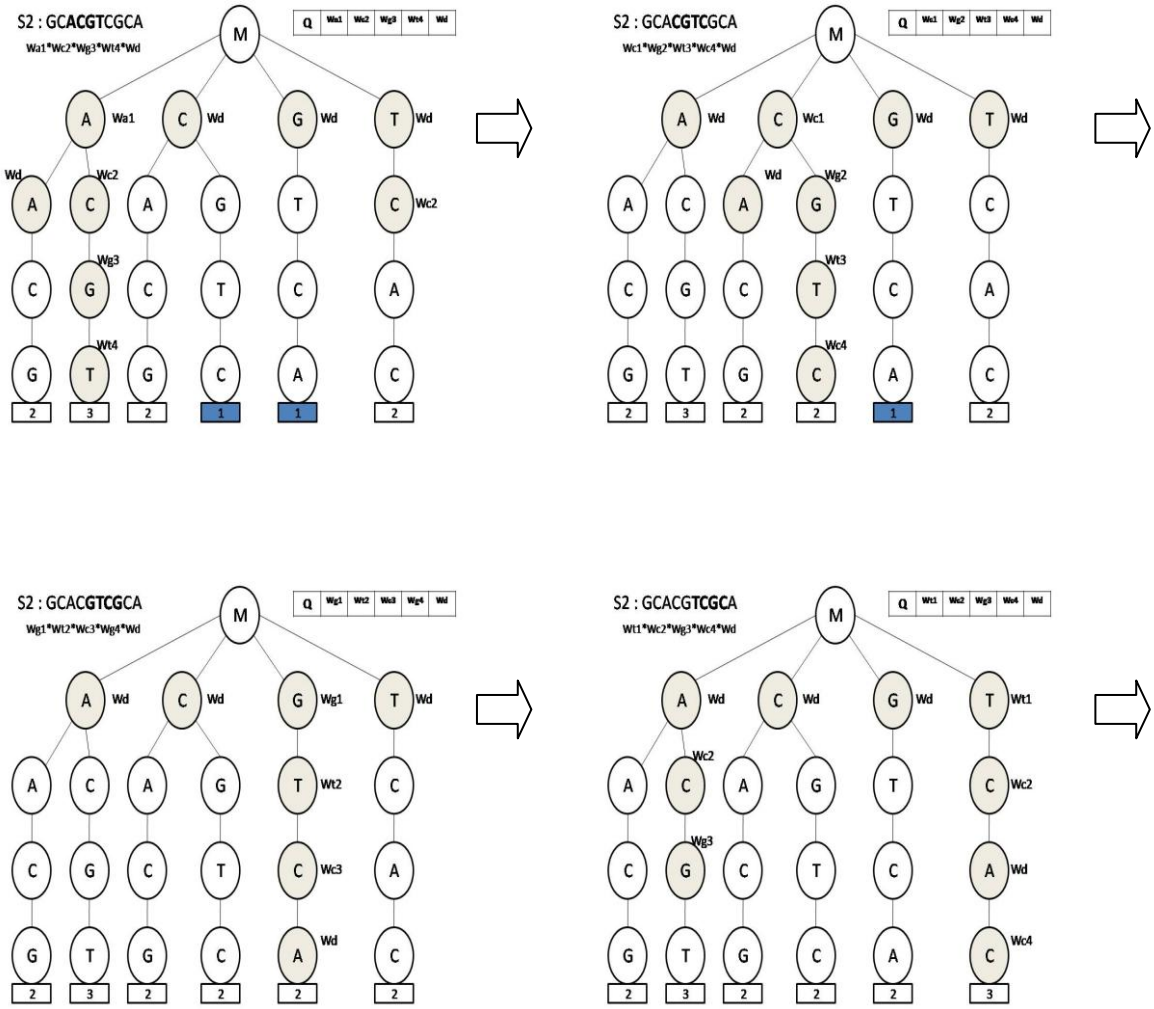


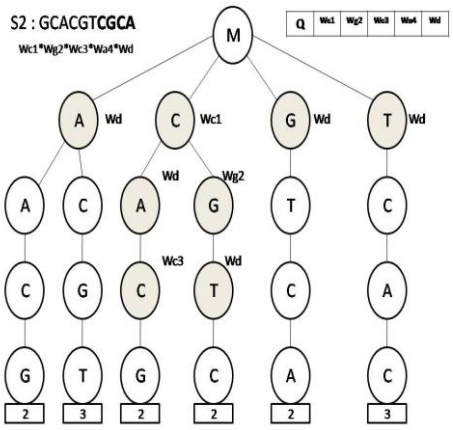
نقوم الآن بقراءة السلسلة S2 وفق ما يلي:



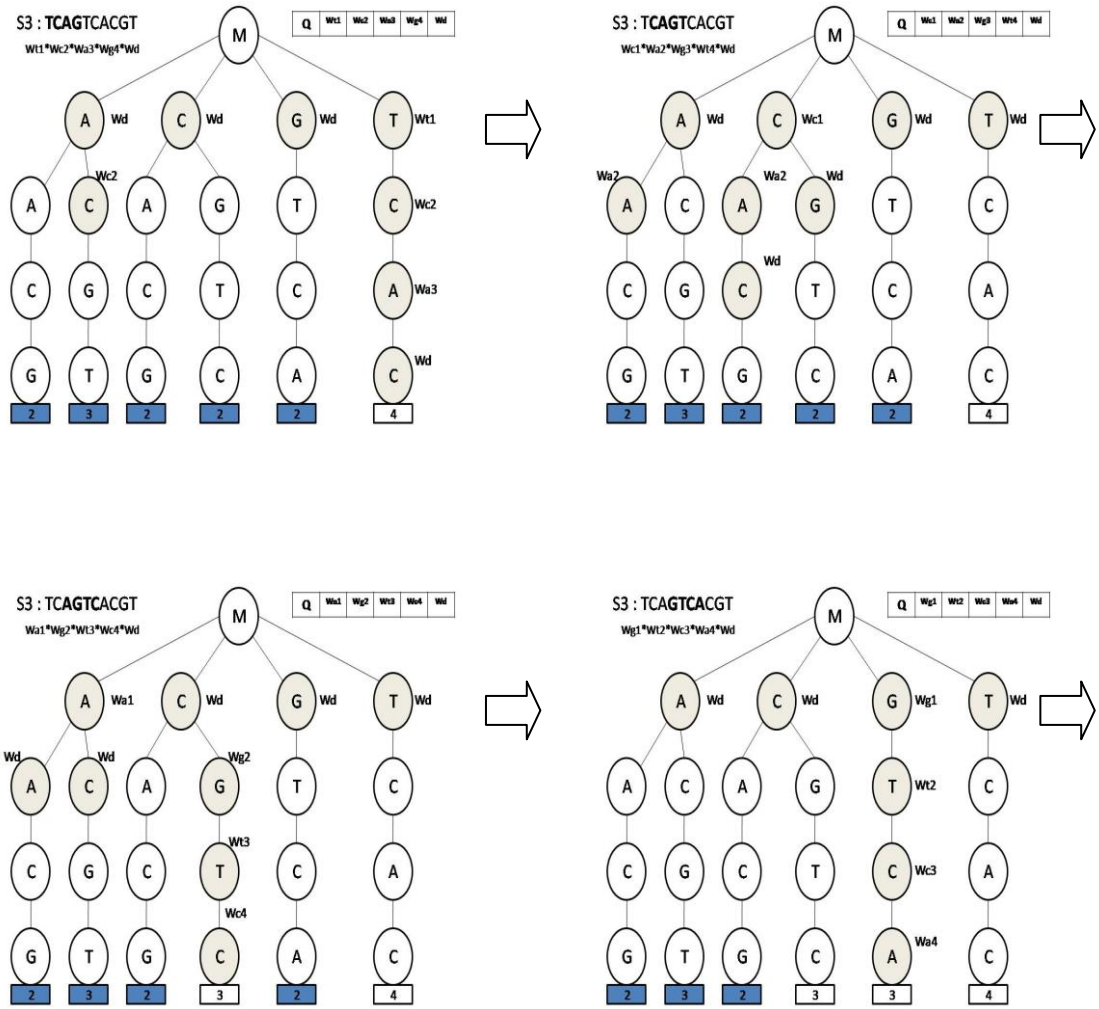
عند السلسلة الجزئية الأولى GCAC: تم تخزين قيم الأعداد الأولية المقترنة لظهور نيكليويتد معين ضمن ترتيب خانات الرتل مع تخزين عدد أولي واحد Wd (يعبر عن d) أصغر من كل الأعداد، ومن ثم فقد تم حساب جداء الأعداد المخزنة ضمن هذا الرتل، ليتم بعدها تحديد العقد الممكن المرور بها ابتداء من الجذر مع حذف المعاملات التي تسمح بالمرور عبر عقدة معينة واختبار القيمة الجديدة عبر الأبناء وهكذا وصولا للأوراق، فنلاحظ أن هذه السلسلة وصلت إلى إحدى الأوراق عبر العقدة T فقط.

