آرثر م. ليسك

سلسلة كراسات «عروض»

إجتهادات حديثة حول العلم والمستقبل

مقدمة فى البيومعلوماتية

عرض: د./ محمد باسم عاشور

ISO 9002

المكتبة الأكاديمية

شركة مساهمة مصرية - القاهرة

مقدمة فى البيومعلوماتية

Introduction to Bioinformatics

تأليف

Arther M. Lesk

الناشر

Oxford Univ. Press, 2002

عــرض

د. محمد باسم عاشور

أستاذ علم السمية – جامعة الزقازيق

الناشر

المكتبة الأكاديمية

ش.م.م

٢٠٠٦

هذه الكراسة تقدم عرضًا تفصيليًا لكتاب :

Introduction to Bioiutormatics

Arther M. Lesk

Oxford Univ. Press, 2002

حقوق النشر

الطبعة الأولى ٢٠٠٦م - ١٤٢٥هـ

المكتبة الأكاديمية

شركة مساهمة مصرية

راس المال المصدر والمدفوع ١٨,٢٨٥,٠٠٠ جنيه مصرى

١٢١ شارع التحرير - الدقى - الجيزة

القاهرة - جمهورية مصر العربية

تليفون : ٧٤٨٥٢٨٢ - ٣٣٦٨٢٨٨ (٢٠٢)

فاكس : ٧٤٩١٨٩٠ (٢٠٢)

هذه السلسلة

هى الثالثة فى مشروع "الكراسات"، الذى تصدره "المكتبة الأكاديمية". والكراسات تعنى بمحورين كبيرين: العلم والمستقبل. لذلك فقد حملت السلسلة الأولى عنوان "كراسات مستقبلية"، وقد بدأ ظهورها عام ١٩٩٧. وفى عام ١٩٩٨ ظهرت السلسلة الثانية تحت اسم "كراسات علمية". وقد فكرنا فى البداية أن تضم السلسلتين، بجانب التأليف والترجمة، عروضاً مطولة لبعض الإصدارات المهمة، التى لا تلاحقها حركة الترجمة. إلا أن أنشط أعضاء أسرة الكراسات، وللكراسات أسرة ممتدة ترحب دائماً بالأعضاء الجدد، أقول أن أنشط الأعضاء الصديق الدكتور محمد رؤوف حامد، الأستاذ بهيئة الرقابة الدوائية، اقترح أن تصدر العروض فى سلسلة خاصة بها. وقد كان اقتراحاً موفقاً كما أرجو أن يوافقنى القارئ.

والكتب المختارة للعرض فى السلسلة لاتأتى فقط من اقتراحات هيئة التحرير، حيث قدم أعضاء الأسرة مقترحاتهم التى حظيت بالترحيب، والباب مفتوح لكل من يرغب فى المشاركة. وإذا كانت السلسلة قد بدأت بمجموعة من الكتب الصادرة بالإنجليزية، فإننا نطمح أن تشمل العروض القادمة كتبا تصدر فى لغات أخرى، لاتشملها عادة خطط الترجمة كاليابانية والروسية والصينية، بالإضافة إلى الفرنسية والألمانية. فرغم أن الأخيرتين أكثر حظاً نسبياً، إلا أن كم المترجم والمعروض لايقارن بما يتم بالنسبة للإنجليزية.

والحديث عن "العروض" يذكرنا بالجهود السابقة، التى لاننكرها، بل نحاول أن نكمل مسيرتها. فبالنسبة للعروض الموسعة، تذكر جهود الهيئة العامة للإستعلامات بالنسبة للمجالات التى تهمها. كما أن العروض المتوسطة، التى أصدرتها هيئة الكتاب فى التسعينيات، ضمن سلسلة "تراث الإنسانية" لايمكن إغفالها. وهما مثلات يقصد بهما الإعتراف بفضل السبق، دون أن ندعى الحصر. وإن كنا، فى الوقت نفسه، نظن أن السلسلة الحالية هى الأولى التى تعنى بالعرض التفصيلى للكتب.

هذه الكراسة

هى الأولى التى يمكن للقارئ المتخصص أن يطالعها وهو متصل بالإنترنت ليمارس بعض تطبيقات البيومعلوماتية الموضحة بها. لكنها أيضاً موجهة للقارئ العام، الذى يريد أن يلم بهذا المجال الجديد، الذى ذكر أنه قد حول البيولوجيا (علم الحياة) إلى علم معلوماتى. ولأن المجال يجمع بين البيولوجيا والمعلوماتية فالبيولوجى المهتم بتطبيقاته يحتاج إلى الإلمام بالقاعدة المعرفية والمهارية المعلوماتية التى تمكن من ذلك. والمشتغل بالمعلومات يحتاج بدوره إلى مزيد من المعرفة عن البيولوجيا التى سيوظف تخصصه فيها. والقارئ العام، الذى يسعى إلى تثقيف نفسه علمياً، يهمه أن يفهم كيف تساهم البيومعلوماتية فى تشكيل صورة المستقبل فى الطب والزراعة والبيئة... إلخ. لذلك رحبت السلسلة بالعرض المتميز لكتاب أرثر ليسك "مقدمة فى البيومعلوماتية"، الذى أعده الدكتور محمد باسم عاشور، أستاذ علم السمية ووكيل كلية الزراعة جامعة الزقازيق، الذى سبق أن قدمنا له كراسة مستقبلية عن التكنولوجيا الحيوية الزراعية.

أ. د. أحمد شوقى

يناير ٢٠٠٦

تقديم

البيومعلوماتية مجال علمى تطبيقى جديد، زادت أهميته والحاجة اليه مع توفر الكم الهائل من المعلومات والتى نتجت عن مشروع الجينوم البشرى، وغيره من جينوم كائنات عديدة، وكذلك ضرورة تشخيص فيروسات تسبب أمراض خطيرة فى الكائنات ومايتطلبه من تحديد تتابعات والتعرف على تراكيب، وانتاج أمصال، وتصميم دواء، وحلول لمشاكل المقاومة للمضادات الحيوية والدواء والمبيدات.

تقدم البيومعلوماتية مجموعة من المهارات المتعلقة بأدوات جمع المعلومات والتنقيب عن البيانات وبناء المعرفة فى البحوث الجارية والتطبيقات الصناعية والأكلينيكية وغيرها. وذلك باستخدام برامج كمبيوتر لعمل استنتاجات من أرشيف بيانات البيولوجيا الجزيئية الحديثة، والربط بينها والوصول الى تنبؤات مفيدة. ويتحقق ذلك ليس فقط عن طريق تطوير وزيادة السعة التخزينية ولكن أيضا بابتكار النماذج الحسابية والصور التقديرية لمعالجة البيانات.

ومجال البيومعلوماتية هو نتاج التزاوج بين علوم البيولوجيا وعلوم الكمبيوتر وهندستها ويحظى باهتمام المشتغلين بتلك العلوم. لذلك من المهم بالنسبة لعلماء البيولوجيا دراسة الكمبيوتر وكذلك تدريس البيولجيا للمتخصصين فى علوم الكمبيوتر. حيث يتطلب هذا المجال تكامل وفهم للخلفية البيولوجية مع اكتساب مهارات ضرورية فى مجال الكمبيوتر.

ويهدف كتاب "مقدمة فى البيومعلوماتية الى بناء وتطوير خلفية عن مجال البيومعلوماتية بدون الاعتماد على طرق معقدة فى علوم الكمبيوتر أو مهارات برمجة متقدمة. وذلك بدعم وتشجيع استخدام الأدوات الحسابية والصور التقديرية بطريقة مرضية وشيقة. وكذلك يتيح الكتاب للقارىء التفاعل مع موضوع الدراسة، وتطويع المادة العلمية طبقا للحاجة مع توفير مصادر لمواقع الكترونية متنوعة.

دكتور محمد باسم عاشور

الفصل الأول

مقدمة

علم البيولوجى من العلوم التى تعتمد على المشاهدة وحديثا أصبح من العلوم الاستنتاجية حيث تغيرت طبيعة النتائج. ومن المعروف للجميع أن جميع المشاهدات البيولوجية الأساسية تشتمل على سرد لروايات بدرجات متفاوتة من الدقة. الا أنه فى الفترة الأخيرة أصبحت البيانات ليست فقط أكثر كما ودقة، ولكن متحفظة وحكيمة كما هو الحال فى بيانات تتابعات االنيكليوتيدات فى الأحماض النووية.

ومن الممكن الآن تقدير تتابع الجينوم لكائن منفرد أو نسخة منه ليس فقط كاملا بل على نحو صحيح ودقيق. ومع أنه ليس من الممكن تجنب الخطأ التجريبى الا أنه منخفض للغاية بالنسبة للتتابع الجينومى.

وكذلك من الخصائص الواضحة لبيانات البيومعلوماتية أنها ذات كم هائل جدا. فمثلا يحتوى بنك البيانات لتتابع النيكليوتيد على X١٦ ١٠*٩ من القواعد (Gbp ١٦). لو استخدم الحجم التقريبى للجينوم البشرى (٣,٢ *١٠ ٩ حرفا) كوحدة فان ذلك يعادل (buges, an apt name ٢). وللمقارنة المرجعية فان buge ١ يقابل عدد الحروف التى ظهرت فى اعداد نيويورك تايمز لمدة ٦ سنوات. تحتوى قاعدة البيانات لتراكيب الجزيئات الكبيرة Macro molecules (الأبعاد الثلاثية للبروتينات بمتوسط طول ٤٠٠ حمض أمينى) على ١٦٠٠٠ قيد.

شجع الكم الهائل المتاح من البيانات العلماء الى السعى نحو أهداف قياسية وطموحة مثل:

- التأكيد على قدرة العلماء على الرؤية الواضحة والكاملة للحياة. بمعنى الفهم المتكامل لبيولوجيا الكائنات كأنظمة مترابطة ومعقدة.

- ايجاد وربط علاقة التتابع والتركيب ثلاثى الأبعاد والتداخلات والوظيفة لبروتينات وأحماض نووية منفردة وكذلك لمعقد البروتين – الحامض النووى.
- استخدام بيانات لكائنات معاصرة كأساس للحركة الزمنية للخلف والأمام للاستدلال الزمنى على وقائع منذ تاريخ النشوء ووصولا دراسات العلماء المتأنية لتطور الأنظمة البيولوجية.
- دعم التطبيقات فى مجال الطب والزراعة وغيرها من المجالات العلمية.

سيناريو A Scenario:

كمقدمة سريعة لدور طرق الحساب الآلى فى البيولوجيا الجزيئية، دعنا نتخيل أن هناك مأساة فى المستقبل حيث ظهور فيروس جديد يسبب مرضا وبائيا للانسان أو الحيوان. فى هذه الحالة سوف يقوم العلماء بعزل المادة الوراثية معمليا وتحديد التتابع وحينئذ تستخدم برامج الكمبيوتر.

وباستعراض هذا الجينوم الجديد ومضاهاته بالبيانات الوراثية المعروفة والموجودة فى بنك البيانات سوف يتم تعريف وتشخيص ذلك الفيروس وكذلك تحديد علاقته بالفيروسات التى سبق دراستها (١٠). تستمر بعد ذلك الدراسات بهدف انتاج علاج مضاد للفيروس. وبما أن الفيروسات تحتوى على جزيئات بروتين والتى تعتبر كأهداف مناسبة للعقاقير التى تتداخل مع تركيب ووظيفة الفيروس وأن تتابع الأحماض الأمينية للبروتينات عبارة عن رسائل مكونة من ٢٠ حرفا من الحروف الهجائية، لذا باستخدام تتابع الحمض النووى دنا DNA يمكن بالاستعانة ببرامج الحاسب الآلى استنتاج تتابعات الحمض الأمينى فى واحد أو أكثر من بروتينات الفيروس والتى لها دورا حاسما فى التكرار والمضاهاة (٠١).

ومن خلال تتابع الحمض الأمينى سوف تقوم البرامج بالحساب الآلى لتراكيب تلك البروتينات على أساس أن تتابع الأحماض الأمينية يحدد التركيب ثلاثى الأبعاد لها وبالتالى خصائصها الوظيفية. فى بداية الأمر

سوف يتم مراجعة البيانات الموجودة فى بنك المعلومات والخاصة بالبروتينات قريبة الصلة والمعروف تركيبها (١٥). اذا تم التعرف على أحد التراكيب فان مشكلة التنبؤ بالتركيب الجديد سوف تختزل الى أدنى درجة ممكنة (التنبؤ بالتغيرات فى التتابع) وهنا يمكن التنبؤ بتركيب الأماكن المستهدفة للبروتين باستخدام نماذج التماثل Homology modeling (٢٥). أما فى حالة عدم وجود تراكيب ذات علاقة ويبدو أن بروتين الفيروس جديد تماما فان عملية التنبؤ بالتركيب يجب أن تجرى بالكامل *ab initio* (٥٥). وسوف يقل فى المستقبل تكرار مثل هذه الحالات حيث أنه بمرور الوقت سيحدث نمو فى حجم البيانات التى يمكن تزويد البنك بها وبالتالى تزداد المقدرة على الكشف عن تراكيب عديدة.

بمعرفة تركيب بروتين الفيروس يصبح من الممكن تصميم مادة للعلاج. وحيث تمتلك البروتينات أماكن على سطح الجزيء تتناسب مع وظيفة البروتين والتى يمكن تعطيلها، فانه من الممكن تعريف وتصميم جزيء صغير متوافق من حيث الشكل والشحنة مع المكان المستهدف على سطح البروتين ليعمل كدواء مضاد للفيروس (٥٠). وهناك طريقة بديلة تعتمد على تصميم أجسام مضادة لبروتين الفيروس ثم تصنيعها واستخدامها لمعادلة الفيروس (٥٠).

يقوم هذا السيناريو على أسس ثابتة وليس هناك مجالا للشك أنه فى يوم ما سوف يطبق كما هو.

هناك سبب وحيد يحول دون تطبيق ذلك السيناريو لموجهة الفيروس المسبب لمرض الايدز وهو أن تلك الفيروسات تمتلك قدرة ما على الحماية الذاتية. وعلماء الكمبيوتر عند قراءتهم لهذا الكتاب يدركون أن الأرقام المذكورة بين أقواس فى هذا الكتاب للم تستخدم للاستدلال على مراجع ولكنها تتبع نظام D. E. Knuth لفهرسة درجة صعوبة مشكلة تحت البحث كما ورد فى كتابه فن برمجية الحاسب الآلى The art of computer

programming حيث تدل الأرقام أقل من ٣٠ على مشـــاكل ذات حلـــول موجودة بالفعل أما الأرقام الأعلى تدل على موضوعات قيد البحث.

