الفصل الخامس عشر

مقدمة فى المعلوماتية الحيوية
Bioinformatics

يمكن تعريف المعلوماتية الحيوية بأنها الفرع الحسابى Computational للبيولوجيا الجزيئية. وقبل عصر المعلوماتية الحيوية كانت هناك طريقتين فقط لإجراء التجارب البيولوجية إما داخل الكائن الحى (*والتى سميت In Vivo أى فى الحى*) أو فى بيئة صناعية (*التى سميت In Vitro أى فى الأنبوب*) وقياسا على ذلك فإنه يمكنا القول بأن المعلوماتية الحيوية هى فى الحقيقة *In Silico BIOLOGY* أى البيولوجيا فى السليكون كناية عن أن مكونات الحاسب الآلى ناتجة من رقائق السليكون.

على مدى الحقبتين الأخيرتين، حدثت تطورات هائلة فى كفاءة أجهزة الكمبيوتر سواء الكمبيوتر الشخصى أو الكمبيوتر العملاق Super Computer نتيجة إستنباط برامج متقدمة ساعدت الباحثين فى مجال البيولوجيا على الحصول على تفسيرات وإجابات عن أسئلة لم يكن بمقدورهم التوصل إليها ناهيك عن السرعة الفائقة فى تحليل النتائج وإستنباط التفسيرات الصحيحة لها.

وفى تعريف آخر للمعلوماتية الحيوية بأنها تطبيق تقنية المعلوماتية لإدارة Management وتنظيم البيانات البيولوجية ، وتعد من المجالات العلمية الحديثة ذات التطور السريع.

قواعد البيانات Data Bases:

فى السنوات العشرين الماضية أصبح شائعا تخزين البيانات البيولوجية فى قواعد البيانات العامة Public Databases مما أدى إلى نمو هائل فى كمية المعلومات المخزونة فى هذه القواعد كما زادت أعداد هذه القواعد نفسها لكى تسع الكم الهائل من هذه البيانات التى تتدفق بمعدلات كبيرة جدا كل يوم تقريبا إن لم يكن كل ساعة على مدار الساعة. ونتيجة لذلك فقد أصبح لزاما على الباحثين أينما كانوا لكى يسترجعوا مايهمهم من قواعد البيانات هذه الإستعانة بالكمبيوتر وشبكة المعلومات الدولية world wide web (www).

وفى ضوء ذلك يمكن إضافة تعريف آخر للمعلوماتية الحيوية بأنها التقنية التى تختص بإستخدام الكمبيوتر فى جمع وتنظيم وتحليل وإستخدام وعرض البيانات البيولوجية والمشاركة فى الإستفادة من هذه البيانات مع المجتمع العلمى وذلك من خلال شبكات المعلومات والكمبيوتر الشخصى لكل باحث حيث يمكن فتح نافذه لانهائية على العالم بمجرد الضغط بأطراف الأصابع على Keyboard وهو جالس أمام الكمبيوتر الشخصى المتصل بتلك الشبكات.

وتعد قواعد البيانات بمثابة القلب للمعلوماتية الحيوية، وتوجد أنواع كثيرة ومختلفة من هذه القواعد حسب طبيعة المعلومات المخزونة (مثل تتابعات القواعد، والتراكيب وشرائح الجيل.....الخ) أو حسب طريقة التخزين

(دوسيهات Flat-Files أو جداول ...الخ) وتزداد أعداد قواعد البيانات بطريقة سريعة جدا ، ففى خلال عام ٢٠٠٠ تم إستحداث ٥٥ قاعدة معلومات جديدة بحيث أصبح العدد الكلى للقواعد فى نهاية السنة ٢٨١ قاعدة وهذه القواعد متاحة للجميع.

وتعد قاعدة بيانات Gene Bank فى الولايات المتحدة الأمريكية من المصادر الأساسية لتتابعات د ن أ والبروتينات ، كذلك توجد قاعدة بيانات يابانية Data Bank of Japan (DDBj) ، كما يشرف معمل البيولوجيا الجزيئية الأوربية على قاعدة بيانات مختصة بتتابعات النيوكلتيدات (EMBL). ويمكن تخزين أى تتابعات جديدة فى أى من هذه القواعد الثلاثة حيث توجد بينها مشاركة أوتوماتيكية يومية فى هذه البيانات. وإذا كان الغرض من البحث عن جين فى تتابعات د ن أ مميزة النواة فإنه من الأفضل أستخدام dbEST وهى قاعدة بيانات مكونة من EST's.

تحليل التتابعات:

بعد تحديد تتابعات قطعة طويلة من د ن أ، فإن أول مهمة هى التعرف على الجينات التى قد توجد فى هذه القطعة وفى حالة غير مميزة النواة تكون الكثافة الجينية Gene Density عادة مرتفعة نظرا لأن معظم الجينات المشفرة لبروتينات لاتوجد بها انترونات مما يسهل كثيرا فى عملية تحليل التتابعات ومعرفة الجينات الموجودة فى قطعة د ن أ تحت الدراسة. ولكن الأمر يختلف إذا كانت تتابعات د ن أ خاصة بمميزة النواة الراقية حيث نجد أنه من الصعب التعرف على الجين نظرا لأنه قد يكون مكونا من عدد من الإكسونات يتخللها عدد مماثل أو يزيد من الإنترونات. فمثلا فى الجينوم البشرى يكون متوسط

طول الإكسون ١٥٠ زوج من القواعد ومتوسط طول الإنترون عدة كيلو قاعدة وقد يصل طول الجين إلى مئات من كيلو قاعدة.