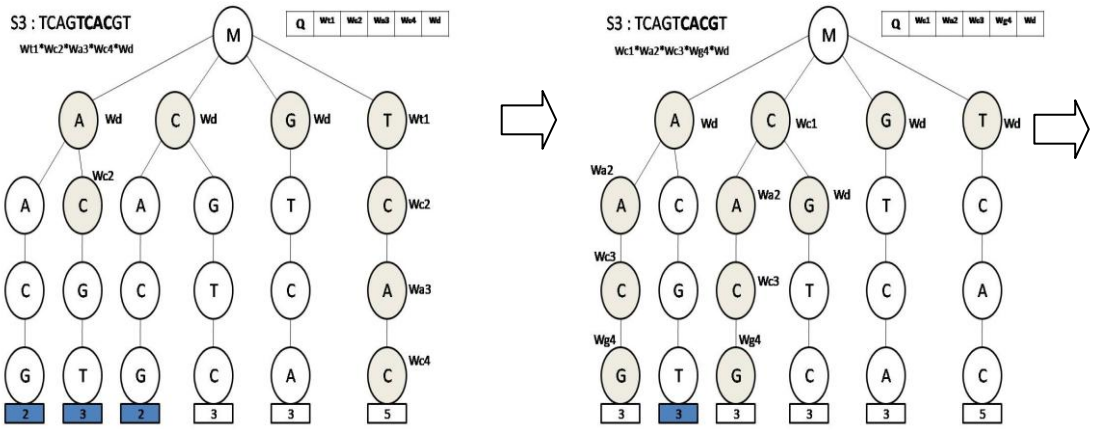
وبذات الطريقة نجد أن السلسلة الجزئية الثانية CACG مرت من الجذر إلى ورقتين، نكمل قراءة بقية السلاسل كما يلي:

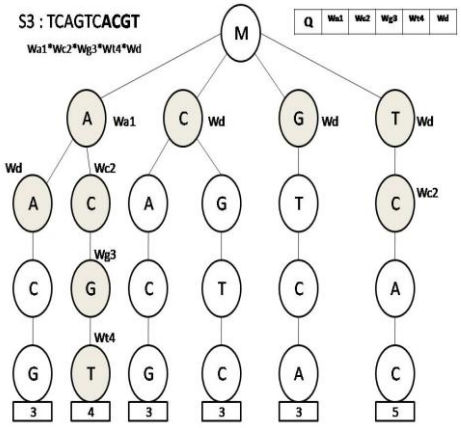




نلاحظ أنه قد تم المرور إلى جميع الأوراق بموجب هذه السلسلة وبالتالي فلا نحذف أي ورقة، وبخطوات متشابهة نقرأ السلسلة S3 كما يلي:







وعليه فإن المحفز المطلوب هو TCAC.

الفصل السابع

النتائج والأعمال المستقبلية

بمقارنة الخوارزمية المقترحة مع أحدث خوارزميتين في هذا المجال ألا وهما MDWB و MCES فإننا نجد أن هذه الخوارزمية تتفوق عليهما بما يلي:

١ –وصول مرة واحدة إلى كل نيكليوتيد في السلسة.

٢ –عالية الدقة حيث لا مجال للخطأ.