وأخيرا فانه يجب أن يكون من المعلوم أن الطرق التجريبية البحتة لايجـلد مواد مضادة للفيروس سوف تظل ولسنوات عديدة قادمة أكثر نجاحا مـــن التوجهات النظرية.[٦]

الدوجما: مركزيا وطرفيا
Dogmas: Central & Peripheral

تعتبر مادة الوراثة دنا DNA وفى بعض الفيروسات رنا RNA هى سـجل المعلومات فى الكائنات (طبيعة التطور والنشاط لكل فـــرد). وكمـــا هـــو معروف فان جزيئات دنا عبارة عن سلسلة طويلة خطية تحتـــوى علـــى رسالة مكونة من أربعة حروف هجائية. والرســـالة طويلـــة حتـــى فـــى الكائنات الدقيقة تتكون من ٦*١٠ حرفا تماما. ويتضمن تركيـــب الدنـــا على آليات للنسخ الذاتى وتكوين البروتينات. يؤدى الـــتركيب الحلزونـــى المزدوج والمتماثل داخليا الى نسخ دقيق. ومع أن النسخ الدقيق ضـــروري لثبات الصفة الوراثية آلا أن بعض عمليات النسخ غير الدقيق أو آليـــات نقل مادة وراثية غريبة تكون أحيانا ضرورية لحدوث النشوء فى الكائنات اللاجنسية.

شرائط الحلزون المزدوج تكون فى وضع متوازيـــا وعكســـيا ويعـــرف بالاتجاه ٣-٥ وذلك بالنسبة لأماكن حلقة الداى أوكســـى ريبـــوز وعنـــد ترجمتها الى بروتين يقرأ تتابع الحمض النووى فى الاتجاه ٥-٣. ويتـــم تنفيذ المعلومة الوراثية من خلال تخليق الرنا والبروتينات. والبروتينـــات هى الجزيئات المسئولة عن غالبية التراكيب والأنشـــطة فـــى الكائنـــات. فالشعر والعضلات والانزيمات والمستقبلات والأجسام المضادة جميعـــها بروتينات وكل من الأحماض النووية والبروتيتات جزيئات تتكـــون مـــن سلسلة طويلة خطية. تتكون الشفرة الوراثية من ثلاثة حروف متتالية مـــن تتابع الدنا تحدد أحماض أمينية متتالية وبامتداد تتابعات الدنا يشفر تتابعات

الأحماض الأمينية فى البروتينات. كذلك تتكون البروتينات من ٢٠٠ – ٤٠٠ حمض ــ أمينى تتطلب من ٦٠٠ – ١٢٠٠ حرفا من رسائل الدنا لتحديدها. تخليق جزيئات الرنا أيضا يتحكم فيه تتابعات الدنا بالرغم من أنه فى معظم الكائنات ليس كل الدنا يترجم الى بروتينات ورنا. بعض المناطق فى تتابع الدنا يختص بالتحكم فى آليات محددة كما توجد كميات كبيرة من الجينوم فى كائنات راقية يبدو أنه لا فائدة منها Junk بمعنى أنه حتى الآن لم يتم التعرف أو فهم وظائفها.

فى جزيئات الدنا التماثل فى الحروف الهجائية يؤدى الى التشابه الكيميائى والتوحد فى الشكل. وعلى العكس – تظهر البروتينات اختلافات كبيرة فى التوزيع ثلاثى الأبعاد. وذلك ضرورى لدعم التنوع التركيبى والوظيفى الكبيرين لها. يحدد تتابع الحمض الأمينى فى البروتين التركيب ثلاثى الأبعاد له. يوجد لكل تتابع طبيعى للحمض الأمينى حالة فطرية ثابتة مميزة والمتأقلمة تلقائيا تحت الظروف المناسبة. عندما يتم تسخين بروتين منقى أو يتعرض لظروف مغايرة للبيئة الفسيولوجية الطبيعية ينتج تركيب غير حلزونى ذو ترتيب مختلف وغير نشط بيولوجيا. لذلك تمتلك الثدييات آليات للحفاظ على ثبات درجة الحرارة داخل أجسامها. وتستعيد جزيئات البروتين تركيبها الفطرى عند العودة للظروف الطبيعية وبصورة مماثلة للحالة الأصلية.

الطى التلقائى فى جزيئات البروتينات لتكوين حالتها الفطرية هى النقطة التى عندها تقوم الطبيعة بالقفزة الضخمة من البعد الأحادى لعالم تتابعات الوراثة والبروتين الى العالم ثلاثى الأبعاد الذى نعيشه. وهنا يوجد تناقض ظاهرى فترجمة تتابعات الدنا الى تتابعات حمض أمينى من السهل أن توصف منطقيا بواسطة الكود الوراثى بينما من الصعب جدا الوصف المنطقى للالتفاف الدقيق للسلسلة متعددة الببتيدات الى تركيب ثلاثى الأبعاد. فبينما تتطلب عملية الترجمة الآلية الضخمة المركبة للريبوسوم

– الرنا الناقل وجزيئات مصاحبة – الا أن عملية طى البروتين تحدث تلقائيا.

تعتمد وظائف البروتينات على التركيب الثلاثى الأبعاد الفطرى. على سبيل المثال يحتوى تركيب أى انزيم على تجويف على سطح الجزئ يقوم بالارتباط بجزئ صغير ليجاور أحماض أمينية حفازة. لذلك نذكر النموذج التالى:

- يحدد تتابع الدنا تتابع البروتين
- يحدد تتابع البروتين تركيب البروتين
- يحدد تركيب البروتين وظيفة البروتين

تركز معظم أنشطة المعلوماتية الحيوية المنظمة على تحليل البيانات ذات الصلة بتلك العمليات.

الى هذا الحد لم يتضمن هذا النموذج على مستويات أعلى من المستوى الجزيئى للتركيب والتنظيم وتشتمل على سبيل المثال على أسئلة كتلك المتعلقة بكيفية تخصص الأنسجة أثناء التطور أو أكثر تعميما كيف تمارس التأثيرات البيئية تحكمها فى الأحداث الوراثية. فى بعض الحالات البسيطة يكون من المفهوم على المستوى الجزيئى كيف أن زيادة كمية مادة تفاعله تسبب زيادة انتاج الأنزيم المحفز لتحوله. والأكثر تعقيدا هى برامج التطور خلال فترة حياة الكائن. تلك المشكلات الساحرة المتعلقة بتدفق المعلومات فى الكائن والتحكم فيها أصبحت الآن تأتى من خلال مجال البيومعلوماتية.

المشاهدات وسجلات البيانات Observables and Data Archives:

يتضمن بنك البيانات على سجل للمعلومات وتنظيم أو كيان منطقى لتلك المعلومات وأدوات للاتصال به. يغطى بنك البيانات للبيولوجيا الجزيئية تتابعات الحمض النووى والبروتين وتراكيب ووظائف الجزيئات الكبيرة. وتشتمل على:

- بنك بيانات سجلية للمعلومات البيولوجية:
  - تتابعات الدنا والبروتين متضمنة الشرح والتفسير.
  - تراكيب الحمض النووى والبروتين وشرحها. بنوك بيانات لطرز تعبير البروتين.
- بنوك بيانات فرعية: تحتوى على المعلومات المجمعة من بنوك البيانات السجلية والناتجة عن تحليل محتواها. فعلى سبيل المثال:
  - بواعث تتابع (خصائص عائلات البروتينات)
  - الطفرات والاختلافات فى تتابعات الدنا والبروتين
  - تصنيف أو علاقات (الصلات والخصائص المشتركة للمدخلات فى السجلات. مثل بنك بيانات لمجموعة من عائلات تتابع البروتين أو تصنيف متسلسل لطرز التفاف البروتين.
- بنوك بيانات مرجعية
- بنوك بيانات لمواقع الوب
  - بنوك بيانات لبنوك البيانات المحتوية على معلومات بيولوجية
  - روابط بين بنوك البيانات

تساؤلات قاعدة البيانات تبحث تعريف مجموعة من المدخلات (على سبيل المثال: تتابعات أو تراكيب) على أساس صفات محددة أو على أساس التشابه مع مجس للتتابع أو التركيب. أكثر التساؤلات شيوعا هو: عند تقدير تتابع أو تركيب جديد ما هو درجة التشابه بينه وبين المجود فى بنوك البيانات؟ بمجرد التوصل الى مجموعة من التتابعات أو التراكيب من قاعدة بيانات ملائمة تتشابه مع المجس يصبح الباحث قادرا على تعريف وبحث خصائصها العامة.

آليات الاتصال ببنك للبيانات هى مجموعة من الأدوات للاجابة على الأسئلة الآتية:

- هل يحتوى بنك البيانات على المعلومات المطلوبة ؟ (مثال: فى أى من بنوك البيانات يمكن الحصول على تتابعات الحمض الأمينى للكحول ديهيدوروجينيزز؟)
- كيف يمكن جمع معلومات منتقاة من بنك البيانات فى صورة مفيدة؟ (مثال: كيف يصنف قائمة لتتابعات الجلوبين أو جدول لمصفوفة تتابعات الجلوبين؟)
- فهارس بنوك البيانات تفيد عند التساؤل: أين يمكن أن يوجد بعض المعلومات المحددة ؟ (مثال: ما هى بنوك البيانات التى تحتوى تتابع الحمض النووى لتربسين بوركيوبين؟) وبالطبع اذا تم معرفة وتحديد ما هو المطلوب بدقة عندئذ تبدأ خطوات حل المشكلة.

وبنك البيانات بدون طرق فعالة للاتصال والتعامل تكون بمثابة مقبرة للبيانات. كيف تحقق اتصال فعال مع قاعدة بيانات مصممة على أن تظل محجوبة عن المستفيدين. أصبح من الواضح أن الاتصال الفعال لا يمكن تزويده بنظام تساؤلات فى أرشيف غير مشبع. بدلا من ذلك يجب أن يصمم التنظيم المنطقى لتخزين المعلومات مع وجود تصور لطرق الاتصال والتعامل – ماهى نوعية الأسئلة التى يريد المستخدم طرحها – كما ينبغى أن ينسجم تركيب الأرشيف بسلاسة مع برامج استدعاء المعلومات.

تتضمن مختلف أنواع استعلام قاعدة البيانات الممكنة فى مجال البيومعلوماتية الآتى:

١) تتابع مفترض، أو جزء من تتابع، وايجاد تتابعات فى قاعدة البيانات مماثلة له. هذه مشكلة مركزية فى البيومعلوماتية. يشارك المعلوماتية الحيوية مجالات عديدة من علم الحاسبات فى سلسلة من المشاكل المتماثلة. على سبيل المثال، برامج معالجة الكلمات والتحرير التى تدعم سلسلة من وظائف البحث.

٢) تركيب بروتين مفترض، أو جزء منه , وايجاد تراكيب مماثلة فى قاعدة البيانات. ذلك هو التعميم لسلسلة المشاكل المتشابهه للأبعاد الثلاثية.

٣) تتابع مفترض لبروتين غير معروف تركيبه، ايجاد تراكيب فى قاعدة البيانات والتى تتبنى تراكيب ثلاثية الأبعاد مشابهة. يحث ذلك على البحث فى بنوك بيانات التتابع عن بروتينات لها تتابعات مماثلة لتتابع المجس: من المتوقع فى حالة وجود اثنان من البروتينات لهما تتابعات متماثلة لدرجة كافية فسوف يكون لهما تراكيب متشابهة. الا أن هذا القول غير حقيقى. ونأمل فى الوصول الى تقنيات بحث أكثر قوة والتى ستستطيع التعرف على بروتينات متشابهة التركيب حتى ولو كان تتابع كل منهما ينحرف عن النقطة التى تقر التشابه بينهما على أساس مقارنة التتابع.

٤) تركيب مفترض لبروتين لايجاد تتابعات فى بنك البيانات والتى تقابل تراكيب مماثلة. مرة أخرى يمكن استخدام هذا التركيب كمجس لتركيب فى بنك البيانات لكن ذلك سوف يحقق فقط نجاحا محدودا لأن هناك العديد من التتابعات المعروفة أكثر من التراكيب. ولذا فانه من المرغوب فيه وجود طريقة يمكنها التقاط التركيب من التتابع.

أرقام (١) و (٢) عبارة عن أمثلة محلولة لعمليات بحث تجرى آلاف المرات كل يوم. بينما (٣) و (٤) عبارة عن مجالات نشطة للبحث.

تتأتى أهداف غاية فى الرقة عند الرغبة فى دراسة العلاقات بين معلومات بنوك بيانات منفصلة. حيث يتطلب ذلك وجود روابط تسهل الاتصال المتزامن مع عدة بنوك بيانات. مثال على ذلك: هل يوجد فى الخميرة بروتين مماثل لذلك البروتين معلوم التركيب والذى يساهم فى حدوث أمراض تخليق البيورين فى الإنسان؟ والخلفية هنا يحددها: تركيب معلوم ووظيفة معينة واكتشاف الصلة والارتباط بالمرض ونوع معين من

الكائنات. أدى الاهتمام بالنمو المتزايد بطرق الاتصال المتزامن ببنوك البيانات الى بحث التفاعل داخل بيانات البنك بمعنى – كيف يتم تبادل البيانات يبين بنك وآخر بدون تضحية كبيرة فى حرية كل بنك فى بناء بياناته بطرق تلائم الصفات الفردية المميزة للمادة التى تحتويها تلك البيانات.

والمشكلة التى لم تظهر بعد فى مجال البيولوجيا الجزيئية هى عملية تحديث السجلات. ففى نظام قاعدة البيانات للحجز بشركات الطيران يمنع حدوث بيع المكان الواحد لأكثر من مسافر مع وجود منافذ للحجز متعددة. فى مجال المعلوماتية الحيوية يمكن للمستخدم قراءة واستخلاص المعلومات من سجلات بنوك البيانات أو تقديم مواد يتم معالجتها بواسطة القائمين بالعمل فى سجل ما ولكنه لا يستطيع اضافة أو تغيير المدخلات مباشرة. قد يتغير هذا الوضع. من منظور عملى تتزايد كمية البيانات الناتجة بسرعة كبيرة لدرجة تعوق قدرة مشروعات السجلات على استيعابها. ويوجد بالفعل تحرك نحو مشاركة أكبر للعلماء فى المعامل لتجهيز بيانات للسجلات.

بالرغم من جدل بخصوص تحكم متميز للسجلات ألا أن هناك حاجة الى الحد من الطرق العامية (غير المتخصصة) للتعامل معها – تصميم front ends –. ويمكن لمجتمعات المستخدم المتخصص أن تستخلص تحت مجموعات من البيانات أو دمج بيانات من مصادر مختلفة والوصول الى طرق متخصصة للاتصال والتعامل. وتعتمد مثل – دكاكين قواعد البيانات هذه – على السجلات الأولية كمصدر للمعلومات التى تحتويها ولكن مع اعادة تصميم التنظيم والعرض بالطريقة التى يرونها أكثر ملاءمة. وفى الحقيقة يمكن لمختلف قواعد البيانات الفرعية أن تصنف ذات المعلومة. يقترح الاستقراء المعقول مفهوم قواعد البيانات الواقعية المتخصصة virtual data bases المبنى على السجلات وفى نفس الوقت يقدم مفهوما ووظيفة فردية فصلت لتلبى احتياجات مجموعات البحث الفردية أو حتى علماء منفردين.

الرعاية والتفسير والتحكم فى الجودة:
Curation Annotation, and Quality Control:

تعتمد المجتمعات العلمية والطبية على جودة بنوك البيانات. ومؤشرات الجودة حتى وان كانت لا تسمح بتصحيح الأخطاء الا أنها قد تتجنب التوصل الى استنتاجات خاطئة.

تضمن مدخلات بنك البيانات النتائج التجريبية الخام ومعلومات معاونة أو تفسيرات. وتمتلك كل واحدة من تلك مصادرها الخاصة من الخطأ.

