الجـــــــــــــــــــــــــــمهـوريـــــــة الجـــــــــــــــزائـريـة الديمـــــــقـراطيـة الشـــــــــعبيـة

وزارة التـــــــــــــــــــــــــــــعـليـم العـــــــالي و البـــــــــحث العـــــلـمـــي

جـــــــــــــــامعة غـردايـــــــــــــــــــة

كـليـة العـلـوم والتـكـنـلـوجيـا

قسـم الرياضيات و الاعلام الآلـي

مذكرة

للحصول على شهادة ماستر

في : الإعلام الآلي

تخصص : الأنظمة الذكية لإستخراج المعارف **(SIEC)**

الموضوع :

محاذاة السلاسل الجينومية

Alignement Des Séquences Génomiques

من اعداد الطالبين:

بالأعور عبد الرحيم
رواري عبد الكريم

الأستاذ رئيس المناقشة	**زيادي جلول**
الأستاذ المشرف	**بالأعور سليمان**
الأستاذ الممتحن	**أولاد النوي سليمان**
الأستاذ الممتحن	**كشيدة خالد**

السنة الجامعية 2017/2016

شـــكر و عـــرفان

الشكر لله قبل الجميع و الحمد لله الذي أنار لنا درب العلم و المعرفة و أعاننا على أداء هذا الواجب و وفقنا الى انجاز هذا العمل

نتوجه بجزيل الشكر و الامتنان الى الأستاذ المشرف بالأعور سليمان الذي لم يبخل علينا بتوجيهاته و نصائحه القيمة التي كانت عونا لنا طيلة فترة إنجاز هذه المذكرة . شاكرين له كل مجهوداته وتضحياته وتفكيره الراقي في مصلحة الطالب . داعين المولى أن يكون مثواه الأخير الفردوس الأعلى ولنا أيضا ولجميع المسلمين .

كما نشكر الأساتذة الممتحنين الأستاذ زيادي جلول و الأستاذ أولاد النوي سليمان و الأستاذ كشيدة خالد لقبول فحص مذكرتنا

كما نشكر جميع الأساتذة الذين درسونا أو عرفناهم منذ أن كنا براعما إلى يومنا هذا. شاكرين لهم نصائحهم تضحياتهم ومجهوداتهم فلهم الفضل بعد الله فيما نحن عليه الآن. وبالأخص جنود الخفاء الذين سعو بكل شغف من أجل تقديم هذه الرسالة السامية وبكل ما تحمله الكلمة من معنى. داعين المولى أن يجزيهم عنا ألف خير.

كما نشكر كل من كان له الفضل فيما نحن عليه اليوم دون ذكر أسماء . من أبسطهم إلى أعلاهم درجة من أقربهم إلينا حتى أبعدهم . شكرا جزيلا بما تحمله الكلمة من معنى

و في الأخير، أتقدم الشكر لأصدقائي وزملائي رفقاء الدراسة .

إهـــداء

اهدي عملي المتواضع هذا

الى والدي الغاليين

شاكرا لهما تضحيتهما و حبهما العظيم لي وثقتهما بي , متمنيا ان يكون نجاحي سبب في سعادتهما

الى اخوتي بدون استثناء

الذين اتمنى لهم التوفيق والنجاح والسعادة في حياتهم

الى اصدقائي بدون استثناء

الذين جعلوا لمعظم لحظات حياتي معنى وغمروها بالسعادة

الى جميع افراد العائلة الكبيرة

الى كل من احببت يوما ما أو ومن احبني

الى كل اولئك الذين تمنو لي الخير والسعادة

رواري عبد الكريم

إهـــداء

أهدي هذه المذكرة إلى

إلى التي لم تبخل علينا بالدعاء للتوفيق في العمل و النجاح في الحياة الى رمز الوفاء وفيض السخاء و جود العطاء إلى من تسكن في ثنايا قلبي الى من وهبني الأمل وعلمتني الصبر و النضال إلى التي روتني بفبض حنانها وعطفها ورعتني حسن المبادئ والأخلاق إلى التي شجعتني و دعمتني لأكمل مسيرتي العلمية إلى من جعلت الجنة تحت قدميها إليك أمي العزيزة أطال الله في عمرك

الى رمز الفخر والأصالة من كان منيرا لدربي و مرشدا لي والى من أدين له بتربيتي إلى من كان ساهرا على راحتي إلى الذي أتشرف بحمل اسمه وتعلمت منه أسمى الأخلاق و أفضلها إلى من علمني معنى الصبر و القناعة الى من هو سندي بعد الله و مصدر قوتي إلى القلب الكبير أبي الحبيب حفضك الله

إلى عائلتي جميعهم من أكبرهم إلى أصغرهم خاصة الكتكوت الصغير محمد علي

إلى جميع من يحبني ويعرفني دون إستثناء، إلى جميع أصدقائي كل واحد بإسمه و خاصة عبود، زهير، فاروق، منصف، عبد الله و الأصدقاء الذين جمعت بيننا الدراسة

إلى كل من فتح هذه الوريقات و تصفحها أهديكم جميعا ثمرة جهدي

بالأعور عبد الرحيم

الفهـــــــــرس

قائمة الجداول

قائمة الأشكال

ملخص

إن المعلوماتية الحيوية تعتبر من أهم العلوم الحالية، لما لها من تأثير مباشر في العديد من العلوم كالطب وعلم الأحياء والعديد من التطبيقات. مقارنة ومحاذاة السلاسل الجينومية هو المجال الأكثر شيوعا واهتماما لدى المختصين في علم الأحياء أو المعلوماتية الحيوية، لانه يعتبر الأساس والمنطلق بالنسبة للفروع الأخرى. السلسلة الجينومية هي عبارة عن تسلسل للنيكليوتيدات. هذا التسلسل يحدد طريقة ترابط الأحماض الأمينية من أجل تشكيل البروتين الذي بدوره يشكل الكائنات الحية. يمكننا تمثيل هذه السلاسل الجينومية بسلاسل نصية ذات أربعة أحرف. في حالة السلاسل الطويلة، تفشل الخوارزميات التقليدية في محاذاة هذه السلاسل في وقت تطبيقي.
إقترحنا مقاربتين مختلفتين لمعالجة تعقيد الوقت لمحاذاة السلاسل الجينومية. هاتان المقاربتان تبحثان عن الشبه الدقيق بطريقتين مختلفتين ثم تقوم بدمج مقاطع الشبه الدقيق للحصول على تنقيط أفضل إن وجد. المقاربة الأولى تعتمد على تحويل السلسلة الأطول إلى شجرة لواحق وبوادئ. بينما تعتمد الثانية على إهمال المناطق غير المتشابهة في السلسلتين. وكلتا المقاربتين لهما تعقيد وقت خطي.

كلمات مفتاحية

المعلوماتية الحيوية، محاذاة السلاسل، سلسلة جينومية، حمض نووي، بروتين، برمجة ديناميكية، خوارزميات تجريبية.

Abstract

Bioinformatics is one of the most important current sciences, it has an effect on sciences like medicine, biology and many applications.
The genome comparison and the sequence alignment are the most well-know and the most important to the biologists or biosciences because it is a basis of many branches.
The genomic sequence (DNA) is a sequence of nucleotides, this sequence defines the amino acid form in ordre to form the protien which form living organisms. The genomic sequence can be represented by alphabet of four symbols.
Traditional algorithms fail on alignment of long sequence in reasonable time.
We have proposed two different methods to deal with time complexity of genomic sequence alignment that search for the exact similarity, and merge the sections of exact similarity after, in ordre to obtain the best score if it exists.
The first depends on transform the long sequence to suffix and prefix tree, while the second ignores non-similar zones (K-mers). both approaches have linear time complexity.

key words: Bioinformatics, Sequence alignment, Genomic sequence, Deoxyribonucleic acid (DNA), Protien, Dynamic programing, Heuristic algorithms.

Résumé

La bio-informatique est l'une des principales sciences actuelles, en raison de son impact direct sur de nombreuses sciences telles que la médecine, la biologie, et de nombreuses applications.
Les biologistes et les bio-informaticiens s'intéressent le plus souvent à l'alignement et la comparaison des séquences génomiques, parce qu'ils présentent la base et le départ pour autres branches.
Le génome, ou bien une séquence génomique, est une séquence de nucléotides. Cette dernière détermine la disposition des acides aminés pour former une protéine, qui à son tour, constitue l'organisme vivant. On peut représenter ces séquences génomiques par des mots de 4 lettres.
Dans le cas où les séquences sont longues, les algorithmes traditionnels échouent dans l'alignement de ces séquences dans un temps raisonnable.
Nous avons proposé deux approches différentes pour répondre au problème de la complexité temporelle de l'alignement des séquences génomiques. Ces deux approches ont pour but de rechercher la similitude exacte entre deux séquences de deux façons différentes, et puis de fusionner les régions de similarité précise pour obtenir un meilleur score dans le cas échéant.
La première approche est basée sur la conversion de la séquence la plus longue à un arbre des suffixes et préfixes. Tandis que la seconde repose sur la négligence des régions non similaires dans les deux séquences. Les deux approches ont une complexité temporelle linéaire.

Mots clés :

La bio-informatique, Alignement des séquences, Séquence génomique, Acide désoxyribonucléique (ADN), Protéine, Programmation dynamique, Algorithmes heuristiques.

مقدمة

إن عالم المعلوماتية الحيوية حديث النشأة ليجمع بين علوم أهمها علم الحاسوب وعلم الأحياء و يعتبر من أهم العلوم الحالية، لما له من تأثير مباشر في العديد من العلوم كالطب وعلم الأحياء والعديد من التطبيقات التي هي كفيلة بحل العديد من المشاكل. و من أهم فروعه و التي تعتبر ضرورة ملحة لبقية الفروع، تحليل السلاسل. و من أهم فروع تحليل السلاسل، محاذاة السلاسل وبالأخص الثنائية، فهي الأساس بالنسبة للمحاذاة المتعددة وأيضا بالنسبة للبحث في قواعد البيانات و هذا ما يجعلها الأساس و نقطة الإنطلاق لفهم ما ينجر عليها من عمليات. والسلاسل التي نتعامل معها المحاذاة الثنائية ما هي إلا سلاسل جينومية وهي عبارة عن تسلسل نيكليوتيدات. هذا التسلسل يحدد طريقة ترابط الأحماض الأمينية من أجل تشكيل البروتين الذي بدوره يشكل الكائنات الحية. يمكننا تمثيل هذه السلاسل الجينومية بسلاسل نصية ذات أربعة أحرف.

إن الطول الهائل للسلاسل الجينومية يضعنا أمام مشكل الوقت غير تطبيقي الذي نستغرقه لمحاذاة هذه السلاسل. فعلى سبيل المثال يقدر طول الجينوم البشري بحوالي ثلاثة ملايير نيكليوتيدة. بغية حل هذا الإشكال نقترح في هذه المذكرة مقاربتين جديدتين. المقاربة الأولى تعتمد على شجرتي اللواحق والبوادئ للسلسة الأطول. بينما تعتمد المقاربة الثانية على إهمال المناطق غير متشابهة في السلسلتين. وتعد المقاربتان فعالتان في حالة السلاسل الطويلة إذ أنهما تمتلكان تعقيد وقت خطي.

للتعامل مع السلاسل الجينومية، وجب علينا الخوض في المفاهيم البيولوجية وتوضيح المهم منها وذلك لتحديد طريقة المحاذاة والمقارنة والقيود التي يجب أن نتبعها وفقا لهذه المفاهيم. كما أن الأبحاث السابقة قد تطرقت إلى معالجة مفهوم المحاذاة بواسطة طريقة المصفوفة النقطية التي تعتبر طريقة رسومية. وأيضا بواسطة إستعمال البرمجة الديناميكية للتعبير عن المحاذاة المحلية والعامة. أو عن طريق الخوارزميات التجريبية التي تتضمن خوارزمية FASTA.

سنشرح في هذه المذكرة كيف يمكن للمقاربتين اللتين إقترحناهما حل مشكل السلاسل الطويلة التي تعجز فيه البرمجة الديناميكية. حيث أن المقاربة الأولى

تعتمد على تحويل السلسلة الأطول إلى شجرة لواحق وبوادئ. وتعمل على مستوى بعد واحد، هو الذي تمثله السلسلة الأقصر. بحيث تقارنها مع شجرتي اللواحق والبوادئ للحصول على مقاطع الشبه الدقيق. وتنقيط هذه المقاطع و إختيار أفضل k مقطع ودمجها بغية الحصول على تنقيط أفضل إن وجد. إن تعقيد هذه الخوارزمية هو خطي وهو تطبيقي للسلاسل الطويلة. أما فيما يخص المقاربة الثانية فإنها تعتمد على إهمال المناطق غير المتشابهة بين السلسلتين في عملية المحاذاة، هذا لأننا خلال عملية المقارنة نحتاج المقاطع المتشابهة فقط وذلك من خلال تقسيم السلستلين إلى مقاطع ذات طول k ثم تحول مقاطع كل سلسلة إلى أرقام بالترتيب إنطلاقا من الرقم 1. ثم تخزن كل واحدة على حدا في جدول قوائم تحتوي المقطع وعدد تكراره في السلسلة بالإضافة إلى مواضع هذا المقطع. وبتصفح العمود الأول من الجدول نحدد المقاطع المتشابهة. ونهمل غير المتشابهة ونقوم بعملية دمج لهذه المقاطع ذات الشبه الدقيق وحساب أفضل تنقيط. وتعقيد الوقت لهذه المقاربة هو خطي وتطبيقي للسلاسل الطويلة.

فيما تبقى، نسرد محطات مذكرتنا في فصول هي كالتالي: نطرح في الفصل الأول المفاهيم البيولوجية، كالخلية والحمض النووي وأهميته في تحديد الكائن الحي، من بنية و وظائف البروتين و علاقته بالوظائف الحيوية.

ويناقش الفصل الثاني مفهوم المعلوماتية الحيوية و مجالاتها و الهدف منها. ثم يحدد لاحقا مفهوم محاذاة السلاسل والهدف منها. بعد هذا نتطرق إلى الخوارزميات التي تتبنى هذا المفهوم آخذة بعين الإعتبار المعنى البيولوجي. وسنختتم الفصل بسرد ايجابيات وسلبيات الخوارزميات المطبقة في محاذاة السلاسل.

بينما نطرح في الفصل الثالث مقاربتين جديدتين بغية حل مشكل الوقت غير التطبيقي الذي تستغرقه عملية محاذاة السلاسل الطويلة.

ونختتم مذكرتنا بأهم المفاهيم التي عالجناها والخوارزميات التي تطرقنا لها والتذكير بالحل الذي طرحناه خلال مساهمتنا لحل مشكل محاذاة السلاسل الطويلة و ووضع آفاق مستقبلية للعمل الذي جاء في هذه المذكرة.

الفصل 1
مدخل للبيولوجيا والمعطيات البيولوجية

1.1 مقدمة

لا بد و أنه مر على خاطر بعض منا وهو يتأمل في أحداث حياته اليومية تساؤل كيف لبذرة متناهية في الصغر أن تصبح نخلة بذلك الشموخ؟ بل وتنتج لنا التمور ما لذ منها وما طاب. وكيف للبذور أيضا أن تقدم لنا أنواعا لا تكاد تعد ولا تحصى من الخضروات والفواكه والأشجار من طعم طيب وشكل متناه في الإتقان؟ لنبقى في حالة من الذهول، كيف لشيء يكاد أن يكون مهملا في أذهاننا أن يتحول إلى نظام هندسي متقن يحتوي الحياة لكائنات أخرى؟

كيف لنقطة من سائل منوي متناهية في الصغر أن تصبح ذلك الإنسان المفكر الذي يرى، يسمع، يتكلم و يخترع الطائرات. لتجعل منه نظاما هندسيا عظيما. وهذا النظام يحتوي على العين، القلب، الأذن، الأرجل، والعديد العديد من المكونات. حتى الإنسان بحد ذاته قد يكون يجهلها و حتى العلم الحديث يراها معقدة جدا.

هذا يدعنا نتأمل في الموضوع بذهول ونتساءل ما الشيء الموجود بداخلها. حتى تشكل لنا هذا النظام الذي تعجز ألسنتنا حتى عن وصفه و الذي يعجز العلم الحديث عن فهمه الكامل. ما بالك بمحاولة صنع شبيه لذلك النظام. لا بد و أن بداخلها نظاما يحكمها، يسيرها، فلا شيء يصنع من العدم.

حسنا الفكرة و بإختصار أن الوحدة الأساسية لبناء كل هذه الكائنات الحية هي الخلية. ويمكننا أن نشبهها بالمصنع العظيم بشكل مصغر. هذا المصنع يحتوي على عضيات، من بينها النواة. يمكن تشبيه العضيات بالآلات وعمال هذا المصنع، أما النواة فيمكن تشبيهها بغرفةالتحكم. فهي تحتوي على الحمض النووي الذي يحتوي بدوره طريقة البناء و الصنع وكيفية تشكيل هذا النظام بهندسته ووظائفه. ونوه أن هناك أنواع معدودة لهذه للخلايا حسب نوع الكائنات الحية.

لا بد أن نكون قد لاحظنا أن أهم شيء هو ذلك المكون الذي يحتوي طريقة البناء و الصنع و كيفية تشكيل هذا النظام بهندسته ووظائفه. إنه الحمض النووي، فهو يحتوي كل المعلومات التي تحدد أشكالنا، لون البشرة، و الطول ولون العين و لون الشعر. ليس هذا فقط إنه هوالذي يحدد كل شيء فينا من صفات ووظائف ما ظهر وما بطن. طبعا ليس على سبيل الحصر الإنسان فقط، بل جميع الكائنات الحية، فهو يستخدم مجموعة من الأحماض الأمينية

كمادة أولية ثم يعطيها تعليمات للتماسك بطريقة معينة من أجل تكوين المنتج (البروتين) ذو وظيفة معينة واستخدام معين. فالحمض النووي هو الوحيد القادر على تحديد كيفية التماسك لإعطاء وظائف فعالة. لذا وجب علينا رؤيته ومعرفته عن قرب ومحاولة معرفة كيف يسير هذا النظام ومعرفة ما وصل له العلم في هذا المجال. كما يجب علينا أن ننقب بداخله عن المعلومات التي تجعل منه أجهزة تحتاج للحل. ليس هذا وفقط بل الولوج الى عالم جديد يسمى المعلوماتية الحيوية. هذا المجال الذي يربط بين الإعلام الآلي وتقنياته و علم الأحياء و معارفه لتحقيق تطبيقات تفيد البشرية بشكل أو بآخر. وسنتطرق في هذا الفصل إلى مفاهيك بيلوجية كالخلية والحمض النووي والبروتين.

2.1 الخلية

إن ما قد يجعلنا نطلق تسمية الكائن الحي هو مصطلح الخلية. فهو عبارة عن وظائف و عن تفاعلات وتكامل بين هذه الوظائف. كالنطق و السمع والرؤية و نقل الدم و دقات القلب وكل وظيفة في جسم الإنسان. أو أي كائن حي آخر، فهو نتاج هذه الخلايا والمعلومات التي تحويها بداخلها والبروتينات التي تنتجها.