كما أن م ر ن أ لبعض الجينات قد يحدث له تعديلات بطرق متعددة مما يؤدى إلى تكوين متراكبات مختلفة من نفس الجزئ بحيث يتم بناء سلاسل ببتيدية مختلفة من جين واحد.

وهناك مشكلة إضافية للعثور على الجين وهى نسبة الإشارة Signal إلى الضوضاء Noise أو التشويش، ففى جينوم البكتريا تشكل الجينات نسبة ٨٠-٨٥ % من د ن أ الكلى ، وفى الخميرة تنخفض هذه النسبة إلى ٧٠% وفى حشرة الدروسفلا ودودة النتيماتودا تكون النسبة ٢٥% فى حين تصل هذه النسبة إلى أدنى مستوى لها فى الجنيوم البشرى لتكون ٣ إلى ٥٪ فقط من د ن أ الكلى.

لذلك نجد أن تحديد موقع بداية ونهاية جين ما وطراز التراكب Splicing فى إكسونات من بين جميع التتابعات غير الشفرية يكون فى منتهى الصعوبة وبمجرد التعرف على الجين يتم تحويل التتابعات النيوكلتيدية إلى تتابعات من الأحماض الأمينية المقابلة. ويكون السؤال الآن هو ما هى الوظيفة المحتملة لهذا البروتين؟

وبإجراء مسح لجميع المعلومات المتاحة فى قواعد المعلومات المختلفة فقد يكون من الممكن تحديد بروتينات آخرى بتتابعات مشابهة مما قد يساعد فى تحديد وظيفته ، ويمكن أيضا إستخدام مقارنات التتابعات لتعريف بعض العناصر Motifs فى البروتين مثل عناصر الإرتباط بجزئ ATP أو بجزئ دن أ وقد يفيد ذلك فى إعطاء معلومات عن وظيفة الجين.

تحديد إطارات القراءة المفتوحة (ORF):

بعد الحصول على التتابعات النيوكلتيدية لشظية طويلة نسبيا من د ن أ فإن أول مهمة هى تحديد إطار القراءة الصحيح وحيث أن هناك ثلاث إطارات قراءة محتملة على كل سلسلة من جزئ د ن أ فإن ذلك يتطلب إجراء ترجمة لستة إطارات Six Translation Frames وينتج عن ذلك ست تتابعات بروتين محتملة (الشكل ١٥-١) ويفترض فى إطار القراءة الصحيحة أن يكون الأطول وأن يكون مستمر وغير متقطع نتيجة لوجود أى من كودونات الإنهاء (TGA,TAA,TAG) وكلما زاد طول ORF كلما زاد إحتمال أن تمثل جين لأن ORF الطويلة لا تحدث عادة بالصدفة. إن تحديد نهاية مثل هذه ORF يكون أسهل من تحديد بدايتها. إن النهاية الأمينية N_Terminal للبروتين تكون عادة مكونة من حامض الميثوثين لذلك فإن وجود كودون ATG يمكن أن يدل على النهاية 5′ للجين إلا أن الميثونين ليس هو الحامض الأول دائما فى تتابع البروتين ويمكن أن يتواجد فى أماكن آخرى فى السلسلة ويتطلب ذلك إستخدام تقنيات أخرى لتحديد بداية ORF مثل تفضيل إستخدام كودونات معينة Codon Bias لكل كائن (الجدول ١٥-١) ووجود تتابع CpG (CpG Islands). إلا أنه يجب الحرص عند مسح تتابعات النيوكلتيدات لتحديد ORF لأن حدوث أى خطأ قد يعطى سلسلة من الأخطاء فى التتابع النهائى خاصة إذا أدى هذا الخطأ إلى إضافة أو حذف غير صحيح لكودون إنهاء ، كما أن الخطأ قد يحدث نتيجة إضافة أو حذف قاعدة واحدة غير صحيحة مما يجعل التحديد الصحيح للـ ORF أكثر صعوبة.

```
Query Sequence:
            10          20          30          40          50
  0  TCCATTGAGC  CTTATACCAG  TAACATCTAC  ACTCGAAGAT  CTTGTCAGGG
 50  GAATTTCAGA  TTGTGAATCC  TCACTTACTG  AAAGATCTTA  CTGAGCGGGG
100  CTTGTGGAAT  GAAGAGATGA  AAAATCAGAT  TATTGCATGC  AATGGCTCCA
150  TTCAGTTTTC  CTTTTTCAGA  GCATACCAGA  AATTCCTGAT  GACCTGAAGC
200  AACTCTATAA  GACCGTGTGG  GAAATCTCTC  AGAAGACTGT  TCTCAAGATG

Six-Frame Amino Acid Translation:

Forward 0
            10          20          30          40          50
  0  SIEPYTSNIY  TRRSCQGNFR  L!ILTY!KIL  LSGACGMKR!  KIRLLHAMAP
 50  FSFPFSEHTR  NS!!PEATL!  DRVGNLSEDC  SQD

Forward 1
            10          20          30          40          50
  0  PLSLIPVTST  LEDLVRGISD  CESSLTERSY  !AGLVE!RDE  KSDYCMQWLH
 50  SVFLFQSIPE  IPDDLKQLYK  TVWEISQKTV  LKM

Forward 2
            10          20          30          40          50
  0  H!ALYQ!HLH  SKILSGEFQI  VNPHLLKDLT  ERGLWNEEMK  NQIIACNGSI
 50  QFSFFRAYQK  FLMT!SNSIR  PCGKSLRRLF  SR

Reverse 0
            10          20          30          40          50
  0  HLENSLLRDF  PHGLIELLQV  IRNFWYALKK  EN!MEPLHAI  I!FFISSFHK
 50  PRSVRSFSK!  GFTI!NSPDK  IFECRCYWYK  AQW

Reverse 1
            10          20          30          40          50
  0  ILRTVF!EIS  HTVL!SCFRS  SGISGML!KR  KTEWSHCMQ!  SDFSSLHSTS
 50  PAQ!DLSVSE  DSQSEIPLTR  SSSVDVTGIR  LNG

Reverse 2
            10          20          30          40          50
  0  S!EQSSERFP  TRSYRVASGH  QEFLVCSEKG  KLNGAIACNN  LIFHLFIPQA
 50  PLSKIFQ!VR  IHNLKFP!QD  LRV!MLLV!G  SM
```