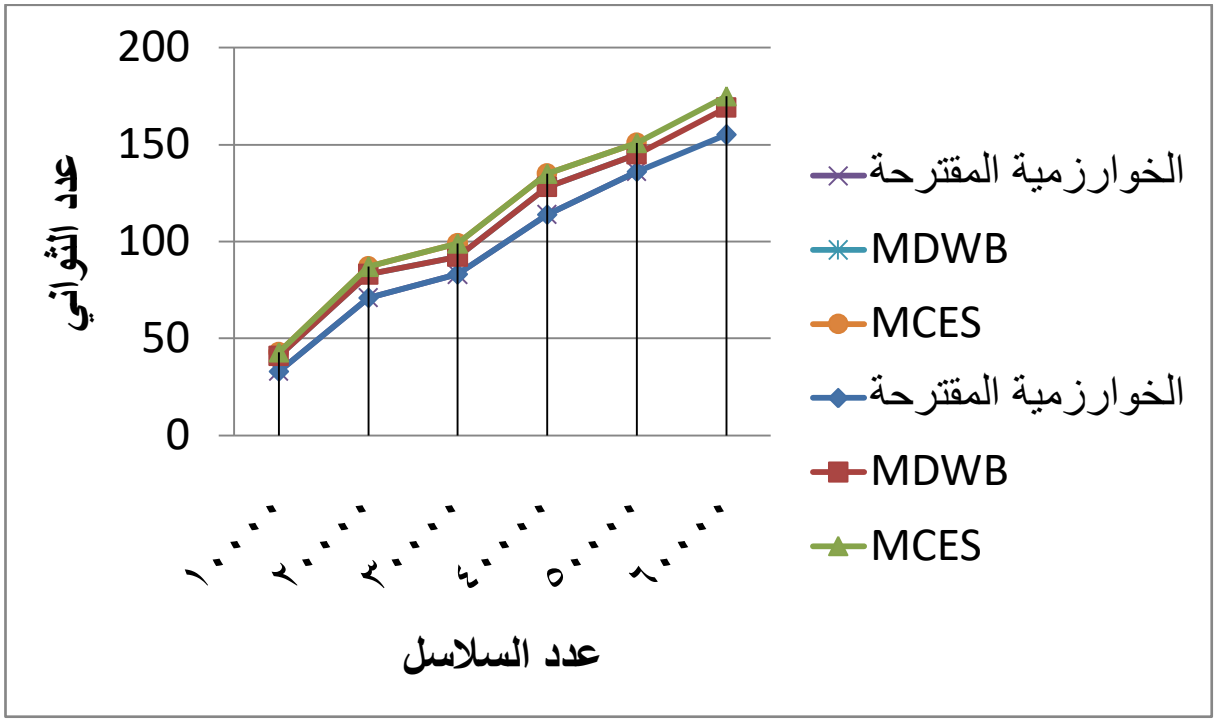
٣ –تعتمد على المفهوم العددي الدقيق بدلا من علم الاحتمالات المستخدم في الخوارزميات السابقة.

ويبين الجدول 1.7 مقارنة زمن التنفيذ بين الخوارزمية المقترحة و خوارزمتي MCES و MDWB

الجدول 1.7 مقارنة زمن التنفيذ بين الخوارزمية المقترحة و خوارزميتي MCES و MDWB

عدد		MDWB	الخوارزمية
10000	43	41	33
20000	87	83	71
30000	99	92	83
40000	135	128	114
50000	151	145	136
60000	175	169	155

ويتضمن الشكل 1.7 مخططاً بيانياً يظهر زمن تنفيذ الخوارزمية المقترحة مقارنة بالخوارزميات السابقة.



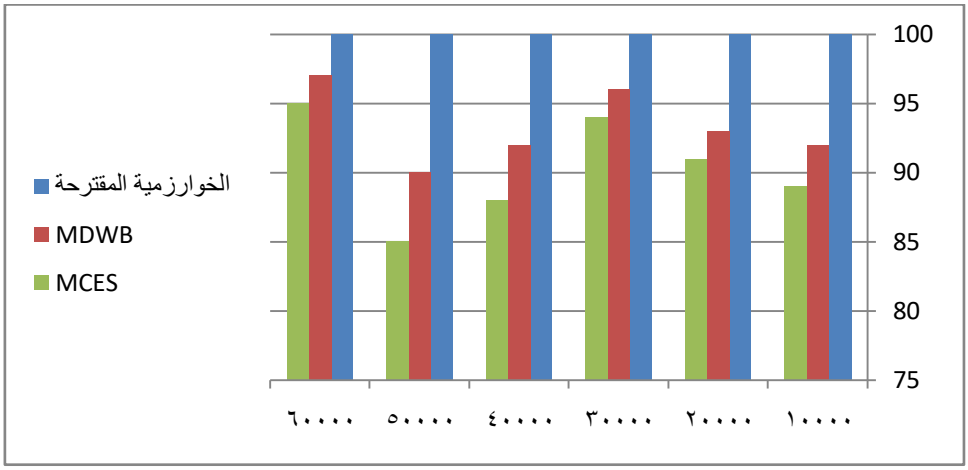
الشكل 1.7 مخطط بياني يظهر زمن تنفيذ الخوارزمية المقترحة مقارنة بخوارزميتي MCES و MDWB

كما يبين الجدول 2.7 مقارنة النسبة المئوية للدقة بين الخوارزمية المقترحة و خوارزمتي MCES وMDWB

الجدول 2.7 مقارنة النسبة المئوية للدقة بين الخوارزمية المقترحة و خوارزميتي MCES و MDWB

عدد السلاسل	MCES	MDWB	الخوارزمية المقترحة
10000	89	92	100
20000	91	93	100
30000	94	96	100
40000	88	92	100
50000	85	90	100
60000	95	97	100

ويتضمن الشكل 2.7 مقارنة بين دقة الخوارزمية المقترحة و خوارزميتي **MCES** و **MDWB** بيانياً.