فالسؤال الذي يتبادر الى أذهاننا كيف لهذه الخلايا أن تقدم الحياة؟ وهذه الأنظمة ذات الدقة العالية التي يعجز العلم على أن ينجز خلية مشابهة لواحدة منها. لنفهم هذا توجب علينا فهم الخلية ومحتوياتها وكيف تتم كل هته الآليات. وما هي المعلومات التي تحتويها.

1.2.1 مفهوم الخلية

الخلية هي الوحدة الأساسية لجميع الكائنات الحية. حيوانية كانت أو نباتية أو غير ذلك وتسمى بالوحدة التركيبية أو الوظيفية. فكل الكائنات الحية تتركب من خلية واحدة أو أكثر وهي تنقسم حسب نوع الكائن الحي الذي تشكله. أماتكاثرها ما هو إلا نتيجة إنقسام الخلية الأم وتسمى مجموعة الخلايا المتشابهة في التركيب والتي تؤدي معاًوظيفة معينة في الكائن الحي عديد الخلايا بالنسيج. وتحتوي الخلية على أجسام أصغر منها تسمى عضيات أهمها النواة التي تحمل في داخلها الشيفرة الوراثية (حمض نووي) [4, 9] .

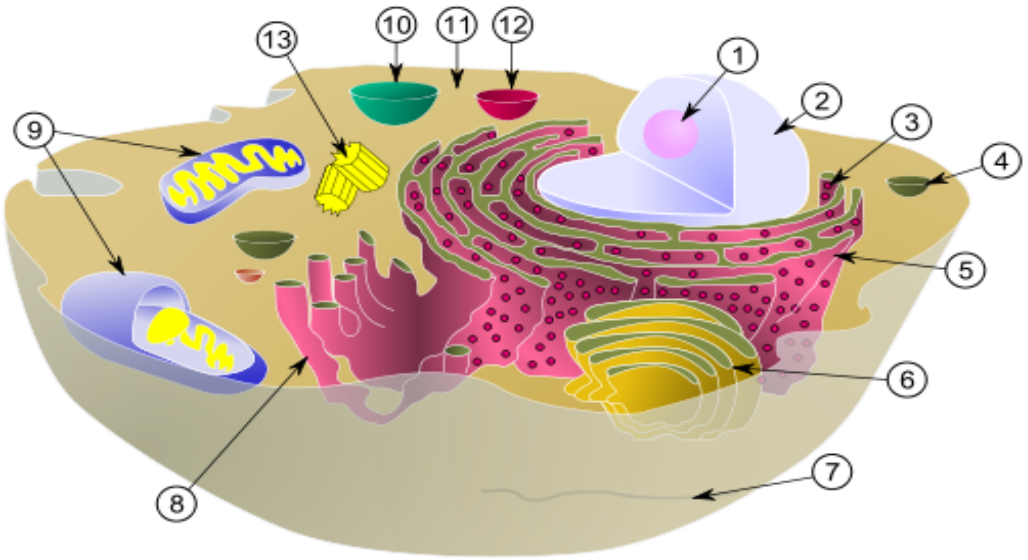
فعلى سبيل المثال جسم الإنسان عبارة عن بناية مبنية من الطوب. أي أن الطوب هو الوحدة الأساسية التي يبنى بها ذلك الجسد. يسمى العلماء هذا الطوب خلايا ولكن هذا

الطوب (الخلايا) ليس مصنوعاً من الأسمنت، بل من مادة تسمى بروتين. يحصل جسمنا على هذه المادة من الغذاء اليومي بعد أن تهضمه المعدة ويتحلل إلى أحماض امينية. بالطبع يوجد أعداد كثيرة من الخلايا في الجسم (تعد بالبلايين وليس الملايين) ولكنها متنوعة. منها الخلايا الجلدية و الخلايا العصبية و الخلايا العضلية و الخلايا الجنسية. جميع خلايا الجسم تموت ولكن أجسامنا وباستمرار تنتج خلايا جديده على مدار الساعة لتعويض النقص ويستثنى من ذلك الخلايا العصبية [12] . فأحيانا مثلا نلاحظ وجود قشور وتسلخات في الجلد بعد التعرض للشمس. هذه هي خلايا ميتة ويحل محلها خلايا جديدة من طبقة الجلد السفلية. وتحتوي الخلية داخل غلافها الخارجي على مادة الحياة "البروتوبلازما" Protoplasma وتتكون من مقطعين Proto أي أولي، وPlasma أي شكل، وتتركب من الناحيَّة الكيميائية من عدد كبير من العناصر المعروفة أهمها: الأكسجين، الكربون، الأيدروجين، النيتروجين، الفسفور، البوتاسيوم، الصوديوم ، الكبريت، الكلور، المغنسيوم، الكوبالت و اليود وغيرها.

ويوجد بالخلية "النواة" التي تعتبر العقل المُنظِم للتفاعلات الحيوية في الخلية. فهي المركز الواعي، فإذا فصلت النواة عن البروتوبلازم لا يمكن لأحدهما أن يعيش بمفرده. و يحيط بالنواة غلاف رقيق يتحكم فيما يمر داخل النواة، وبه الكروموسومات أو الصبغيات، وهي جسيمات عضوية دقيقة تحمل الصفات الوراثية و عددها ثابت في كل نوع من أنواع الكائنات الحيَّة. ففي الإنسان يوجد 48 من الكروموسومات بينما تبلغ لدى الأرنب 44. فالجدول رقم 1.1 يبين عدد كروموسومات بعض الكائنات الحية [12] .

أما فيما يخص التفاعلات الكيميائية بالخلية. فإنه توجد جزئيات بروتينية معقدة التركيب جدا تُعرف باسم "الإنزيمات" وهذه تساعد على إحداث التفاعلات الكيميائية دون أن تدخل فيها فهي بمثابة عامل مساعد. وتقوم الخلية بعملية التنفس وعملية التمثيل الكلوروفيلي (في الخلية النباتية) والحركة والنمو والتكاثر الذي يتم عن طريق تضاعف الحمض النووي. وقد تكون هذه البروتينات التي صنعتها شيفرة الحمض النووي، بروتينات بنائية، مثل العضلات أو الشعر أو الأظافر. وقد تكون هذه البروتينات أنزيمات أو هرمونات تتحكَّم في العمليات الحيوية في الكائن الحي.

الشكل 1.1 يمثل شكل ومكونات الخلية الحيوانية



1.النويّة 2.النواة 3.الجسيم الريبي 4.حويصل 5.الشبكة الإندوبلازمية الخشنة 6.جهاز جولجي 7.الغشاء الخلوي 8.الشبكة الإندوبلازمية الملساء 9.الميتوكوندريا 10.فجوة عصارية 11السيتوبلازم 12.الجسيم الهاضمّ 13.مريكزات

شكل 1.1: خلية حيوانية

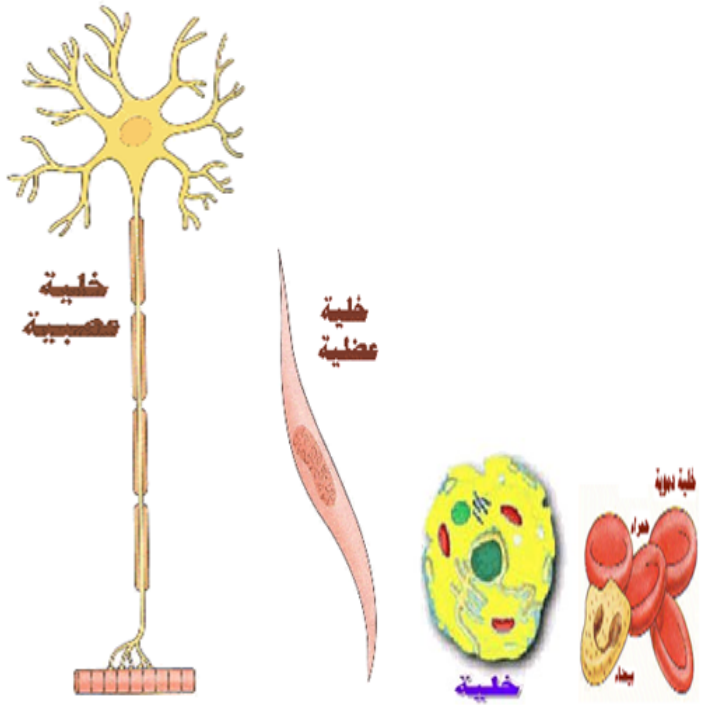
2.2.1 تخصص الخلية

كل خلية تحتوي بداخلها نفس عدد الكروموسومات الموجودة في بقية الخلايا. لذلك فإن كل خلية يوجد بداخلها نفس الوصفات الوراثية (المعلومات) لتحضير جميع البروتينات. أي أن كل خلية لديها القدرة لإنتاج جميع البروتينات من غير استثناء، ولكن في الحقيقة لا تقوم كل خلية بإنتاج جميع البروتينات. ليس لأنها لا تستطيع ولكن لأنها لا تحتاج جميع البروتينات. فلذلك على حسب تخصص الخلية ومكانها في الجسم تنتج المواد التي تحتاجها أما بقية المواد الأخرى فلا تقوم بتصنيعها. فمثلا خلايا الكبد تنتج فقط المواد التي تحتاجها وكذلك خلايا المخ تقوم بإنتاج المواد التي تحتاجها خلايا المخ فقط حتى وان كان لديها القدرة على إنتاج جميع المواد. فلكل عضو وظيفة خاصة به، فالكبد لها وظيفة معينة والعين لها وظيفة معينه وكذلك لبقية الأعضاء.

إذا فهمنا هذا الأمر فانه يسهل علينا معرفة لماذا يصاب عضو واحد أو عدة أعضاء محدودة في الجسم عندما يصاب أحد المورثات بعطب. مع أننا نعرف أن هذا العطب موجود في جميع

الخلايا. لان خلايا الكبد مثلا لا تتأثر بوجود العطب في المورث حتى وان كان موجودا. لأن خلاياها لاحتاج وجود المورث في الأصل. بينما تصاب فقط خلايا المخ لان المورث المعطوب مهم جدا لقيام المخ بوظائفه الطبيعية، وقد يصاب اكثر من عضو في آن واحد إذا كان المورث المعطوب مهم لجميع الأعضاء التي ظهر فيها المرض.

الشكل 2.1 يوضح بعض تخصصات الخلايا المختلفة.



شكل 2.1: تخصص الخلايا

3.1 الحمض النووي

أو ما يسمى بالشيفرة الوراثية، عبارة عن جزيء يحتوي كل المعلومات الخاصة ببناء الكائن الحي وهندسته وكل وظائفه المعقدة. يتواجد في بعض الكائنات في نواة الخلية. فهو من يحدد كل الصفات الجسمانية والبنائية للكائن الحي. على سبيل المثال اللإنسان من شكله، طوله، لون عينيه، لون شعره، لون بشرته و كل شيء موجود فيه من صفات.

الحمض النووي عبارة عن تتابع لنيكليوتيدات بترتيب معين. هذا الترتيب هو الذي يحدد الصفة أو الشكل الموجود في الكائن الحي. أجسامنا ووظائفها العديدة والمعقدة ليست إلا نتاج المعلومات التي تحملها مورثاتنا الموجودة في الحمض النووي.

1.3.1 مفهوم الحمض النووي

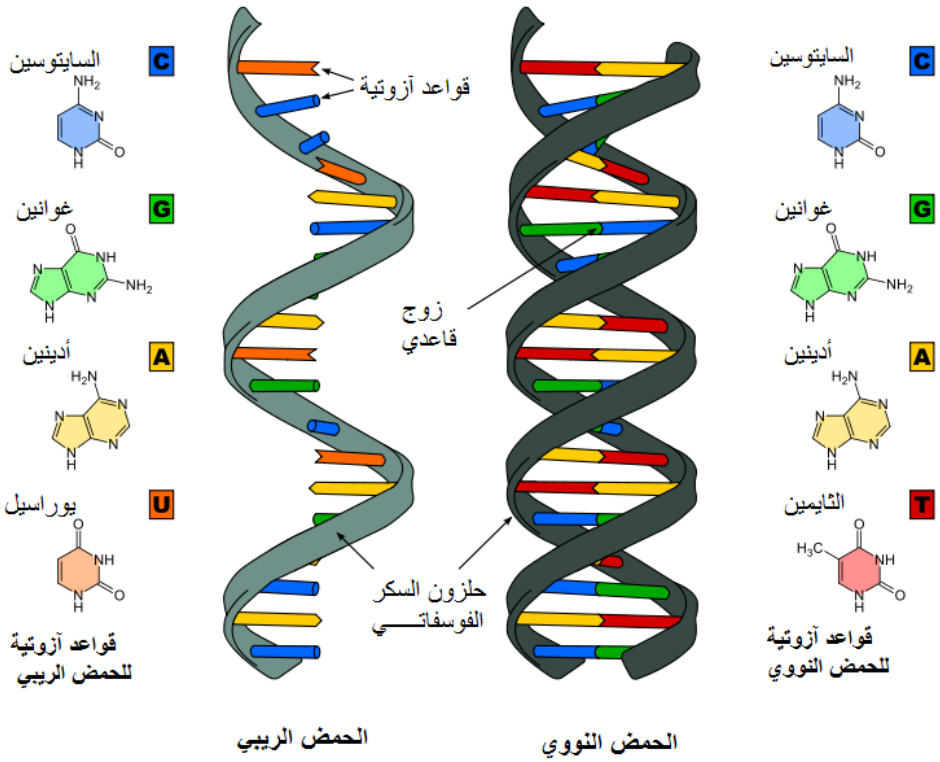
الحمض النووي Acid Deoxyribonucleic DNA في الرئيسيات، عبارة عن سلسلتين يلتف كل منهما حول الآخر بحيث يشبهان السلم الملتوي. تتكون السلسلتان المتوازيتان من جزيئات سكر خماسي والفوسفات، والسلسلتان مرتبطتان عرضيا بواسطة جزيئات كيميائية تحتوي على النتروجين. تلك الجزيئات حلقية و تسمى القواعد النتروجينية (نكليوتيدات) ويرمز إليها اختصاراً A و T و C و G [2, 5] . اختصارات الجزيئات المذكورة التي توصل بين السلسلة اليمنى للحمض النووي بالسلسلة اليسرى هي:

- أدينين Adenine وتختصر A
- غوانين Guanine وتختصر G
- الثايمين Thymine وتختصر T
- السايتوسين Cytosine وتختصر C

تتكرر هذه القواعد مليارات المرات في جميع أجزاء الجينوم (مجموعة الكروموسومات). يحتوي الجينوم البشري على سبيل المثال على 3 مليارات زوج من هذه القواعد. لو رمزنا لكل قاعدة بحرف من حروف الكتابة لملأت 3 آلاف كتاب يحتوي كل كتاب على 500 صفحة، أي لشكل مجموعها كتاب كبير يبلغ ارتفاعه ارتفاع ناطحة السحاب، كل ذلك في نواة خلية بشرية واحدة. يحتوي الجسم البشري على نحو 80 تريليون خلية.

تربط بين السلسلتين دائما جزيئين، إما A - T أو T - A أو G - C أو C - G ولا تصلح رابطات أخرى، ميزة تلك الطريقة البارعة هو أنه إذا ضاع أحد الجزيئين الرابطين فإن الآخر يستطيع جذب الجزيء قرينه بالذات ليشغل الفراغ، أي أن الحمض النووي يستطيع أن يصلح نفسه بنفسه. ليس هذا فقط وإنما في استطاعة الحمض النووي مضاعفة نفسه. لنتخيل أننا شققنا الحمض النووي من وسطه طوليا فأصبحت كل سلسلة منفصلة عن الأخرى. حينئذ تستطيع كلا السلسلتين بما هو متصل بها من الجزيئات القاعدية تكوين سلسلة جديدة عليها الجزيئات القاعدية المناسبة: مثل T في السلسلة الأولى تكوّن A في السلسلة الجديدة و C في السلسلة الأولى تكوّن G على السلسلة الجديدة. أي تصبح السلسلة الجديدة "قالبا" يمكن أن يكوّن بنفس المبدأ سلسلة مناظرة للسلسلة الأولى. ومنه يمكن لكل سلسلة مضاعفة نفسها وبالتالي يمكن للحمض النووي مضاعفة نفسه.

الشكل 3.1 يوضح شكل وبنية الحمض النووي.



شكل 3.1: شكل الحمض النووي

يعد الترتيب المحدد للقواعد A و T و C و G في غاية الأهمية. فهذا الترتيب يحدد جميع أوجه التنوع الحيوي. ففي هذا الترتيب تكمن الشفرة الوراثية. فكما أن ترتيب الحروف التي تتكون منها الكلمات هو الذي يجعلها ذات معنى. فترتيب هذه الجزيئات يحدد كون هذا الكائن الحي إنسانا أو ينتمي إلى نوع آخر من الأحياء كأن يكون شجرة أو سمكة أو نمرا أو نحلة على سبيل المثال. والتي يمتلك كل منها الشفرة الوراثية الخاصة به والذي ركزت عليه أبحاث وراثية خاصة عدة.

يكون الحمض النووي محمولا على أجسام صغيرة جدا(كالعصي القصيرة) تسمى صبغيات وراثية (ومعروفة بشكل أوسع بالكروموسومات). وتحمل هذه الصبغيات الوراثية التعليمات الكاملة لخلق الكائن الحي (الإنسان على سبيل المثال). عدد الصبغيات الوراثية في كل خلية من خلايا جسمنا 46 كروموسوم. وهذه 46 كروموسوم عبارة عن 23 زوج، كل زوج منها عبارة عن كروموسومين متشابهين بشكل كبير (وقد نقول تجاوزا انهما متطابقان). واحد من هذه الكروموسومات أعطته لنا أمهاتنا والآخر أعطاه لنا آباؤنا. وكل زوج من هذه الأزواج المتطابقة يعطيه علماء الوراثة رقما يميزه عن الآخر ابتداء برقم واحد لزوج الأول إلى الزوج الأخير رقم 23 [12] .

الكائن الحي	عدد الكروموسومات
الإنسان	46
الأرنب	44
البصل	16
الدجاج	78
الضفدع	39
الكلب	78
الفأر	40
التفاح	34 أو 51
الثعلب	38

جدول 1.1: جدول يوضح بعض الكائنات الحية وعدد كروموسوماتها

يبدأ خلق الإنسان مثلا من خليه واحدة، تحتوى هذه الخلية على 46 كروموسوم مكونه من 23 زوج.

تنتقل الكروموسومات من الأبوين عن طريق البويضة (الأم) والحيوان المنوي (الأب) فوظيفة البويضة والحيوان المنوي حقيقة هو نقل الكروموسومات من الأم والأب لتكوين الجنين. والفرق بين البويضة والحيوان المنوي (وتسمى مجتمعة خلايا جنسية) و الخلايا العادية في باقي الجسم (وتسمى خلايا غير جنسية) هو أن عدد الكروموسومات في البويضة والحيوان المنوي هو 23 كروموسوم فقط. بينما الخلايا العادية هو 46 كروموسوم (عبارة عن 23 زوج). عندما يلقح الحيوان المنوي بالبويضة (أي تندمج مع بعضها) فإن العدد الكامل للكروموسومات يكتمل فيصبح دخل الخلية الجديدة هذه 46 كروموسوم. من هنا يبدء تكون الإنسان وعبر سلسلة طويلة ومحكمة من انقسامات لهذه الخلية والخلايا الأخرى الناتجة منها ليصبح إنسانا كاملا. لذلك فكل إنسان يبدأ حياته بخلية واحدة فيها 46 كروموسوم إلى أن يكتمل البناء.

