الجمهورية العربية السورية
وزارة التعليم العالي
جامعة تشرين
كلية الهندسة الميكانيكية والكهربائية
هندسة الحاسبات والتحكم الآلي

خوارزميات محاذاة السلاسل البيولوجية وتطبيقاتها

Sequence Alignment Algorithms

إعداد

رؤى سلمان سليمان

عمار محمد الشيخ حسن عواد

إشراف

د. م سامر سليمان

العام الدراسي 2017-2018

الشكر والعرفان:

الإهداء:

الفهرس:

فهرس الأشكال:

مقدمة

المعلوماتية الحيوية Bioinformatics:

هو علم يقوم على استخدام أحدث تقنيات الرياضيات التطبيقية، المعلوماتية، الإحصاء وعلوم الحاسب لحل مشكلات بيولوجية حيوية وبكلام آخر يقصد بمصطلح المعلوماتية الحيوية تحليل المعلومات البيولوجية باستخدام الحاسب والتقنيات الإحصائية أو هو العلم الذي يسعى لاستخدام وتطوير قواعد البيانات والخوارزميات الحاسوبية لتوسيع وتعزيز الأبحاث البيولوجية.

اتجهت جهود الأبحاث الرئيسية في هذا المجال إلى عدة أقسام ومنها:

- محاذاة السلاسل البيولوجية Sequence alignment.
- إيجاد المورثات.
- محاذاة السلاسل البروتينية.
- تنبؤ البنية البروتينية.
- التنبؤ بالتعبير الجيني.
- نمذجة التطور.

أهداف المشروع:

١. تطوير خوارزميات جديدة وتقنيات إحصائية تساعد في تحصيل المعلومات من مجموعة ضخمة من البيانات.
٢. تحليل وتفسير الأنماط المختلفة من البيانات التي تتضمن سلاسل الحموض الأمينية والسلاسل البروتينية.
٣. تطوير وتنفيذ أدوات تساعد على إدارة فعالة للأنماط المختلفة من المعلومات.

أهمية المشروع العلمية والمشاكل التي يحلها:

١. يعتبر من الصعب جداً تحديد مدى التطابق بين سلسلتين بيولوجيتين بالعين المجردة ولذلك تم تطوير خوارزميات تعتمد على البرمجة الديناميكية (التنبؤ) للحصول على دقة أكبر في عملية المطابقة.
٢. تطبيق خوارزميات حاسوبية تقوم بعملية تصنيف للأمراض حسب التسلسل الاميني للحموض المكونة لها بالإضافة إلى بناء أصناف لأمراض جديدة غير مكتشفة سابقاَ.
٣. تطبيق الخوارزميات المختلفة على سلاسل متعددة يساعد على الحصول على نتائج أفضل للتصنيف وذلك عن طريق نتيجة (رصيد) يعبر عن مدى كفاءة الخوارزمية المستخدمة.

الفصل الأول

محاذاة السلاسل البيولوجية sequence alignment

١-١مفاهيم نظرية:

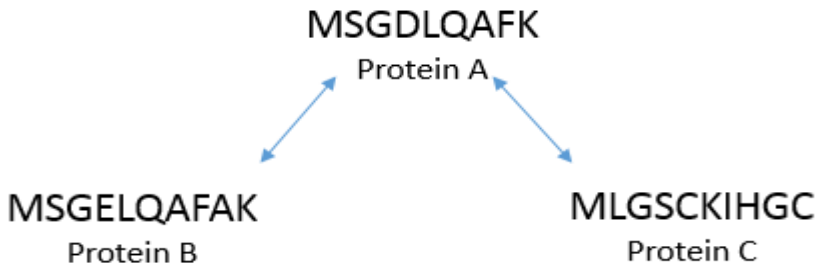
السلسلة البيولوجية:

هي عبارة عن تتالي من الحموض الأمينية أو النيكليوتيدية ويمكن أن تكون سلاسل بروتينية بحيث يقودنا هذه التسلسل إلى معرفة البنية الأساسية المكونة لمركب ما وبالتالي الوظيفة التي يقوم بها هذا المركب.

المحاذاة:

أو المطابقة هي طريقة أو منهجية يتم من خلالها اكتشاف مناطق التشابه بين سلاسل متعددة بحيث يتم معرفة مدى صلة السلسلة الأولى مع السلسلة المراد المطابقة معها.

مثال ١-١:



الشكل(١-١) محاذاة سلاسل بروتينية افتراضية

يظهر المثال السابق سلسلة بروتينية مفترضة اسمها A ونود معرفة أياَ من السلسلتين B وC هي الأكثر مشابهة (مطابقة) للسلسلة المفروضة A.

الشكل(١-٢) نتيجة محاذاة سلسلتي بروتين

نلاحظ أن التشابه بالتسلسل بين A وB هو 80% بينما التشابه بين A وC هو30% وبالتالي نقول إن السلسلة B أكثر مطابقة للسلسلة A من السلسلة C وبالتالي هذا يقودنا إذا تعمقنا قليلاً في البيولوجيا أن ل A وC نفس البنية بحد كبير جداً ونستنتج من هذه البنية أن لهاتين السلسلتين نفس الوظيفة تماماً.