الشكل (١٥-١): يوجد لكل جين ستة إطارات مفتوحة للقراءة ORF

إستخدام التناظر لتعريف الجينات:

Using Homology to Find Genes:

يمكن تسهيل عملية تحديد هوية الجينات الموجودة فى تتابعات طويلـــة إذا بحثنا عن تتابعات مماثلة (متناظرة) معروف نواتج نسخها (مثل cDNA أو EST

أو حتى لجين فى نوع آخر) وقد أدى النمو السريع فى بيانات EST إلى تغير جذرى فى معدل تحديد الجينات إذ يمكن مسح Screen التتابعات الجينية بسرعة لوجود مواقع EST لتعريف الجينات المحتملة. وتشتق كلونات EST من تتابعات النهاية 3′ غير المترجمة 3′ UTR والتى يتم الحصول عليها بإستخدام بادئ من متعدد T إذا كان م ر ن أ محتويا على ذيل متعدد الأدينين Poly A. وتتميز تتابعات 3′ UTR بخاصيتين حيث يندر وجود أنترونات بها كما أنها تحتوي عادة على تتابعات محفوظة أقل مما فى المناطق الشفرية وتؤدى الخاصية الأولى إلى الحصول على نواتج PCR قصيرة بحيث يمكن إكثارها فى حين تؤدى الظاهرة الثانية إلى سهولة التمييز بين أفراد العائلات الجينية المتشابهة جدا فى مناطقها الشفرية. إلا أن تتابعات EST كثيرا ما تمتد فى إتجاه النهاية 5′ لتصل إلى التتابعات الشفرية و بذلك تتداخل Overlap مع إكسونات محتملة إلا أنه لا يمكن توقع أن تقوم بتحديد و تعريف جميع الإكسونات الشفرية لهذا الجين. ويجب التنويه إلى أنه لا يمكن إفتراض أن جميع EST's تعتبر دلائل موثوق بها لجين أو لجزئ م ر ن أ ناضج ففى بعض الحالات قد تكون مشتقة من تتابعات أنترونات غير معدلة Unprocessed بادئة من مسار Poly A جينومى أو من جينات كاذبة معدلة. ويعد برنامج (BLAST) Basic Local Alignment Search Tool من أكثر البرامج إستخداما لإجراء مسح لمدى التشابه بين التتابع النيوكلتيدى محل الدراسة Query Sequence مع تلك الموجودة فى قواعد بيانات التتابعات. ويبحث البرنامج BLASTN عن درجة التشابه مع قواعد تتابعات النيوكلتيدات فى حين يقوم برنامج BLASTX بترجمة التتابع النيوكلتيدى محل الدراسة (Query) فى الإطارات الستة المحتملة والبحث عن

درجة التشابه مع قواعد بيانات البروتين، ويستخدم برنامج BLASTP الخاص بتتابع البروتين للبحث عن درجة التشابه مع قواعد بيانات البروتينات.

الجدول (١٥-١) النسب المئوية لاستخدام الكودونات الستة للحامض الأميني سيرين في الكائنات المختلفة (تحيز الكودون Codon bias)

الكودون	بكتريا القولون	الدروسفلا	الإنسان	الذرة	الخميرة
AGT	٣	١	١٠	٣	٥
AGC	٢٠	٢٣	٣٤	٣٠	٤
TCG	٤	١٧	٩	٢٢	١
TCA	٢	٢	٥	٤	٦
TCT	٣٤	٩	١٣	٤	٥٢
TCC	٣٧	٤٨	٢٨	٣٧	٣٣

وقبل البدء فى البحث عن التشابه فإنه من المهم التخلص من التتابعات المتكررة Repetitive ومن أى تتابعات خاصة بناقل الكلونة التى قد تكون موجودة ضمن الشظية وذلك بإستخدام برامج متخصصة مثل Vec Screen وRepeat Masker التى تقوم بهذه المهمة بالإضافة إلى التعرف على الجين ويستخدم برامج BLASTP, BLASTN كثيرا فى دراسة الجينوميا المقارنة Comparative Genomics.

ومن أهم خواص مطابقة التتابعات قياس ما إذا كانت نتائج المقارنة تمثل دليلا فعليا للتناظر أى ما هو إحتمال أن يكون التطابق الناتج يمكن توقعه بمحض الصدفة؟

أسس البحث عن التشابه Principles of Similarity Searching:

لا يعطى تتابع د ن أ أو البروتين فى حد ذاته معلومات كافية لتحديد هوية جين ما ولكن لابد من تحليل تلك التتابعات بطرق مقارنة مقابل التتابعات المتاحة فى قواعد البيانات حتى يمكن إستنباط فرضيات تتعلق بدرجات التشابه (القرابة) والوظيفة الممكنة. ويتم ذلك بإستخدام أحد البرامج المشار إليها لمقارنة التتابع محل الدراسة (Query) مع جميع التتابعات الآخرى فى قواعد البيانات وتتم هذه المقارنات فى أزواج Pairwise وتأخذ كل مقارنة درجة Score تعكس درجة التشابة بين التتابع محل الدراسة (Query) والتتابعات التى يجرى مقارنتها به وكلما زادت الدرجة كلما دل ذلك على زيادة درجة التشابه.