تنتقل المعلومات من خلية إلى أخرى عند انقسام الخلية. فكيف يحدث الانقسام وكيف تنتقل المعلومات (على شكل حمض نووي) عند انقسام الخلية؟ فقبل الإنقسام تتضاعف كمية الحمض النووي إلى ضعفين ويحدث هذا التضاعف عن طريق كسر الرابط الذي يربط الشريطين الملتصقين. ثم يقوم كل شريط بصنع شريط جديد مكمل له أي إن كل شريط ينشئ شريط أخر ليرتبط به كما ذكرنا سابقا. وفي هذه المرحلة نجد أن الخلية تحتوي على أربعة أشرطة من الحمض النووي. كل شريطين ملتصقين مع بعضهما البعض وهذا الوضع

غير طبيعي بالنسبة للخلية لذلك تبدء الخلية في الانقسام فيذهب كل زوج من الأشرطة إلى طرف الخلية ثم يحدث قص لها من الوسط. فينتج عن هذا خليتين متشابهتين وفي كل واحدة نسخة من الحمض النووي مطابقة تمامًا لما هو موجود في الخلية الأم وبنفس ترتيب القواعد النيتروجينية. ولو تمعنا جيدا لوجدنا أن كل شريطين عبارة عن شريط قديم من الخلية الأم. وشريط مماثل له تم نسخه أثناء عملية الانقسام [12] .

فالحمض النووي عبارة عن قطع من السكر والفسفور والقواعد النيتروجينية (النيكليوتيدات). إن هذه النيكليوتيدات مرتبة بشكل متقن، ويقسم هذا الشريط المتراص من النيكليوتيدات إلى أجزاء ووحدات تسمى بالمورثات أو الجينات. وكل مورث يحمل صفة معينة تعطي التعليمات المطلوبة لصناعة نوع معين من البروتين، فالبروتينات هي المواد الخام التي يصنع منها أنسجه الجسم، وكذلك الإنزيمات المطلوبة لوظائف الجسم الحيوية والتفاعلات الكيمائية. كان يعتقد أن عدد المورثات الموجودة في كل خلية يتراوح بين 50 ألفًا إلى 100 ألف مورث. ولكن تبين للمختصين بعلم الوراثة أن مجموع المورثات في الخلية اقرب إلى 50 ألف بل إنها اقل من ذلك العدد ويعتقد أنها حوالي 20 ألف إلى 30 ألف مورث. لذا فإن الحمض النووي يتحكم في صفات الخلية ومن ثم الصفات البشرية عن طريق التحكم في بناء البروتينات.

لقد وجد أن كل الكائنات الحية تتكون من البروتينات. وأن الإنسان يمكن أن يصنع حوالي 30 ألف نوع مختلف من البروتينات. وهي عبارة عن جزيئات كبيرة ومعقدة ومكونة من سلسلة طويلة من القطع التي تسمى بالأحماض الأمينية. ويوجد 20 نوع فقط من الأحماض الأمينية المختلفة في جميع المخلوقات الحية. وإختلاف تتابع هذه الأحماض الأمينية هو الذي يكوِّن البروتينات المختلفة. وفي المورث (الجين) ترمز كل ثلاث قواعد نيتروجينية إلى حمض أميني معين. فمثلا تتابع القواعد النيتروجينية الثلاثة (ادينين – الجوانين – الثيمين) AGT يرمز إلى الحمض الأميني المسمى بالميثايونين، وهكذا بقية الأحماض الأمينية.

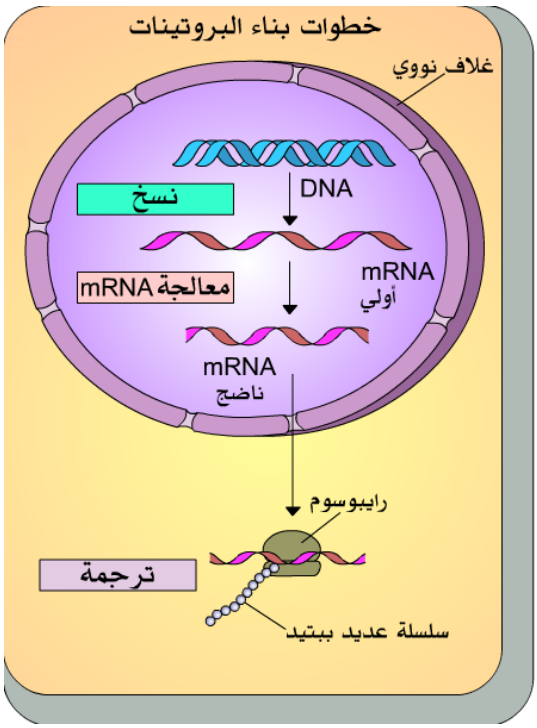
2.3.1 صناعة البروتينات انطلاقا من الجينات

البروتين هو مجموعة من الأحماض الأمينية المربوطة معا. إن تتابع القواعد النيتروجينية A و T و C وG على طول جزيء الحمض النووي DNA يحدد للخلية أنواع الأحماض الأمينية التي يجب أن تستخدم، وكذلك ترتيبها وترابطها لتكوين بروتين معين. كما أن الجين الواحد يمتلك تتابعاً من القواعد النيتروجينية لجزيء بروتين واحد. على سبيل المثال لأحرف CAA وهي

تتابع قاعدي (سايتوسين، أدنين، أدنين) يخص إنتاج الحمض الأميني المعروف باسم فالين. وبمفهوم آخر فإن المعلومات المشفرة مكتوبة في كلمات مكونة من ثلاثة أحرف (مجموعات يتكون كل منها ثلاثة قواعد نتروجينية) [12, 5] . وكل كلمة تحدد للخلية أي حمض أميني يجب أن يرتبط مع حمض أميني آخر. ويمكننا ترتيب خطوات صناعة البروتين إلى:

1. ينفك الرباط بين خيطين من الحمض النووي DNA ليكشف عن الجين اللازم لتصنيع بروتين من نوع معين.
2. يتم نسخ القواعد النيتروجينية الموجودة على الجين.
3. ينتقل الحمض النووي الرايبوزي الرسول (mRNA) من النواة إلى الرابوزوم.
4. يحمل الحمض النووي الرابوزي الناقل (tRNA) الأحماض الأمينية إلى الرابوزوم.
5. تتصل الأحماض الأمينية لصنع جزيء البروتين المعين.

ويوضح الشكل 4.1 مراحل الترجمة بالتسلسل



شكل 4.1: مراحل الترجمة

3.3.1 إحصائيات حول الحمض النووي والبيانات البيلوجية

- إن المعلومات التي يحتوي عليها الجينوم البشري (نحو 3.3 مليار زوجا من القواعد)، تحتاج هذه المعلومات لكتابتها 3300 كتابا كل كتاب يبلغ عدد صفحاته 100 صفحة. وكل صفحة بها 1000 حرف وهذا كله في نواة خلية بشرية واحدة.
- لدى الكائن البشري، يختلف الحمض النوي DNA من فرد لآخر بنسبة 2% فقط ,أو 1 من كل 50 حرفاً.
- عند قراءة الجينوم البشري بسرعة حرف واحد في الثانية لمدة 24 ساعة يومياً, فسيستغرق الأمر قرناً كاملاً من الزمان للانتهاء من قراءة كتاب الحياة.
- إذا بدأ شخصان مختلفان في قراءة كتاب الحياة الخاص بكل منهما بسرعة حرف واحد في الثانية، فسيستغرق الأمر نحو ثماني دقائق ونصف الدقيقة (500 ثانية) قبل أن يصلا إلى أول اختلاف في ترتيب حروف كتابيهما.
- يتشابه الحمض النووي الخاص بالبشر مع مثيله في الشمبانزي بنسبة 96%.
- يبلغ العدد التقديري للجينات في الإنسان 22.000 من الجينات. خلايا الخميرة يبلغ عدد الجينات 6.000 تقريباً , بينما يبلغ عدد جينات أحد أنواع الجراثيم 4.000.
- تظل وظيفة الغالبية العظمى (98%) من الحمض النووي الموجودة في الجينوم البشري غير معروفة لدينا حتى الآن، حيث تشكل الجينات نحو 2% منه فقط وهي الجينات التي تنتج بروتينات وهرمونات. ويبدو أن 98% من سلسلات القواعد الموجودة بين الجينات وبعضها ليست لها وظيفة وتسمى بالنفايات أو المهملات. ولكن البحث العلمي الحديث يبين أن لها وظيفة إدارية بالنسبة لتنشيط أو تهدئة عمل الجينات.
- كان أول كروموسوم (chromosome) بشري تم فك شفرته بالكامل هو الكروموسوم رقم 22. وقد تم ذلك في المملكة المتحدة في ديسمبر 1999، وتحديداً في مركز (سانجر) بمقاطعة كمبردج.
- يبلغ طول الحمض النووي الموجود في كل من خلايانا 1.8 متر .
- إذا تم فرد جميع الحمض النووي الموجود في الجسم البشري طرفا لطرف , يمكن للخيط الناتج أن يصل من الأرض إلى الشمس بل و أكثر بكثير.

• يقوم الباحثون في مشروع الجينوم البشري بفك شفرة 12.000 حرف من الحمض النووي البشري في الثانية الواحدة. [2, 5]

4.3.1 تطبيقات الحمض النووي

بنوع من التفصيل في مجالات الدراسة والبحث في هذا المجال نذكر : [1]

• الطب الجزيئي medicine Molecular والذي يشمل :

- تحسين تشخيص الأمراض.
- الاكتشاف المبكر للاستعداد للإصابة بالأمراض الوراثية.
- تصميم الأدوية بصورة أكثر ملاءمة.
- المعالجة بالجينات وأنظمة التحكم للأدوية.
- علم الأدوية المجيني ، تصميم أدوية تستهدف أمراضاً وراثية بعينها.

• الجينوميات الجرثومية genomics Microbial والذي يشمل :

- مصادر جديدة للطاقة (الوقود الحيوي).
- مراقبة البيئة لاكتشاف الملوّثات.
- الوقاية من الحرب البيولوجية والكيميائية.
- التخلص من النفايات السامة بطرق مأمونة وفعّالة في الوقت نفسه.
- فهم القابلية للتعرض للأمراض والكشف عن الأهداف .

• الدراسات السكانية وعلم الإنسان والذي يشمل :

- دراسة التطور عبر طفرات الخط الجنسي في السلالات البشرية المختلفة.
- دراسة أنماط هجرة المجموعات السكانية المختلفة استنادا إلى التوريث الجيني للإناث.
- دراسة طفرات الكروموسوم (Y) لتتبع السلالات وأنماط هجرة الذكور.

• استخدامات الحمض النووي في مجال الطب الشرعي والذي يشمل :

- التعرف على المشتبه بهم المحتملين التي قد يطابق الحمض النووي الخاص بهم الأدلة الموجودة في مسرح الجريمة.

- تبرئة الأشخاص المتهمين بالخطأ في الجرائم.
- التحقق من علاقات البنوّة وغيرها من قضايا النسب.

• الزراعة وتربية الحيوان والمعالجة البيولوجية

- تحديد نسب البذور أو الماشية في برامج التهجين.
- المحاصيل المقاومة للأمراض, والحشرات، والجفاف.
- حيوانات المزرعة الصحيحة، والأكثر إنتاجاً, والمقاومة للأمراض.
- منتجات زراعية أكثر فائدة غذائية.
- المبيدات الحشرية البيولوجية .
- اللقاحات التي يمكن أكلها, ومن ثم دمجها في المنتجات الغذائية.

• التعرف على الأنواع الحية المهددة بالانقراض والمحمية كمساعدة لمسئولي هيئات حماية الحياة البرية (ويمكن استخدامها في ملاحقة منتهكي قوانين حماية الحياة البرية).

• التعرف على البكتريا وغيرها من الجراثيم التي قد تلوّث الهواء أو الماء أو التربة أو الغذاء.

• تحقيق التوافق النسيجي بين المتبرع والمتلقي في برامج زراعة الأعضاء.

4.1 البروتين

1.4.1 مفهوم البروتين

البروتين هو مركب عضوي معقّد التركيب ذو وزن جزيئي عال يتكون من أحماض أمينية مرتبطة مع بعضها بواسطة رابطة ببتيدية و بترتيب معين يحدده الحمض النووي. هذا الترتيب هو من يحدد شكله ووظيفته. البروتين ضروري في تركيب ووظيفة كلّ الخلايا الحية وحتى الفيروسات [9] .

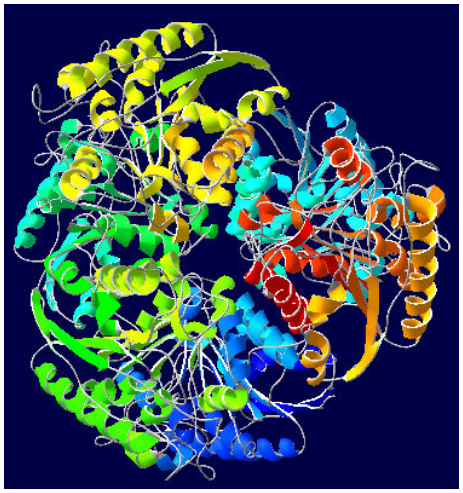
ومن أهم ما يميز البروتين نذكر:

• يشارك البروتين تقريبا في جميع العمليات الخلوية بما فيها تنظيم الوظائف الخلوية.

• يتميز كل بروتين ببنية مختلفة عن البروتينات الأخرى، تدعى هذه البنية بالحالة الأصلية للبروتين وتتحدد حسب ترتيب الأحماض الأمينية في عملية الترابط التي تشكل السلاسل البروتينية.

- كل بروتين له وظيفة معينة وله شكل وبنية فراغية فريدة تحدد وظيفته.
- كل جين مسؤول عن إنتاج بروتين أو أكثر.
- الخلل في الجين يسبب تشوه في بنية البروتين ووظيفته.
- التشابه في التسلسل الجيني يعني التقارب البنيوي و منه التقارب الوظيفي.

الشكل 5.1 يعبر عن البنية الثلاثية الأبعاد للبروتين.



شكل 5.1: بنية البروتين ثلاثية الأبعاد

2.4.1 وظائف البروتين

للبروتين وظائف عدة نذكر منها : [9]

- وظائف الأعضاء
- تخزين الأحماض الأمينية
- نقل مواد داخل الكائن الحي
- تنسيق أنشطة الكائن الحي
- استجابة الخلايا للمؤثرات الكيميائية
- الحركة
- الحماية ضد المرض
- التفاعلات الكيميائية للكائن الحي

البنية الفراغية للبروتين هي عبارة عن شكل ثلاثي الأبعاد و ثابت لبروتين معين. إن ترتيب الأحماض الأمينية الذي يحكم فيه بدوره الحمض النووي هو من يحدد بنية البروتين والتي بذاتها تحدد الوظيفة حيث تنشأ بين جذورها روابط كيميائية في مواقع معينة.

البروتين ينتقل من شكل إلى آخر للحصول على بنيته النهائية ووظيفته الفعالة. فخلل واحد في هذه الأحماض الأمينية قد يفقد بشكل كبير وظيفة هذا البروتين وإحداث خلل في شكله، نتيجة لتتفكك بعض الروابط الكيميائية أو وجودها في غير محلها. فتتغير بنيته الفراغية ويفقد تخصصه الوظيفي. أي أن حدوث طفرة في المعلومة الوراثية يحدث خللا في تسلسل الأحماض النووية وبالتالي الوظيفة البروتينية. إذ أن الحمض النووي يحدد طريقة ترتيب الأحماض الأمينية المشكلة للبروتين. وبدورها هي من تحدد بنيته التي هي بذاتها من تحدد وظيفة هذا البروتين داخل الخلية أو خارجها.

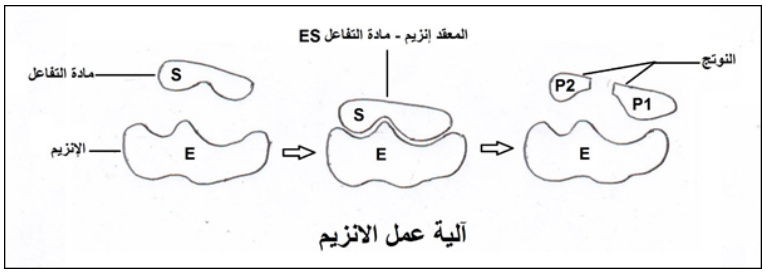
وعليه الرغبة في دراسة الشبه بين وظائف البروتينات أو الشبه بين الكائنات الحية هو نفسه دراسة الشبه بين تسلسل الأحماض الأميني أو سلاسل الحمض النووي. فهما وجهان لعملة واحدة. لكن الوجه الثاني أسهل وذو جانب تطبيقي أكثر من أن نقارن البنى والوظائف التي يكون العمل عليها مباشرة أكثر تعقيدا من التعامل مع السلاسل مباشرة.

ومن بين أهم الوظائف التي تؤديها هذه البروتينات بعد أن تنضج وتصبح فعالة: [9]

1. الأنزيمات :

فالأنزيم عبارة عن محفز بيولوجي. فهو فقط عبارة عن مركب بروتيني وظيفته بشكل عام في الكائنات الحية هي تسريع التفاعلات الكيميائية الحيوية والتحكم بالبنية الفراغية للناتج. بحيث يحتوي على موقع فعال يميز بعدد محدود من الأحماض الأمينية يتميز بالتكامل البنيوي مع مادة التفاعل ومهمته تحفيز التفاعل.

الشكل 6.1 يوضح آلية عمل الإنزيم



شكل 6.1: آلية عمل الأنزيم

بحيث يرتبط الانزيم بالركيزة في مستوى الموقع الفعال لوجود تكامل بنيوي ويتشكل المعقد انزيم-مادة التفاعل ثم تتحول الركيزة إلى ناتج P بواسطة موقع التحفيز للموقع الفعال. ثم يتحرر الناتج عن الانزيم. أنواع التفاعلات التي يشارك فيها الأنزيم هي تفاعلات البناء والهدم والتحويل و لا يستهلك أثناء التافعل فهو عبارة عن وسيط محفز فقط.

2. الدفاع عن الذات:

بواسطة جزيئات ذات طبيعة بروتينية تساهم في مناعة الكائن و حمايته من تمييز بين الذات واللاذات، وكذلك التعرف على المستضدات بواسطة مستقبلات غشائية نوعية ذات طبيعية بروتينية، و أيضا الإتصال بين الخلايا المناعية و التخلص من المستضدات والخلايا المصابة بواسطة جزيئات ذات طبيعة بروتينية تتمثل في الأجسام المضادة فدورها جلي في حماية مناعة الكائن.