لنفترض أن السلسلة البروتينية كانت مكونة لخلية ما وحصل تغير في هذه الخلية أدى إلى انقسامها إلى خليتين ابن وظهرت سلاسل جديدة شبيهة إلى حد كبير بالسلسلة الأساسية في الخلية الأم.

إن التغير في السلسلة الابن يمكن أن يكون تغير في أحد البروتينات المكونة للسلسلة الأم أو حذف له أو إضافة بروتين جديد وبالتالي لحل هذه المشكلة لا بد أن نقوم بعملية محاكاة للتغير الذي طرأ على السلسلة الابن ومحاولة إعادتها إلى شكلها الأساسي.

عملية المحاكاة الافتراضية تكون بإضافة فراغ للسلسلة الأبن يعوض عن البروتين المحذوف من السلسلة الأم، أو إضافة الفراغ للسلسلة الأم لمحاكاة عملية الإضافة. بتطبيق الخوارزمية البسيطة على المثال السابق نلاحظ ما يلي:

الفصل الثاني

التطبيق العملي

٣-١: الأدوات المستخدمة في المشروع:

٣-١-١ لغة البرمجة بايثون Python:

هي لغة برمجة عالية المستوى تتميز ببساطة كتابتها وقراءتها تستخدم أسلوب البرمجة الكائنية مفتوحة المصدر وتستخدم لبناء البرامج المستقلة باستخدام الواجهات الرسومية وفي عمل تطبيقات الويب ولإنجاز المشاريع البرمجية الضخمة كأي لغة برمجية أخرى.

لها الكثير من المكتبات البرمجية ذات الأغراض العامة والخاصة مثلاً مكتبة Sys ومكتبة Numby وهي لغة محمولة تعمل على العديد من المنصات دون أن يتطلب ذلك أي تغييرات وقد تم استخدامها في بيئة Linux.

٣-١-٢ لغة Html:

هي لغة نصوص تشعبية Hypertext markup language تستخدم في إنشاء وتصميم صفحات ومواقع الويب وتعتبر الهيكل الرئيسي لأي صفحة أو موقع على الويب، لا تعتبر لغة برمجة لكنها تستخدم في إعطاء الأوامر لمتصفح الانترنت وترشده إلى طريقة عرض الصور والروابط والنصوص والأشياء الأخرى المحتواة في الصفحة وأماكن عرض كل منها داخل الصفحة كما تقوم بإمداد المتصفح بالمعلومات الخاصة بالصفحة مثل عنوان الصفحة ووصفها الكلمات الدلالية الخاصة بها.

٣-١-٣ لغة Css:

تستخدم للفصل بين تصميم الموقع ومحتوى صفحات الويب المكتوبة بلغة Html وتسمح لمصمم المواقع بالتحكم في الألوان والخطوط والتصميم بأكمله.

٣-١-٤ لغة Css3:

قدمت خاصية جديدة وهي الحركات Animation بحيث يمكن للأشياء أن تتفاعل مع المستخدم إضافة إلى إضافة التدرجات والتي يمكن إضافتها بكود بسيط بدل من صنع صورة خاصة ووضعها.

تسمح بإضافة صورة بمثابة إطار للعناصر بدل الإطار الخطي وإضافة أكثر من خلفية لنفس العنصر.

٣-١-٥ Java script:

هي لغة سهلة التحكم وهي جزء يتم وضعه داخل لغة Html لزيادة فاعليته، تحول موقع الويب إلى موقع يتفاعل مع المستخدم من خلال إضافة أزرار ونماذج تأخذ بيانات من المستخدم وتحولها إلى نماذج أخرى

أو تجري عليها عمليات حسابية ليست بسيطة أي باختصار تحول الصفحة إلى ما يسمى صفحات الويب الديناميكية أو التفاعلية وهو مالا تقدمه لغة Html.

إن تنفيذ البرنامج المكتوب بهذه اللغة هو من اختصاص المتصفح التي ينفذها سطر سطر وليس عن طريق ترجمتها تجميعياً.

٣-٢: الخوارزميات المستخدمة ومميزاتها:

٣-٢-١: خوارزمية Smith Waterman:

تقدم هذه الخوارزمية محاذاة أحادية محلية بين سلسلتين Local single sequence alignment من خلال تطابق أو عدم تطابق أو إدخال أو حذف علماً أن كل من عمليات الإدخال والحذف تقدم فجوات وفقاً ما تمثله الشروط.

تعتبر واحدة من خوارزميات البرمجة الديناميكية وهي تستخدم لإيجاد المحاذاة المحلية المثلى مع الأخذ بعين الاعتبار نظام مجموع النقاط بالاعتماد على مصفوفة Substitution matrix والتي تحدد نسب التقابل وعدم التقابل إضافة إلى وجود الGap الذي يعبر عن التغير البيولوجي بالسلسلة (الذي تم شرحه سابقاً).