ويمكن أن تتم دراسة التشابه Alignment إما بطريقة شاملة Global أو محلية Local. وتتضمن الطريقة الشاملة تحديد درجة التشابه على مستوى تتابعات جزئ د ن أ بأكمله مع ما يقابله من تتابعات فى قواعد البيانات. أما الطريقة المحلية فتعنى بتجزئة التتابعات إلى مناطق صغيرة ودراسة أقصى درجات التشابه لكل منطقة محلية. ويتم التمييز بين التطابقات الحقيقية وتلك الناتجة عن الصدفة بإستخدام تقديرات إحصائية لتحديد إذا كان التطابق راجع للصدفة. وتستخدم معظم برامج مقارنة التتابعات مثل BLAST الطريقة المحلية لتحديد تطابق التتابعات وتبدأ العملية بتقطيع (كسر) التتابع محل الدراسة Query (المجهول) وكذلك تتابعات قواعد البيانات إلى شظايا (كلمات) والبدء فى البحث عن الكلمات المتطابقة Word Matches.

ويكون البحث المبدئى لكلمة بطول W والتى تعطى درجة لا تقل عن T عند مقارنتها بالتتابع محل الدراسة (Query) بإستخدام ماتركس معين من التقييم.

ويستمر تحديد التطابق بين الكلمات ممتدا فى أى من الإتجاهين فى محاولة لتكوين درجة تطابق لا تقل عن خمسة درجات كحد أدنى. وتحدد قيم T سرعة وحساسية البحث ويتم التعبير عن كل زوج من التطابقات بدرجة Score ويتم ترتيب الدرجات المختلفة فى رتب Ranks. ويستخدم ماتركس الدرجات لحساب درجة التطابق بين كل قاعدة وقاعدة أو بين كل حامض أمينى وحامض أمينى ويستخدم ماتركس الوحدة فى حالة د ن أ بحيث تعطى قيمة (+١) إذا كان متطابقا و القيمة صفر إذا لم يكن هناك تطابق. أما فى حالة مقارنة تتابعات الأحماض الأمينية فيستخدم ماتركس الإستبدال والمبنى على حساب قدرة الأحماض الأمينية النسبية على الطفور Relative Mutability.

ويتم تقييم كل حامض أمينى مستبدل بدرجة تتناسب مع مدى إمكانية أن يكون قريبا فى الوظيفة للحامض الأمينى المقابل فى التتابع البروتينى محل الدراسة Query ولا يمثل إلا تغييرا محدودا بحيث لا يؤثر على تركيب ووظائف البروتين ويطلق على مثل هذه الحالة Conservative Change. ومن جهة آخرى إذا أدى الإستبدال بين حامض أمينى صغير مثل الفالين بحامض آخر كبير محتويا على سلسلة جانبية قطبية مثل حامض الأسبارتيك فإن ذلك سيؤدى إلى تأثير كبير على تركيب و وظيفة البروتين مما يعرف Non-Conservative Substitution ولذلك يجب مقارنة تتابعات الأحماض الأمينية ليس على أساس النسبة المئوية للتطابق Identity فقط ولكن أيضا على أساس النسبة المئوية للتشابه Similarity. ويحسب التشابه على أساس القدرة النسبية للأحماض الأمينية على الطفور كما سبق القول وعموما فإن التغيرات المحافظة Conservative Changes تكون أكثر شيوعا على مدار التطور وتكون الدرجة النهائية للتطابق عبارة عن مجموع درجات كل موقع. وتسمى المواقع التى تقابل

الحرف (الممثل لرمز الحامض الأميني) بفراغ Null بالفجوات Gaps وتكون درجة الفجوة سالبة وحيث أن حدث طفرى واحد قد يؤدى إلى إضافة أو حذف حامض أمينى أو أكثر فإن مجرد وجود الفجوة يكون عادة أكثر أهمية عن طول هذه الفجوة. تحسب المعنوية لكل تطابق كدرجة إحتمال (p) أو درجة توقع Expectation (E) وهما مجرد طريقتان لتمثيل معنوية التطابق لأن $P=1^{-e}$ وكلما زادت قيمة P لتقترب من (١) كلما زادت معنوية التطابق فى جين كلما إنخفضت قيمة E كلما دل على زيادة المعنوية. ويبين المثال التالى بعض التفاصيل الخاصة بهذه العملية.

إستخدام برنامج BLAST لتحديد تشابه تتابعات البروتين:

لإيجاد تتابعات البروتينات فى قواعد البيانات التى تتشابه مع تتابع البروتين محل الدراسة فإن BLAST يعتبر أفضل برنامج للتنقيب عن البيانات Data Mining ، ويوجد نوعان من هذا البرنامج يمكنهما التعامل مع تتابعات البروتينات وهما:

١- Blastp للمقارنة بين تتابعات بروتين ما مع قواعد البيانات للبروتينات.

٢- tblastn للمقارنة بين تتابعات البروتين مع قواعد البيانات للتتابعات النيوكلتيدية.