من أهم المحددات لجودة البيانات ذاتها هو مدى دقة التجارب. البيانات القديمة محدودة بتقنيات قديمة. فعلى سبيل المثال: تتابعات الحمض الأمينى للبروتينات التى تم تقديرها بواسطة تحديد التتابع الببتيدى يتم الآن ترجمتها جميعا من تتابعات الدنا. أحد عواقب انفجار البيانات هو أن غالبية البيانات جديدة وتم الحصول عليها بتكنولوجيا حديثة والتى فى معظم الحالات تؤدى عملا جيدا.

تشتمل التفسيرات معلومات حول مصادر البيانات والطرق المستخدمة فى تقديرها. كما أنها تقدم روابط مع المعلومات ذات الصلة فى بنوك المعلومات الأخرى. فى بنوك بيانات التتابع تتضمن التفسيرات جداول توصيف: قوائم بأجزاء التتابعات التى لها معنوية بيولوجية – على سبيل المثال مناطق من تتابع الدنا الخاصة بشفرة البروتينات. يظهر ذلك فى تصميم مناسب للتعامل مع الحاسب الآلى ويكون محتوياتها مقيدة بمفردات لغوية محكومة. حتى وقت قريب ادخال تتابع نمطى لدنا يتم بواسطة مجموعة بحثية منفردة تبحث الجين ونواتجه بطريقة مترابطة منطقيا. تستند التفسيرات على البيانات التجريبية والمكتوبة بواسطة متخصصون. على النقيض، لا تقدم مشروعات تتابع الجينوم تأكيد تجريبى للتعبير فى معظم الجينات المفترضة ولا تشخيص نواتجها. يقيم الأمناء فى بنوك البيانات تفسيراتهم على أساس تحليل التتابعات بواسطة برامج الحاسب الآلى.

التفسيرات هى أضعف مكون فى مؤسسة الجينوم وميكنة التفسيرات تكون ممكنة فقط بدرجة محدودة. وللحصول عليها بطريقة صحيحة تكون كثيفة

العمالة وتقسيم المصادر يكون غير كافيا. ولكن لا يمكن الاستخفاف بالتفسيرات الصحيحة. وقد علق بورك بأن تلك الأخطاء فى دراسة الجين تفسد جودة بيانات التتابع. سوف يؤدى نمو بيانات الجينوم الى تحسين جودة التفسير كما أن الطرق الاحصائية تزيد من الدقة وتسمح باعادة محسنة لتفسير المدخلات. التحسن فى التفسيرات شيء جيد ولكن التلازم الحتمى – التفسير سيكون متدفقا – سيكون مزعجا. هل يجب الاطلاع دوريا على البحوث المكتملة والأخذ فى الاعتبار خلاصة ما توصلت له من نتائج؟ وقد تفاقمت المشكلة بوجود العديد من مواقع الوب مع تزايد شبكات الربط الكثيفة. ويتيح ذلك طرق مفيدة للتطبيقات المتنوعة. والوب أيضا ناقل لعدوى تضخم الأخطاء فى البيانات الخام والتفسيرات المختلفة ويتم تباعا تصويب الأخطاء فى البيانات فى صورتها الأولية ولكن الحاجة الى التصويب لا نهاية لها.

والحل الوحيد الممكن هو تصويب موزع وديناميكى لمعالجة الخطأ والتفسير. سوف يقوم الأخصائيون الموزعون بدور الأمناء حيث أن العاملين ببنك البيانات ليس لديهم لا الوقت ولا الخبرة للقيام بتلك المهمة. كما تسمح ديناميكية التقدم فى ميكنة تعريف وتصحيح الخطأ والتفسير باعادة التفسير لبنوك البيانات. ولسوف نضطر أن نسلم بالفكرة الآمنة لبنك بيانات مستقر يتكون من مدخلات سليمة من بداية توزعها وتظل كذلك. وسوف تصبح بنوك البيانات مصدر غال وثمين للمعلومات والتى تنمو فى الحجم وتزداد نضجا ونأمل أن تكون بجودة مناسبة.

الشبكة العنكبوتية العالمية واسعة الإنتشار The World Wide Web (www):

تستخدم الشبكة العنكبوتية للحصول على مواد مرجعية و أخبار وللتعامل مع قواعد البيانات في البيولوجيا الجزيئية أو لمجرد التصفح. وتعتبر الوب الآن وسيلة من وسائل الاتصال بين الأشخاص وبين الحاسبات عبر الشبكات. وهى بمثابة قرية عالمية تشتمل على ما يقابل المكتبة ومكتب البريد والأسواق والمدارس وغيرها.

ويجرى المستخدم برنامج للتصفح على الحاسب الخاص به. ومن برامج التصفح الشائعة:

Netscape and Internet Explorer ويمكن بواسطة تلك البرامج قراءة وعرض مواد من أي مكان في العالم. تقدم برامج التصفح معلومات للتحكم كما تتيح نقل المعلومات إلى حاسب محلى. ومن الأشياء الرئيسية لبدء استخدام الشبكة بفاعلية إيجاد نقاط دخول مفيدة حيث تأخذك روابط الوصل (links) الى المكان الذي تريده بمجرد بدء التشغيل. ومن بين أكثر المواقع أهمية تلك الخاصة ببرامج البحث والتنقيب search engines والتي تفهرس الوب بأكمله وتتيح استرجاع المعلومات باستخدام كلمات رئيسية key words. ومن الممكن إدخال واحدا أو أكثر من المصطلحات مثل 'phosphorylase', 'llosteric change', 'rys tal structurre' وسوف يظهر برنامج البحث قائمة بروابط الوصل لمواقع على الوب تحتوى على تلك المصطلحات وعندئذ يمكن التعرف على المواقع محل الاهتمام.

أثناء جلسة تصفح الوب يمكن الاحتفاظ بالمستندات التى نحتاج الرجوع اليها فى جلسات تصفح قادمة وذلك بحفظ روابط الوصل الخاصة بها فى ملف bookmarks أو favourites ومن ثم يمكنك فى جلسات لاحقة الرجوع الى أى موقع مباشرة دون الحاجة إلى اتباع محاولات روابط الوصل المستخدمة في أول مرة.

الوب ليس طريقا ذو اتجاه واحد بمعنى أن العديد من مستندات الوب تتضمن نماذج يمكن ادخال معلومات فيها وإجراء برنامج للحصول على النتائج خلال نفس الجلسة. وتعتبر برامج البحث والتنقيب خير مثال على ذلك. وتنطلق الآن العديد من العمليات الحسابية فى البيومعلوماتية عبر أجهزة الخدمة servers. وفى حالة العمليات الحسابية الطويلة ربما لا تتأتى النتائج أثناء نفس الجلسة ولكن ترسل بالبريد الإلكتروني e-mail.

محددات مواقع المصدر The hURLy-bURLy (Uniform Resource Locators):

تحدد تلك الحروف شكل المادة ومكانها حيث ينبغى أن يكون لكل مستند على الوب ملف فى مكان ما على حاسب معين. مثال لURL:

http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/FindInfo.html

وهو موقع لدرس حول كيفية الحصول على معلومات من الإنترنت. ويشير مختصر http:// الى أن المستند فى صيغة بروتوكول نقل النص المحورى hypertext transfer protocol أما www.lib.berkely.edu فهو اسم الحاسب: المكتبة المركزية فى جامعة كاليفورنيا بيركيلى. وباقية الأجزاء تشير الى مكان واسم الملف فى الحاسب.

النشر الالكترونى Electronic publication:

هناك العديد والعديد من المواد المنشورة على صفحات الوب. وفى المجلات العلمية قد ينشر جدول المحتويات فقط أو جدول المحتويات مع ملخصات المقالات أو المقالات كاملة. والآن يظهر العديد من المطبوعات المؤسسية والنشرات الدورية والتقارير الفنية على الوب. كما يحتوى الكثير من المطبوعات على مراجع لروابط وصل تحتوى على مواد مساعدة لا يمكن ظهورها على ورق. كما أن المواد المطبوعة على ورق يمكن أن تتضمن عناوين لمواقع على الوب وللبريد الإلكتروني.

الحاسبات وعلم الحاسب Computers and computer science:

لم يكن من المحتمل ظهور مجال البيومعلوماتية بدون التقدم الذى تحقق فى المكونات المادية Hardware وبرامج software الحاسبات. كما أن وسائط الحفظ السريعة ذات الكفاءة العالية ضرورية للإبقاء على السجلات. وتتطلب عملية استرجاع وتحليل المعلومات برامج بعضها بسيط وسلس والبعض الآخر متطورة ومعقدة. كما يتطلب توزيع

المعلومات توافر إمكانيات من شـبكات الحاسـب computer networks وكذلك بالنسبة للشبكة العنكبوتية العالمية.www.

علوم الحاسب مجال صغير ومزدهر يستهدف تعظيم الاسـتخدام الفعـال لمكونات تكنولوجيا المعلومات. مناطق معينة من علـم الحاسـب تمـس البيومعلوماتية مسا وثيقا ومباشرا. دعنا نتصور مشكلة بيولوجيـة مثـل استرجاع كل التتابعات المماثلة لمجس التتابع probe sequence من قـاعدة بيانات سيكون الحل الصواب هو اللجوء الى علم الحاسب فى:

- **التحليل العددى (الخوارزميات) Analysis of algorithms:**

وهو الوصف الكامل والدقيق لطريقة حل المشكلة. بالنسبة لاسـترجاع تتابعات مماثلة نحتاج قياس مدى تشابه تتابع المجس لكل تتابع مجود فى قاعدة البيانات. ومن المحتمل أن يكون ذلك أفضل بكثير من تلك الاتجاه الساذج لفحص كل زوج من الأماكن فى كل تجاور محتمل وهى طريقـة حتى بدون السماح بفراغات فانها تحتاج الى وقت يتناسب مـع نـاتج ضرب عدد حروف مجس التتابع فى عدد حـروف تتابعـات قـاعدة البيانات. ويركز تخصص الحاسب والمعروف عامة 'stringology'علـى ايجاد طرقا ذات كفاءة للتعامل مع هذا النوع من المشاكل وتحليل فاعليـة أدائها.

- **تراكيب البيانـات واسـترجاع المعلومـات Data structures, and information retrieval:**

كيف يتم تنظيم البيانات بطريقة تتيح استجابة كفء للتساؤلات؟. علـى سبيل المثال: هل هناك طرق لفهرسة أو اعداد معالجة للبيانـات لجعـل بحث تماثل التتابع أكثر كفاءة؟ كيف يمكننا تقديم حدودا مشـتركة مـن شأنها مساعدة المستخدم على تصميم و تنفيذ التساؤلات؟.

- **هندسة البرامج Software engineering:**

لم يعد من الصعب بتاتا كتابة برامج بلغة الحاسبات الأصليـة. يعمـل المتخصصون فى عمل البرامج بلغات عالية المستوى مثل

C, C++, PERL ('Practical Extraction and Report Language')

أو حتى لغة FORTRAN. يعتمد اختيار لغة البرمجة على طبيعة الحساب وتركيب البيانات المصاحبة. وبالطبع فمعظم البرامج المعقدة المستخدمة فى البيومعلوماتية تكتب بواسطة متخصصون.

البرمجة Programming:

البرمجة بالنسبة لعلم الحاسب كالقائم بالبناء فى فن العمارة كلاهما مبدع أحدهما فن والآخر حرفة.

يستفسر العديد من طلاب البيومعلوماتية هل من الضرورى أن يتعلموا كتابة برامج حاسب معقدة ؟ والاجابة أنه ليس من الضرورى ذلك إلا إذا كانت هناك رغبة فى التخصص فى هذا المجال. ويتطلب العمل فى مجال البيومعلوماتية اكتساب خبرات فى استخدام الأدوات المتاحة على صفحات الوب. كما أنه من الأشياء الأساسية هو تعلم كيفية إنشاء موقعا على الوب وكذلك الابقاء عليه وبالطبع هناك حاجة الى إمكانيات لاستخدام نظام تشغيل الحاسب الشخصى. ومهارة كتابة نصوص بسيطة بلغة مثل PERL تعتبر من ضمن أساسيات نظام التشغيل.

وحيث يجب أن يؤخذ فى الاعتبار حجم سجلات البيانات والنمو المتزايد فى درجة التعقيد فى التساؤلات المطروحة لذلك من الأفضل أن يترك الابداع الحقيقى للبرمجة فى هذا المجال للمتخصصين ذى الخبرة الجيدة فى علم الحاسب.

وينصح بتعلم المهارات الأساسية للغة PERL لأنها أداة قوية تجعل من السهل جدا القيام بالعديد من العمليات البسيطة والمفيدة. وتمتاز لغة PERL بأنها متاحة فى غالبية أنظمة الحاسب.

كيف يمكنك تعلم PERL بدرجة كافية لاستخدامها فى البيومعلوماتية ؟ تقدم العديد من المعاهد دروسا لهذه اللغة كما يمكن تعلمها بمساعدة

الزملاء وكذلك بالرجوع الى الكتب المتوفرة. ومن الطرق المفيدة أيضا البحث عن دروس على صفحات الوب بالاستعانة ببرامج البحث حيث يوجد مواقع لذلك. يمكن الرجوع الى موقع مشروع بيوبيرل Bioperl project : (http://bio.perl.org/) والذى يتيح مصدرا لبرامج PERL ومكوناته المستخدمة فى مجال البيومعلوماتية.

قوة PERL فى تناول سلسلة الحروف جعلها تلائم عمليات تحليل التتابع فى علم البيولوجى. وفيما يلى مثال لاستخدام برنامج PERL بسيط لترجمة تتابع نيكليوتيدة الى تتابع حمض نووى طبقا لكود وراثي قياسي. السطر الأول: #!/usr/bin/perl هو اشارة الى نظام تشغيل UNIX (or LINUX) وما يليه هو برنامج PERL. خلال البرنامج جميع النصوص التى تبدأ ب # ليست الا تعليق. ويشير السطر_END_ الى انتهاء البرنامج وأن ما يليه هو عبارة عن بيانات تم ادخالها. (يمكن الحصول على المواد والبرامج بالرجوع الى موقع الكتاب على الوب وهو:

http://www.oup.com/uk/lesk/bioinf

ويعرض هذا المثال البسيط صور عديدة للغة PERL. و يحتوى هذا الملف على بيانات مصاحبة (جدول ترجمة الكود الوراثى)، وعبارات تخبر الحاسب لعمل شئ معين بالمدخلات (مثل التتابع المطلوب ترجمته)، والبيانات التى يتم ادخالها (وهى تظهر بعد سطر _END_). كما تلخص التعليقات أقسام من البرنامج وتوصف تأثير كل عبارة.

يتركب البرنامج كبلوكات داخل أقواس متعرجة : { ... } مما يفيد فى انسياب الأداء. وداخل البلوكات عبارات فردية (كل منها تنتهى ب ;). والبلوك الخارجى عبارة عن لوب:

```
while ($line = <DATA>) {
    ...
}
```

تشير <DATA> الى سطور ادخال البيانات (والتى تظهر بعد _END_). ويتم اجراء البلوك مرة واحدة لكل سطر من المدخلات ويستمر ذلك حتى نهاية كل السطور.