الشكل 2.7 مخطط بياني يظهر الدقة بين الخوارزمية المقترحة و خوارزميتي MCES و MDWB

ملاحظات:

١ –تم التنفيذ على البيانات الطبية المحفوظة ضمن قواعد بيانات المعهد القومي لبحوث الجينوم البشري (Human Genome Center) والمتوفرة على الموقع الإلكتروني http://hgc.jp.

٢ –إن ملفات البيانات محفوظة بالصيغة FASTA، والمخصصة لتخزين هذا النوع من البيانات.

٣ –تم التنفيذ على جهاز من النوع ASUS مزود بمعالج (Intel Core I5-8250U) وذاكرة وصول عشوائي GB8 وبنظام تشغيل (Windows10 - 64 bit).

٤ –أن النتائج الواردة أعلاه قد تتغير من **حيث زمن التنفيذ فقط وليس من حيث الدقة**، وذلك باختلاف ملف FASTA المستخدم، ويعود الاختلاف إلى سلاسل البيانات المخزنة ضمن هذه الملفات والتي تؤثر على الأغراض المشكلة من الصف Tuples المستخدم في كلتا الخوارزميتين MCES و MDWB.

الآفاق المستقبلية:

١ –إيجاد طريقة أكثر تفاعلية للتعامل مع اختلاف طول المحفز L.

٢ –تطبيق المنهجية العامة للخوارزمية لتشمل كل أنواع البيانات الضخمة بأبجدياتها المختلفة.

المراجع:

1) Hilbert, Martin; López, Priscila (2011). "The World's Technological Capacity to Store, Communicate, and Compute Information". Science. 332 (6025).

2) O'Neil, Cathy (2017). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.

3) Karolin Kappler; Jan-Felix Schrape; Lena Ulbricht; Johannes Weyer (2018). "Societal Implications of Big Data". KI – Künstliche Intelligenz.

4) Kitchin, Rob; McArdle, Gavin (17 February 2016). "What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets.

5) "Big data and analytics: C4 and Genius Digital". *Ibc.org*. Retrieved 8 October 2017.

6) Han, Kamber, Pei, Jaiwei, Micheline, Jian (June 9, 2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.

7) Óscar Marbán, Gonzalo Mariscal and Javier Segovia (2009); A Data Mining & Knowledge Discovery Process Model.

8) Cabena, Peter; Hadjnian, Pablo; Stadler, Rolf; Verhees, Jaap; Zanasi, Alessandro (1997); Discovering Data Mining: From Concept to Implementation.

9) Nisbet, Robert; Elder, John; Miner, Gary (2009); Handbook of Statistical Analysis & Data Mining Applications.

10) Lukasz Kurgan and Petr Musilek (2006); A survey of Knowledge Discovery and Data Mining process models.

11) Günnemann, Stephan; Kremer, Hardy; Seidl, Thomas (2011). "An extension of the PMML standard to subspace clustering models". Proceedings of the 2011 workshop on Predictive markup language modeling.

12) Brockmann, Erich N.; Anthony, William P. (December 2016). "Tacit knowledge and strategic decision making".

13) Triantaphyllou, Evangelos (2000). Multi-criteria decision making methods: a comparative study.

14) Brunton, Bingni W.; Botvinick, Matthew M.; Brody, Carlos D. (April 2013). "Rats and humans can optimally accumulate evidence for decision-making.

15) Gardner, Margo; Steinberg, Laurence (July 2005). "Peer influence on risk taking, risk preference, and risky decision making in adolescence and adulthood: an experimental study".

16) Mashaghi A, Katan A (2013). "A physicist's view of DNA". De Physicus.

17) Lesk, A. M. (26 July 2013). "Bioinformatics".

18) Sim, A. Y. L.; Minary, P.; Levitt, M. (2012). "Modeling nucleic acids". Current Opinion in Structural Biology.

19) Ay, Ferhat; Noble, William S. (2 September 2015). "Analysis methods for studying the 3D architecture of the genome".

20) Nisbet, Robert (14 May 2009). "BIOINFORMATICS". Handbook of Statistical Analysis and Data Mining Applications.