3. الإتصال العصبي :

لعل أهم الوظائف تتمثل في النقل المشبكي الذي يعتمد على المستقبلات البروتينية و توليد و انتشار كمون العمل
و الحفاظ على استقطاب العصبون عن طريق مضخات ذات طبيعة بروتينية

5.1 خاتمة

لقد أدركنا أن الخلية تحتوي على الحمض النووي الذي يحمل بداخله هندسة الكائنات الحيية وجميع وظائفها الحيوية. وكيف له أن يترجم إلى أحماض أمينية مشكلة بروتينات التي بدورها تعبر عن التخصص الوظيفي للكائنات الحية. أي أن أي تشابه في الكائنات الحية سيكون على مستوى الوظائف والبروتينات و علاقة متعدية يعني التشابه في الحمض النووي أي أنه كل ما كثر التشابه في البروتينات أو الحمض النووي لكائنين كان التشابه جليا في الوظائف والهندسة

والشكل. هذا ما يقودنا للتساؤل كيف يمكننا مقارنة البروتينات والحمض النووي للكائنات الحية اذا كنا نستطيع تمثيلهم على شكل تسلسل نصي. و هل يمكننا الوصول لتحديد نسبة الشبه بين الكائنات الحية إنطلاقا من الجينوم الخاص بها.

الفصل 2
محاذاة السلاسل و خوارزمياتها

1.2 مقدمة

كنا قد وضحنا في الفصل السابق كيف أن الخلية تحتوي على الحمض النووي الذي يحمل بداخله هندسة الكائنات الحيية وجميع وظائفها الحيوية. وكيف له أن يترجم إلى أحماض أمينية مشكلة بروتينات التي بدورها تعبر عن التخصص الوظيفي للكائنات الحية. أي أن أي تشابه في الكائنات الحية سيكون على مستوى الوظائف والبروتينات و علاقة متعدية يعني التشابه في الحمض النووي أي أنه كل ما كثر التشابه في البروتينات أو الحمض النووي لكائنين كان التشابه جليا في الوظائف والهندسة والشكل. وفي هذا الفصل سنجيب عن التساؤل كيف يمكننا مقارنة ومحاذاة البروتينات والحمض النووي للكائنات الحية اذا كنا نستطيع تمثيلهم على شكل تسلسل نصي و كيف يمكننا الوصول لتحديد نسبة الشبه بين الكائنات الحية إنطلاقا من الجينوم الخاص بها.

سنمهد لهذا الفصل بالتطرق للمعلوماتية الحيوية ذاك العلم الحديث الذي يجمع بين مفاهيم البيولوجيا وتقنيات وخوارزميات الإعلام الآلي. بغية الاستفادة من نتائجها في تطبيقات طبية وغيرها. وسنركز بصفة خاصة في مجال تحليل السلاسل وبالظبط محاذاة السلاسل. وسنقدم الخوارزميات التي تعالج هذا المفهوم كما سنتطرق لإيجابيات وسلبيات كل منها سواء التي تعتمد المفهوم الرسومي أو البرمجة الديناميكية أو الخوارزميات التجريبية.

2.2 المعلوماتية الحيوية

1.2.2 تعريف

المعلوماتية الحيوية (bioinformatics) هي إستخدام تكنولوجيا المعلومات في مجال التكنلوجيا الحيوية بغية تخزين البيانات وتحليلها وتحليل سلاسل الحمض النووي. وهي تحتاج عدة مجالات مثل علم الأحياء و الرياضيات وعلم الحاسوب وقوانين الفيزياء والكيمياء والمعرفة الجيدة لتكنلوجيا المعلومات. لتحليل ودراسة المعلومات الحيوية [3]. أغلب النشاطات في المعلوماتية الحيوية من أجل تحسين برامج وأدوات مفيدة لتحليل وتخزين واسترجاع المعلومات الحيوية.

2.2.2 الهدف من المعلوماتية الحيوية

الهدف الأساسي للمعلوماتية الحيوية متمثل في الفهم بشكل أفضل للخلية الحية وكيف تعمل على المستوى الجزيئي و من خلال تحليل السلاسل الجزيئية. والسبب في ذلك أن وظائف

الخلية يمكن أن تفهم بشكل أفضل من خلال تحليل سلسلة البيانات التي تحتويها. لأن كل التدفق المعلوماتي والعمليات محتواة في الخلية من أول وظيفة لآخر عملية فيها [10, 7].

3.2.2 مجالات المعلوماتية الحيوية

تنقسم مجالات المعلوماتية الحيوية إلى ثلاثة فروع وهي [3]:

1.3.2.2 التحليل الهيكلي (الفراغي أو البنيوي)

ويتضمن:

- التنبؤ ببنية الحمض النووي
- التنبؤ ببنية البروتين
- تصنيف بنى البروتين
- مقارنة بنى البروتين

2.3.2.2 تحليل الوظائف

ويتضمن:

- التعبير الجيني
- التنبؤ بتفاعل البروتين
- نمذجة المسار الأيضي (عبارة عن سلسلة من التفاعلات الكيميائية التي تحدث في داخل الخلية)

3.3.2.2 تحليل السلاسل

وهو المجال الذي سنخوض في سياقه ويتضمن:

- مقارنة ومطابقة الجينوم
- العائلة الشجرية الجينية
- المورثات والتنبؤ بالبادئة
- إكتشاف العناصر
- البحث في قاعدة بيانات السلاسل
- محاذاة السلاسل

4.2.2 تطبيقات المعلوماتية الحيوية

إن تطبيقات المعلوماتية الحيوية متعددة وتمس العديد من المجالات نذكر أهمها و هي كالتالي [10]:

- الطب الجزيئي
- الطب الوقائي
- العلاج الجيني
- تطوير الأدوية
- المقاومة للمضادات الحيوية
- تنظيف النفايات
- التكنولوجيا الحيوية
- دراسات تغير المناخ
- مصادر الطاقة البديلة
- تحسين المحاصيل الزراعية
- تحليل الطب الشرعي
- إنشاء الأسلحة الحيوية

3.2 محاذاة السلاسل

الفكرة البديهية لعملية المحاذاة هي(alignment) إزاحة داخل إحدى السلسلتين ومقارنة عدد التطابقات في الأحرف وعدد الاختلافات. و أفضل محاذاة هي تلك التي تعطينا أكبر عدد من التشابه.

1.3.2 تعريف

بشكل عام هو طريقة لترتيب ومطابقة السلاسل. وعلى وجه التحديد بالنسبة للمعلوماتية الحيوية فهو يهتم بسلاسل الحمض النووي أو البروتين والهدف من ذلك هو تحديد مناطق التشابه التي قد تكون نتيجة علاقات للتشابه بين الوظائف الحيوية أو البنى أو علاقة تطورية بين السلاسل [10].

2.3.2 الهدف من محاذاة السلاسل

إن الهدف الأساسي من عملية المحاذاة هو تحديد مناطق الشبه في الوظائف والبنية بين السلاسل وفق بعض الشروط والقيود. و وفقا لقوانين بيولوجية وفيزيائية و كيميائية [6]. المثال الموالي يوضح عملية المحاذاة

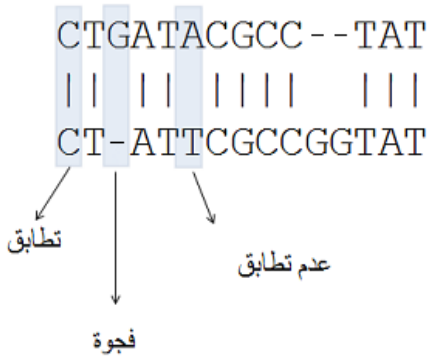
وعلى سبيل المثال لدينا سلسلتي حمض نووي ونود محاذاتهما بطريقة ما

السلسلة 1 : CTGATACGCCTAT

السلسلة 2 : CTATTCGCCGGTAT

شكل 1.2: سلسلتين نصيتين للحمض النووي

بعد القيام بعملية مطابقة تكون النتيجة كالتالي:



شكل 2.2: مثال عن محاذاة سلسلتي حمض نووي

ففي المثال أعلاه لدينا ثلاث حالات ممكنة نجدها في المحاذاة عموما بين سلسلتين:

- تطابق: وهو تماثل نفس الحرفين في السلسلتين
- عدم تطابق: وهو عدم التماثل في الحرفين المتقابلين
- الفجوة : وهي عملية إزاحة نتيجة لإضافة أو حذف. الغرض منها الحصول على تطابق أكثر وهي على المستوى البيولوجي تعني طفرة على السلسلة إما اضافة نكليوتيدة أو حذفها. ففي لحظة ما كانت السلسلتين متشابهيتن تماما فندخل مصطلح الفجوة للتعبير عن التغير الذي حدث على سبيل الإفتراض. وهي أيضا تمثل عملية إزاحة بإحدى السلسلتين غرض الحصول على توافق .

على سبيل المثال لدينا بروتين A. و من جهة أخرى بروتين B وبروتين C. ونود أن نعرف أيهما أقرب للبروتين A من حيث الوظائف والبنية. و أيهما له نسبة تشابه أكبر مع A. فهذا هو هدفنا من خلال عقد المقارنة بين البروتينات وهو البحث عن نسبة التشابه بينها.

ويوضح الشكل 3.2 مقارنة لثلاثة بروتينات

شكل 3.2: مقارنة البروتينات

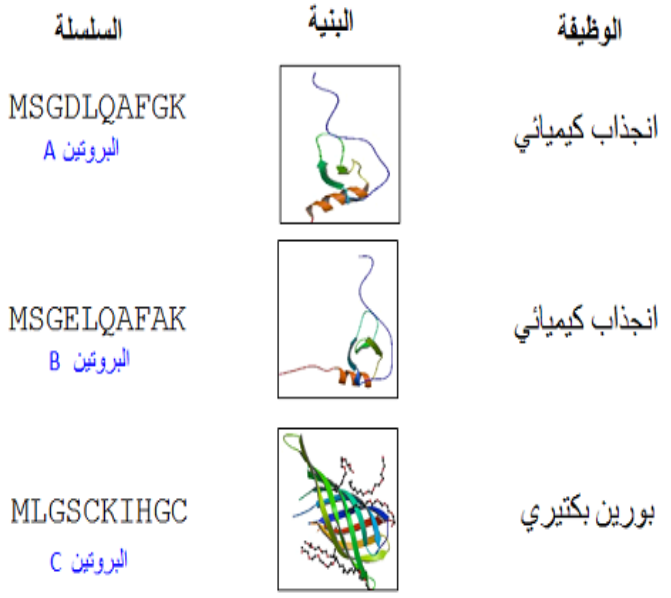
وبعد عقد المقارنة سنصل إلى نتيجة التي تتمثل في نسبة الشبه بين البروتينات والتي تعبر بدورها عن التقارب في وظيفة البروتين أو بنيته. وهذه النسبة هي حاصل قسمة عدد التطابقات على العدد الكلي لأحرف إحدى السلسلتين.

شكل 4.2: نسب تشابه البروتينات

ففي الشكل 4.2 نجد ان البروتين B أقرب بكثير للبروتين A بنسبة 80% منه للبروتين C بنسبة 30%. وذلك من خلال محاذاة الأحماض الأمينية لكل بروتين. كما يمكننا أيضا عقد المقارنة على مستوى سلاسل الاحماض الأمينية DNA أو RNA.

طبعا الأمر ليس مجرد مقارنة بغية المقارنة بل لمعرفة التقارب في الوظائف البيولوجية. فكلما كان التشابه أكبر على مستوى الأحماض الأمينية للبروتينات كلما كان التشابه على مستوى الوظائف الحيوية أكبر.

فكما هو موضح في الشكل 5.2 يكون التوافق والتشابه في السلسلة أكبر كلما كانت نسبة التشابه بين البنية أكبر. والعكس صحيح بحيث يكون التباعد في البنية والشكل واضح بشكل كبير.



شكل 5.2: التشابه في السلسلة و أثره على البنية والوظيفة

من هذا المثال يمكننا أن نستنتج أن عملية المحاذاة لها عدة تطبيقات من بينها [6, 10]

- معرفة الوظائف الحيوية المشتركة بين البروتينات أو سلاسل الأحماض الأمينية.
- معرفة البنى المتقاربة بين البروتينات.
- حساب المسافة التي تفصل البروتينات أو سلاسل الأحماض الأمينية.

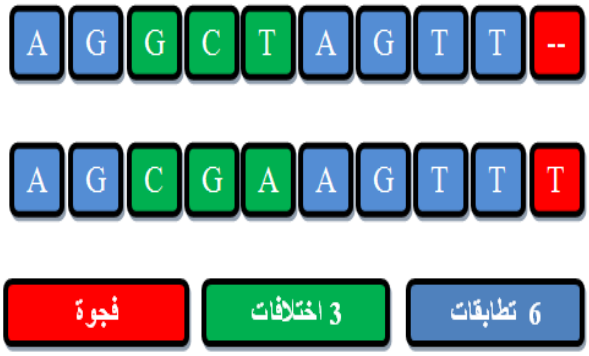
إن عملية المحاذاة لديها احتمالات عديدة. فهل يتوجب علينا مقارنة جميع الاحتمالات لمعرفة أفضل واحدة. وهل يمكننا اضافة الفجوات بشكل غير محدود أم أن هناك قيود بيولوجية. ولتوضيح ذلك نقترح الأمثلة التالية

في المثال التالي لدينا سلسلتي حمض نووي AGGCTAGTT و AGCGAAGTTT بحيث يمكن أن نقوم بعملية المحاذاة باحتمالات مختلفة ولكل نتيجتها.

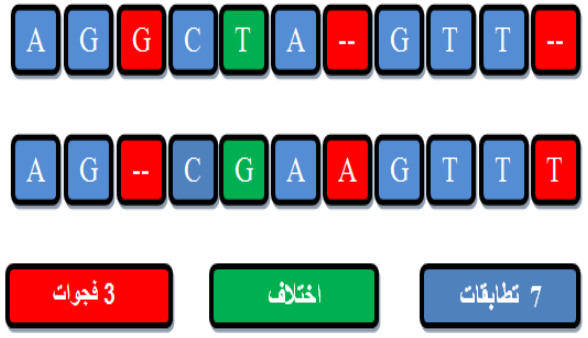
فالإحتمال الأول موضح بواسطة الشكل 6.2 :

والاحتمال الثاني موضح في الشكل 7.2 :

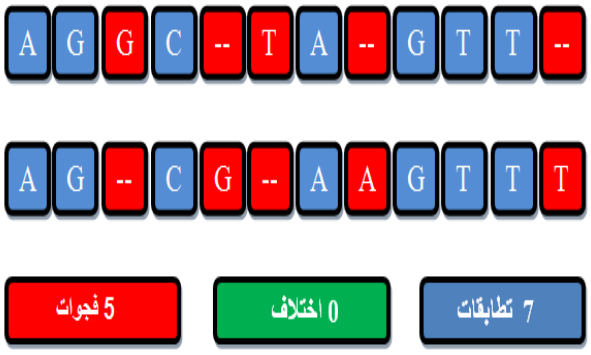
والإحتمال الثالث موضح في الشكل 8.2 :



شكل 6.2: أحد إحتمالات محاذاة السلسلتسن



شكل 7.2: أحد إحتمالات محاذاة السلسلتسن



شكل 8.2: أحد إحتمالات محاذاة السلسلتسن

لمعرفة أي محاذاة أفضل. نقوم بعملية التنقيط لكل محاذاة ونقارن بينها. التي لها أكبر قيمة تنقيط هي الأفضل. ونعطي كل عملية تطابق أو عدمه أو إضافة فجوة، نقطة محددة.

- على سبيل المثال لكل تطابق في المحاذاة قيمة 1 +
- لكل إختلاف قيمة 0
- لكل فجوة (إزاحة) قيمة 1-

وبهذا من خلال عملية رياضية بسيطة يمكننا حساب النقاط لكل عملية محاذاة . فعبارة حساب التنقيط لمحاذاة ما هي كالتالي: $S = \sum(Matches + Mismatches + Gaps)$

حيث S هي مجموع النقاط للمحاذاة، أي أن المحاذاة التي يكون لديها مجموع النقاط أكثر هي الأفضل. بينما تعبر Matches عن عدد التطابقات في أحرف السلسلتين. وتعبر Mismatches عن عدد الإختلافات في أحرف السلسلتين. كما تعبر Gaps عن عدد الفجوات التي تظهر عند محاذاة السلسلتين. و لكن هل سنقوم بحساب جميع الإحتمالات والتنقيط الخاص بها لتحديد المحاذاة الأفضل؟

أسهل طريقة: حساب وتنقيط كل إحتمالات المحاذاة الممكنة لكن عدد الإحتمالات الموجود أسي وهو يقدر ب: $\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \approx \frac{(2)^{2n}}{\sqrt{\pi n}}$ وهو حل غير تطبيقي مطلقا حتى للسلاسل القصيرة ما بالك بالسلاسل المتوسطة والطويلة. و لتوضيح ذلك نقترح المثالي التالي:

لدينا سلسلتين بنفس الطول الذي يقدر ب 100، إذن عدد إحتمالات المحاذاة الممكنة هو $10^{77} \approx$.

3.3.2 أنواع المحاذاة

و تنقسم إلى نوعين محاذاة زوجية(ثنائية) و متعددة:

1.3.3.2 محاذاة زوجية للسلاسل

محاذاة زوجية للسلاسل pair sequence alignment وهي عبارة عن عملية محاذاة مدخلاتها سلسلتين بغية تحديد مناطق الشبه بينها [10] .

كما هو موضح في المثال الموالي:

السلسلة 1 : T Y I W M R E A Q Y E

السلسلة 2 : T C I W M R E A - Y E

شكل 9.2: صورة توضيحية للمحاذاة الزوجية للسلاسل

2.3.3.2 محاذاة متعددة السلاسل

محاذاة متعددة السلاسل multiple sequence alignment وهي عبارة عن عملية محاذاة مدخلاتها أكثر من سلسلتين بغية تحديد مناطق الشبه المشترك لهذه السلاسل [10] .

كما هو موضح في المثال الموالي:

```
T Y I W M R E A Q Y E : 1 السلسلة
T C I W M R E A - Y E : 2 السلسلة
Y - I W M Q E V Q V E : 3 السلسلة
Y - I W M R E - Q Y E : 4 السلسلة
```

شكل 10.2: صورة توضيحية للمحاذاة المتعددة للسلاسل

4.3.2 طرق وخوارزميات المحاذاة

سنتناول في هذه المذكرة طرق وخوارزميات المحاذاة الزوجية للسلاسل. لأنها تمثل الحالة العامة للمحاذاة وأيضا الحالة الأساسية لأنواع المحاذاة [11] .

1.4.3.2 طريقة المصفوفة النقطية

طريقة المصفوفة النقطية DOT PLOT وهي الخوارزمية الأساسية في المحاذاة. تستخدم لمطابقة السلاسل الزوجية. وهي عبارة عن طريقة مطابقة رسومية بحيث في مصفوفة ذات بعدين السطر الأول منها يحتوي السلسلة الأولى والعمود الأول يحتوي السلسلة الثانية. ونجد داخل المصفوفة كل إحتمالات المطابقة الممكنة. و يمكننا تصفية الاحتمالات حسب طول الكلمات المرجوة [3] .

ويمكن وصف طريقة المصفوفة النقطية كالتالي [3] :

- نحدد طول الكلمة الأدنى.
- نقوم بعملية تصفية النقاط التي لا تحقق طول الكلمة المرجوة.
- نقوم بعملية ربط للسلاسل الجزئية الناتجة من عملية التصفية من أجل الحصول على محاذاة للسلسلتين.

ونقترح المثال التالي لتوضيح كيفية محاذاة السلسلتين ACGGAGCACGT و ACGGACACGT بطريقة المصفوفة النقطية .

الشكل 11.2 يوضح الحالة الإبتدائية للمصفوفة النقطية .