ويعتمد إختيار أى من النوعين على الغرض من الدراسة ، فإذا كان المطلوب معرفة وظيفة البروتين محل الدراسة فيفضل إستخدام Blastp لمقارنة هذا البروتين مع البروتينات الآخرى الموجودة فى قواعد البيانات ، أما إذا كان الهدف هو إكتشاف جينات جديدة تشفر لبروتينات بسيطة فيستخدم tplastn لمقارنة المدروس مع تتابعات د ن أ فى قواعد البيانات والتى يتم ترجمتها إلى

بروتينات فى ستة إطارت مفتوحة للقراءة ORF. ويقوم برنامج Tplastn بهذه الترجمة أوتوماتيكا بدون تدخل من الباحث وكل ما على الباحث التأكد منه هو أن تكون تتابعات د ن أ محل الدراسة Query فى الإتجاه 3'→5' أو إذا كانت بروتين أن تكون التتابعات من النهاية الأمينية إلى النهاية الكربوكسيلية.

تفسير نتائج BLAST:

Understanding BLAST Output:

بعد الإنتهاء الدقيق من تنفيذ خطوات إدخال بيانات التتابع المطلوب (المثال هنا بروتين النيوكليولين Nucleolin) إلى الكمبيوتر الشخصى المزود بالبرامج المناسبة والمتصل بشبكة المعلومات ، سنحصل على ثلاث مخرجات (أقسام) تمثل النتائج الرئيسية التى يمكن من خلالها الحكم على درجة التشابه بين البروتين محل الدراسة مع أى من البروتينات الموجودة فى قواعد البيانات.

ويتم ظهور هذه الأقسام بالتتابع كما يلى:

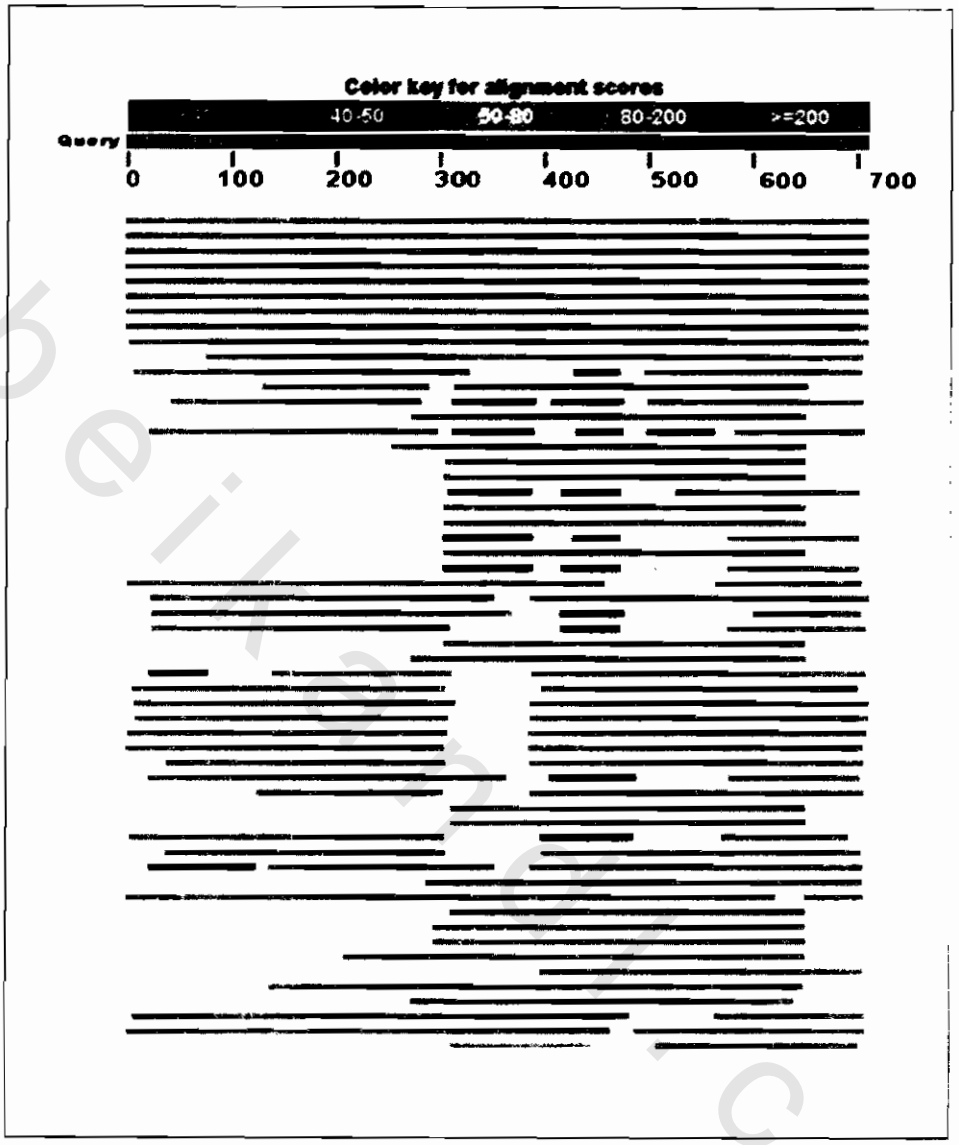
١- عرض بيانى Graphic Display: بين مناطق التشابه بين التتابع محل الدراسة مع التتابعات الأخرى.

٢- قائمة إصابة الهدف Hit List: وتعطى أسماء التتابعات المشابهة للتتابع تحت الدراسة مرتبة حسب درجات التشابه.

٣- المحاذاة Alignment (التطابق): وتبين جميع درجات التوازى بين التتابع تحت الدراسة وتلك المشابهة لها (النجاحات).