21) AlbeGhosh A, Bansal M (April 2003). "A glossary of DNA structures from A to Z".

22) Tropp BE (2012). Molecular Biology (4th ed.). Sudbury, Mass.: Jones and Barlett Learning.

23) Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2014). Molecular Biology of the Cell (6th ed.).

24) Gregory SG, Barlow KF, McLay KE, Kaul R, Swarbreck D, Dunham A, et al. (May 2006). "The DNA sequence and biological annotation of human chromosome 1".

25) Halford E.S; Marko J.F (2004). "How do site-specific DNA-binding proteins find their targets?". Nucleic Acids Research.

26) Stormo GD (2000). "DNA binding sites: representation and discovery". Bioinformatics.

27) Bird A (January 2002). "DNA methylation patterns and epigenetic memory". Genes & Development.

28) Wang Milo, Ron; Philips, Rob. "Cell Biology by the Numbers: What is faster, transcription or translation?". book.bionumbers.org. Archived from the original on 20 April 2017.

29) Bailey T.L (2008). Discovering sequence motifs. Methods in Molecular Biology. Methods in Molecular Biology™.

30) Elnitski L, Jin VX, Farnham PJ, Jones SJ (2006). "Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques".

31) Lawrence CE, Reilly AA. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins. 1990.

32) Pesole G, Prunella N, Liuni S, Attimonelli M, Saccon C. WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences. Nucleic Acids Res. 1992.

33) Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science. 1993.

34) Bailey TL, Elkan C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine Learning. 1995.

35) Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics. 1999.

36) Marsan L, Sagot M. Algorithms for extracting structured motifs using a Suffix tree with an application to promoter and regulatory site consensus identification. J Comput Biol. 2000.

37) Eskin E, Pevzner P. Finding composite regulatory patterns in DNA sequences. Bioinformatics. 2002.

38) Carmack CS, McCue LA, Newberg LA, Lawrence CE. PhyloScan: identification of transcription factor binding sites using cross-species evidence. Algorithms for Molecular Biology. 2007.

39) Qiang Yu, Hongwei Huo. "An Efficient Algorithm for Discovering Motifs in Large DNA Data Sets ". IEEE TRANSACTIONS ON NANOBIOSCIENCE. 2015.

40) Mohamed Divan Masood, Manjula. "A Novel Approaches For Motif Discovery Using Data Mining Algorithm ". International Research Journal of Engineering and Technology (IRJET). 2018.

41)PhD.Mohsen Hussien, Eng.Kotyba Barakat. "Developing DNA Mining Algorithm for Motif Discovery ".Al_Baath University Magazine .2019. Electronic copy available at:

http://magazine.albaath-univ.edu.sy/1/pages/2019/01/83.pdf.

Abstract

The effectiveness of decisions depends on the efficiency of the information on which they depend, and the importance of information increases as the more accurate the means of the mining of data containing this information between the huge folds.

The most important functional element in DeoxyriboNucleic Acid (DNA) sequences is the Transcriptin Factor (TF). The position of this factor within these sequences is called Transcription Factor Binding Sites (TFBS), these sites are detected by motifs, which are defined as the most frequent substrings in DNA sequences.

The identification of motifs is one of the most challenges faced by researchers in biology, as their methods have diversified and their efforts have intensified to find the best way for identifying them due to their primary role in the medical domain.

The challenge is even more greater due to the fact that DNA is one of the most huge data growing in size, so the researcher interested in this challenge must develop an algorithm for searching the desired factor within this huge data, navigating the curves of their structural sequences to reach the desired target, achieving the accuracy and speed measurements.

In this paper, we have developed an algorithm based on trees and queues concept for representing motifs within DNA sequences.

we have also used prime numbers to reflect the repetition of motifs within DNA sequences, and finally have compared the performance of the proposed algorithm with previous algorithms in this domain of research.

Keywords: Decision-Making Process - Data Mining - Big Data - DeoxyriboNucleic Acid – DNA – Motifs – Transcription Factor Binding Sites – TFBS – Transcription Factor – TF – Text Mining.

Improving Decision-Making Process Using Big Data Mining (Medical Data)

A thesis submitted in partial fulfillment of the requirements for the Master's Degree in Software Engineering and Information Systems

Prepared By

Eng. Kotyba Barakat

Supervised By

PhD. Mohsen Hussien

Professor in Software Engineering and Information Systems

Faculty of Informatics Engineering

Al Baath University

1441 A.H. – 2019 A.D.