ويظهر فى البرنامج ثلاثة أنواع من تراكيب البيانات. سطر ادخال البيانات ويشار اليه line $ وهو سلسلة من الحروف البسيطة والتى تجزأ الى منظومة متعددة البيانات array أو حامل لثلاثيات triplets. وتخزن المنظومة بيانات لموضوعات عديدة فى ترتيب خطى. ويمكن استرجاع بيانات كل موضوع على حدي من أماكنها فى المنظومة. لتسهيل التقاط الكود الثلاثى لحامض أمينى فانه يتم تخزين الكود الوراثى كمنظومة ترتيب. ومنظومة الترتيب أو جدول التكرار عبارة عن تعميم لمنظومة بسيطة أو تسلسلية. اذا كانت عناصر المنظومة البسيطة مفهرسة بواسطة أرقام متتابعة فان عناصر منظومة الترتيب تفهرس باستخدام سلاسل من حروف وهى فى هذه الحالة تكون ٦٤ ثلاثية. يتم معالجة الثلاثيات المدخلة طبقا لترتيب ظهورهم فى تتابع النيكليوتيدة مع الاستعانة بعناصر جدول الكود الوراثى بترتيب تحكمى كما يقرأ فى الثلاثيات المتتالية. المنظومة البسيطة أو حامل سلاسل الحروف يكون مناسبا لمعالجة ثلاثيات متتالية بينما تلائم المنظومة التسلسلية التقاط الأحماض الأمينية المقابلة للثلاثيات.

مثال : برنامج بيرل لترجمة تتابع حامض نووى الى تتابع حامض أمينى

Translate.pl - PERL program to translate nucleic acid sequence to amino acid sequence:

```
#!/usr/bin/perl
#translate.pl    --  translate nucleic acid sequence to protein
                     sequence
#                according to standard genetic code

#   set up table of standard genetic code

%standardgeneticcode = (
 "ttt"=> "Phe",   "tct"=> "Ser", "tat"=> "Tyr",   "tgt"=> "Cys",
 "ttc"=> "Phe",   "tcc"=> "Ser", "tac"=> "Tyr",   "tgc"=> "Cys",
 "tta"=> "Leu",   "tca"=> "Ser", "taa"=> "TER",   "tga"=> "TER",
 "ttg"=> "Leu",   "tcg"=> "Ser", "tag"=> "TER",   "tgg"=> "Trp",
 "ctt"=> "Leu",   "cct"=> "Pro", "cat"=> "His",   "cgt"=> "Arg",
 "ctc"=> "Leu",   "ccc"=> "Pro", "cac"=> "His",   "cgc"=> "Arg",
```

```
 "cta"=> "Leu",   "cca"=> "Pro", "caa"=> "Gln",   "cga"=> "Arg",
 "ctg"=> "Leu",   "ccg"=> "Pro", "cag"=> "Gln",   "cgg"=> "Arg",
 'att"=> "Ile",   "act"=> "Thr", "aat"=> "Asn",   "agt"=> "Ser",
 'atc"=> "Ile",   "acc"=> "Thr", "aac"=> "Asn",   "agc"=> "Ser",
 'ata"=> "Ile",   "aca"=> "Thr", "aaa"=> "Lys",   "aga"=> "Arg",
 'atg"=> "Met",   "acg"=> "Thr", "aag"=> "Lys",   "agg"=> "Arg",
 "gtt"=> "Val",   "gct"=> "Ala", "gat"=> "Asp",   "ggt"=> "Gly",
 "gtc"=> "Val",   "gcc"=> "Ala", "gac"=> "Asp",   "ggc"=> "Gly",
 "gta"=> "Val",   "gca"=> "Ala", "gaa"=> "Glu",   "gga"=> "Gly",
 "gtg"=> "Val",   "gcg"=> "Ala", "gag"=> "Glu",   "ggg"=> "Gly"
}

#   process input data


while ($line = <DATA>) {                                    # read in
line of input
    print "$line";                                          #
transcribe to output
    chop();                                                 # remove
end-of-line character
    @triplets = unpack("a3" x (length($line)/3), $line); # pull out
successive triplets
    foreach $codon (@triplets) {                            # loop
over triplets
        print "$standardgeneticcode{$codon}";               # print
out translation of each
    }                                                       # end loop
on triplets
    print "\n\n";                                           # skip
line on output
}                                                           # end loop
on input lines


#   what follows is input data


__END__
atgcatccctttaat
tctgtctga
```

Assemble.pl - PERL program to assemble overlapping fragments of strings:

```
#!/usr/bin/perl
#assemble.pl -- assemble overlapping fragments of strings


#  input of fragments
while ($line = <DATA>) {                #   read in fragments, 1
per line
   chop($line);                         #   remove trailing
carriage return
   push(@fragments,$line);              #   copy each fragment
into array
}
#  now array  @fragments  contains fragments


#  we need two relationships between fragments:
#  (1) which fragment shares no prefix with suffix of another
fragment
#       * This tells us which fragment comes first
```

```
#  (2) which fragment shares longest suffix with a prefix of
another
#       * This tells us which fragment  follows  any fragment


#  First set array of prefixes to the default value
"noprefixfound".
#       Later, change this default value when a prefix is found.
#       The one fragment that retains the default value must be come
first.


#  Then loop over pairs of fragments to determine maximal overlap.
#       This determines successor of each fragment
#       Note in passing that if a fragment has a successor then the
#           successor must have a prefix


foreach $i (@fragments) {                #   initially set  prefix
of each fragment
    $prefix{$i} = "noprefixfound";       #       to
"noprefixfound"
    }                                    #   this will be
overwritten when a prefix is found


#  for each pair, find longest overlap of suffix of one with prefix
of the other
#       This tells us which fragment FOLLOWS any fragment


foreach $i (@fragments) {                #   loop over fragments
    $longestsuffix = "";                 #   initialize longest
suffix to null


    foreach $j (@fragments) {            #   loop over fragment
pairs
        unless ($i eq $j) {              #   don't check fragment
against itself


            $combine = $i . "XXX" . $j;  #   concatenate fragments,
with fence XXX
            $combine =~ /([\S ]{2,})XXX\1/;   #   check for
repeated sequence
            if (length($1) > length($longestsuffix)) {    # keep
longest overlap
              $longestsuffix = $1;       #   retain longest suffix
              $successor{$i} = $j;       #   record that $j follows
$i
            }


        }
    }
    $prefix{$successor{$i}} = "found";   #  if $j follows $i then
$j must have a prefix
}


foreach (@fragments) {                   #  find fragment that has
no prefix; that's the start
    if ($prefix{$_} eq "noprefixfound") {$outstring = $_;}
}
```

```
$test = $outstring;                          #   start with fragment
without prefix
while ($successor{$test}) {                  #   append fragments in
order
     $test = $successor{$test};              #   choose next fragment
     $outstring = $outstring."XXX". $test; # append to string
     $outstring =~ s/([\S ]+)XXX\1/\1/;      #   remove overlapping
segment
}


$outstring =~ s/\\n/\n/g;                    #   change signal \n to
real carriage return
print "$outstring\n";                        #   print final result


__END__
the men and women merely players;\n
one man in his time
All the world's
their entrances,\nand one man
stage,\nAnd all the men and women
They have their exits and their entrances,\n
world's a stage,\nAnd all
their entrances,\nand one man
in his time plays many parts.
merely players;\nThey have
```

وهنالك اصدار بديل من البرنامج لمضاهاة الأجزاء المتداخلة assemble

: overlapping fragments

```
# /usr/bin/perl


$. = "";
@fragments = split("\n",<DATA>);


foreach (@fragments) { $firstfragment{$_} = $_; }


foreach $i (@fragments) {
    foreach $j (@fragments) { unless ($i eq $j) {
        ($combine = $i . "XXX" . $j) =~ /([\S ]{2,})XXX\1/;
        (length($1)   <= length($successor{$i})) || { $successor{$i}
= $j };
    }                          }
    undef $firstfragment{$successor{$i}};
}


$test = $outstring = join "", values(%firstfragment);
while  ($test   = $successor{$test}) { ($outstring .= "XXX" . $test)
=~ s/([\S ]+)XXX\1/\1/; }


$outstring =~ s/\\n/\n/g; print "$outstring\n";


__END__
the men and women merely players;\n
one man in his time
```

```
All the world's
their entrances,\nand one man
stage,\nAnd all the men and women
They have their exits and their entrances,\n
world's a stage,\nAnd all
their entrances,\nand one man
in his time plays many parts.
merely players;\nThey have
```

التصنيف والتسمية البيولوجية

Biological Classification and nomenclature:

تعتمد التسمية البيولوجية على تقسيم لكائنات الحية الى مملكة وقبيلة وأقسام وأجناس وأنواع وذلك على أساس أوجه التشابه المشاهدة. وتقدم نتائج تحليل التتابع دلائل قاطعة على العلاقات بين الأنواع.

استخدام التتابع لتحديد علاقات القرابة: Use of sequence to determine phylogenetic

توضح الأمثلة التالية تطبيقات استرجاع التتابعات من بنوك المعلومات ومقارنات التتابع فى تحليل العلاقات البيولوجية:

المثال الأول: استرجاع تتابع الحمض الأمينى لانزيم الريبونيوكليز فى بنكرياس الحصان باستخدام ExPASY server فى المعهد السويسرى للبيومعلوماتية وعنوانه:

http://www.expasy.ch/cgi-bin/sport-search-ful.

١- اكتب فى المكان المخصص ل key words: horse pancreatic ribonuclease

٢- ثم اضغط على مفتاح ENTER

٣- اختار RNP_HORSE ثم FASTA format سوف تحصل على التالى:

```
>sp|P00674|RNP_HORSE  RIBONUCLEASE  PANCREATIC (EC 3.1.27.5) (RNASE
1) ...
KESPAMKFERQHMDSGSTSSSNPTYCNQMMKRRNMTQGWCKPVNTFVHEP
LADVQAICLQKNITCKNGQSNCYQSSSSMHITDCRLTSGSKYPNCAYQTS
QKERHIIVACEGNPYVPVHFDASVEVST
```

وهنا يمكنك أن تقوم بعملية قص ولصق الى برامج أخرى حيث يمكن استرجاع عدة تتابعات وعمل sequence alignment وهذا يفيد فى تقدير درجة القرابة والعلاقات.

المثال الثانى: حدد باستخدام تتابعات انزيم الريبونيوكليز البكتريا للحصان والحوت والكنجر النوعين الأكثر قرابة:

Pancreatic ribonuclease sequences from horse *(Equus caballus)*, minke whale *(Balaenoptera acutorostrata)* and red kangaroo *(Macropus rufus)*:

```
>RNP_HORSE
KESPAMKFERQHMDSGSTSSSNPTYCNQMMKRRNMTQGWCKPVNTFVHEP
LADVQAICLQKNITCKNGQSNCYQSSSSMHITDCRLTSGSKYPNCAYQTS
QKERHIIVACEGNPYVPVHFDASVEVST
>RNP_BALAC
RESPAMKFQRQHMDSGNSPGNNPNYCNQMMMRRKMTQGRCKPVNTFVHES
LEDVKAVCSQKNVLCKNGRTNCYESNSTMHITDCRQTGSSKYPNCAYKTS
QKEKHIIVACEGNPYVPVHFDNSV
>RNP_MACRU
ETPAEKFQRQHMDTEHSTASSSNYCNLMMKARDMTSGRCKPLNTFIHEPK
SVVDAVCHQENVTCKNGRTNCYKSNSRLSITNCRQTGASKYPNCQYETSN
LNKQIIVACEGQYVPVHFDAYV
```

يستخدم برنامج multiple - sequence alignment program CLUSTAL W: وعنوانه:

http://www.ebi.ac.uk/clustalw/

وهناك بديل آخر وهو :T-coffee:

http://www.ch.embnet.org/software/TCoffee.html

وسوف تصل الى تطابق مواقع فى تتابعات كلا من الحصان والحوت.

البحث عن التتابعات المتماثلة باستخدام قواعد بيانات PSI-BLAST:

من أشهر الأمثلة هو البحث عن قاعدة بيانات لموضوعات تتشابه مع مجس. ففى حالة تحديد تتابع جين جديد أو التعرف من خلال الجينوم البشرى على جين مسئول عن مرض معين يكون هناك رغبة لمعرفة وجود جينات مماثلة فى أنواع أخرى. والطريقة المثلى لتحيق ذلك يجب أن تكون حساسة بحيث تتعرف على كل العلاقات واختيارية بأن تكون تلك العلاقات حقيقية.

وتشتمل طرق البحث فى قواعد البيانات التناوب بين الحساسية والاختيارية. هل تسطيع الطريقة ايجاد كل أو معظم التطابقات الموجودة بالفعل أم أنها تفقد أجزاء كبيرة؟. وعلى النقيض كم من التطابقات الواردة

تكون غير صحيحة؟. بافتراض أن قاعدة بيانات تحتوى على ١٠٠ تتابع جلوبين وأن عملية البحث فى تلك القاعدة للجلوبين أعطت ٩٠٠ تتابع، وكان منها ٧٠٠ جلوبين حقيقى و ٢٠٠ خطأ. اذا يمكن القول أن هذه النتائج تحتوى على ٣٠٠ مفقودة (نتيجة خادعة سلبية) و ٢٠٠ نتيجة خادعة ايجابية. وسينتج عن تخفيض الحد الحرج للأمان زيادة لكلا النوعين. وهنا يكون الحرص على العمل بحدود حرجة منخفضة للتأكد من عدم فقد أى شئ ويتطلب ذلك فحص دقيق للنتائج للتخلص من النتائج الخادعة الايجابية.

من الأدوات الفاعلة للبحث فى قواعد بيانات التتابع بالاستعانة بمجس للتتابع استخدام PSI-BLAST

(Position Sensitive Iterated - Basic Linear Alignment Sequence Tool)

وهوبرنامج من المركز القومى الأمريكى لمعلومات التكنولوجيا الحيوية (NCBI). ويعمل برنامج BLAST على التعرف على مناطق التشابه بدون فراغات ثم ضمها معا. ويشير PSI الى تحسين وتهذيب نمط التعرف داخل التتابع فى المراحل الأولية للبحث فى قاعدة البيانات. يؤدى اقرار انماط متحفظة الى تعظيم كل من حساسية واختيارية البحث. ويتضمن PSI-BLAST عمليات متكررة حيث يتحسن تعريف الأنماط الناشئة من خلال المراحل المتعاقبة للبحث.

مثال: تماثل الجين البشرى PAX-6 وهى جينات تتحكم فى تطور العين فى أنواع عديدة من الكائنات. وقد وجد تماثل هذا الجين فى الانسان وذبابة الدروسفيلا.

ويمكن اجراء البحث عن التماثل كما يلى:

١- الحصول على تتابع الحمض الأمينى للبروتين بالرجوع الى SWISS-PROT entry P26367.

٢- اجراء PSI-BLAST من خلال الموقع:

http://www.ncbi.nlm.nih.gov/blast/psiblast.cgi.

٣- ادخل التتابع واستخدم الخيارات لتحديد قاعدة بيانات للبحث وجـــدول التشابه المستخدم.

سوف يظهر البرنامج قائمة مشابهة لتتابع المجس ومرتبة تنازليـــا طبقـا لدرجة المعنوية الاحصائية. وفيما يلى نموذج طبق الأصل لأحد الأســطر فى القائمة:

Pir 11 I 45557 eyeless, long form - fruit fly (Drosophila melano
255 7e-67

حيث Pir عبارة عن مصدر تعريف البروتين وهو 145557 entry وهـــو التماثل لعين الدروسفيلا. والرقم ٢٥٥ مقياس لدرجـــة التطـــابق. و7e-67 تدل على مدى معنوية التطابق.