ومن ثم يتم تشكيل مصفوفة النتيجة وذلك بتهيئة مصفوفة ثنائية البعد أعمدتها عناصر السلسلة المدخلة مضافاً إليها عمود وأسطرها عناصر السلسلة المراد المقارنة معها من قاعدة البيانات مضافاً إليها سطر ويتم إضافة السطر والعمود الإضافيين من أجل أول عملية مقارنة نقوم بها لملء أول خانة حيث تتم تهيئة المصفوفة بأصفار ويتم ملء المصفوفة من اليسار إلى اليمين ومن الأعلى إلى الأسفل أي بدءاً من الخانة في اليسار الأعلى باتجاه اليمين على التتالي سطر سطر بملء كل خانة بناءاً على المقارنة بين أربع قيم هي قيمة التهيئة (الصفر) وقيمة الخانة التي فوقها وقيمة الخانة على يسارها والقيمة السابقة لها قطرياً (في القطر الرئيسي).

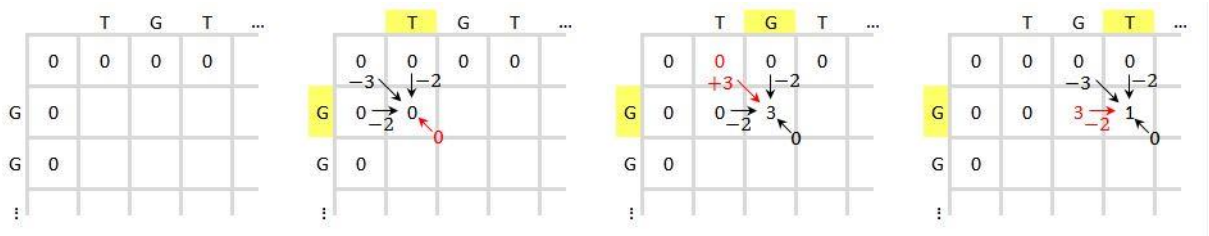
ثم تقوم بما يعرف بالتتبع الخلفي Backtracking بدءاً من القيمة العظمى max value التي تنتج عن مصفوفة النتيجة Scoring matrix وتتبع القيمة العظمى التي أتت منها كل قيمة عظمى خلفياً (مصدر هذه القيمة بشكل متكرر) ضمن الScoring matrix للحصول على أفضل محاذاة محلية بينما إذا أردنا الحصول على ثاني أفضل محاذاة محلية نبدأ بتطبيق التتبع الخلفي بدءاً من ثاني أعلى قيمة والمثال التالي يوضح مراحل عمل الخوارزمية.

$$\text{Substitution matrix: } S(a_i, b_j) = \begin{cases} +3, & a_i = b_j \\ -3, & a_i \neq b_j \end{cases}$$

$$\text{Gap penalty: } W_k = kW_1, \quad W_1 = 2$$

الشكل () يوضح التوابع المعتمد عليها في مثالنا

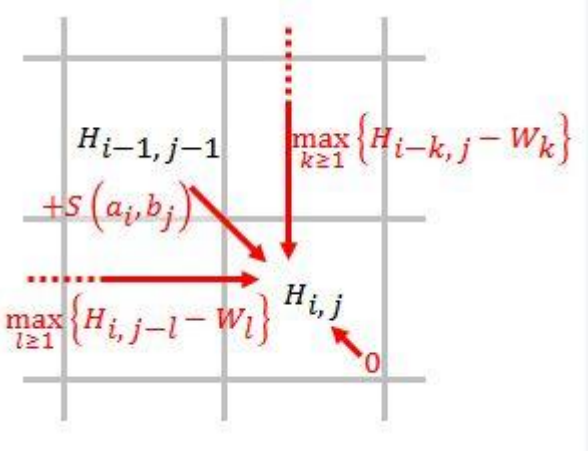
حيث تبين هذه التوابع قيم مصفوفة ال substitution matrix حيث أن قيمة التقابل (Match) هي 3+ وقيمة عدم التقابل (Mismatch) هي 3- وقيمة الGap هي 2-.



الشكل () يوضح آلية تهيئة مصفوفة النتيجة

مصفوفة النتيجة: نهيئ مصفوفة عدد أسطرها يساوي عدد محارف السلسلة الأولى 1+

وعدد أعمدتها يساوي عدد محارف السلسلة الثانية 1+ ونقوم بملء السطر الأول والعمود الأول أصفاراً.