أولاً: العرض البيانى The Graphic Display:

يساعد العرض البيانى (الشكل ١٥-٢) على فهم نتائج البحث. ويشتمل العرض البيانى على التتابع المدروس الذى يكون فى القمة (أعلى خط أفقى) ويمثل كل خط Bar جزء من تتابع آخر مشابه للتتابع محل الدراسة وتحديد المنطقة التى يحدث بها هذا التشابه. ويوجد كود للألوان للدلالة على درجة التشابه حيث تدل الخطوط الحمراء على أعلى درجات التشابه ، فى حين تدل الخطوط الوردية Pink على درجة تشابه أقل جودة ، وتشير الخطوط الخضراء على إنخفاض حاد فى درجة التشابه ، أما الخطوط الزرقاء أو السوداء فتعنى إنعدام التشابه تقريبا.



شكل (١٥-٢): عرض بياني Graphic Display لنتائج بحث المطابقة ببرنامج BLAST بين مناطق التشابه بين التتابع محل الدراسة مع تتابعات قواعد البيانات المقارنة (اللون الأحمر والوردى يمثلان أعلى درجة تشابه فى حين أن اللون الأزرق والأسود يمثلان اقل تشابه أو إنعدام التشابه)

وعموما يمكن إعتبار الخطوط الحمراء والوردية Pink والخضراء دلالات جيدة للتشابه. أما الخطوط السوداء فتدل على إنعدام التشابه مع التتابع المدروس ويقال أنها فى "المنطقة الرمادية" "Twilight Zone" أى تلك التى تكون نسبة التشابه فيها أقل من ٢٥% كما سيأتى بعد (يمثل ٢٥% بالنسبة لتتابعات الأحماض الأمنية فى البروتين و ٧٠% لتتابعات النيوكلتيدى فى د ن أ الحد الأدنى للتناظر Homology).

ثانيًا: قائمة إصابة الهدف (النجاحات) The Hit List:

تعطى هذه القائمة مؤشرات مباشرة للحكم على ما إذا كان التتابع تحت الدراسة يشبه أى من التتابعات الموجودة بالفعل فى قواعد البيانات ومدى الثقة فى مصداقية وجودة هذا التشابه (الشكل ١٥-٣) ويحتوى كل خط على أربعة مكونات هامة:

١- الرقم الكودى للتتابع والإسم The Sequence Accession Number and Name:

مما يعطى الفرصة للوصول إلى قاعدة البيانات المحتوية على هذا التتابع للحصول على المزيد من المعلومات عن هذا التتابع.

٢- التوصيف Description:

يعطى وصف كل تتابع فكرة سريعة عما إذا كان هذا التتابع ذا قيمة مقارنة بالتتابع محل الدراسة.

٣- نقط النجاح The Bit Score:

وهى مقياس للمعنوية الإحصائية للتشابه. وكلما زادت قيمتها (مقدارها) دل ذلك على درجة تشابه عالية. وإذا إنخفضت الدرجة عن ٥٠ نقطة فإن معنى ذلك أنه ليس هناك تشابه يذكر.

٤- قيمة The E–Value E (قيمة التوقع Expectation Value):

وتعطى تقديرا لعدد المرات المتوقعة للحصول على نفس التشابه بمحض الصدفة. وكلما إنخفضت قيمة E كلما دل ذلك على درجة تشابه عالية بين التتابعات مما يعطى ثقة أكبر بأن هذا التتابع مناظر بالفعل للتتابع تحت الدراسة ، إذ أن قيمة E قريبة جدا من الصفر تعنى أن هذه التتابعات متطابقة ، وعموما فإن أى تشابهات بقيمة E أعلى من (10^{-4}) لا تؤخذ فى الإعتبار حيث أنها تقع فى المنطقة الرمادية Twilight Zone وتدل على إنعدام التشابه تقريبا بين التتابعات.

```
                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

gi|128843|sp|P09405|NUCL_MOUSE  Nucleolin (Protein C23)            676  0.0
gi|128844|sp|P13383|NUCL_RAT  Nucleolin (Protein C23)              627  e-179
gi|128842|sp|P08199|NUCL_MESAU  Nucleolin (Protein C23)            598  e-171
gi|128841|sp|P19338|NUCL_HUMAN  Nucleolin (Protein C23)            509  e-144
gi|128840|sp|P15771|NUCL_CHICK  Nucleolin (Protein C23)            326  1e-88
gi|464252|sp|P20397|NUCL_XENLA  Nucleolin (Protein C23)            308  2e-83
gi|12229875|sp|Q13310|PAB4_HUMAN  Polyadenylate-binding prot...    101  6e-21
gi|310354|sp|P11940|PAB1_HUMAN  Polyadenylate-binding prote...     96  2e-19
gi|417441|sp|P04147|PABP_YEAST  Polyadenylate-binding protei...    96  3e-19
gi|129535|sp|P29341|PAB1_MOUSE  Polyadenylate-binding protei...    96  3e-19
gi|12229876|sp|Q15097|PAB2_HUMAN  Polyadenylate-binding prot...    93  1e-18
gi|1171978|sp|P42731|PAB2_ARATH  Polyadenylate-binding prote...    92  3e-18
gi|1352709|sp|P20965|PABP_XENLA  Polyadenylate-binding prote...    89  2e-17
gi|3123239|sp|P31209|PABP_SCHPO  Polyadenylate-binding prote...    87  1e-16
gi|128576|sp|P27476|NSR1_YEAST  Nuclear localization sequenc...    81  6e-15
gi|1171979|sp|Q05196|PAB5_ARATH  Polyadenylate-binding prote...    79  4e-14
gi|585638|sp|P21187|PABP_DROME  Polyadenylate-binding protei...    78  7e-14
gi|12229883|sp|Q9ZQA8|PABX_ARATH  Probable polyadenylate-bin...    74  8e-13
gi|133249|sp|P19684|ROC5_NICSY  33 kDa ribonucleoprotein, ch...    74  1e-12
gi|126436|sp|Q64380|PAB3_ARATH  Polyadenylate-binding prot...    74  1e-12
gi|417556|sp|P32588|PUB1_YEAST  Nuclear and cytoplasmic poly...    73  1e-12
gi|133247|sp|P28644|ROC1_SPIOL  28 kDa ribonucleoprotein, ch...    66  3e-10
gi|13124200|sp|Q12926|ELV2_HUMAN  ELAV-like protein 2 (Hu-an...    64  8e-10
gi|123734|sp|P26378|ELV4_HUMAN  ELAV-like protein 4 (Paraneo...    64  8e-10
gi|13124206|sp|Q60899|ELV2_MOUSE  ELAV-like protein 2 (Hu-an...    64  8e-10
gi|2500580|sp|Q61701|ELV4_MOUSE  ELAV-like protein 4 (Parane...    64  9e-10
```