كما يمكن استخلاص اسماء الأنـــواع مـــن نتـــائج PSI-BLAST وذلـك باستخدام برنامج PERL كما يلى:

PERL program for extraction of species names from PSI-BLAST output:

```
#!/usr/bin/perl
#extract species from psiblast output


# Method:
#    For each line of input, check for a pattern of form [Drosophila
melanogaster]
#   Use each pattern found as the index in an associative array
#   The value corresponding to this index is irrelevant
#   By using an associative array, subsequent instances of the same
#          species will overwrite the first instance, keeping only a
unique set
#   After processing of input complete, sort results and print.


while (<>) {                                  # read line of input
    if   (/\[([A-Z][a-z]+ [a-z]+)\]/) {    # select lines containing
strings of form
                                              #               [Drosophila
melanogaster]
     $species{$1}   = 1;                      # make or overwrite entry
in
   }                                               #         associative
array
}

foreach (sort(keys(%species))){               # in alphabetical order,
    print "$_\n";                             #    print species names
}
```

وقد وجد أن هناك تماثل مع ٥٢ نوعا.

التنبؤ بتركيب البروتينات وهندستها: Protein structure and engineering

يحدد تتابع الحمض الأمينى لبروتين ما التركيب ثلاثى الأبعاد له. باحتواء تتابعات الحمض الأمينى على معلومات كافية لتحديد التراكيب ثلاثية الأبعاد للبروتينات يصبح من الممكن استنباط نظام حسابى للتنبؤ بتركيب تركيب بروتين من تتابع الحمض الأمينى. بالاضافة الى التنبؤ بالتركيب فقد حدد العلماء عددا من الأهداف أقل طموحا يمكن تحقيقها:

١- التنبؤ بالتركيب الثانوى Secondary structure prediction:

تحديد ماهى أجزاء التتابع التى تكون الشكل اللولبى والأخرى التى تكون الشرائط.

٢- تمييز الطى Fold recognition:

بعمل مكتبة لتراكيب بروتينات معروفة ولتتابعات الحمض الأمينى لها هل يصبح فى استطاعتنا الحصول من تلك المكتبة على تركيب يشابه لدرجة كبيرة نظام طى للبروتين المطلوب معرفة تركيبه؟.

٣- نماذج التماثل Homology modeling:

نفترض أن هناك بروتين معروف تتابع الحمض الأمينى له وغير معروف التركيب ولكنه متماثل مع واحد أو أكثر من البروتينات المعروف تركيبها. هنا يمكننا أن نتوقع أن البروتين الأكثر تماثلا يمكن استخدامه كأساس لنموذج للبروتين المجهول التركيب. ويعتمد كمال ودقة النتائج على مدى تشابه التتابع. وعموما وجد انه فى حالة ما اذا كان فىاثنين من البروتينات قريبة الصلة تطابق ٥٠% فى التتابع عند عمل المحازاة يكون هناك تماثل ٩٠% فى تركيبهما.

فيما يلى التتابعات المصفوفة والتراكيب المتطابقة لاثنين من البروتينات قريبة الصلة وهما ليسوزيم الأبيض فى بيض الدجاج وألفا لاكتوألبيومين

فى البابون. والتتابعات فيهما شديدة التقارب (٣٧% تطـــابق كـــامل فـــى التتابعات المتراصة) كما أن هناك تماثل فى التركيب. وكل بروتين يمكـــن استخدامه كنموذج جيد للبروتين الآخر:

Chicken lysozyme :

KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNT
DGS

Baboon ? - lactalbumin :

KQFTKCELSQNLY- -DIDGYGRIALPELICTMFHTSGYDTQAIVEND - ES

Chicken lysozyme :

TDYGILQINSRWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVS

Baboon ? - lactalbumin :

TEYGLFQISNALWCKSSQSPQSRNICDITCDKFLDDDITDDIMCAKKILD

Chicken lysozyme :

DGN- GMNAWVAWRNRCKGTDVQA- WIRGCRL-

Baboon ? - lactalbumin :

I - - KGIDYWIAHKALC - TEKL - EQWL - - CE - K

التقويم الحرج للتنبـــؤ بـــالتركيب Critical Assessment of Structure Prediction (CASP):

يتطلب الحكم على تقانات التنبؤ بـــتراكيب البروتينـــات الـــى اســـتخدام اختبارات مصممته blind tests. لهذا الغرض صمم جاى مولـــت برامـــج CASP. العلماء المنشغلون بتقدير تركيب البروتين باستخدام طرق القيـلس البللورى والرنين النووى المغناطيسى مدعوون الـــى: (١) نشـــر تتـــابع الحمض الأمينى عدة شهور قبل التاريخ المتوقع الى اكتمـــال التجـــارب، (٢)الاحتفاظ بسرية النتائج حتى تاريخ متفق عليه. يقدم القائمون بـــالتنبؤ بنماذج تنبؤ يحتفظ بها حتى التاريخ المتفق عليه لاعلان النتائج التجريبية. وهنا يتم مقارنة نتائج التنبؤ والتجريب. ولقد سجلت نتائج التقييم باســتخدام برامج CASP تقدما فى فاعلية التنبؤ والـــذى يرجـــع الـــى نمـــو بنـــوك المعلومات وأيضا بسبب التحسن فى الطرق المستخدمة. والباب الخـــامس من الكتاب يناقش التنبؤ بتركيب البروتين.

هندسة البروتينات Protein engineering:

كان هناك تشابه بين طبيعة عمل كلا من علمـــاء البيولوجيـــا الجزيئيـــة وعلماء الفلك من حيث القدرة على ملاحظة الأشياء دون اجراء تعديـــلات عليها. الا أن هذا لم يعد حقيقيا الآن. ففى المعمل يمكننا عمل تعديل فـــى الأحماض النووية والبروتين واحداث طفـــرات لمعرفـــة التـــأثير علـــى وظائفهم. فمن الممكن أن يصبح لبروتين قديم وظيفة جديدة مثـــل انتـــاج أجسام مضادة حفازة ومحاولة تحضير بروتينات جديدة.

تشتق العديد من قواعد تركيب البروتينـــات مـــن ملاحظـــات للـــبروتين الطبيعى. وليس ضروريا تطبيق تلك القواعد فى البروتينـــات المهندســـة. تمتلك البروتينات الطبيعية صفات مرتبطـــة بالأســـس العامـــة للكيميـــاء الطبيعية وآليات نشوء البروتين. ويجب أن تخضع البروتينات المهندســـة

لقوانين الكيمياء الطبيعية وليس لقيود النشوء. وتسطيع هندسة البروتينات أن تستكشف آفاق ومجالات جديدة.

التطبيقات الاكلينيكية:

هناك اجماع في الرأى على أن معرفة التتابعات فى الجينوم البشرى وجينوم العديد من الكائنات الأخرى سوف يؤدى الى تحسن كبير فى صحة البشر. وتطبيقات ذلك يمكن أن تتضمن الآتى:

١- تشخيص الأمراض ومخاطرها:

يمكن أن يكشف تتابع الدنا عن غياب جين معين، أو طفرة. يؤدى تعريف تتابعات جين محدد مصاحبا لمرض ما الى سرعة ودقة تشخيص الحالة.

غالبا ما تكون العلاقة بين طبيعة الجين ومخاطر المرض من الصعوبة تحديدها. تعتمد بعض الأمراض مثل الحساسية على تداخلات بين جينات عديدة بالاضافة الى عوامل بيئية. وفى حالات أخرى يكون الجين موجودا وسليما الا أن حدوث طفرة فى مكان ما قد يغير من تعبير ذلك الجين وتوزيعه بين الأنسجة. مثل تلك الأشياء غير السوية يجب اكتشافها بواسطة قياسات نشاط البروتين. كما أن قياس طرز تعبير البروتين تعتبر وسيلة هامة لقياس مدى الاستجابة للعلاج.

٢- وراثة الاستجابة الى العلاج المتخصص:

نظرا لاختلاف الناس فى قدرتها على تمثيل الأدوية فان المرضى المختلفة قد تحتاج الى جرعات مختلفة لعلاج نفس الحالة. يسمح تحليل التتابع اختيار الدواء والجرعة التى تتناسب مع كل مريض وذلك فى اطار مجال ينمو بسرعة يعرف باسم Pharmacogenomics. ويصبح فى استطاعة الأطباء تفادى تجريب طرق علاجية مختلفة والتى لها آثار جانبية خطرة بالاضافة الى التكلفة الباهظة.

٣- تعريف الأماكن المستهدفة للدواء:

المكان المستهدف قد يكون بروتين يتم تعديل وظيفته بالتداخل مع الدواء للتأثير على أعراض المرض ومسبباته. ويلقى التعرف على المكان المستهدف الضوء على الخطوات المتتالية لتصميم الدواء. والأماكن المستهدفة لمعظم الأدوية المستخدمة الآن نصفها مستقبلات وحوالى الربع انزيمات والربع المتبقى هرمونات وحوالى ٧% تعمل على أهداف غير معلومة.

أدى التزايد فى ظاهرة مقاومة البكتريا للمضادات الحيوية الى خلق مأساة بخصوص علاج الأمراض. والحاجة الماسة لايجاد أدوية جديدة مرهونة بقاعدة البيانات اللازمة لانتاجها. ونتائج دراسة الجينوم قد تؤدى الى اقتراح أماكن مستهدفة. الاختلافات فى الجينوم ومقارنة طرز تعبير البروتين بين سلالات مسببات الأمراض الحساسة والمقاومة للأدوية يمكن أن تحدد البروتين المسئول عن مقاومة الدواء. ويأمل استطاعة دراسة الاختلافات الوراثية بين الخلايا السرطانية والعادية فى التعرف على تعبير بروتينات مختلفة تستخدم كأهداف لأدوية مضادة للسرطان.

٤- العلاج بالجينات:

وهو امكانية استبدال جين - أو على الأقل مد الجسم بناتجه - محل جين غائب او مصاب بخلل. كما يتضمن العلاج بالجينات العمل على خفض النشاط الزائد لجين ما. وهناك بعض الحالات المرضية فى الانسان والتى أظهرت نتائج مشجعة بالعلاج الجينى.

ومن التوجهات التى تستخدم لايقاف تعبير الجين هو مايطلق عليه "antisense therapy". وهو عبارة عن انتاج شريط قصير من الدنا أو الرنا والذى يرتبط بأسلوب تتابع متخصص بمنطقة بالجين. وهذا الارتباط يتداخل على التوالى مع عمليات النسخ والانتقال. لقد أظهر

هـذا النـوع مـن العـلاج كفـاءة ضـد بعـض المـراض مثـل cytomegalovirus; and Crohn disease. كما أنه ذهب مباشـرة مـن تتابع المستهدف الى اختصار مراحل عديدة من عمليات تصميم الدواء.

المستقبل The future:

سوف يشهد القرن الحالى تطورا مذهلا في انتاج وتوزيع وسائل العـلاج والعناية بصحة البشر وتتلاشى الحواجز بين قـلاع البحـث والممارسـة الاكلينيكية. ويصبح بامكان قارىء هذا الكتاب أن يتوصـل الـى عـلاج لمرض قاتل. كما يمكن اكتشاف عقاقير ووسائل من شأنها منـع حـدوث الأورام السرطانية وليس فقط العمل على التحكم فى نموها وانتشارها.

مصادر على الانترنت يمكن الرجوع اليها Web Resources:

Human Genome Project Information :
http://www.ornl.gov/hgmis/project/info.html

Genome Statistics:
http://bioinformatics.weizmann.ac.il/mb/statistics.html

Taxonomy Sites :
Species : http://www.sp2000.org
Tree of life : http://phylogeny.arizona.edu/tree

Database of genetics of disease :
http://www.ncbi.nlm.nih.gov/omim/
http://www.geneclinics.org/profiles/all.html

Lists of databases :
http://www.infobiogen.fr/services/dbcat/
http://www.ebi.ac.uk/biocat/

List of tools for analysis :
http://www.ebi.ac.uk./Tools/index.html

Debate on electronic access to the scientific literature :
http://www.nature.com/debates/e-access/

تمرينات محلولة Solved Problems:

Problem 1. For what of following sets of fragment strings does the PERL program mentioned before work correctly ?

(a) Would it correctly recover :

Kate, when France is mine and I am yours, then yours is France and you are mine .

From :

Kate, when France
France is mine
is mine and
and I am\nyours
yours then
Then yours is France
France and you are mine\n

Sample input strings for assembly:

(a) Input data:

Kate, when France
France is mine
is mine and
and I am\nyours
yours then
Then yours is France
France and you are mine\n

(a) Correct answer:

Kate, when France is mine and I am
yours, then yours is France and you are mine.

(b) Would it correctly recover :

One women is fair , yet I am well; another is wise, yet I am well; another virtuous, yet I am well; but till all graces be in one woman, one woman shall not come in my grace .

from :

One woman is
woman is fair,
is fair, yet I am
yet I am well;
I am well; another
another is wise, yet I am well;
yet I am well; another virtuous,
another virtuous, yet I am well;
well; but till all
all graces be

be in one woman,
one woman, one
one woman shall
Shall not come in my grace.

(b) Input data:

One woman is
wo nan is fair,
is f ir, yet I am
yet I am well;
I an well; another
anc ther is wise, yet I am well;
yet I am well; another virtuous,
anc ther virtuous, yet I am well;
we l; but till all
all graces be
be in one woman,
one woman, one
one woman shall
Shall not come in my grace.

(b) Correct answer:

One woman is fair, yet I am well;
another is wise, yet I am well;
another virtuous, yet I am well;
but till all graces be in one woman,
one woman shall not come in my grace.

(c) Would it correctly recover :

That he is made, "tis true: "tis true "tis pity; And pity "tis "tis true.

from :

That he is
is mad, 'tis
'tis true
true: 'tis true 'tis
true 'tis
'tis pity;\n
pity;\nAnd pity
pity 'tis
'tis 'tis
'tis true.\n

(c) Input data:

That he is
is mad, 'tis
'tis true
true: 'tis true 'tis
true 'tis

'tis pity;\n
pity;\nAnd pity
pity 'tis
'tis 'tis
'tis true.\n

(c) Correct answer:

That he is mad, 'tis true: 'tis true 'tis pity;
And pity 'tis 'tis true.

الفصل الثانى

نشوء وتنظيم الجينوم
Genome organization and evolution

الجينوميك والبروتيوميك
Genomics and Proteomics:

الجينوم البكتيرى يتأتى من جزئ دنا منفرد والذى يصل طوله عند فرده الى نحو ٢ ملليمتر (يبلغ قطر الخلية حوالى ٠,٠٠١ ملليمتر). وينظم الدنا فى الكائنات الراقية على كروموسومات حيث تحتوى الخلية العادية فى الانسان على ٢٣ زوجا من الكروسومات. والكمية الكلية للمعلومات الوراثية لكل خلية – تتابع النيكليوتيدات فى الدنا – تكون ثابتة لدرجة كبيرة لكل أفراد النوع الواحد بينما تختلف بصورة واسعة بين الأنواع المختلفة. وبم أن ليس كل الدنا يشفر الى بروتينات فان كم معلومات تتابع البروتين فى الخلية لا يمكن تقديره بسهولة من حجم الجينوم.