الشكل () يوضح آلية ملء كل خانة

آلية ملء كل خانة من مصفوفة النتيجة:

يتم المقارنة بين أربع قيم هي: قيمة الخانة الابتدائية، قيمة الخانة التي فوقها مباشرة مع اعتبار وجود Gap في السلسلة الممثلة لأسطر المصفوفة، قيمة الخانة على يسارها مع اعتبار وجود Gap في السلسلة الممثلة لأعمدة المصفوفة، قيمة الخانة السابقة للخانة الحالية قطرياً مع ملاحظة تماثل المحرفين الممثلين لإحداثيات الخانة الحالية أو عدم تماثلهما ففي حال التماثل يتم إضافة 3 إلى قيمة الخانة السابقة قطرياً وفي حال عدم تماثلهما يتم طرح 3 من قيمتها.

ومن ثم يتم اختيار القيمة العظمى بين القيم الأربعة السابقة.

Initialize the scoring matrix

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0								
G	0								
T	0								
T	0								
G	0								
A	0								
C	0								
T	0								
A	0								

الشكل () تطبيق الخوارزمية مرحلة ١

مراحل تطبيق الخوارزمية:

يتم في المرحلة الأولى ملء أول سطر وأول عمود من مصفوفة النتيجة أصفار.

Fill the scoring matrix

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0								
T	0								
T	0								
G	0								
A	0								
C	0								
T	0								
A	0								

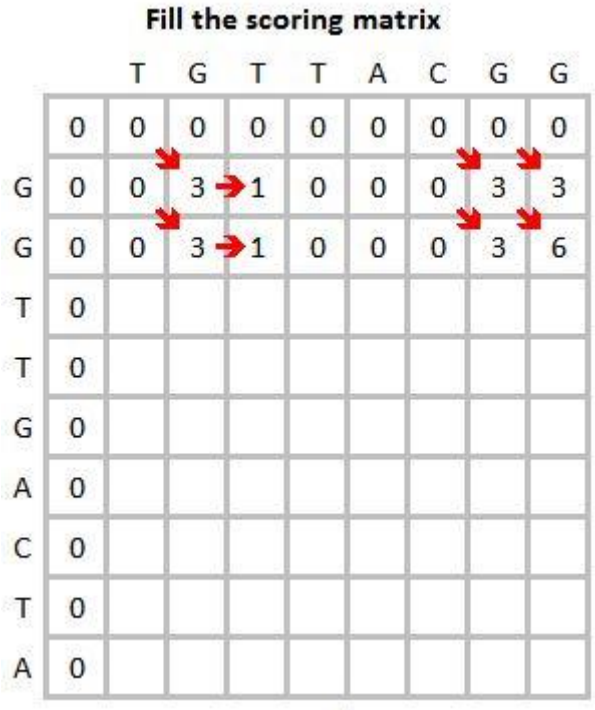
الشكل () تطبيق الخوارزمية مرحلة ٢

يتم في المرحلة الثانية ملء الخانات من اليسار إلى اليمين ومن الأعلى إلى الأسفل بدءاً من الخانة (G,T)

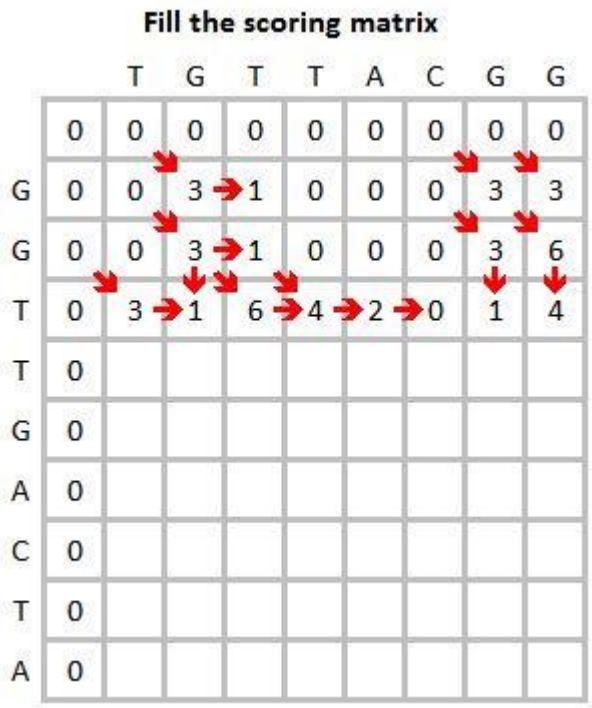
حيث تمت المقارنة بين قيمتها الابتدائية (0) والخانة التي فوقها مطروحة منها ال Gap(2-=2-0) والخانة على يسارها مطروحة منها الGap (2-=2-0) والخانة التي فوقها قطرياً مع ملاحظة عدم تماثل المحرفين T،G والتي نطرح منها 3 لتصبح قيمتها (3-=3-0) ونأخذ القيمة العظمى بين 3-,2-,2-,0 لتكون القيمة المختارة هي 0 .