أما عملية التصفية حسب طول الكلمة المرجوة (1، 2، 3، 5) للمصفوفة النقطية موضحة في الأشكال 12.2 - 16.2 .

	A	C	G	G	A	C	A	C	G	T
A	X				X		X			
C		X				X		X		
G			X	X					X	
G			X	X					X	
A	X				X		X			
G			X	X					X	
C		X				X		X		
A	X				X		X			
C		X				X		X		
G			X	X					X	
T										X

شكل 11.2: شكل توضيحي للحالة الإبتدائية للمصفوفة النقطية

	A	C	G	G	A	C	A	C	G	T
A	X				X		X			
C		X				X		X		
G			X	X					X	
G			X	X					X	
A	X				X		X			
G			X	X					X	
C		X				X		X		
A	X				X		X			
C		X				X		X		
G			X	X					X	
T										X

طول الكلمة = 1

X نقاط معزولة

X نقطتين

X ثلاث نقط

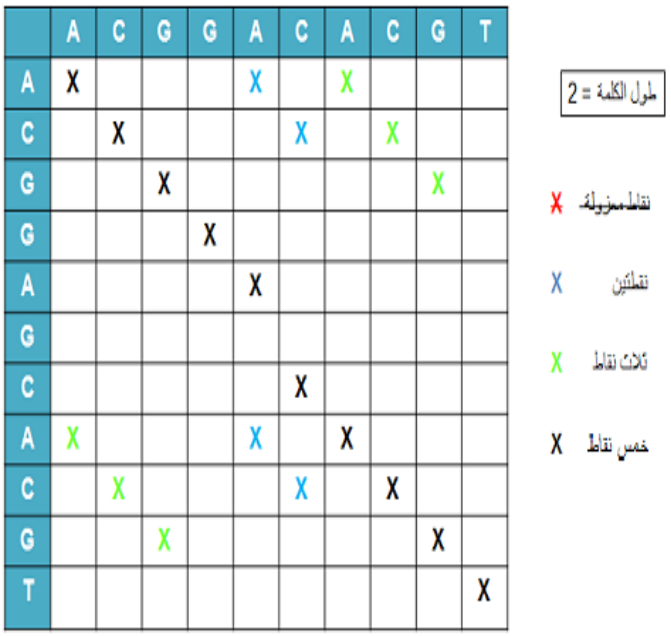
X خمس نقاط

شكل 12.2: شكل توضيحي لتصفية المصفوفة النقطية بكلمة ذات طول يساوي 1

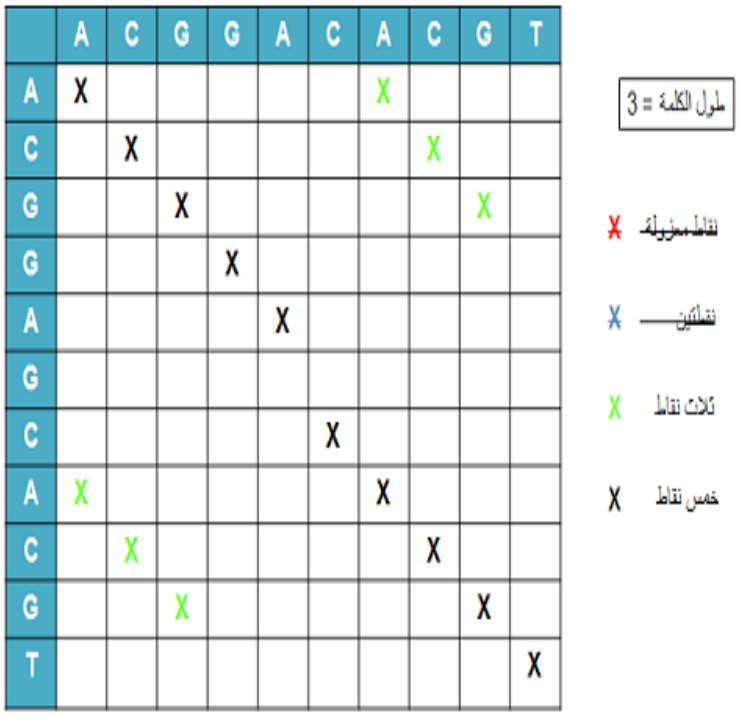
أما الشكل 16.2 يوضح عملية المحاذاة الرسومية حسب طول الكلمة الذي يختاره الخبير. في هذه الحالة طول الكلمة يساوي 5.

من بين إيجابيات هذه الطريقة نذكر [8] :

- مفيد جدا في عملية التجارب وعملية الملاحظة الرسومية بالعين المجردة
- مفيد في حالة التشابه الكبير والواضح
- يمكنه إظهار كل من التشابه العام والتشابه المحلي



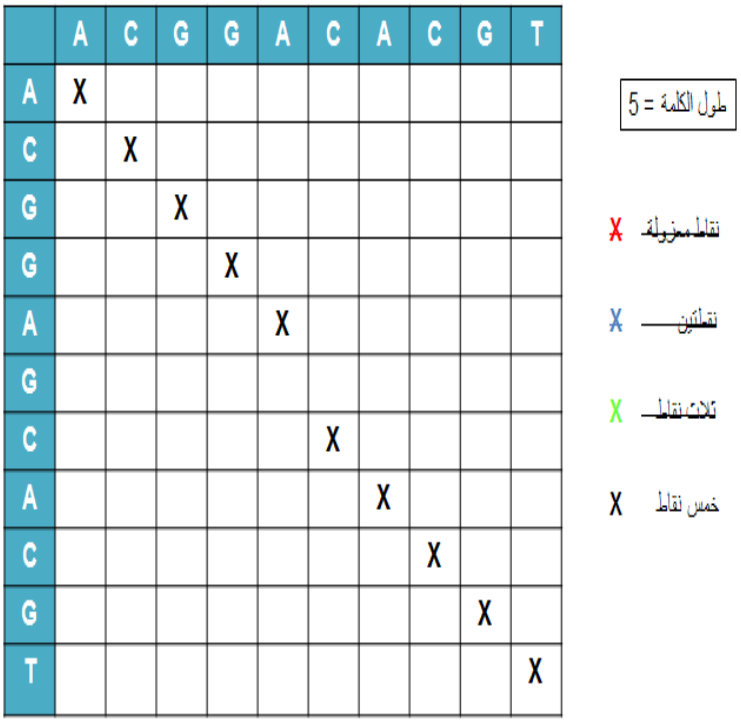
شكل 13.2: شكل توضيحي لتصفية المصفوفة النقطية بكلمة ذات طول يساوي 2



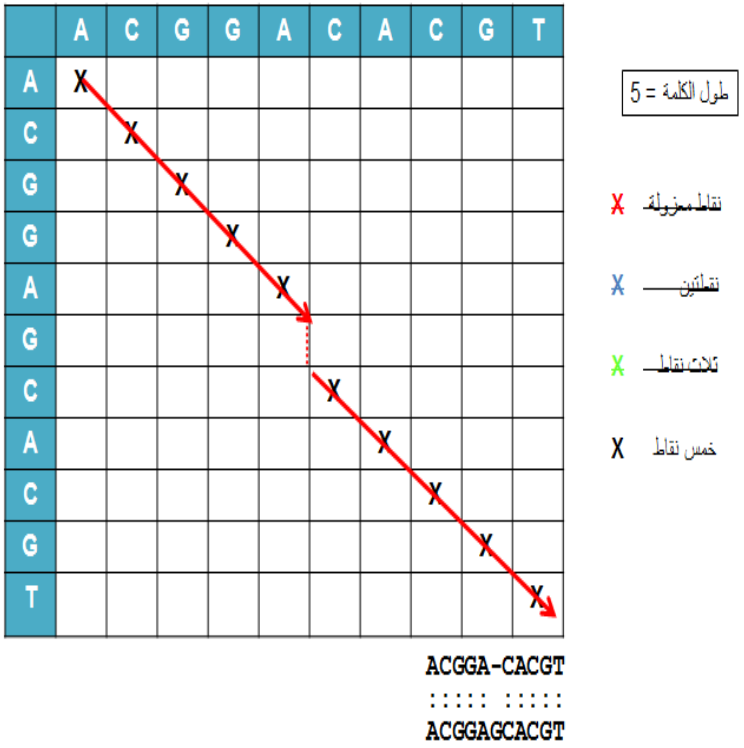
شكل 14.2: شكل توضيحي لتصفية المصفوفة النقطية بكلمة ذات طول يساوي 3

بينما في الجانب الآخر لها سلبيات نذكر منها:

- غير مفيد في السلاسل الطويلة و التداخل والتشويش
- هو طريقة رسومية ولا يعطينا طريقة محددة لتحديد التطابق بين السلسلتين



شكل 15.2: شكل توضيحي لتصفية المصفوفة النقطية بكلمة ذات طول يساوي 5



شكل 16.2: شكل توضيحي لتصفية المصفوفة النقطية بكلمة ذات طول يساوي 5 والقيام بعملية محاذاة للكلمات ذات الطول 5

- لا يقوم بالتحليل بل إعطاء نتيجة رسومية والعين المجردة هي من تقوم بعملية التحليل

5.3.2 طريقة البرمجة الديناميكية

تستعمل طريقة البرمجة الديناميكية في عملية محاذاة السلاسل بغية تحديد التنقيط الأمثل لمحاذاة سلسلتين.

1.5.3.2 البرمجة الديناميكية

هي طريقة لحل مسائل معقدة عن طريق تقسيمها لمسائل فرعية أبسط في مجال الرياضيات أو الحاسوب.
و الفكرة وراء البرمجة الديناميكية هي أننا لحل مسألة ما، نحن بحاجة إلى حل أجزاء مختلفة من المسألة (مسائل فرعية). ومن ثم جمع حلول المسائل الفرعية للحصول على حل شامل في كثير من الأحيان. كثير من هذه المسائل الفرعية هي في الواقع متشابهة [10] .

الهدف من البرمجة الديناميكية هو البحث عن حل كل مسألة فرعية مرة واحدة فقط. وبالتالي تقليل عدد الحسابات فكلما تم حساب حل مسألة فرعية ما، يتم حفظه في المرة القادمة عند الحاجة للحل نفسه يتم ببساطة استرجاعه.

عندما تطبق هذه الطريقة فإنها تستغرق وقت أقل مما تستغرقه الطرق الأخرى التي ليس لها ميزة حل المسائل الثانوية المتداخلة.

2.5.3.2 إستعمال البرمجة الديناميكية في المحاذاة

نستعمل البرمجة الديناميكية كطريقة لتحديد التشابه الأمثل بين سلسلتين عن طريق مطابقتهما من أجل جميع الأزواج الممكن للحروف بين السلسلتين.
في مصفوفة ذات بعدين السطر الأول منها يحتوي السلسلة الأولى والعمود الأول يحتوي السلسلة الثانية. وبتحويلها إلى مصفوفة تنقيط لتحديد التطابق في أحرف السلسلتين والإختلاف بينهما و في نهاية المطاف التنقيط الأعلى هو من يحدد التطابق الأمثل.

فليكن لدينا سلسلتين CACGA و CGA على سبيل المثال. لدينا ثلاث إحتمالات للوضعية الأولى من المطابقة. وذلك باستعمال سلم التنقيط التالي :

- التطابق = 1+
- الإختلاف = 1-
- عقوبة الفجوة = 2-

السلسلة المتبقية	التنقيط	أول وضعية
ACGA GA	+1	C C
CACGA GA	-2	- C
ACGA CGA	-2	C -

شكل 17.2: إحتمالات الوضعية الأولى للمحاذاة في البرمجة الديناميكية للسلسلتين CACGA و CGA

ويجدر الإشارة أنه يمكن التمييز بين نوعين من طرق البرمجة الديناميكية ألا وهي المحاذاة العامة والمحلية.

6.3.2 المحاذاة العامة

تندرج ضمن المحاذاة الثنائية وهدفها الحصول على تشابه أمثل لسلسلتين من خلال طولهما الإجمالي. ويجب أن يمتد من بداية إلى نهاية السلسلتين بحيث يكون الطول النهائي للسلسلتين متساو مع تحقيق أعلى درجات التنقيط [3] .

ويمكن ترجمة ذلك بواسطة العلاقة التراجعية التالية:

$$F(i,j) = max \begin{cases} F(i-1,j-1) + s(x_i, y_i) & (1) \\ F(i-1,j) + g & (2) \\ F(i,j-1) + g & (3) \end{cases}$$

بحيث $F(i,j)$ تعبر عن التنقيط الأمثل لمحاذاة السلسلتين في الموضع (i,j)، ففي حالة التطابق أو عدمه للحرفين في الموضع (i,j)، الممثل في المعادلة الأولى، فإننا نقوم بإضافة $s(x_i, y_i)$ للتنقيط الأمثل في الموضع $(i-1,j-1)$، وكما ذكرنا سابقا فإن التطابق بين لحرفين له تنقيط معين وكذلك عدم التطابق.

بينما تمثل المعادلة الثانية عملية إضافة في السلسلة x. المتمثلة في العمود الأول من المصفوفة.

فيما تمثل المعادلة الثانية عملية اضافة على مستوى السلسلة y. المتمثلة في السطر الأول من المصفوفة.

بينما تعبر g عن معامل الفجوة في المحاذاة العامة.

في النهاية نختار أكبر قيمة ل $F(i,j)$

للمعادلات الثلاث.
وفيما يلي نلخص خطوات خوارزمية المحاذاة العامة [3] :

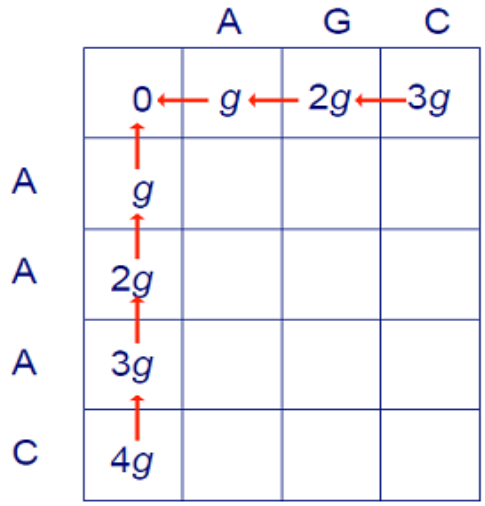
- لدينا n حرف في السلسلة x و m حرف في السلسلة y
- ننشأ مصفوفة F ذات البعد (n+1) * (m+1)
- $F(i,j)$ = تنقيط أحسن محاذاة ل x[1...i] و y[1...j]
- تهيئة السطر الأول والعمود الأول من المصفوفة
- ملء بقية قيم المصفوفة وفقا للعلاقة التراجعية من الأعلى للأسفل، من اليسار لليمين
- من أجل كل $F(i,j)$ نقوم بحفظ مسار الخلايا المسببة لأفضل تنقيط
- F(m,n) تحتوي تنقيط أفضل محاذاة، ثم بإتباع مساراالخلايا رجوعا من F(m,n) نحو F(0,0)

ولتوضيح طريقة عمل الخوارزمية نقترح المثال التالي:

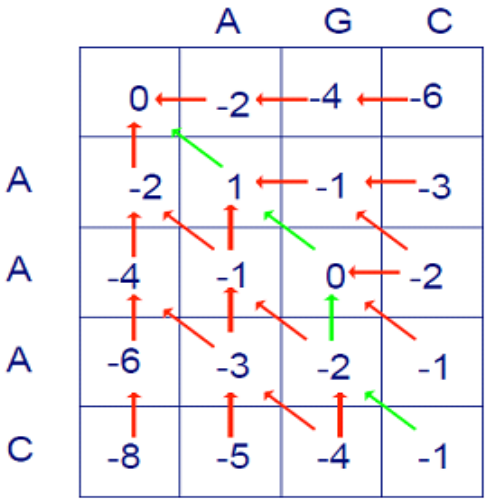
نفترض لدينا سلم التنقيط الآتي:

- التطابق = 1+
- الإختلاف = 1-
- عقوبة الفجوة = 2-

الأشكال 18.2 و 19.2 يوضحان طريقة عمل خوارزمية المحاذاة العامة .



شكل 18.2: الحالة الإبتدائية لخوارزمية المحاذاةالعامة



شكل 19.2: الحالة النهائية لخوارزمية المحاذاة العامة

فتكون أفضل محاذاة بعد تطبيق الخوارزمية كالتالي:

X=AAAC

Y=AG-C

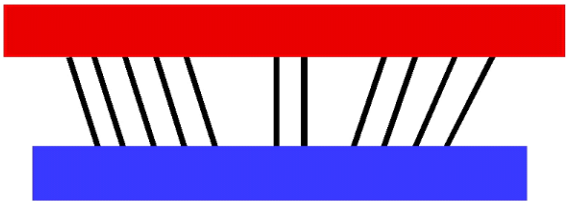
نذكر من أهم إيجابيات هذه الخوارزمية:

- جيدة للسلاسل المتقاربة في الطول
- سهلة نوعا ما بحيث تعتمد على الفجوات

كما نذكر أهم سلبياتها وتتمثل في:

- غير مناسبة للسلاسل ذات الطول المختلف بشكل واضح
- في بعض الحالات لا يمكنها كشف بعض أنواع التشابه على سبيل المثال بداية السلسلة الأولى تشبه نهاية السلسة الثانية ونهاية السلسلة الثانية تشبه بداية السلسلة الأولى
- مطابقة السلاسل وفق قيد الإتجاه
- أحيانا لا يعكس المعنى البيولوجي

كما نوضح في الشكلين 20.2 و 21.2 شكل المحاذاة العامة .



شكل 20.2: صورة توضيحية للمحاذاة العامة

```
1 TGTCGATTAAGCGGTCGTAGCTGACCTGAGATTGCCCGATGGCGTAGTAGCTGACC 56
  ||||||||| |||||||||| |||||||| || ||||||||||||||| |||||
1 TGTCGATTATGCGGTCGTAG--GACCTGAGTTTCCCCGATGGCGTAGTAGGTGACC 54
```

شكل 21.2: مثال توضيحي للمحاذاة العامة

كما نلاحظ في الشكل 21.2 فالمحاذاة العامة جيدة في حالة الطول المتقارب والشبه الكبير بإضافة بعض من الفجوات.

7.3.2 المحاذاةالمحلية

يندرج ضمن المحاذاة الثنائية وهو يهتم بالحصول على أفضل تشابه لمناطق جزئية من السلسلتين وأهم مميزتين له [3]

- الكشف عن التشابه الأعظم لمناطق محلية من السلسلة.
- لا يهتم لا بإتجاه الشبه ولا بالطول الإجمالي.

و يمكن ترجمة ذلك بواسطة العلاقة التراجعية التالية:

$$F(i,j) = max \begin{cases} F(i-1,j-1) + s(x_i, y_i) \\ F(i-1,j) + g \\ F(i,j-1) + g \\ 0 \end{cases}$$

وفيما يلي نلخص خطوات خوارزمية المحاذاة المحلية [3] :

- تهيئة السطر الأول والعمود الأول وملئها بقيمة 0.
- ملء بقية قيم المصفوفة وفقا للعلاقة التراجعية من الأعلى للأسفل من اليسار لليمين.
- إيجاد أكبر قيمة في المصفوفة F(i,j) حيثما كانت.
- إنطلاقا من هاته الخانة نتتبع المسار ونتوقف عند أول خانة تحمل قيمة 0.