البروتيوم Proteomes:

يقدم جينوم الكائن مجموعة كاملة لصفات الفرد. وتعتمد حالة تطور الكائن ونشاطه على المستوى الجزيئى فى أى لحظة على كميات وتوزيع البروتينات. ويعتبر برنامج مشروع البروتيوم من البرامج الضخمة التى تعتنى – باسلوب تكاملى – بطرز تعبير البروتينات فى الأنظمة البيولوجية بشكل يتطابق مع مشروعات الجينوم ويعظمها.

تمييز الجينات فى الجينوم:

تتعرف برامج الكمبيوترلتحليل الجينوم على أطر قراءة مفتوحة open reading frames (ORFs). ويعرف ORF بأنه منطقة من تتابع الدنا تبدأ

بكود بادئ (ATG) وينتهى بكود ايقاف. وهـــى منطقـــة فاعلـــة لتشـــفير البروتين.

وعمليات التعريف على تشفير البروتين تتم بواسطة واحد أو الجمع بيـــن اثنان من التوجهات المحتملة التالية:

١- الكشف عن مناطق مشابهة لمناطق تشفير من كائنات أخرى. وتقـــوم تلك المناطق بتشفير تتابعات حمض أمينى مماثل لبروتين معروف أو مماثل علامات تتابعات ناتج الجيـــن Expressed Sequence Tags (ESTs) وبم أن ESTs تشتق من الرنا الرسول فانها تناظر جينـــات معروفة. ومن الضرورى أن يتم تحديد التتابع لعدة مئات فقـــط مـــن قواعد cDNA لتعطى معلومات كافية للتعرف على الجين. ويشـــابه تحديد الجينات بواسطة ESTs عمليـــة فهرســـة القصـــائد الشـــعرية والأغانى بالاستعانة بأول بيت فيها.

٢- طرق *Ab initio* والتى تسعى الى التعرف على الجينات مـــن خـــلال خصائص تتابعات الدنا.

مصادر على الانترنت Web Resources:

HUMAN GENOME INFORMATION

Interactive access to DNA and protein sequences :
http://www.ensemble.org/

Images of chromosomes, maps, loci :
http://www.ncbi.nlm.nih.gov/genome/guide/

Gene map 99 :
http://www.ncbi.nlm.nih.gov/genemap99/

Overview of human genome structure :
http://hgrep.ims.u-tokyo.ac.jp

Single-nucleotide polymorphisms :
http://snp.cshl.org/

Human genetic disease :
http://www.ncbi.nlm.nih.gov/Omim/
http://www.geneclinics.org/profiles/all.html

Ethical, legal, social issues :
http://www.nhgri.nih.gov/ELSI/

DATABASES OF ALIGNED GENE FAMILIES

Pfam: protein families database :
http://www.sanger.ac.uk/software/pfam/

COG: Clusters of orthologous groups :
http://www.ncbi.nlm.nih.gov/COG/

HBAGENE : Homologous Bacterial Genes Database :
http://pbil.univ-lyon1.fr/database/hobacgen.html

HOVERGEN: Homologous Vertebrate Genes Database :
http://pbil.univ-lyon1.fr/databases/hovergen.html

TAED: The Adaptive Evolution Database :
http://www.sbc.su.se/~liberles/TAED.html

GENOME DATABASES

List of completed genomes :
http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/allorg.html
http://www.ebi.ac.uk/genomes/mot/index.html
http://pir.georgetown.edu/pirwww/search/genome.html

Organism-specific databases :
http://www.unl.edu/stec-95/ResTools/biotools/biotools10.html
http://www.-fp.mcs.anl.gov/~gaasterland/genomes.html
http://www.hgmp.mrc.ac.uk/GenomeWeb/genome-db.html
http://www.bioinformatik.de/cgi-bin/browse/Catalog/Databases/Genome_projects/

الفصل الثالث

السجلات واسترجاع المعلومات

يتناول هذا الفصل من الكتاب مهارات استرجاع المعلومات والتى تتيح الاستخدام الفعال لبنوك البيانات.

فهرسة قواعد البيانات وخصائص مصطلحات البحث:

الفهرس عبارة عن مجموعة من المؤشرات للمعلومات فى قواعد البيانات. عند البحث على شبكة المعلومات الدولية أو من خلال قاعدة بيانات للبيولوجيا الجزيئية يتم ادخال واحد أو أكثر من مصطلحات البحث ويقوم برنامج بالبحث عنها فى جداول الفهارس. وهنا يتعرف برنامج الاسترجاع على الموضوعات ذات المحتويات المتعلقة بمجال الاهتمام. على سبيل المثال عند ادخال كلمة حصان horse سوف تحصل على معلومات تختص بموضوعات عديدة تتعلق بالحصان مثل البيولوجيا الجزيئية والتربية والسلالات والقصائد الشعرية حول الحصان وغيرها. وللتركيز على موضوع محدد يسمح نظام استرجاع المعلومات باستخدام كلمات استرشادية. فمثلا بالبحث عن انزيم ديهيدروجينيز للكحول فى كبد الحصان horse liver alcohol dehydrogenase سوف يتم التعرف فقط على الموضوعات المحتوية على الكلمات الاسترشادية الأربع التى تم ادخالها وهى horse و liver و alcohol و dehydrogenase.

لذلك من المهم استخدام قواعد بيانات متخصصة بما فى ذلك قواعد بيانات البيولوجيا الجزيئية والتى تفرض تركيب معين على المعلومات لفصل فئات مختلفة من المعلومات.

تحليل البيانات المسترجعة:

أحيانا يكون هناك حاجة الى الحصول على برنامج يسمح باستخدام النتائج – المسترجعة من عمليات بحث فى قواعد البيانات – كمدخلات. على سبيل المثال، تستخدم نتائج تتابع بروتين ما كمدخلات للتنقيب بواسطة برنامج PSI- BLAST. وفى هذه الحالة يجب تغذية البرنامج بنتيجة التتابع يدويا. الا أنه كما هو الحال فى التنقيب فى قواعد البيانات المتعددة، تتيح نظم المعلومات المسترجعة – فى البيولوجيا الجزيئية – تسهيلات لبدء تلك العمليات.

السجلات Archives:

بالرغم من أن معرفتنا ببيانات التتابع والتركيب البيولوجى لم تكتمل بعد، الا أن حجم تلك البيانات ينمو سريعا. ويعمل العديد من العلماء على توفير البيانات أو تنفيذ مشروعات بحثية لتحليل النتائج. كما تقوم هيئات معنية بقواعد البيانات بأرشفة وتوزيع البيانات.

وتجرى عمليات أرشفة البينات المتعلقة بالمعلوماتية الحيوية بواسطة مجاميع بحثية مهتمة بالعلوم ذات الصلة.

وتتضمن عملية الجمع الأولى للبيانات المتعلقة بالجزيئات البيولوجية الكبيرة مايلى:

- تتابعات الحمض النووى بما فى ذلك مشروعات الجينوم.
- تتابعات الحمض الأمينى للبروتينات.
- تراكيب البروتين والحمض النووى.
- التراكيب البللورية للجزيئات الصغيرة.
- وظائف البروتين.
- طرز تعبير الجينات.
- المراجع والمؤلفات.

قواعد بيانات تتابع الحمض النووى:

يتكون الأرشيف العالمى لتتابع الحمض النووى من شراكة ثلاثية وهى: المركز القومى لمعلومات التكنولوجيا الحيوية بالولايات المتحدة الأمريكية، ومكتبة البيانات لمعهد البيومعلوماتية الأوروبى بالمملكة المتحدة (EMBL)، وبنك اليابان لمعلومات الدنا (المعهد القومى للوراثة باليابان). ويتم تبادل البيانات بين المجموعة يوميا. ولذلك فان البيانات الأولية تكون متطابقة بالرغم من اختلاف نماذج تخزين البيانات وطبيعة تفسيرها. وتقوم قواعد البيانات هذه بمعالجة وأرشفة وتوزيع تتابعات الدنا والرنا المجمعة من مشروعات الجينوم والمنشورات العلمية وتطبيقات براءات الاختراع. وتطلب المجلات العلمية ايداعات جديدة لتتابعات جديدة للنيكليوتيدات فى قاعدة البيانات كشرط لنشر المقالة وذلك للتأكد من حرية توفير تلك البيانات الأساسية. وكذلك الحال بالنسبة لتتابعات الحمض الأمينى وتراكيب الحمض النووى والبروتين.

تتكون قواعد بيانات تتابع الحمض النووى من مجموعات من المدخلات. لكل مدخل ملف نص يحتوى على البيانات والتفسيرات لكل تتابع مجلور. العديد من المدخلات تكون مجمعة من عدة أوراق علمية منشورة تقرر أجزاء متداخلة لتتابع كامل.

فيمايلى مثال لمدخلات تتابع دنا من مكتبة بيانات EMBL متضمنة تفسيرات وبيانات تتابع جين مثبط التربسين البنكرياسى:

The EMBL Data Library entry for the Bovine pancreatic trypsin inhibitor gene.

```
ID   BTBPTIG    standard; DNA; MAM; 3998 BP.
XX
AC   X03365; K00966;
XX
SV   X03365.1
XX
DT   18-NOV-1986 (Rel. 10, Created)
DT   20-MAY-1992 (Rel. 31, Last updated, Version 3)
XX
DE   Bovine pancreatic trypsin inhibitor (BPTI) gene
XX
KW      Alu-like  repetitive  sequence; protease inhibitor; trypsin
inhibitor.
XX
OS   Bos taurus (cow)
```

```
OC         Eukaryota;    Metazoa;  Chordata;  Craniata;  Vertebrata;
Euteleostomi; Mammalia;
OC      Eutheria;  Cetartiodactyla;  Ruminantia;   Pecora; Bovoidea;
Bovidae; Bovinae;
OC   Bos.
XX
RN   [1]
RP   1-3998
RX   MEDLINE; 86158754.
RA   Kingston I.B., Anderson S.;
RT    "Sequences encoding two trypsin inhibitors occur in strikingly
similar
RT   genomic environments";
RL   Biochem. J. 233:443-450(1986).
XX
RN   [2]
RX   MEDLINE; 84070725.
RA   Anderson S., Kingston I.B.;
RT       "Isolation of a genomic clone for bovine pancreatic trypsin
inhibitor by
RT   using a unique-sequence synthetic dna probe";
RL   Proc. Natl. Acad. Sci. U.S.A. 80:6838-6842(1983).
XX
DR   SWISS-PROT; P00974; BPT1_BOVIN.
XX
CC   Data kindly reviewed (08-DEC-1987) by Kingston I.B.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..3998
FT                   /db_xref="taxon:9913"
FT                   /organism="Bos taurus"
FT   misc_feature    795..800
FT                   /note="pot. polyA signal"
FT   misc_feature    835..839
FT                   /note="pot. polyA signal"
FT   repeat_region   837..847
FT                   /note="direct repeat"
FT   misc_feature    930..945
FT                            /note="sequence homologous to Alu-like
consensus seq."
FT   repeat_region   1035..1045
FT                   /note="direct repeat"
FT   misc_feature    2456..2461
FT                   /note="pot. splice signal"
FT   CDS             2470..2736
FT                   /db_xref="SWISS-PROT:P00974"
FT                   /note="put. precursor"
FT                   /protein_id="CAA27062.1"
FT
/translation="PSLFNRDPPIPAAQRPDFCLEPPYTGPCKARIIRYFYNAKAGLCQ
FT                   TFVYGGCRAKRNNFKSAEDCMRTCGGAIGPWGKTGGRAEGEGKG"
FT   misc_feature    2488..2489
FT                   /note="pot. intron/exon splice junction"
FT   misc_feature    2506..2507
FT                   /note="pot. intron/exon splice junction"
FT   CDS             2512..2685
FT                   /db_xref="SWISS-PROT:P00974"
FT                   /note="trypsin inhibitor (aa 1-58)"
FT                   /protein_id="CAA27063.1"
FT
/translation="RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNF
FT                   KSAEDCMRTCGGA"
FT   misc_feature    2698..2699
FT                   /note="pot. exon/intron splice junction"
FT   misc_feature    3690..3695
FT                   /note="pot. polyA signal"
```

```
FT   misc_feature    3729..3733
FT                   /note="pot. polyA signal"
XX
SQ   Sequence 3998 BP; 1053 A; 902 C; 892 G; 1151 T; 0 other;
     aattctgata   atgcagagaa   ctggtaagga   gttctgattg   ttctgcttga
ttaaatgggt        60
     tgtaacagga   tagtgtcttg   tcctgatcct   agcattcata   tggtgtgtgt
tctggggcaa       120
     gtcatctgca   gtttcttcac   ctgaacaggg   ggaccaggtt   acatgagttt
cttaaaagat       180
     taccagtcat   gagtatgaag   agtttacact   ttcctgatca   atgacgtcca
tttcccatca       240
     aaatatttta   gtccaaaaga   ctcatctatc   taatgtagat   cattttctca
ccacccctct       300
     aaaaaattta   tctttcagat   atgatcattt   ctctattatg   aaattaatca
gagagttgag       360
     tgacagctga   gtgtcttccc   tccaaaggca   actgcaggaa   gagcaagaaa
tgcaatactt       420
     ttctatgagt   ttgctcgtgg   ggccaagact   gctttttcca   ggctggtaca
atagtaatca       480
     aatctcaaag   atattcttct   ttcctcctgg   ccagactatt   attttatttt
cctatcaaga       540
     tatagaaagt   tagaagtaga   ctcataatta   tataggcagg   cctcatcatc
aaatagacta       600
     acaagaattt   tattttatct   gccttttcaa   tgactgtgca   cttggcatga
ggatgaaatg       660
     ggagatttat   tcccttgata   aatattcatg   aaatacttat   gctttttgtc
cctaaaaagc       720
     atatttcttg   atataggaaa   acagctgtaa   acaaaaggta   gtaaaataat
atgccttcta       780
     agagggatac   agacaataaa   gacgggggcg   gattcctata   ccaggtcatg
atgagtgtca       840
     tgaggaaggt   gagttatggg   gttcaggatg   ctgtagagga   tcagggaaac
cctctgtgat       900
     gaggagacat   taagcagaag   ctgccaaaaa   ggagcctggt   gtgtttgagc
agcagccagg       960
     accagggtgg   ctggagctga   gtgggtgagg   ggaggggagt   ggaaggggat
gaagcagaga      1020
     ggccatgggg   gcaggtcatg   aggaaccttc   taggacttta   taaggataaa
aatttgactc      1080
     tgagagagct   gggaaaccac   tgagggactg   gtcgttggaa   caacgagata
ggactggagt      1140
     tttaacaggg   tccttgcgac   tgcagtgtgg   agcgtggcct   ctaggggcg
aaaagcaggg      1200
     acaggggcc    cctgggcagg   tggctgcagg   ggtccagtga   gctatgatgg
agaatatata      1260
     cctgtgttgt   tccctgggtt   gattccagtt   ctcttgaata   acctgaatga
ctgctatat      1320
     taagatatct   ggggaggctt   catcacaaat   catgattcat   taaatcttta
gtcatttgtg      1380
     attgattcaa   cctccaatcc   ctctcccctt   cctgagatgt   ggatacaatg
aaagaagtag      1440
     gaatgaaaat   tcccacaccc   aactcaggca   gttgtttccc   ctgacaactt
atccccattc      1500
     ttggctttgc   ttgaggcttt   acaaaactca   tctccctcac   atgataaagg
actccccttt      1560
     gctctcatct   cttaggaaat   tccaatgttt   taggagctct   gtggcaggaa
tgggatgcag      1620
     accaagttaa   tatttctttt   ataagtcaca   gtatcaatat   ttcctcaata
ttctattatt      1680
     ccagtctcca   tgaggtaacc   aaagtaaaca   ccggtgtgtt   ttctaccatg
cctttctcca      1740
     tttatggcat   gatttcctca   cacttttgta   atagtgacgg   gtcacgcagg
cctatcaacc      1800
     attggctggc   atccaggtgg   gcaccttcat   caagggataa   ctgtaaatga
gcaacccttg      1860
```

```
     gtggtccagt    tcagccattg   ccactgtcgt   agccacagtg   ggctcttcgc
cctcccgttc      1920
     ttttgtataa    aaggaacagg   aatttatact   gtgggaagat   ggttttcttg
agacagtagc       1980
     atgctatcat    ctccgtgggc   ccaatttcca   attaaaaatg   ttattcctag
ttccagcaac       2040
     tcttctccgg    attattggct   ggccctgagg   tgagcagaat   gagactgggc
tcagtgatgc       2100
     tttcttaacg    gtggaagttt   ccaccacaca   catacataga   aagcatagta
ttaaaaaagc       2160
     cgtggatcca    ttgtccagct   ccagtaattt   ctatacatgg   agagtatttt
tatatgtgtc       2220
     cctcttttgt    gttacttttg   aaactcatca   gtagcatcat   gctaattaat
gcataaacat       2280
     tcatcaatgg    catgtaatta   tttataatat   tgccctgtca   ttgtcacacc
taacaacatt       2340
     aataataatg    tcctggaaag   cagggtgtca   aaaggccttt   tcacgtttca
cacttctgcc       2400
     caccccccat    cactctctat   cacaaactgg   tggttttagt   tgttcatctt
gtagactgag       2460
     ctgtgatgac    cttccctctt   taaccgagat   cctcccatcc   ctgcagccca
gcggcctgac       2520
     ttctgcctag    agcctccata   tacgggtccc   tgcaaggcca   gaattatcag
atacttctac       2580
     aacgccaagg    ctgggctctg   ccagaccttt   gtatatggcg   gctgcagagc
taaaagaaac       2640
     aatttcaaga    gcgcagagga   ctgcatgagg   acctgtggtg   gtgctattgg
gccctggggt       2700
     aagacagggg    gcagggcaga   gggagagggg   aagggctagg   gaaagtggtg
gcgctcagaa       2760
     ggccacacac    ctttccaaaa   aagtgatttt   tttcccttgt   tgcctcccaa
gagaagttgc       2820
     agaagtatcc    gtggattgag   catgtcctcc   atggaccagc   ttggtgaaag
gccacccccc       2880
     agaagccctg    tcatcataat   ctgagcctac   tcacatgctc   ccatttttca
gatgggaaca       2940
     ctgagtcagt    cactctgcag   agcaagtctg   gagtgctcct   cagtgcccca
cctcagcctg       3000
     gaaaactccc    ttgtttattg   ttggttatcc   tggtcctggg   aggactgtgg
ttgcgcattt       3060
     ctgggatggt    ctaggacctg   tcagggtgga   cagtgtccag   gtctgggcct
tcagagatgt       3120
     cattcagcaa    gttcctttct   ttttacagag   aacctgtgaa   ctgtgctccc
ctgagatgct       3180
     gaagtatgag    gaggacccac   ccaaggctgg   cctctatctg   cttctgaaaa
atttcagcct       3240
     ccttttattt    cttctcaacc   ctcccctcct   cagcagaaat   ctgtctcttt
cctcctccca       3300
     caggtccact    tactttagcc   ctatctcatc   cagtttgctc   taagcaccat
gaaagcaaat       3360
     cttccctttg    tccctcacac   ttcccacaat   ttctggcaca   aaggagaagg
tccagaaata       3420
     ttggaggaag    gaaggaatga   agttccccat   gactggagca   tctgtagagt
ctgagattta       3480
     aatctggatt    cttgtcctaa   tcttcctcct   cacggcatcc   ttaccttcat
cctccacccc       3540
     accatcactg    ctctcccctct  actggcgaaa   gtagaatttc   catcatcgag
ttttcagctc       3600
     agtggtggga    gaggtctttt   catgaacgaa   acctcctcct   cacattgatt
tgaaggtctg       3660
     tggcttcaaa    gagtctggcc   ttatctttaa   ataaattcat   attttaatta
aactaactgg       3720
     agtggattgt    gttgtttgca   actaagaacc   ttaacccata   ggttccatgg
aaacggtggt       3780
     ctttctcatt    ttatgcagat   gggtgggcag   ctctccatca   cctctcctca
gactcagccc       3840
```

```
     taccaagtag    aaggagccaa   ccccttacac   tgacatctac   ctcttatggc
cgtgccagtg      3900
     tacatgaaaa    actggatgag   agacacctca   acaagaaaac   ttttgtcctt
cacttcttgg      3960
     gccaggtcaa actttggggt gtgttatttc cctgaatt            3998
//
```