ثم يتم الانتقال لملء الخانة G،G حيث تمت المقارنة بين قيمتها الابتدائية (0) والخانة التي فوقها مطروحة منها ال Gap(2-=2-0) والخانة على يسارها مطروحة منها الGap (2-=2-0) والخانة التي فوقها قطرياً مع ملاحظة تماثل المحرفين G,G والتي نضيف إليها 3 لتصبح قيمتها (3+=3+0) ونأخذ القيمة العظمى بين 3+,2-,2-,0 لتكون القيمة المختارة هي 3+.

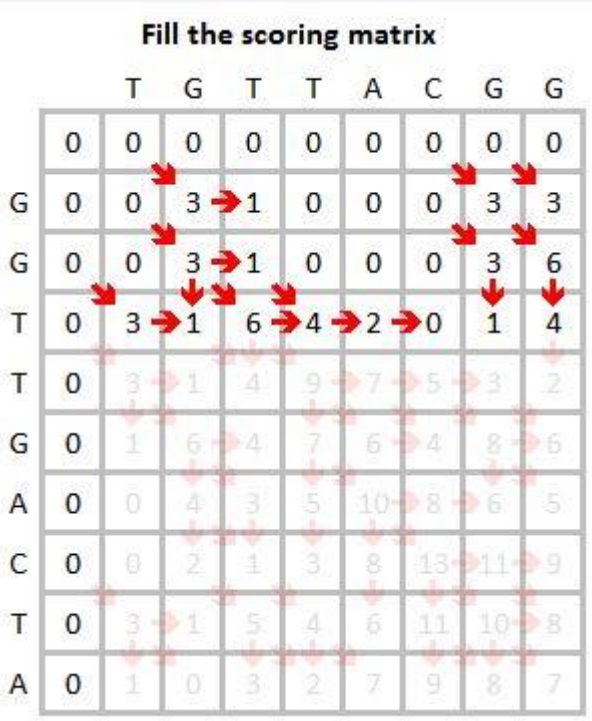
وهكذا يتم ملء جميع خانات مصفوفة النتيجة على التتالي مع ملاحظة الخانة المصدر لكل خانة تم ملأها في هذه المصفوفة كما هو مبين بالمراحل التالية:



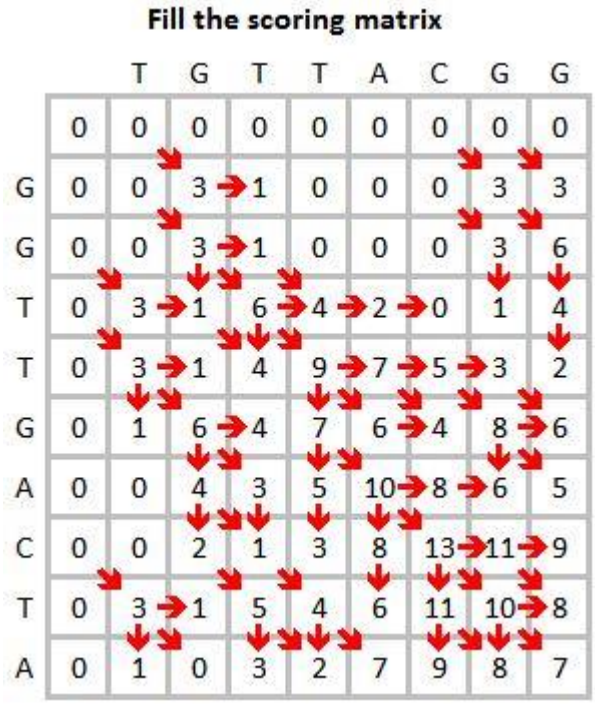
الشكل () تطبيق الخوارزمية مرحلة ٣



الشكل () تطبيق الخوارزمية مرحلة ٣

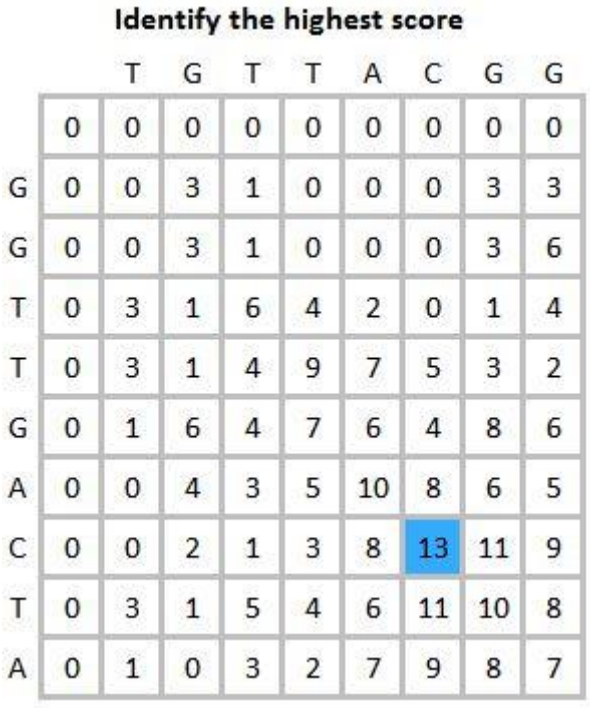


الشكل () تطبيق الخوارزمية مرحلة ٤



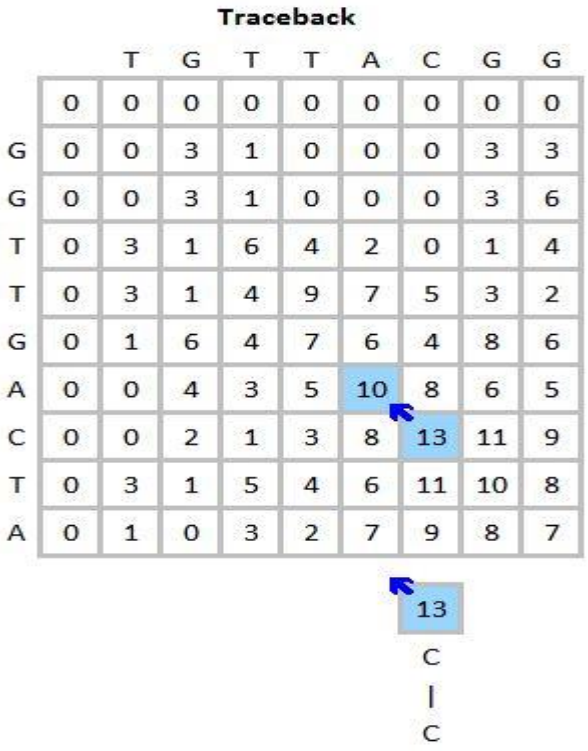
الشكل () تطبيق الخوارزمية مرحلة٥

بعد الانتهاء من ملء مصفوفة النتيجة مع الانتباه إلى مصدر قيمة كل خانة فيها تبدأ مرحلة التتبع الخلفي Backtracking.



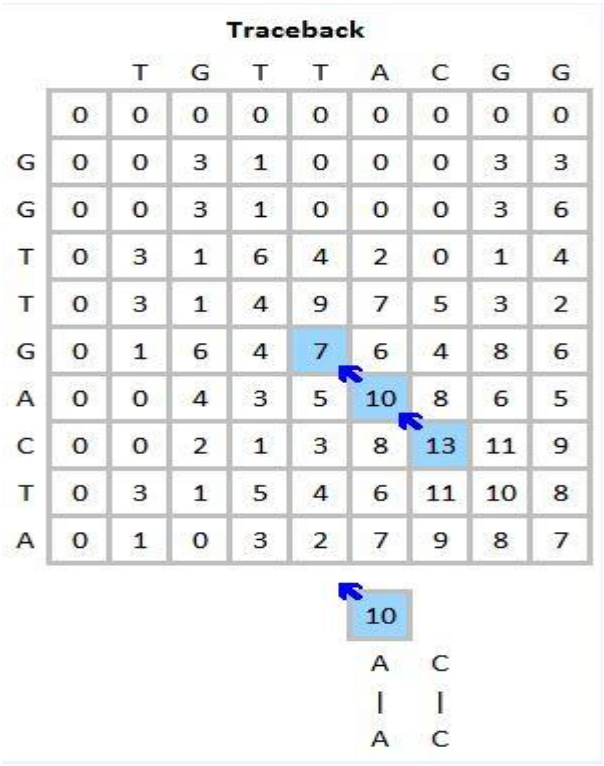
الشكل () تطبيق الخوارزمية مرحلة التتبع Backtracking ١-

في المرحلة الأولى من التتبع الخلفي يتم اختيار القيمة العظمى من مصفوفة النتيجة وهي القيمة 13 في مثالنا هذا وملاحظة الخانات فوقها وعلى يسارها والسابقة لها قطرياً حيث القيمة العظمى بين قيم هذه الخانات هي 10 وهي فعلاً الخانة المصدر للخانة ذات القيمة 13.



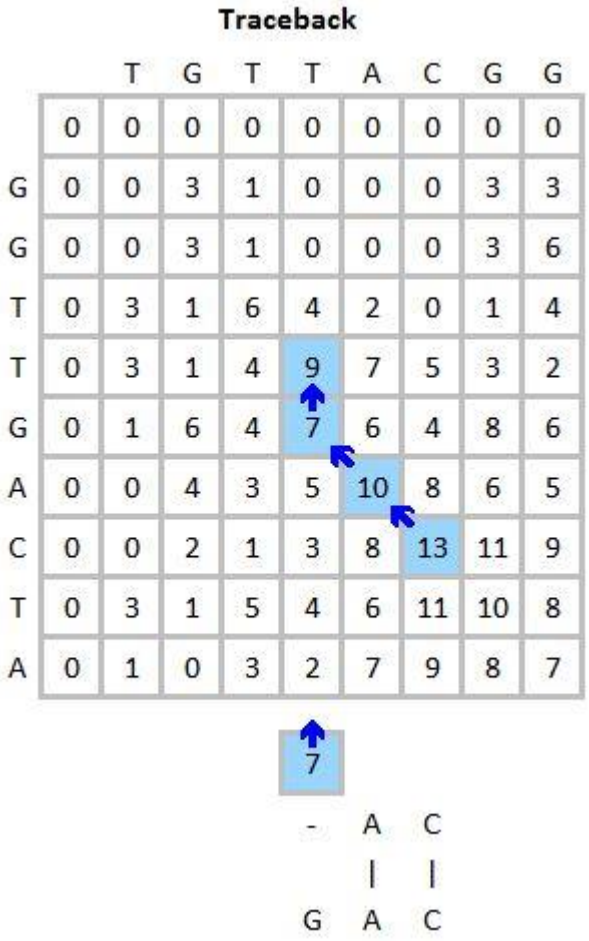
الشكل () تطبيق الخوارزمية مرحلة التتبع Backtracking –٢