نفترض لدينا سلم التنقيط الآتي:

- التطابق = +1
- الإختلاف = -1
- عقوبة الفجوة = -2

الشكل 22.2 يوضح طريقة عمل خوارزمية المحاذاة المحلية .

		A	A	G	A
	0	0	0	0	0
T	0	0	0	0	0
T	0	0	0	0	0
A	0	1	1	0	1
A	0	1	2	0	1
G	0	0	0	3	1

شكل 22.2: آلية عمل خوارزمية المحاذاة المحلية

فبعد تطبيق هذه الخوارزمية نحصل على أفضل مطابقة محلية كالآتي:

X=AAG

y=AAG

فلهذه الخوارزمية إيجابيات نذكر منها:

- جيدة للكشف عن التشابه في مناطق محلية مختلفة عبر السلسلتين
- جيدة للسلاسل المختلفة في الطول
- تحرر من محاذاة السلاسل وفقا للإتجاه
- يصف التشابه بين الوظائف الحيوية للبروتينات

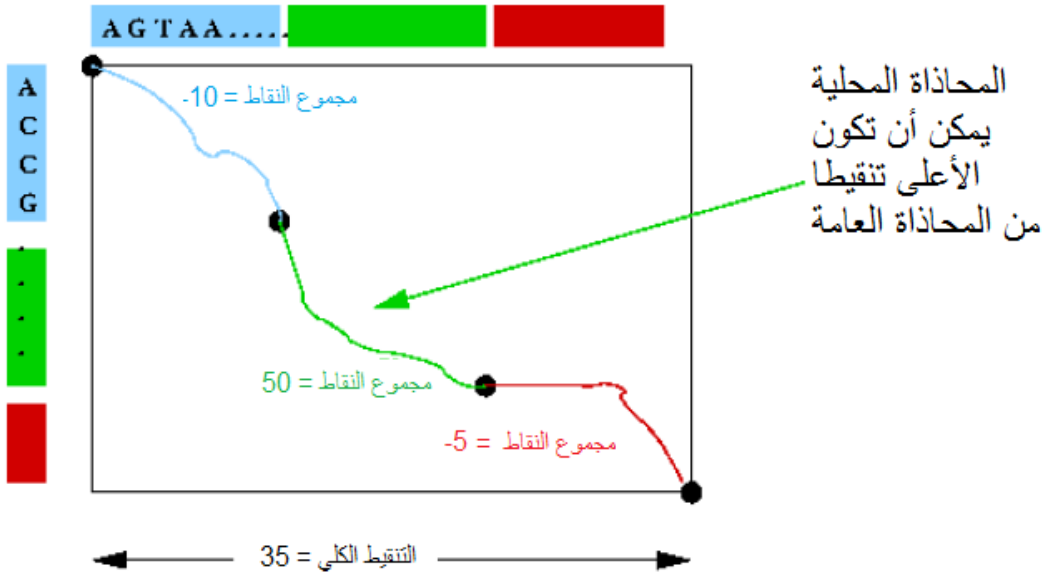
كما نوضح في الشكل 23.2 تحرر المحاذاة من قيد الإتجاه .

شكل 23.2: شكل يبرز تحرر الشبه المحلي من قيد الإتجاه

شكل 24.2: شكل يوضح المناطق الوظيفية

و في الشكل 24.2 بخصوص الشكل العام. يمكننا أن نجزم أنه لا شبه بين المفاتيح لكن بخصوص المكان الوظيفي الذي يفتح القفل فهناك تشابه. بحيث هذا المثال يوضح أن المحاذاة المحلية تكشف عن المناطق الوظيفية للسلاسل. وهو أيضا يعطينا فكرة عن المناطق الوظيفية في البروتينات التي يمكن للمطابقة المحلية الكشف عنها بين بروتينين.

فالشكل 25.2 يبرز أن إستعمال المحاذاة العامة قد لا يعطي التنقيط الأمثل، وقد يهمل تنقيطا أمثل بين أجزاء في السلسلتين. و أن المحاذاة المحلية تعطي تنقيطا أفضل من المحاذاة العامة.

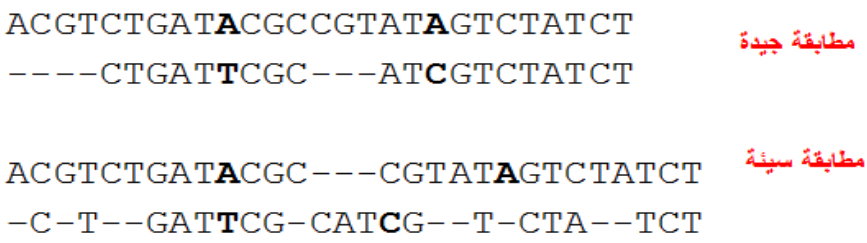


شكل 25.2: شكل يبرز مدى أهمية الشبه المحلي

إن خوارزمية المحاذاة المحلية عند اسقاطها على التفسير البيولوجي لها عدة مميزات :

- مفيدة جدا في حالة مقارنة البروتينات لمعرفة النطاق أو المقطع المشترك والذي يمثل لإشتراك في وظيفة حيوية للبروتين.
- مفيد في حالة مقارنة الحمض النووي بحيث يمكنه إكتشاف المقاطع المشتركة بين السلاسل.
- يعبر عن المعنى أو الشبه البيولوجي بشكل أوضح.

بالإضافة إلى هذا يمكننا أن نلاحظ بأن الفجوات شيء حساس للغاية في المحاذاة فهو كالملح في الأكل قليل منه يجعل الأكل طيب والكثير منه يفسده وتفسيره أنه يلغي المعنى البيولوجي للمطابقة. فمن غير المعقول أن تصاب سلسلة ما بطفرات مبالغ فيها.
فالكثير من الفجوات يفرغها من معناها البيولوجي ويجعلها عملية لحصول على أكثر عدد من المطابقات وبالتالي أكبر قيمة من عدد النقاط.
الشكل 26.2 يبرز مدى تأثير الفجوات على نتيجة المحاذاة .



شكل 26.2: تأثير الفجوات على المحاذاة

إن لكل محاذاة شروطها التي تعمل عليها فالعامة من بين أهم شروطها التقارب في الطول و الشبه. بينما المحلية تعالج المختلفة في الطول والأقل تشابها.

ففي الشكل 27.2 بعد تطبيق نوعي المحاذاة السابقتين (العامة والمحلية) على زوجي سلاسل مختلفة، فإننا نلاحظ أن المحاذاة المحلية تعطي دوما أفضل تنقيط. وذلك لأن المحاذاة العامة تجبر السلسلتين المختلفتين في الطول أو الشبه على التساوي في الطول في النتيجة النهائية عكس المحاذاة المحلية التي تحدد أكبر سلسلة جزئية. وبهذا فإن المحاذاة العامة في حالة الإختلاف في الطول أو الشبه فإنها تضيف فجوات من أجل ذلك. وبذلك يفقد المعنى البيولوجي للشبه بين السلسلتين. وإن كان هناك تنقيط أعظم لمناطق جزئية فإن المحاذاة العامة تجعله غير ظاهر في نهاية المطاف عكس المحاذاة المحلية.

8.3.2 تعقيد خوارزميات البرمجة الديناميكية

بالنسبة للمحاذاة العامة و المحلية فإن كلتاهما لهما نفس تعقدي الذاكرة بحيث اذا كان n طول السلسلة الأولى و m طول الثانية و كل خلية من المصفوفة لديها s بت كقدرة تخزين. فإننا بحاجة إلى $s \times n \times m$ بت لتخزين المصفوفة. أي أن تعقيد الذاكرة لكلتا الخوارزميتين هو $O(n \times m)$.

اذا كانت لدينا سلسلتي حمض نووي لنوعين من البكتيريا كل سلسلة فيها مليون قاعدة نيتروجينية. فنحن بحاجة إلى ذاكرة بسعة $s \times 10^6 \times 10^6 = s \times 10^{12}$ بت, حيث s سعة تخزين الخلية.

لكن إذا اخذنا بعين الإعتبار جينوم الانسان فإن عدد قواعده فهو أضعاف قواعد الجينوم البكتيري حوالي 3000 مرة أي $9 \times s \times 10^{18}$ بت .

PIR Entry				Similarity Score		
				Global		Local
				End Penalty	No End Penalty	
HBHU	vs	HBHU	Hemoglobin beta-chain—human	725	725	725
		HAHU	Hemoglobin alpha-chain—human	314	320	322
		MYHU	Myoglobin—Human	121	164	166
		GPYL	Leghemoglobin—Yellow lupin	8	28	43
		LZCH	Lysozyme precursor—Chicken	−107	16	32
		NRBO	Pancreatic ribonuclease—Bovine	−124	16	31
		CCHU	Cytochrome c—Human	−160	10	26
MCHU	vs	MCHU	Calmodulin—Human	671	671	671
		TPHUCS	Troponin C, skeletal muscle	395	430	438
		PVPK2	Parvalbumin beta—Pike	−57	103	115
		CIHUH	Calpain heavy chain—Human	−2085	89	100
		AQJFNV	Aequorin precursor—Jelly fish	−65	48	76
		KLSWM	Calcium binding protein—Scallop	−89	45	52
QRHULD	vs	EGMSMG	Epidermal growth factor precursor	−591	475	655

شكل 27.2: مقارنة التنقيط النهائي للخوارزميتين

يعني نحن بحاجة إلى ذاكرة ضخمة لكن لا يمكن التعامل معها مباشرة لذلك توجب علينا التعامل مع ملفات وليس الذاكرة مباشرة.

أما التعقيد الخوارزمي بالنسبة للوقت لخوارزميتي المحاذاة العامة والمحلية فإنه اذا كان n طول السلسلة الأولى و m طول الثانية وكل خلية تحتاج s عملية فإننا بحاجة إلى $s \times n \times m$ من العمليات. أي أن تعقيد الوقت لكلتا الخوارزميتين هو $O(n \times m)$.

اذا كانت لدينا سلسلتي حمض نووي لنوعين من البكتيريا كل سلسلة فيها مليون قاعدة نيتروجينية. فنحن بحاجة إلى وقت يقدر ب : $s \times 10^6 \times 10^6 = s \times 10^{12}$.

لكن إذا اخذنا بعين الإعتبار جينوم الانسان فإن عدد قواعده هو أضعاف قواعد الجينوم البكتيري حوالي 3000 مرة أي $9 \times s \times 10^{18}$ وحدة.

لو إعتبرنا أننا نملك معالجا يمكنه تنفيذ 10^6 عملية في الثانية فإننا بحاجة 10^{12} ثانية لإنهاء عمل هذه الخوارزمية أي ما يعادل 31688 سنة.

رغم أن تعقيد الخوارزمية من ناحية التوقيت كحل تربيعي يعتبر مقبولا في أغلب مسائل الإعلام الآلي. إلا أنه في حالتنا يعتبر حلا غير تطبيقي نهائيا.

لكن في حالة أخذنا سلسلتين بطول 10^5 أي أنها تستغرق 10^4 ثانية بنفس المعالج السابق أي 2,78 ساعة تقريبا. ومنه هذه الخوارزميات جيدة في حالة السلاسل المصفوفات ذات الحجم الصغير وغير تطبيقية في حالة السلاسل الطويلة.

رغم أن خوارزميات المحاذاة العامة والمحلية لديها مزايا جيدة ومفيدة لتحديد الشبه بين السلاسل لكننا أمام عائق أنها غير تطبيقية البته. لذلك سنحاول اللجوء إلى ما يعرف بالحلول التقريبية لأن الحل الأمثل غير تطبيقي من ناحية الوقت. الذي سنتطرق له لاحقا.

9.3.2 محاذاة سلاسل البروتين

في الحقيقة التطابق له قيمة ثابة في حالة الحمض النووي. لكن في حالة سلاسل البروتين ليس بالأمر البسيط. فكما ذكرنا سابقا يوجد عشرون حمضا أمينيا ولكل منها خصائصه الكيميائية. قد يكون هناك إختلاف في الحرف لكن هناك تشابه في الخصائص الكيميائية فنحن لسنا نتعامل مع كتاب أو سلسلة أحرف جافة بل مع منظومة حية و تفاعلات وخصائص كيمياية. لذلك وجب علينا الأخذ بعين الإعتبار هذه الخائص البيولوجية للأحماض الأمينية والبروتينات في عملية محاذاة البروتينات.

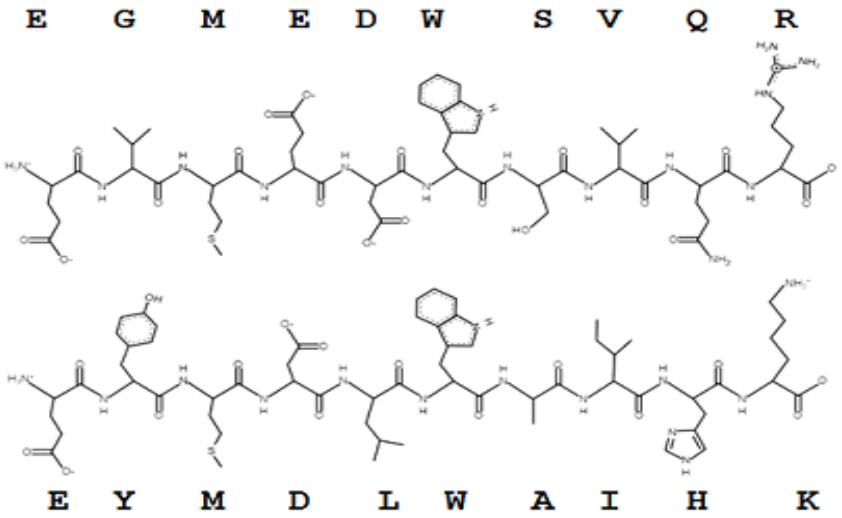
فليس كل التطابقات أو الإختلافات لها نفس القيمة والتنقيط. لذلك لجأ علماء الأحياء لإنشاء مصفوفة التنقيط النموذجية فيما بين جميع الأحماض الأمينية والإعتماد عليه في حالة البحث عن أفضل مطابقة في حالة سلاسل الأحماض الأمينية.

الأشكال 28.2 و 29.2 تعبر عن الفرق بين محاذاة البروتين على شكل أحرف جافة وعلى شكل خصائص كيميائية. هذه الخصائص التي يمكننا وصفها ممن خلال الشكل 30.2 الذي بعبر عن تنقيط مطابقات جميع الأحماض الأمينية مع بعضها البعض.

```
EGMEDWSVQR
: :   :
EYMDLWAIHK
```

تطابق = ؟
عدم تطابق = ؟
فجوات = ؟

شكل 28.2: مثال لمحاذاة البروتينات



شكل 29.2: الموضع الكيميائي للأحماض الأمينية في المحاذاة

إن التنقيط يختلف بين أنواع السلاسا البيولوجية فبالنسبة للحمض النووي DNA يكون التنقيط ثابتا. وعلى سبيل المثال التنقيط التالي

- التطابق = 1+
- الإختلاف = 3-
- عقوبة الفجوة = 5-

أما بالنسبة لسلاسل البروتين

- التطابق والإختلاف نستعمل مصفوفة Blossum62 لتنقيط الأحماض الأمينية أو أي نموذج آخر
- عقوبة الفجوة = 11 –

حيث الشكل 30.2 يمثل مصفوفة Blossum62 .

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

شكل 30.2: مصفوفة Blossum62 لمحاذاة البروتينات

إذ أن تقابل الحمضين E و D في التنقيط السابق سيكون 3- كونه اختلافا بين الحرفين لكن إذا أخذنا بعين الإعتبار الخصائص الكيميائية والتقارب في الوظائف والشكل فإن مصفوفة التنقيط النموذجية تنقط هذا التقابل بين الحرفين ب 2 وهنا يكمن الفرق بين هذين النظامين في التنقيط.

و المثال الموالي يعبر عن حساب تنقيط محاذاة سلسلتي بروتين

السلسلة 1 : V D S – C Y

السلسلة 2 : V E S L C Y

مجموع التنقيط = 4+ 2+ 4- 11+ 9+ 7

= 15

شكل 31.2: مثال عن محاذاة سلسلتي بروتين

4.2 الخوارزميات التجريبية

بغية حل مشكل تعقيد الوقت في حالة محاذاة السلاسل الجينومية الطويلة فإننا نستعمل الخوارزميات التجريبية.

1.4.2 مفهوم الخوارزميات التجريبية

إن مفهوم الخوارزميات التجريبية heuristic algorithms يستخدم في الخوارزميات التي تقترح حلا من الحلول الممكنة، لكنها لا تضمن أن يكون الأمثل. ومن هذا المفهوم يمكن أن نعتبرها خوارزميات تقريبية وليست بالدقيقة. في العادة هذه الخوارزميات تجد حلا أقرب للمثالي وغالبا ما تكون هذه الطريقة سريعة وسهلة. لكن في بعض الأحيان يمكن لها أن تجد الحل الأمثل لكن تبقى تجريبية إلى أن يتم إثبات أنه الحل الأمثل. على سبيل المثال الخوارزميات الجشعة greedy algorithms .ومن اسمهما نستنتج أنها تبحث عن أسرع طريقة لايجاد الحل [10].

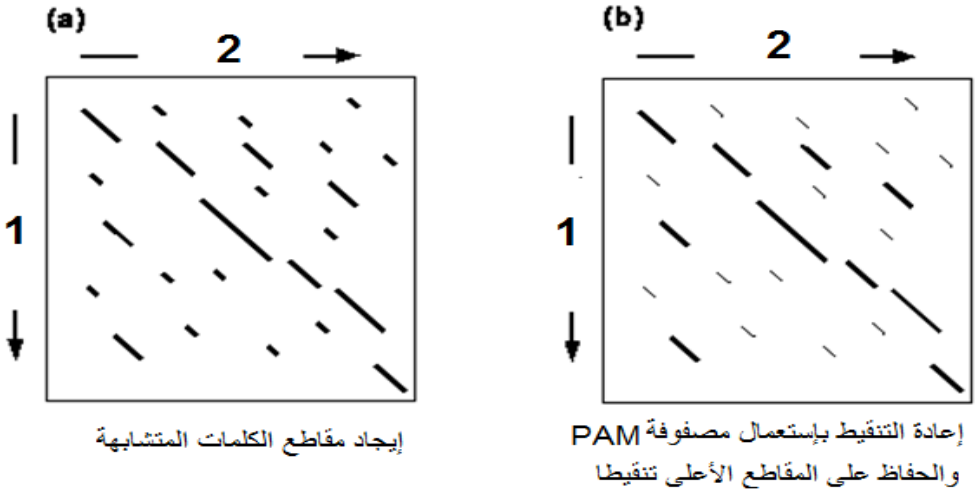
الخوارزميات التجريبية تعتبر طريقة مساعدة في حل المسائل عندما نكون الطرق التقليدية بطيئة. أو لإيجاد حلول تقريبية عندما تفشل الطرق التقليدية في إيجاد حل محدد و الهدف منها هو الحصول على حل في وقت منطقي بحيث يكون جيدا بما فيه كفاية وتطبيقي. ومن أشهر الأمثلة هي مسألة البائع المتجول.