الشكل (١٥-٣): قائمة إصابة الهدف Hit List وتعطى أسماء التتابعات المشابهة للتتابع محل الدراسة Query مرتبة حسب درجات التشابه

ثالثًا: المحاذاه The Alignments (التطابق):

يعتبر عرض نتيجة المحاذاة بمثابة حجر الزاوية فى بحث BLAST لأنها تعطى تفاصيل دقيقة عن حقيقة التشابة وأهميته (الشكل ١٥-٤) ويمكن منه ستنباط الخواص التالية:

١- نسبة التطابق The Percent Identity:

وهى تعطى بديل قوى لقيمة E وتجدر الإشارة مرة أخرى إلى أن نسبة تطابق أعلى من ٢٥% تعتبر نسبة مشجعة. وتعطى الإشارات الموجبة مقياسا لأجزاء التتابعات المتطابقة أو المتشابهة بشكل كبير وممثلة بالعلامة (+) فى الشكل أعلاه ، وتدل أماكن الفجوات على أماكن التتابعات التى تخلو من التشابه.

٢- الطول Length:

وتمثل طول منطقة التشابه والتى تدل على مدى إستمرار التشابه بدون إنقطاع بين التتابع محل الدراسة والتتابع المقارن وتمثل نتيجة تشابه كل منطقة بثلاث سطور:

السطر العلوى: يمثل تتابع الشظية محل الدراسة Query.

السطر السفلي: يمثل التتابع المقارن المشتق من قواعد البيانات The Hit or The Subject.

السطر الأوسط بين التتابعين: : ويحتوى على علامة + تمثل الأحماض الأمينية المتشابهة وحرف يمثل الأحماض المتطابقة أو فراغ يمثل عدم وجود تشابه و يدل وجود مناطق xxxxx فى التتابع محل الدراسة Query على وجود عدد كبير من الأحماض الأمينية المترادفة (من نفس النوع) والتى يقوم BLAST بحجبها أتوماتيكيا Automatically Masked ويطلق عليها المناطق المنخفضة التعقيد Low-Complexity Segments

وقد يؤدى إظهارها إلى حدوث مشاكل فى البحث عن التشابه وإعطاؤها علامات × تجعل BLAST بتجاهل المناطق المقابلة ويحدث هذا فى التتابع محل الدراسة فقط.

وتدل الإعداد Numbers على جانبى القطعة المدروسة من التتابعات على عدد الأحماض الأمينية الذي يشمله كل تتابع فى التتابع المدروس والتتابع المقابل فى قاعدة المعلومات.

ويجب أن لا يحتوي آى تتابع على عدد كبير من الفجوات وأن يتكون من عدد قليل من القطع ذات تشابه مرتفع بدلا من وجود نقط متطابقة متفرقة هنا وهناك ويكون لذلك أهمية خاصة إذا كان التشابه يقع فى المنطقة الرمادية.

```
>gi|464252|sp|P20397|NUCL_XENLA  Nucleolin (Protein C23)
          Length = 651

 Score =  308 bits (789), Expect = 2e-83
 Identities = 203/381 (53%), Positives = 267/381 (69%), Gaps = 22/381 (5%)

Query: 280 KRKKEMTXXXXXXXXXXXXXVEGSEPTTPFNLFIGNLNPNKSVNELKFAISELFAKNDLAV 339
           KRKKEM           +    ++FIGNLN  K  +ELK  A+ E F+K +L +
Sbjct: 209 KRKKEM----PKTIPEAKKTKTDTASEGLSIFIGNLNSTKEFDELKDALREFFSKKNLTI 264

Query: 340 VDVRTGTNRKFGYVDFESAEDLEKALELTGLKVFGNEIKLEKP------KGRDSKKVRAA 393
           D+R G ++KFGYVDF S E++EKAL+LTG K+ G E+K+EK        K  ++KK R +
Sbjct: 265 QDIRIGNSKKFGYVDFSSEEEVEKALKLTGKKILGTEVKIEKAMAFDKNKTAENKKERDS 324

Query: 394 RTLLAKNLSFNITEDELKEVFEDAMEIRL-VSQDGKSKGIAYIEFKSEADAEKNLEEKQG 452
           RTL  KN+ ++ T +EL+E+FE+A +IR+   +DG +KGIAY+EF +E +A K LEEKQG
Sbjct: 325 RTLFVKNIPYSTTVEELQEIFENAKDIRIPTGKDGSNKGIAYVEFSNEDEANKALEEKQG 384

Query: 453 AEIDGRSVSLYYTGEKGQRQERTGKTSTWSGESKTLVLSNLSYSATKETLEEVFEKATFI 512
           AEI+GRS+ + +TGEK Q      +G      G+SK LV++NLSYSAT+++L EVFEKAT I
Sbjct: 385 AEIEGRSIFVDFTGEKSQ---NSGNKKGPEGDSKVLVVNNLSYSATEDSLREVFEKATSI 441

Query: 513 KVPQNPHGKPKGYAFIEFASFEDAKEALNSCNKMEIEGRTIRLELQ-------GSNSRSQ 565
           ++PQN  G+  KG+AFIEF+S EDAK+A++SCN  EIEGR+IRLE         G    S
Sbjct: 442 RIPQN-QGRAKGFAFIEFSSAEDAKDAMDSCNNTEIEGRSIRLEFSQGGGPQGGGRGGSA 500

Query: 566 PSKTLFVKGLSEDTTEETLKESFEGSVRARIVTDRETGSSKGFGFVDFNSXXXXXXXXXX 625
```