بعد معرفة الخانة المصدر للخانة ذات القيمة 13 (وهي الخانة ذات القيمة 10 السابقة قطرياً لها) نلاحظ وجود تطابق بين المحرفين الممثلين للخانة 13 (تقاطع سطر مع عمود) حيث اختيار الخانة القطرية فعلاً يعني وجود تطابق بين المحرفين حسب المرحلة الأولى من الخوارزمية أي أننا بالنتيجة وجدنا تطابق بين المحرفين C,C .



الشكل () تطبيق الخوارزمية مرحلة التتبع Backtracking –٣

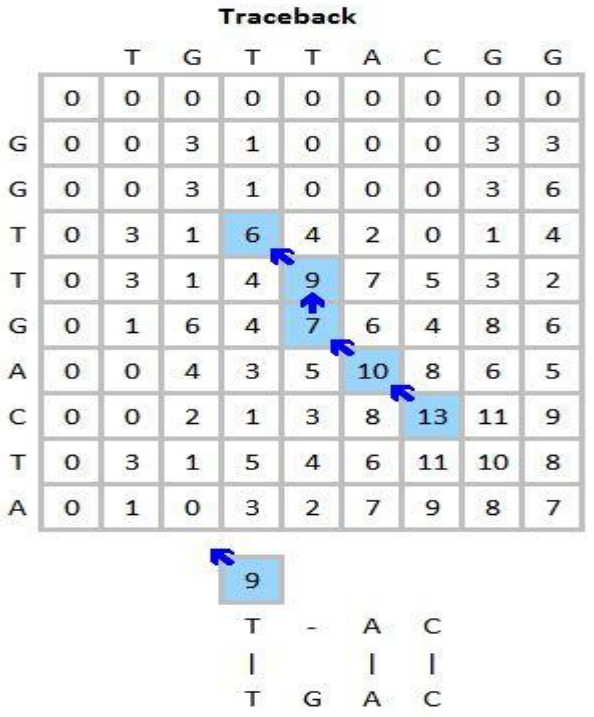
كما في السابق الخانة المصدر للخانة 10 هي الخانة ذات القيمة 7 السابقة لها قطرياً مما يعني وجود تطابق بين محرفي هذه الخانة A,A .



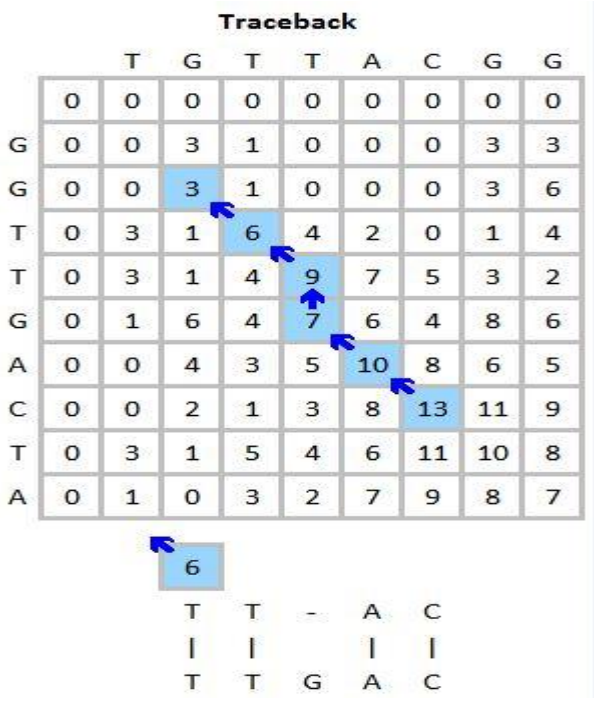
الشكل () تطبيق الخوارزمية مرحلة التتبع Backtracking – ٤

في هذه المرحلة نجد أن الخانة المصدر للخانة ذات القيمة 7 هي الخانة ذات القيمة 9 الواقعة فوقها مما يعني وجود Gap في السلسلة الممثلة لأعمدة المصفوفة فنحصل على التقابل الموضح أسفل الشكل مع وجود ال Gap (G,-) .