السطور التى تبدأ بFT هى عبارة عن مكون لتفسير المدخل الذى يقـرر صفات مناطق متخصصة (تتابعات كودية CDS) بحيث تقـرأ بواسـطة برامج كمبيوتر. على سبيل المثال: لترجمة كود منطقة الى تتابع حمـض أمينى، توجد نماذج متحكم فيها بعناية ومفردات مقيدة. ويكون من المـهم وجود مفردات لغوية محكمة وقواميس وكلمات استرشادية وجداول مميزة حتى يمكن انشاء روابط بين قواعد بيانات مختلفة.

قواعد بيانات الجينوم:

بالرغم من أن تتابعات الجينوم تشكل مدخلات الأرشيفات المرجعية لتتـابع الحمض النووى، فان العديد من الأنواع لها قواعد بيانات خاصة والتـى تجمع تتابع الجينوم وتفسيره مع باقى البيانات ذات العلاقة بالنوع.

مصادر على الانترنت Web Resources لروابط قواعد بيانات للكائنات:

http://www.unl.edu/stc-95/ResTools/biotools/biotools10.html

http://www.-fp.mcs.anal.gov/~gaasterland/genomes.html

http://www.hgmp.mrc.ac.uk/GenomeWeb/genome-db.html

http://www.bioinformatik.de/cgi-bin/browse/Catalog/Database/Genome_projects/

قواعد بيانات تتابع البروتين:

تتأتى بيانات تتابع الحمض الأمينى من ترجمة تتابعات الحمض النـووى. يتعاون المعهد السويسرى للبيومعلوماتية مع مكتبة بيانات EMBL لـتزود قاعدة بيانات تفسيرية لتتابعات الحمض الأمينى تسـمى SWISS-PROT. كما توجد قاعدة بيانات لتتابع البروتين تنتجـها The PIR International والتى تشكل مجموعات فى المؤسسة القومية للبحـوث الطبيـة بجامعـة جورج تاون واشنطن – الولايات المتحدة الأمريكية، ومركـز معلومـات

ميونيخ لتتابعات البروتين (MPIS) بألمانيا، وقاعدة بيانات اليابان الدولية لمعلومات البروتين.

وفيمايلى نموذج لمدخلات تتابع الحمض الأمينى لبروتين لمثبط التربسين البنكرياسى فى قاعدة بيانات PIR:

PIR entry for the amino-acid sequence of Bovine pancreatic trypsin inhibitor

```
ENTRY            TIBO  #type complete    iProClass View of TIBO
TITLE            basic proteinase inhibitor precursor - bovine
ALTERNATE_NAMES aprotinin; basic pancreatic trypsin inhibitor;
BPTI;
                 cationic kallikrein inhibitor; inhibitor IV
ORGANISM         #formal_name Bos primigenius taurus #common_name
cattle
   #cross-references taxon:9913
DATE             24-Apr-1984 #sequence_revision 22-Jul-1994
#text_change
                 16-Jun-2000
ACCESSIONS       S00277; A30333; S10546; S02486; S28197; A90162;
A92023;
                 A90736; A90927; A34658; A93977; S10062; A01205
REFERENCE        S00274
   #authors      Creighton, T.E.; Charles, I.G.
   #journal      J. Mol. Biol. (1987) 194:11-22
   #title        Sequences of the genes and polypeptide precursors
for two
                 bovine protease inhibitors.
   #cross-references MUID:87283904
   #accession    S00277
      ##molecule_type DNA; mRNA
      ##residues 1-100 ##label CR2
      ##cross-references GB:M20934; GB:X05274; NID:g162767;
                 PIDN:AAD13685.1; PID:g162769
REFERENCE        A90926
   #authors      Creighton, T.E.; Charles, I.G.
   #journal      Cold Spring Harb. Symp. Quant. Biol. (1987) 52:511-
519
   #title        Biosynthesis, processing, and evolution of bovine
                 pancreatic trypsin inhibitor.
   #cross-references MUID:88295740
   #accession    A30333
      ##molecule_type DNA
      ##residues 1-100 ##label CRE
      ##cross-references GB:M20934; GB:X05274; NID:g162767;
                 PIDN:AAD13685.1; PID:g162769
REFERENCE        S10546
   #authors      Kingston, I.B.; Anderson, S.
   #journal      Biochem. J. (1986) 233:443-450
   #title        Sequences encoding two trypsin inhibitors occur in
                 strikingly similar genomic environments.
   #cross-references MUID:86158754
   #accession    S10546
      ##molecule_type DNA
      ##residues 34-97 ##label KIN
REFERENCE        S02485
   #authors      Fioretti, E.; Angeletti, M.; Fiorucci, L.; Barra,
D.;
                 Bossa, F.; Ascoli, F.
   #journal      Biol. Chem. Hoppe-Seyler (1988) 369(Suppl.):37-42
```

```
   #title       Aprotinin-like isoinhibitors in bovine organs.
   #cross-references MUID:89076531
   #accession   S02486
      ##molecule_type protein
      ##residues 36-93 ##label FIO
REFERENCE       S28197
   #authors     Ikekita, M.; Jone, C.S.; Kamo, M.; Tsugita, A.;
Kizuki,
                K.; Moriya, H.
   #journal     Protein Seq. Data Anal. (1992) 5:7-11
   #title       Purification and characterization of the major
cationic
                kallikrein inhibitor in bovine pituitary gland.
   #cross-references MUID:93150003
   #accession   S28197
      ##molecule_type protein
      ##residues 36-93 ##label IKE
REFERENCE       A90162
   #authors     Kassell, B.; Laskowski, M.
   #journal     Biochem. Biophys. Res. Commun. (1965) 20:463-468
   #title       The basic trypsin inhibitor of bovine pancreas. V.
The
                disulfide linkages.
   #cross-references MUID:66083012
   #contents    annotation; disulfide bonds
   #accession   A90162
      ##molecule_type protein
      ##residues 36-93 ##label KAS
REFERENCE       A92023
   #authors     Anderer, F.A.; Hornle, S.
   #journal     J. Biol. Chem. (1966) 241:1568-1572
   #title       The disulfide linkages in kallikrein inactivator of
bovine
                lung.
   #cross-references MUID:66171231
   #contents    annotation; disulfide bonds
   #accession   A92023
      ##molecule_type protein
      ##residues 36-93 ##label AN2
REFERENCE       A90736
   #authors     Chauvet, J.; Acher, R.
   #journal     Bull. Soc. Chim. Biol. (1967) 49:985-1000
   #title       La structure covalente d'un inhibiteur
polypeptidique de
                la trypsine (inhibiteur de Kunitz et Northrop).
   #cross-references MUID:68012003
   #contents    annotation; disulfide bonds
   #accession   A90736
      ##molecule_type protein
      ##residues 36-93 ##label CHA
REFERENCE       A90927
   #authors     Dlouha, V.; Pospisilova, D.; Meloun, B.; Sorm, F.
   #journal     Collect. Czech. Chem. Commun. (1968) 33:1363-1365
   #title       Sequence of residues 18-20 in pancreatic trypsin
                inhibitor.
   #accession   A90927
      ##molecule_type protein
      ##residues 36-93 ##label DLO
REFERENCE       A93410
   #authors     Huber, R.; Kukla, D.; Ruhlmann, A.; Epp, O.;
Formanek, H.
   #journal     Naturwissenschaften (1970) 57:389-392
   #title       The basic trypsin inhibitor of bovine pancreas. I.
                Structure analysis and conformation of the
polypeptide
                chain.
   #cross-references MUID:70255230
```

```
   #contents     annotation; X-ray crystallography of basic protease
                 inhibitor, 2.5 angstroms
REFERENCE        A34658
   #authors      Lewis, R.V.; Ray, P.; Coguill, R.; Kruggel, W.
   #journal      Biochem. Biophys. Res. Commun. (1990) 167:543-547
   #title        Presence of pancreatic trypsin inhibitor in adrenal
                 medullary chromaffin cells.
   #cross-references MUID:90211226
   #accession    A34658
      ##molecule_type protein
      ##residues 36-53,55-81,##label LEW
REFERENCE        A93977
   #authors      Anderson, S.; Kingston, I.B.
   #journal      Proc. Natl. Acad. Sci. U.S.A. (1983) 80:6838-6842
   #title        Isolation of a genomic clone for bovine pancreatic
trypsin
                 inhibitor by using a unique-sequence synthetic DNA
probe.
   #cross-references MUID:84070725
   #accession    A93977
      ##molecule_type DNA
      ##residues 'PSLFNRDPPIPA',34-97,'GKTGGRAEGEGKG' ##label AND
      ##cross-references GB:X03365; GB:K00966; NID:g142;
PIDN:CAA27062.1;
                 PID:g1364183
REFERENCE        S00371
   #authors      Siekmann, J.; Wenzel, H.R.; Schroeder, W.;
Tschesche, H.
   #journal      Biol. Chem. Hoppe-Seyler (1988) 369:157-163
   #title        Characterization and sequence determination of six
                 aprotinin homologues from bovine lungs.
   #cross-references MUID:88221840
   #accession    S10062
      ##molecule_type protein
      ##residues 36-66,'P',68-82,'S',84-93 ##label SIE
      ##experimental_source lung
      ##note the authors designated this protein as isoaprotinin 2
COMMENT          Basic proteinase inhibitor is an intracellular
polypeptide
                 found in many tissues, probably located in granules
of
                 connective tissue mast cells.
GENETICS
   #introns      34/1; 98/1
CLASSIFICATION   #superfamily basic proteinase inhibitor; animal
                 Kunitz-type proteinase inhibitor homology
KEYWORDS         serine proteinase inhibitor
FEATURE
   1-20                  #domain signal sequence #status predicted
#label
                         SIG\
   21-35                 #domain propeptide #status predicted
#label PRO\
   36-100                #product basic proteinase inhibitor
#status
                         experimental #label MAT\
   40-90                 #domain animal Kunitz-type proteinase
inhibitor
                         homology #label BPI\
   40-90,49-73,65-86     #disulfide_bonds #status experimental\
   50                    #inhibitory_site Lys (trypsin,
chymotrypsin,
                         kallikrein, plasmin) #status experimental
SUMMARY          #length 100 #molecular_weight 10903


SEQUENCE
```

```
            5        10        15        20        25        30
  1 M K M S R L C L S V A L L V L L G T L A A S T P G C D T S N
 31 Q A K A Q R P D F C L E P P Y T G P C K A R I I R Y F Y N A
 61 K A G L C Q T F V Y G G C R A K R N N F K S A E D C M R T C
 91 G G A I G P W E N L
```

---

in the PIR1 section of the Protein Sequence Database, release 71.00,
31-Dec-2001, assembled and annotated by the PIR-International.