الشكل (١٥-٤): أزواج التتابعات المتشابهة Pair Wise Alignments الناتجة من BLAST

تحليل ر ن أ غير الشفرى وتتابعات د ن أ غير الجينية:

Analysis of Non-Coding RNA and Extragenic DNA:

يتكون الجينيوم فى مميزة النواة متعددة الخلايا من تتابعات أكثر بكثير من تلك الخاصة بالتشفير للجينيات فعلى سبيل المثال ، يمكن بتحليل ر ن أ غير الشفرى ومناطق تنظيم التعبير الحصول على معلومات هامة. ومن السهل الحصول على rRNA (ر ن أ الريبوسومى) وهو من أكبر أنواع ر ن أ غير الشفرية. ويمكن التعرف عليه بتحليل تشابه التتابعات . كما يمكن التعرف على tRNA بإستخدام برنامج tRNAscan-SE والذى يمكنه البحث عن تراكيب مميزة مثل قدرة tRNA على تكوين عروات دبوس الشعر Hairpins. وتعد المناطق المختصة بتنظيم تعبير الجينات بصفة خاصة هامة فى تتابعات الجينيوم وقد تم حتى الآن التعرف على عدد محدود من مواقع الإرتباط بعوامل النسخ وذلك بالطرق التقليدية. ويعد وجود مثل هذه المواقع فى التتابع محل الدراسة Query مؤشرا لمنطقة التنظيم هذه و لكن هذه التتابعات تتميز بالقصر النسبى ويمكن أن توجد بمحض الصدفة. ويمكن الحصول على أدلة أفضل لإثبات وجود هذه المناطق عن طريق عمل دراسة مقارنة للتتابعات المحفوظة التى تسبق نفس الجين Upstream فى جينومين متقاربين مثل الفأر والإنسان.

تحديد وظيفة جين جديد:

Identifying the Function of a New Gene:

إن أبسط طريقة للتعرف على وظيفة جين جديد تتم من خلال البحث عن نظائر Homologoues فى قواعد بيانات البروتينات وإذا أعطى البروتين الجديد المشفر بإطار قراءة ORF غير مميز درجة تشابه معنوية مع بروتين آخر معروف الوظيفة، فإن ذلك يثبت أن ORF محل الدراسة هو فى الحقيقة جين

جديد وغالبا مايتم تحديد وظيفته ويتم ذلك من خلال أحد برامج BLAST مثل BLASTP أو PSI-BLAST أو SWISS-PROT. ويلخص الجدول (١٥-٢) كيفية الإختيار الصحيح لأحد برامج BLAST حسب متطلبات الدراسة.

الجدول (١٥-٢) إختيار برنامج BLAST المناسب للدراسات المختلفة لتشابه التتابعات

نوع الدراسة	البرنامج المناسب
البحث عن وجود جينات فى الجينوم **Finding Genes in a Genome**	يتم قطع تتابعات الجينوم إلى تتابعات متداخلة صغيرة (٢-٥ ك قاعدة) ويستخدم برنامج Blastx لتقدير درجة التشابه مقابل NR (قواعد بيانات البروتين غير المتكررة Non-Redundant) و يعطى ذلك نتائج أفضل إذا لم يكن الجينوم محتويا على أنترونات (مثل البكتريا).
التنبؤ بوظيفة بروتين **Predicting a Protein Function**	يستخدم SWISS-PROT وعند الحصول على نسبة تشابه جيدة (أعلى من ٢٥% تطابق) على مستوى الطول الكلى للبروتين فإن ذلك يعطى تأكيد على أن البروتين محل الدراسة يقوم بوظيفة مشابهة للبروتين المقارن المشتق من SWISS-PROT.
التنبؤ بالتركيب الثالثى 3-D للبروتين **Predicting a Protein 3-D Structure**	يستخدم Blastp مقابل PDB (قاعدة بيانات تركيب البروتين) وعند الحصول على درجة تشابه مناسبة (أعلى من ٢٥% تطابق) فإن معنى ذلك أن البروتين محل الدراسة تركيبه الثالثى مشابه للبروتين المقارن فى قاعدة البيانات.

العثور على أفراد فى عائلة من البروتينات **Finding Protein Family Members**	يستخدم Blastp أو PSI-BLAST ويقارن ببرنامج NR وبعد الحصول على جميع أفراد العائلة يمكن عمل بحث على قواعد البيانات للتشابه بين التتابعات المتعددة Multiple Sequence Analysis ورسم الشجرة التطورية Phylogenetic Tree.