مع التنويه أنه في حال كانت الخانة المصدر واقعة على يسار الخانة الحالية فهذا يعني وجود Gap في السلسلة الممثلية لأسطر المصفوفة.

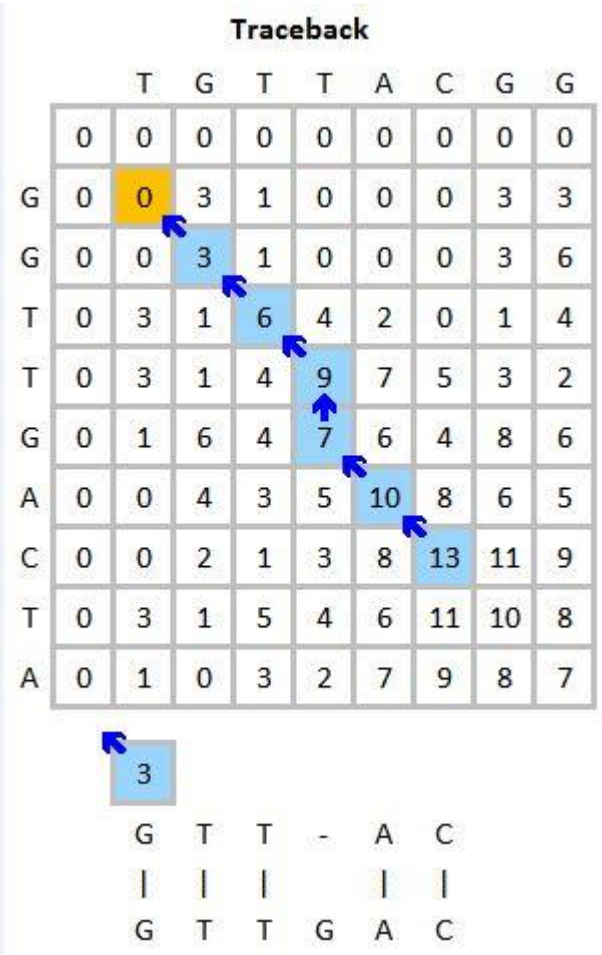


الشكل () تطبيق الخوارزمية مرحلة التتبع Backtracking –٥

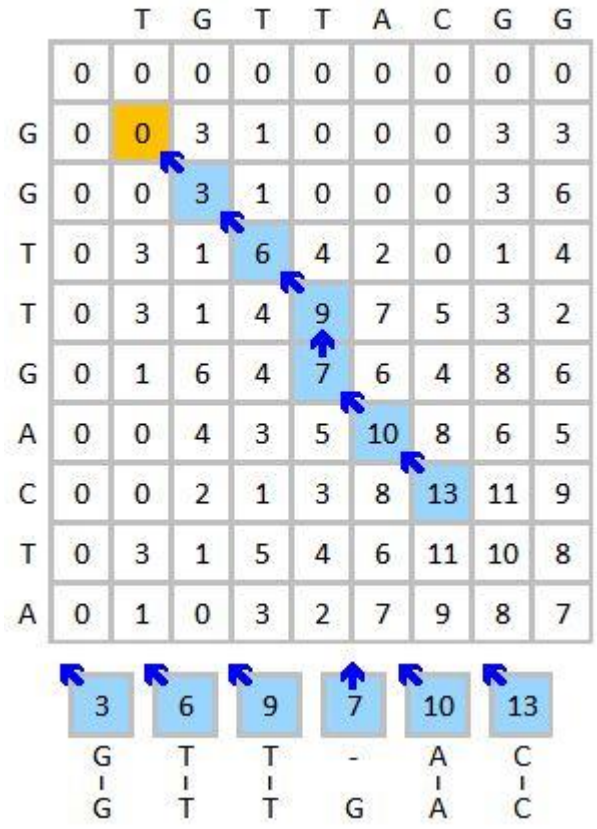


الشكل () تطبيق الخوارزمية مرحلة التتبع Backtracking –٦

في المرحلتين السابقتين نجد أن مصدر الخانة ذات القيمة 9 هي الخانة ذات القيمة 6 الواقعة فوقها قطرياً مما يعني وجود تطابق بين محرفي هذه الخانة T،T ومصدر الخانة ذات القيمة 6 هي الخانة ذات القيمة 3 الواقعة فوقها قطرياً مما يعني وجود تطابق بين محرفي هذه الخانة T،T.



الشكل () تطبيق الخوارزمية مرحلة التتبع Backtracking –٦



الشكل () تطبيق الخوارزمية مرحلة التتبع Backtracking –٧

نتابع التتبع الخلفي كما وضحنا سابقاً حتى نصل إلى الخانة الأولى التي بدأنا منها لنحصل على أفضل محاذاة اعتماداً على القيمة العظمى في مصفوفة النتيجة حيث التقابل يتضمن وجود Gap كما هو موضح في الشكل السابق.