2.4.2 خوارزمية FASTA

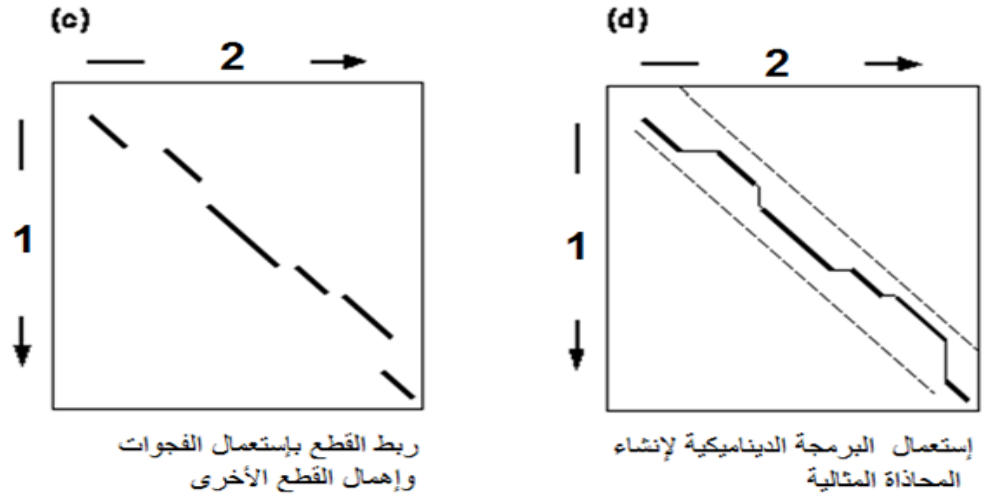
مشتقة من الطريقة الذي تعمل به Dot. plot فهذه الخوارزمية تبحث عن الشبه الدقيق بين السلسلتين غالبا ما تستعمل طول الكلمة للحمض النووي 6 نيكليوتيدات اما بالنسبة للبروتين فهو حمضين أمينيين.
وخطوات الخوارزمية يمكن وصفها بالترتيب التالي [3] :

- تحديد الكلمات المشتركة ذات طول محدد إصطلاحا (k-word) بين السلسلتين.
- تنقيط الأقطار باستعمال الطول المحدد للكلمة للتطابقات وتحديد أفضل 10 تطابقات.
- إعادة تنقيط المناطق الإبتدائية بإستعمال مصفوفة التنقيط التعويضية.
- ربط المناطق الإبتدائية بإستعمال الفجوات والعقوبات لها.
- إستعمال البرمجة الديناميكية للبحث عن الأمثل بينها.

الأشكال 32.2 33.2 توضح عمل خوارزمية FASTA .



شكل 32.2: خطوات خوارزمية FASTA



شكل 33.2: خطوات خوارزمية FASTA ، المرحلتين الأخيرتين

كما يجدر الإشارة أن هذه الخوارزمية تحل مشكل البرمجة الديناميكية بخصوص السلاسل الطويلة.

5.2 تجميع الأنواع والسلاسل

إننا وبعد تطبيق إحدى خوارزميات المحاذاة بغية حساب الشبه بين السلاسل. سنتحصل على مصفوفة الشبه بين السلاسل. وعلى سبيل الفرض سنوضح في هذا القسم كيف يمكن إستعمال نتائج مصفوفة الشبه. حيث يمكن استخدامها من أجل تصنيف السلاسل أو التجميع. ومن بين أشهر تطبيقاتها إكتشاف مصدر الفيروسات أي إلى أي عائلة تنتمي السلالة المجهولة

و ذلك بحساب أقرب السلالات لها عن طريق عملية المحاذاة التي تعبر عن مقياس للشبه. سنوضح من خلال هذا الفصل كيف يمكن استخدامها من أجل عملية التجميع بغية تحديد السلالات والأنواع.

على سبيل الفرض الشكل 34.2 يمثل مصفوفة شبه بين مجموعة من السلالات.

	A	B	C	D	E
سلالة A	------	0.16	0.36	0.53	0.74
سلالة B		--------	0.11	0.44	0.61
سلالة C			--------	0.56	0.12
سلالة D				----------	0.31
سلالة E					---------

شكل 34.2: مصفوفة شبه بين مجموعة من السلالات

بحيث نعتمد خوارزمية K-means لتجميع السلالات. وخطوات الخوارزمية على النحو التالي:

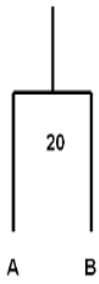
- بناء مصفوفة المسافات بين الكائنات أو البروتينات
- تجميع الكائنين الأقصر مسافة
- حساب المسافة المتوسطة بين المجموعة الجديدة والكائنات المتبقية
- إعادة الخطوات لكن بمعطيات المجموعة الجديدة
- التوقف في حالة مجموعة واحدة أو العدد المحدد

ونقترح المثال التالي لتوضيح الخوارزمية:

بحيث الشكل 35.2 يعبر عن مصفوفة الشبه الإبتدائية بين السلالات

	A	B	C	D	E
سلالة A	------	20	60	100	90
سلالة B		--------	50	90	80
سلالة C			--------	40	50
سلالة D				----------	30
سلالة E					---------

شكل 35.2: مصفوفة الشبه الإبتدائية بين السلالات

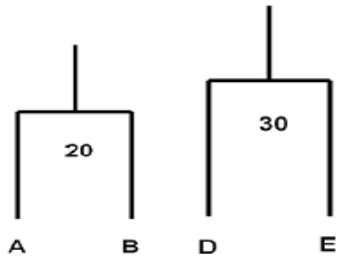


شكل 36.2: حساب المسافة بين السلالة A والسلالة B

أصغر مسافة هي بين الكائن A و B = 20
حساب متوسط المسافة بين AB و C
متوسط المسافة بين AB و C = (المسافة بين A و C + المسافة بين B و C) / 2
متوسط المسافة بين AB و C = (50 + 60) / 2 = 55
متوسط المسافة بين AB و D = (90 + 100) / 2 = 95
متوسط المسافة بين AB و E = (90 + 80) / 2 = 85

E	D	C	AB	
85	95	55	--------	سلالة AB
50	40	--------		سلالة C
30	----------			سلالة D
---------				سلالة E

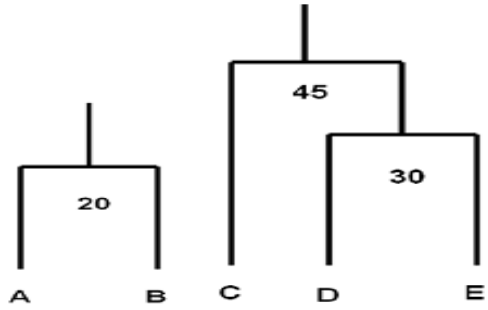
شكل 37.2: حساب مصفوفة الشبه الجديدة بعد تجميع A و B



شكل 38.2: المرحلة الثانية من عملية التجميع

DE	C	AB	
90	55	--------	سلالة AB
45	--------		سلالة C
----------			سلالة DE

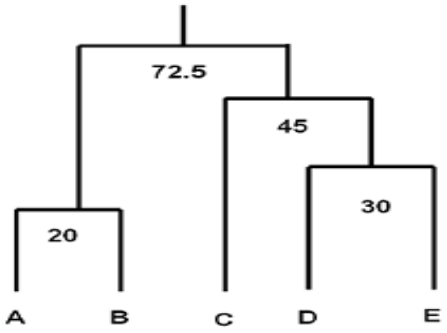
شكل 39.2: المصفوفة الجديدة الناتجة بعد المرحلة الثانية



شكل 40.2: المرحلة الثالثة من عملية التجميع

CDE	AB	
72.5	--------	سلالة AB
----------		سلالة CDE

شكل 41.2: المصفوفة الجديدة الناتجة بعد المرحلة الثالثة



شكل 42.2: النتيجة النهائية لعملية تجميع السلالات

6.2 خاتمة

يجب مراعاة أن تكون المحاذاة ونتائجها بيولوجية لأننا بصدد مقارنة منظومة حية. كما يجب أن تكون خوارزميات المحاذاة تطبيقية. وكما ذكرنا في هذا الفصل أن خوارزميتي المحاذاة العامة والمحاذاة المحلية تفشل تطبيقيا في حالة السلاسل الطويلة. لذلك انتقلنا للخوارزميات التجريبية التي لا تضمن الحل الأمثل وتعطي حلا تقريبيا. ومن هذا المنطلق في الفصل القادم سنطرح خوارزميتين تجريبيتين. بغية حل مشكل الوقت في السلاسل الطويلة.

الفصل 3
تحسين خوارزميات المحاذاة

1.3 مقدمة

كما ذكرنا سابقا فإن الخوارزميات التجريبية تحل مشكلة البطء للخوارزميات التقليدية التي تعتبر دقيقة وتضمن الحل الأمثل. لكنها غير تطبيقية بينما الخوارزميات التجريبية تعطي حلا تقريبيا لكنه سريع وفعال وتطبيقي. و من هذاالسياق سنطرح مقاربتين تندرجان ضمن هذا النوع. واحدة تعتمد على فكرة شجرة البوادئ واللواحق. بينما الأخرى تعتمد على تصفح مناطق الشبه دون المرور لمناطق عدم الشبه بإستعمال تقنية التقسيم للمناطق المتكررة.

2.3 مقاربة شجرة اللواحق و البوادئ

لاحظنا أن خوارزميات المحاذاة في البرمجة الديناميكية تعمل على مستوى بعدين. هذان البعدان يمثلان السلسلتين اللتين نود محاذاتهما. إقترحنا أن نعمل على بعد واحد وهو السلسلة الأقصر. وهذا إعتمادا على أن البحث عن سلسلة جزئية في شجرة اللواحق يستغرق طول هذه السلسلة. إعتبرنا السلسلة الأقصر مجموعة من السلاسل الجزئية وتقوم الخوارزمية بالبحث عنها في السلسلة الأطول التي تخزن على شكل شجرة لواحق. هذه المرحلة يمكنها أن تسبب مشكل الحرمان لبعض المقاطع. نعالج هذا المشكل بشجرة البوادئ.

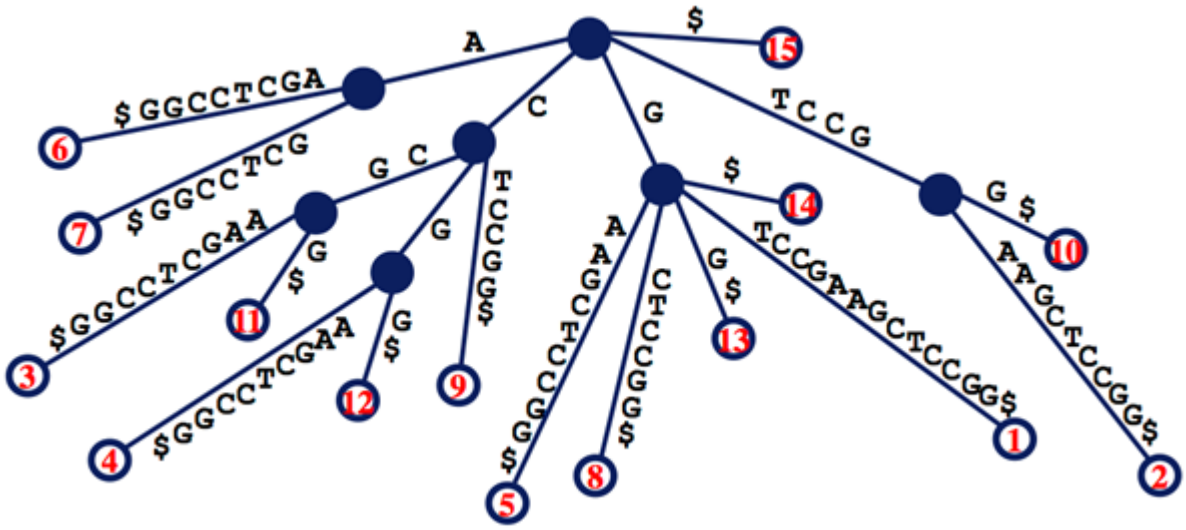
1.2.3 شجرة اللواحق

شجرة اللواحق هي نوع من هياكل البيانات الذي يمكن من الحل السريع للعديد من مشاكل البحث عن الأنماط في السلاسل النصية. فمن أجل البحث عن سلسلة نصية فرعية ذات طول m فهي تستغرق $O(m)$

من أجل بناء شجرة اللواحق $T(s)$ المتعلقة بالسلسلة النصية s ذات طول n فشجرة اللواحق $T(s)$ معرفة بالخواص التالية:

- $T(s)$ هي شجرة ذات جذور لها بالظبط n ورقة (leaves)
- يتم تسمية كل حافة (edge) من $T(s)$ بسلسلة نصية جزئية (substring) من s
- كل عقدة داخلية internal node من $T(s)$ خلاف الجذر لها إبنان على الأقل
- يجب أن تبدأ التسميات الفرعية للحواف المؤدية من عقدة إلى أبنائها برموز مختلفة

فشجرة اللواحق على سبيل المثال للسلسلة التالية GTCCGAAGCTCCGG موضحة في الشكل 1.3 :



شكل 1.3: بناء شجرة اللواحق

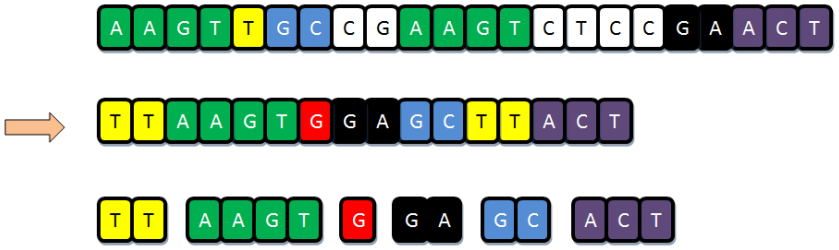
وخطوات الخوارزمية هي بالشكل الآتي:

المدخلات: سلسلتين نصيتين
المخرجات: مواضع مناطق الشبه الدقيق بين السلسلتين

- نقوم بتحويل أطول سلسلة لشجرة لواحق.
- نقوم بمقارنة السلسلة الأصغر مع شجرة اللواحق للسلسة الأطول بدءا من أول حرف تباعا حتى آخر حرف وفقا لاتجاه المقارنة من اليسار الى اليمين.
- في كل مرة نجد التوافق بين مقطعين في السلسلتين نخزن موضع البداية والنهاية لكل مقطع تمت فيه عملية التشابه الدقيق.
- تستغرق هذه المقارنة طول السلسلة الأصغر بحيث تكون النتيجة هي المقاطع المتشابه و غير المتشابه بين السلسلتين.

على سبيل المثال إذا كان لدينا سلسلتين AAGTTGCCGAAGTCTCCGAACT و TTAAGTGGAGCTTACT ونود مطابقتهما بواسطة شجرة اللواحق.

فتكون النتيجة كما هو موضح في الشكل 2.3



شكل 2.3: مقارنة سلسلتي حمض نووي بواسطة شجرة اللواحق

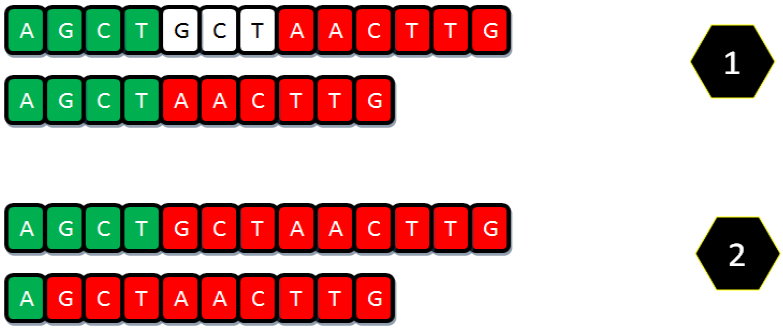
يعيب على هذه الخوارزمية أنها في حالة توالي مقطعين في السلسلة الأُقصر هما من مقاطع التشابه فإن هنالك إحتمالية أن يكون المقطع الثاني ناقصا وذلك لأن المقطع الذي قبله حرمه من البداية بشكل طبيعي.
لحل هذا المشكلة فإننا سنحاول إطالة المقاطع من جهة اليسار للتحقق من أنها لم تحرم أن تكون كاملة.

نقترح المثال التالي

لدينا السلسلتين AGCTGCTAACTTG و AGCTAACTTG كما هو موضح في الشكل 3.3

بعد القيام بعملية مطابقة بواسطة شجرة اللواحق ستكون النتيجة كما هي موضحة في الرقم 1 الشكل 3.3

بينما باستعمال شجرة البوادئ لتمديد المناطق المحرومة فسنحصل على النتيجة الموضحة في الرقم 2 الشكل 3.3



شكل 3.3: مقارنة سلسلتي حمض نووي باستعمال شجرتي اللواحق والبوادئ

بحيث بعد القيام بعملية مطابقة بواسطة شجرة اللواحق نتحصل على مناطق متاشبهة لكن ليست المثالية. بحيث نبحث عن شبه له معنى بيولوجي أكثر.

لذلك نستخدم شجرة البوادئ للسلسلة الأطول ومطابقة المناطق المتشابهة سلفا بعد تطبيق خوارزمية اللواحق بدءا من آخر حرف في المقطع لأول حرف فيها. ونقوم بعملية التمديد من جهة اليسار لتحديد وجه الشبه الذي تم إهماله لضعف المرحلة السابقة.

تستغرق الخوارزمية طول أصغر سلسلة للواحق و مثلها للبوادئ أي أن تعقيد الخوارزمية خطي $O(n)$.
نتائج هذه الخوارزمية عبارة عن مناطق الشبه الدقيق بين السلسلتين.
يمكن ربط كل المقاطع لتحديد الشبه العام أو المحلي وفقا لقيود كل منهما.

يمكن ان يكون هناك تداخل في المقاطع الناتجة بعد التمديد بشجرة البوادئ، لكن هذا التداخل لا يؤثر على النتيجة أي أننا فالمحصلة نقوم بربط هذه المقاطع باضافة فجوات وبخصوص التداخل سيكون مشتركا لا غير.

وسنقوم بدمج هذه المقاطع ذات الشبه الدقيق بغية لحصول على سلسلة جزئية أكبر ذات تنقيط أفضل إن وجد.

2.2.3 تنقيط ودمج مقاطع الشبه الدقيق

بعد إستعمال شجرة اللواحق والبوادئ لمقارنة السلسلتين، تحصلنا على مناطق الشبه الدقيق بين السلسلتين. يمكننا حساب التنقيط لجميع المقاطع ذات الشبه الدقيق. أي أنه يمكننا تحديد أفضل شبه دقيق بين السلسلتين.

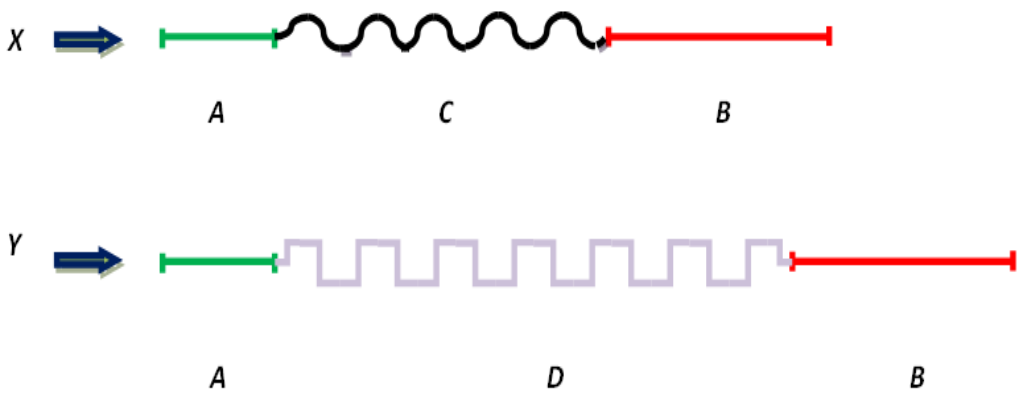
في هذه الحالة لم نكترث للإختلافات ولا الفجوات. لكن يمكن أن تكون هناك سلسلة جزئية أكبر تحتوي على سلاسل جزئية بها شبه دقيق وإختلافات وفجوات. ولها تنقيط أكبر من تنقيط السلاسل الجزئية التي تحتويها ذات الشبه الدقيق.

بغية الكشف عن أكبر تنقيط لمحاذاة محلية بالتقريب. فإننا نقوم بحساب التنقيط للمناطق المتجاورة ذات الشبه الدقيق للسلسلة الأصغر مع السلسلة الأكبر.

بحيث نحسب التنقيط الإجمالي للسلسلتين الجزئيتين اللتان تحويان القطعتين المتجاورتين. إن كان التنقيط الجديد أفضل من تنقيط القطع المشكلة للسلاسل الجزئية على حدا فإننا على يقين أن السلسلة الجزئية لديها أحسن تنقيط.
بغية الإحتفاظ بأعلى قيم التنقيط للسلاسل الجزئية فإننا نحدد أفضل k تنقيط لها. ونقوم بدمج هذه السلاسل الجزئية مع بعضها البعض والبحث عن تنقيط أعلى إن وجد.

هذه المرحلة تبحث عن تنقيط أعلى إن وجد. من خلال دمج السلاسل الجزئية بإضافة فجوات وإختلافات.

الشكل 4.3 يعبر عن كيفية دمج مقطعين متجاورين ذوي شبه دقيق



شكل 4.3: تنقيط قطعتين متجاورتين ذات شبه دقيق

حيث يمكن حساب التنقيط الأفضل بعد دمج السلاسل بالعبارة التالية

$$S(x,y) = max(S(A), S(B), S(K))$$

$$S(K) = S_d \times min(||c||, ||d||) + S_g \times |||c|| - ||d||| + S(A) + S(B)$$

حيث تمثل $S(x,y)$ قيمة تنقيط محاذاة السلسلتين x و y.
S(A) قيمة تنقيط الشبه الدقيق للسلسلة الجزئية A بين السلسلتين.
S(B) قيمة تنقيط الشبه الدقيق للسلسلة الجزئية B بين السلسلتين.
$S(K)$ يعبر عن التنقيط الجديد لمحاذاة السلسلتين بعد دمج السلسلة الجزئية A مع السلسلة الجزئية B بإضافة فجوات وإختلافات تربط قطعتي الشبه الدقيق.
S_d معامل تنقيط الإختلاف الواحد بين السلسلتين.
S_g معامل تنقيط الفجوة الواحدة.

3.3 محاذاة سلسلتين بإهمال المقاطع غير المتشابهة

لاحظنا أن عملية محاذاة سلسلتين على مستوى البرمجة الديناميكية تقوم بتصفح ومقارنة جميع أحرف السلسلتين وفقا لإتجاه محدد.
إن عملية محاذاة أي سلسلتين ينتج عليه مقاطع متشابهة(مناطق الشبه) وأخرى مختلفة.
وتكمن الفكرة التي سنطرحها في هذا المبحث لمحاذاة سلسلتين في الجواب على السؤال: هل يمكننا مسبقا معرفة مناطق عدم الشبه وتفادي البحث فيها. وتصفح مناطق الشبه فقط؟
والإجابة على هذا التساؤل موضحة كالتالي

إن أي سلسلة تتكون من مقاطع وأن إختلافها (عددها) محدود. كما هو الحال في سلاسل الحمض النووي والبروتين بحيث يمكن تقسيمها إلى مقاطع بطول معين فيكون عدد المقاطع محدودا. لأن عدد الأحرف المشكل لها منتهي ويكون أصغر كلما كانت عدد الأحرف المشكلة للسلاسل أقل.

إذا كان هناك مقطع لا تتشابه فيه السلسلتان ويتكرر على طوليهما فلماذا في كل مرة نعيد البحث عنه ونحن نعلم أننا لن نجده فنحن بصدد إضاعة الوقت ليس إلا وإضافة عملية نعلم أن لا فائدة منها.

وماذا لو كانت عدد المقاطع غير المتشابهة بشكل جلي وبعدد أكبر فبالتأكيد سوف نخسر الكثير من الوقت في البحث عليه

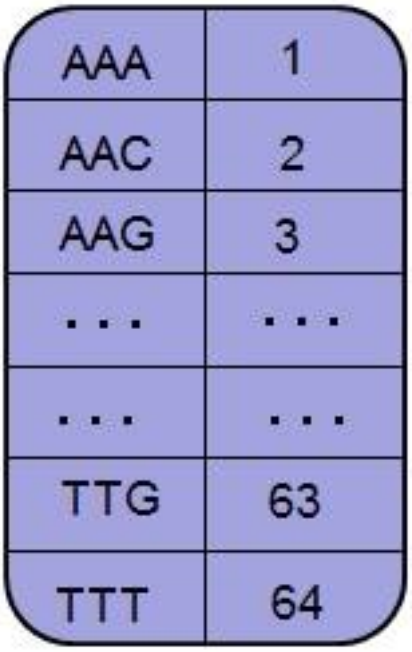
وخطوات المقاربة التي تجيب على التساؤل هي كالتالي:

- نقسم السلسلتين إلى قطع متساوية في الطول، ذا الطول يحدده الخبير مثلا (طول = 3) مع اهمال تكرار نفس المقطع
- نحول المقاطع الى ارقام بواسطة جدول. بحيث تكون المقاطع مرتبة حسب الترتيب الأبجدي ويقابلها أرقام إنطلاقا من الرقم1.
  كما هو موضح في الشكل 5.3
- نضع مقاطع كل سلسلة في جدول. عدد خلاياه هو نفسه عدد المقاطع. كل خلية في الجدول تحتوي جدولا يحتوي تكرار كل مقطع و مواضعه في السلسلة. كما هو موضح في الشكل 6.3 .

- نقوم بتصفح مقاطع السلسلتين في الجدولين والاحتفاظ بالمقاطع المتشابهة واهمال المقاطع المختلفة. أي أننا نخزن الخلية التي تحتوي المقطع والمعلومات المتعلقة بها. كما هو موضح في الشكل 6.3 . بحيث نبدء تصفح المقاطع ابتداءا من أول خلية في الجدول الأول و الجدول الثاني. نقارن محتوى الخليتين، فإذا كان محتوى الخلية الحالية في الجدول الأول أصغر أو يساوي محتوى الخلية الحالية في الجدول الثاني، فإننا ننتقل للخلية الموالية في الجدول الأول. أما إذا كان محتوى الخلية الحالية في الجدول الثاني أصغر من محتوى الخلية الحالية في الجدول الأول، فإننا ننتقل للخلية الموالية في الجدول الثاني. ونعيد نفس العملية كلما تغيرت الخلية الحالية حتى نصل إلى آخر خلية في الجدولين. وكلما تساوى محتوى الخليتين في الجدولين فإننا نخزن الخلية التي تحتوي المقطع والمعلومات المتعلقة بها من تكرار ومواضع.
- تكون نتيجية تصفح مقاطع الجدولين عبارة عن مقاطع الشبه الدقيق بين السلسلتين ذات الطول k نقوم بعملية تنقيط ودمج هذه المقاطع الجزئية بغية الحصول على تنقيط أفضل إن وجد، بنفس طريقة التنقيط والدمج للسلاسل الجزئية التي ذكرناها في المبحث السابق.
- كما يمكن ربط هذه المقاطع باستعمال خوارزمية شجرة اللواحق والبوادئ لتمديد الشبه

إن نتائج هذه الخوارزمية هي مناطق الشبه الدقيق بين السلسلتين. أي أننا أمام نفس المحطة التي صادفتنا في الخوارزمية السابقة. لذلك سنستعمل طريقة دمج السلاسل لتحديد أفضل محاذاة بين السلسلتين.

تعقيد الوقت لهذه المقاربة خطي.
إذا خزنت مقاطع السلاسل مسبقا في جدول كما وضح مسبقا. وهذا الجدول يحتوي مواضع المقاطع في السلسلة. فإننا نحتاج وقت تقسيم المقاطع وتصفح السلسلة كاملة في كل مرة نود محاذاة نفس السلسلة. وحينها حينما نريد تصفح الجدول بغية البحث عن المقاطع المشتركة بين السلسلتين فإن هذه المقاربة تستغرق عدد خلايا الجدول أي عدد مقاطع السلسلة دون تكرار. الذي نعبر عليه بالعبارة التالية L^k
حيث L يمثل عدد الأحرف التي تشكل السلسلة. بينما k يمثل طول المقطع في السلسلة.



AAA	1
AAC	2
AAG	3
...	...
...	...
TTG	63
TTT	64

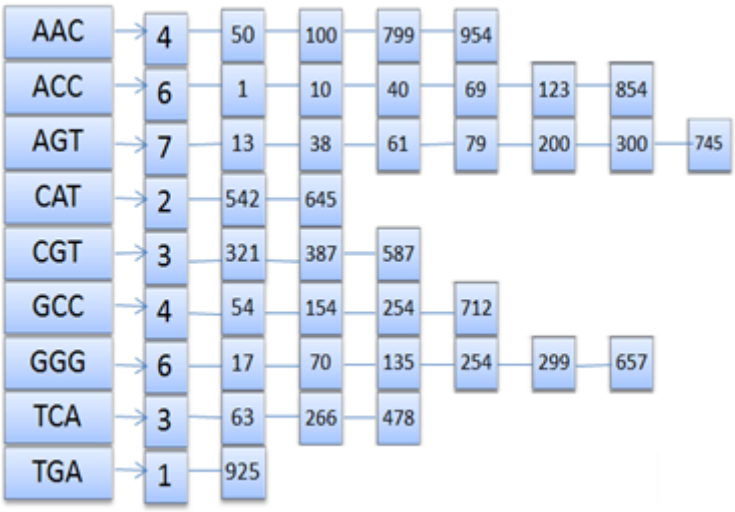
شكل 5.3: تحويل مقاطع سلسلة حمض نووي إلى أرقام

من بين مزايا هذه الخوارزمية نذكر

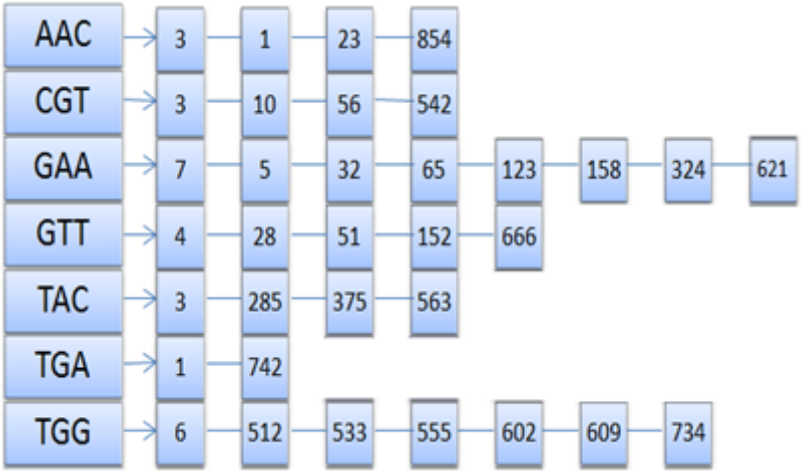
- الخوارزمية فعالة بالنسبة للشبه الضئيل أو أقل من المتوسط وخاصة في حالة السلاسل الطويلة ويكون البحث سريع جدا. وكحالة خاصة الخوارزمية فعالة في حالة المحاذاة المتعددة للسلاسل بحيث تكون المناطق المشتركة بين السلاسل في هذه الحالة قليلة مقارنة بطولها.
- الخوارزمية تبحث خلال مناطق الشبه فقط, عدد المقاطع أصغر بكثير من طول السلسلة في حالة السلاسل المتوسطة والطويلة
- تستغرق هذه العملية طول السلسلة.
- يمكن حساب الحد الأعلى لعدد المقاطع وهو مرتبط بعدد الأحرف وطول المقطع و كلما كان تكرار المقاطع أكثر قل عدد المقاطع.
- يستغرق البحث في جدول المقاطع في اسوء الحالات الحد الأعلي للمقاطع وبوجه التحديد عدد المقاطع في الجدول.

ولتوضيح عمل هذه المقاربة نقترح المثال التالي

فلتكن لدينا سلسلتي حمض نووي ما مختلفيتن وبعد تطبيق مرحلة التقسيم الى قطع متساوية وحفظ عددها ومواضعها كما هو موضح في الأشكال 6.3 و 7.3



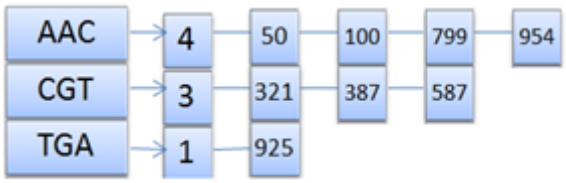
شكل 6.3: قائمة قوائم تعبر عن القطع الموجودة في السلسلة الأولى



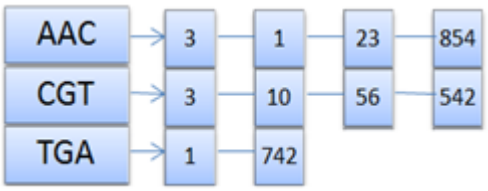
شكل 7.3: قائمة قوائم تعبر عن القطع الموجودة في السلسلة الثانية

بحيث مثلنا كل سلسلة بقائمة متصلة تحتوي القطع المختلفة لسلسلة معينة. وكل خلية من القائمة تحتوي قائمة متصلة. تحتوي الخلية الأولى منها على عدد تكرار مقطع ما في السلسلة. بينما تمثل بقية الخلايا مواضع هذه المقاطع في السلسلة بالترتيب.

بعد الحصول على قائمة المقاطع لكلتا السلسلتين فإننا وبعد تطبيق مرحلة تصفح القائمتين وإهمال القطع غير المتاشبهة و حفظ المتشابهة منها. فتكون النتيجية كما هو موضح في الأشكال 8.3 و 9.3 لكلتا السلسلتين:



شكل 8.3: القطع المتشابهة ومواضعها في السلسلة الأولى



شكل 9.3: القطع المتشابهة و مواضعها في السلسلة الثانية

وإنطلاقا من هذه النتيجة يمكننا عمل محاذاة للقطع المتشابهة بإضافة فجوات وتمديد الشبه.

خاتمة

في هذه المذكرة تطرقنا إلى محاذاة السلاسل الجينومية وحاولنا معالجة مشكل تعقيد الوقت في محاذاة هذه السلاسل التي تتميز بطول كبير.

وتطرقنا خلال هذه المذكرة للمعلوماتية الحيوية التي تتضمن تحليل ومقارنة ومحاذاة السلاسل، والهدف منها. وكيف أن طريقة المصفوفة النقطية تساعد على التحليل غير أنها لا تعطي قرارا حاسما حول المحاذاة. بينما المحاذاة العامة تحدد الشبه من خلال الطول الإجمالي للسلسلتين. ولا يمكنها التعامل مع السلاسل المختلفة في الطول أو السلاسل ذات الشبه الضئيل. فجاءت المحاذاة المحلية لتعالج هذا العائق. ووضحنا أن التنقيط المثالي للمحاذاة المثالية يحدد مدى الشبه بين السلاسل وذلك باستعمال البرمجة الديناميكية التي مرت بنا من تعقيد خوارزمي أسي لتعقيد وقتي تربيعي. لكن رغم ذلك ففي حالة السلاسل الطويلة تصبح هذه الخوارزميات حتى وبإستعمال البرمجة الديناميكية غير تطبيقية . بغية حل هذا العائق تناولنا الخوارزميات التجريبية. التي تعطي حلا تقريبيا للتنقيط الأمثل للمحاذاة لكن في وقت أسرع وهذا ما يميزها. وذكرنا من بينها خوارزمية FASTA .

كما تجدر الإشارة أننا ساهمنا بمقاربتين جديدتين لحل مشكل تعقيد الوقت للسلاسل الجينومية. في المقاربة الأولى إعتمدنا على إستعمال شجرة اللواحق للسلسلة الأطول ومقارنتها بالسلسلة الأقصر وتحديد مناطق الشبه الدقيق. وبغية تمديد هذه المقاطع ومعالجة مشكل الحرمان إستعملنا شجرة البوادئ. ثم إعتمدنا عملية دمج المقاطع المتجاورة ذات الشبه الدقيق من أجل الحصول على تنقيط جديد أعلى من تنقيط السلاسل الجزئية، إن وجد. و تعقيد الوقت لهذه الخوارزمية خطي.

بينما في المقاربة الثانية، قسمنا السلاسل إلى مقاطع متساوية ذات طول k بغية الإحتفاظ بالمقاطع المتشابهة و تجاهل غير المتشابهة قصد الإقتصاد في وقت البحث وخزناها في مصفوفة مواضع لكل سلسلة. يحتوي جدول القوائم مقاطع السلسلة دون تكرار، بالإضافة إلى مواضعها و عدد ظهورها في السلسلة. ومن ثم من خلال تصفح جدول المقاطع نحدد المقاطع المتشابهة فقط. وهو مفيد جدا في حالة الشبه القليل بين سلسلتين. وطبقنا بعدها مرحلة تنقيط ودمج المقاطع

المتجاورة ذات الشبه الدقيق و تعقيد الوقت لهذه الخوارزمية خطي.

وكآفاق لهذا العمل، يمكن الإشارة إلى برمجة المقاربتين المقترحتين مع تجريبهما و تقييمهما . كما نشير إلى أهمية تطوير وتحسين المقاربتين فيما يخص تعقيد الذاكرة.

المصادر

[1] إيهاب عبد الرحيم . مجلة العربي عدد يناير . 2001

[2] عبد الهادي، عائدة وصفي . مقدمة في علم الوراثة . 1998

[3] JIN XIONG . Essential Bioinformatics . 2006

[4] The Language of God . Francis Collins .

[5] Bruce, Alberts, other . Molecular Biology of the Cell . 2002

[6] Brenner, S. E., Chothia, C., and Hubbard, T. J. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships . 1998

[7] Attwood, T. K., and Miller, C. J. Progress in bioinformatics and the importance of beingearnest . 2002

[8] Batzoglou, S. The many faces of sequence alignment . 2005

[9] C.R. Calladine and H.R. Drew. Understanding DNA: The Molecule and How It Works. San Diego, CA: Academic Press, 2nd edition, 1997

[10] N.C. Jones and P.A. Pevzner. An introduction to bioinformatics algorithms. MIT Press., 2004.

[11] W-K. Sung. Algorithms in Bioinformatics: A practical introduction. Chapman and Hall/CRC, 2010.

[12] http://www.werathah.com