PDB structures most related to TIBO:
1CBWD (36-93) 100.0%; 1BZ5E (36-93) 100.0%; 9PTI (36-91) 100.0%
1BZXI (36-93) 100.0%; 1B0CB (36-93) 100.0%; 1CBWI (36-93) 100.0%
1B0CD (36-93) 100.0%; 1B0CE (36-93) 100.0%; 6PTI (36-93) 100.0%
1BHCA (36-93) 100.0%; 1BHCC (36-93) 100.0%; 1MTND (36-93) 100.0%
1BHCE (36-93) 100.0%; 1MTNH (36-93) 100.0%; 1BHCG (36-93) 100.0%
1PIT (36-93) 100.0%; 1BHCI (36-93) 100.0%; 1TPAI (36-93) 100.0%
1BPI (36-93) 100.0%; 2HEXA (36-93) 100.0%; 1BTHQ (36-93) 100.0%
2HEXB (36-93) 100.0%; 1BZ5B (36-93) 100.0%; 2HEXC (36-93) 100.0%
1BZ5D (36-93) 100.0%; 2HEXD (36-93) 100.0%; 1B0CC (36-93) 100.0%
2HEXE (36-93) 100.0%; 1BHCD (36-93) 100.0%; 2KAII (36-93) 100.0%
1BHCH (36-93) 100.0%; 2PTCI (36-93) 100.0%; 1BTHP (36-93) 100.0%
2TGPI (36-93) 100.0%; 1BZ5C (36-93) 100.0%; 2TPII (36-93) 100.0%
1BHCB (36-93) 100.0%; 3TGII (36-100) 100.0%; 1BHCJ (36-93) 100.0%
3TGJI (36-100) 100.0%; 1B0CA (36-93) 100.0%; 3TPII (36-93) 100.0%
1BZ5A (36-93) 100.0%; 4PTI (36-93) 100.0%; 1BHCF (36-93) 100.0%
5PTI (36-93) 100.0%; 1FAN (36-93) 98.3%; 1BPT (36-93) 98.3%
1NAG (36-93) 98.3%; 1BTI (36-93) 98.3%; 4TPII (36-93) 98.3%
8PTI (36-93) 98.3%; 1AALA (36-93) 96.6%; 1AALB (36-93) 96.6%
7PTI (36-93) 96.6%; 1BRBI (36-93) 94.8%; 1QLQA (36-93) 93.1%

SCOP: 1CBW ; 1BZ5 ; 9PTI ; 1BZX ; 1B0C ; 6PTI ; 1BHC ; 1MTN ; 1PIT ; 1TPA
; 1BPI ; 2HEX ; 1BTH ; 2KAI ; 2PTC ; 2TGP ; 2TPI ; 3TGI ; 3TGJ ; 3TPI ; 4PTI
; 5PTI ; 1FAN ; 1BPT ; 1NAG ; 1BTI ; 4TPI ; 8PTI ; 1AAL ; 7PTI ; 1BRB ; 1QLQ

CATH: 1CBW ; 1BZ5 ; 9PTI ; 1BZX ; 1B0C ; 6PTI ; 1BHC ; 1MTN ; 1PIT ; 1TPA
; 1BPI ; 2HEX ; 1BTH ; 2KAI ; 2PTC ; 2TGP ; 2TPI ; 3TGI ; 3TGJ ; 3TPI ; 4PTI
; 5PTI ; 1FAN ; 1BPT ; 1NAG ; 1BTI ; 4TPI ; 8PTI ; 1AAL ; 7PTI ; 1BRB ; 1QLQ

```
FSSP:   1CBW ; 1BZ5 ; 9PTI ; 1BZX ; 1B0C ; 6PTI ; 1BHC ; 1MTN ;
1PIT ; 1TPA
; 1BPI ; 2HEX ; 1BTH ; 2KAI ; 2PTC ; 2TGP ; 2TPI ; 3TGI ; 3TGJ ;
3TPI ; 4PTI
; 5PTI ; 1FAN ; 1BPT ; 1NAG ; 1BTI ; 4TPI ; 8PTI ; 1AAL ; 7PTI ;
1BRB ; 1QLQ


MMDB:   1CBW ; 1BZ5 ; 9PTI ; 1BZX ; 1B0C ; 6PTI ; 1BHC ; 1MTN ;
1PIT ; 1TPA
; 1BPI ; 2HEX ; 1BTH ; 2KAI ; 2PTC ; 2TGP ; 2TPI ; 3TGI ; 3TGJ ;
3TPI ; 4PTI
; 5PTI ; 1FAN ; 1BPT ; 1NAG ; 1BTI ; 4TPI ; 8PTI ; 1AAL ; 7PTI ;
1BRB ; 1QLQ


ALIGNMENTS containing TIBO:
     FA2061 basic proteinase inhibitor - 328.8 1.0
     SA0572 basic proteinase inhibitor superfamily 328.8
     M01603 basic proteinase inhibitor - 1561.0 1.0
Associated Alignments:
     DA1053 animal Kunitz-type proteinase inhibitor homology
```

قواعد البيانات المشاركة لـ SWISS-PROT:

يوجد اثنان من قواعد البيانات المشاركة ل SWISS-PROT وهما ENZYME DB & PROSITE.

يقوم ENZYME DB بتخزين المعلومات التالية عن الانزيمات:

- EC Number وهو وسيلة تعريف رقمية تم اقرارها بواسطة مفوضية الانزيمات المنبثقة من الاتحاد الدولى للكيمياء الحيوية والبيولوجيا الجزيئية، (يمكن الرجوع الى الموقع التالى: http://www.chem.qmw.ac.uk/iubmb/enzyme/)
- الاسم الموصى به.
- الاسم البديل، ان وجد.
- النشاط المحفز.
- المرافقات.
- مؤشرات الى SWISS-PROT وبنوك المعلومات الأخرى.
- مؤشرات الى الأمراض المصاحبة للخلل فى النشاط الانزيمى.

والمثال التالى يوضح المخلات فى ENZYME DB حيث:

ID = التعريف

DE = الوصف = الاسم الرسمى

AN = الاسم البديل

CA = النشاط التحفيزى

CF = المرافقات

CC = التعليقات

DR = مرجع قاعدة البيانات (SWISS-PROT).

A Sample Entry in ENZYME DB

```
ID 1.14.17.3
DE PEPTIDYLGLYCINE MONOOXYGENASE.
AN PEPTIDYL ALPHA-AMIDATING ENZYME.
AN PEPTIDYLGLYCINE 2-HYDROXYLASE.
CA PEPTIDYLGLYCINE + ASCORBATE + O(2) = PEPTIDYL(2-HYDROXYGLYCINE)+
CA DEHYDROASCORBATE + H(2)O.
CF COPPER.
CC -!- PEPTIDYLGLYCINES WITH A NEUTRAL AMINO ACID RESIDUE IN THE
CC PENULTIMATE POSITION ARE THE BEST SUBSTRATES FOR THE ENZYME.
CC -!- THE ENZYME ALSO CATALYZES THE DISMUTATATION OF THE PRODUCT TO
CC GLYOXYLATE AND THE CORRESPONDING DESGLYCINE PEPTIDE AMIDE.
DR P10731, AMD_BOVIN ; P19021, AMD_HUMAN ; P14925, AMD_RAT ;
DR P08478, AMD1_XENLA; P12890, AMD2_XENLA;
```

PROSITE :

تحتوى على طرز لمتبقيات مجاميع البروتينات مثل الطراز (البصمـــة أو القالب) الذى يظهر عادة فى عائلة من البروتينات القريبة بسبب متطلبـــات أماكن الارتباط والتى تحد من التحول فى عائلة بروتين ما.

PIR وقواعد البيانات المشاركة:

يدير PIR عدة قواعد بيانات خاصة بالبروتينات مثل:

- PIR-PSD : وهى قاعدة البيانات الأساسية لتتابع البروتين.
- iProClass : تقسيم البروتينات طبقا للتركيب والوظيفة.

- ASDB : قاعدة بيانات التفسير والمماثلة، وكل مدخل متصل بقائمة من التتابعات المماثلة.
- P/R-NREF : تجميع لأكثر من ٨٠٠٠٠٠ تتابع بروتين من مصادر متاحة.
- NRL3D : قاعدة بيانات لتتابعات وتفسيرات لبروتينات معروفة التركيب وموجودة ببنك معلومات البروتين.
- ALN : قاعدة بيانات لمحازاة تتابع البروتين.
- RESID : قاعدة بيانات لتحورات التركيب التساهمى للبروتين.

وقد أنشأ PIR موقع IESA (Integrated Environment for Sequence Analysis) لاسترجاع المعلومات والمعاملة الحسابية لها. ويمكن الرجوع الى موقع PIR على الوب:

http://pir.georgetown.edu

قواعد بيانات التراكيب:

يقوم أرشيف قواعد بيانات التركيب بتفسير وتوزيع مجاميع من المنسقات الذرية. ويعتبر بنك بيانات البروتين Protein Data Bank (PDA) من أفضل قواعد البيانات الخاصة بتراكيب الجزيئات البيولوجية الكبيرة. ويحتوى البنك على تراكيب بروتينات وأحماض نووية و قليل من الكربوهيدرات. ويدار البنك بواسطة the Research Collaboratory for Structural Bioinformatics (RCSB).

والموقع الالكترونى الأساسى لبنك بيانات البروتين هو: http://www.rcsb.org. وهناك مواقع رسمية للبنك لأماكن عديدة حول العالم منها أوروبا، سينغافورة، اليابان، والبرازيل.

وتحتوى الصفحة الأساسية للبنك على روابط اتصال ملفات البيانات ذاتها، ومواد تعليمية، أخبار، نشرات دورية وبرامج بحث متخصصة لمعالجة التراكيب.

مثال: يمكن الحصول على بيانات بروتين ثيوردوكسين فى بكتريا *E. coli* (Protein data bank entry 2trx, *E. coli* thioredoxin) من خلال الموقع التالى:

http://www.rcsb.org/pdb/cgi/explore.cgi?job=download;pdbId=2TRX;page=0;pid=197301010382789&opt=show&format=PDB

وتضمن المدخلات البيانات التالية:

- نوع البروتين ومصدره.
- مراجع لوصف طرق تحديد التركيب.
- بيانات تفصيلية لتجارب تحديد التركيب متضمنة نتائج الأشعة السينية واحصائيات للشكل الفراغي للتركيب.
- تتابع الحمض الأمينى.
- الجزيئات الاضافية مثل المرافقات، المثبطات، جزيئات الماء.
- التركيب الثانوي: الشكل المطوى، المنفرد.
- المواقع ثنائية الكبريت.
- المنسقات الذرية.

وقد حدد بنك بيانات البروتين كود تعرف رباعى لكل تركيب موجود بالبنك، يبدأ برقم (من ١ الى ٩). وأصبح من السهل استرجاع تركيب ما وذلك باستخدام كود التعريف الخاص به. حيث يمكن ادخال الكود على الصفحة الرئيسية لموقع RCSB ثم اختيار Explorer للحصول على ملخص لبيانات التركيب من صفحة واحدة كما هو مبين:

Protein data bank entry Structure Explorer - 2trx, *E. coli* thioredoxin
http://www.rcsb.org/pdb/cgi/explore.cgi?pid=197301010382789&page=0&pdbId=2TRX

RCSB **PDB**
PROTEIN DATA BANK

Structure Explorer - 2TRX

Summary Information

Summary Information
View Structure
Download/Display File
Structural Neighbors
Geometry
Other Sources
Sequence Details
Crystallization Info

Top of Form
Explore
SearchLite SearchFields
Bottom of Form

Title: **Crystal structure of thioredoxin from Escherichia coli at 1.68 A resolution.**

Compound: **Thioredoxin**

Authors: **S. K. Katti, D. M. LeMaster, H. Eklund**

Exp. Method: **X-ray Diffraction**

Classification: **Electron Transport**

Source: **Escherichia coli**

Primary Citation: Katti, S. K., LeMaster, D. M., Eklund, H.: **Crystal structure of thioredoxin from Escherichia coli at 1.68 A resolution.** ***J Mol Biol*** **212** ***pp.*** **167 (1990)**

Deposition Date: **19-Mar-1990** *Release Date:* **15-Oct-1991**

Resolution [Å]: **1.68** *R-Value:* **0.165**

Space Group: **C 2**

Unit Cell: dim [Å]: *a* **89.50** *B* **51.06** *c* **60.45**
alpha **90.00** *Beta* **113.50** *gamma* **90.00**

Polymer Chains: **A, B** *Residues:* **216**

Atoms: **1842**

Chemical Components: ("HET" groups)

ID (needs Rasmol)	Name	Formula	Retrieve All PDB IDs Containing
CU	COPPER (II) ION	2(Cu)	CU
MPD	2-METHYL-2,4-PENTANEDIOL	$7(C_6H_{14}O_2)$	MPD

CATH: Structural Classification

PDBSum: Summary of PDB Structure

SCOP: Structural Classification

RCSB

والصفحة السابقة بها روابط اتصال للمعلومات التالية:

- مراجع للبيانات الموجودة من خلال قاعدة بيانات المراجع PubMed.
- صور للتراكيب ويحتاج ذلك وجود برنامج لاظهار الصور.
- مداخل لملفات البيانات.
- قوائم بالتراكيب الأخرى ذات الصلة طبقا لتصنيف تراكيب البروتين.
- التحليل الكيميائى الفراغى، وتوزيع أطوال الروابط والزوايا.
- مصادر لمعلومات حول المدخلات.
- التتابع والتراكيب الثانوية.
- تفاصيل حول الشكل البللورى وطرق انتاج البللورت.

وفى حالة عدم معرفة كود التعرف يمكن استخدام أداة بسيطة من على الصفحة الرئيسية لبنك بيانات البروتين وهى Search Lite والتى تسمح باستخدام كلمات استرشادية key words.

مصادر على الانترنت Web Resources لتراكيب البروتين والحمض النووى:

الصفحة الرئيسية لبنك بيانات البروتين:

http://www.rcsb.org

الصفحة الرئيسية لقاعدة بيانات EBI التركيب الجزيئى الماكرو:

http://msd.ebi.ac.uk/

الصفحة الرئيسية ل BioMagResBank:

http://www.bmrb.wisc.edu/

للبحث فى بنك بيانات البروتين:

الصفحة الرئيسية ل SCOP (Structural Classification of Protein):

http://scop.mrc-lmb.cam.ac.uk/scop/

قائمة بأماكن التصفح:

http://pdp-browsers.ebi.ac.uk/browse_it.shtml

أداة البحث OCA:

http://oca.ebi.ac.uk/oca-bin/ocamain

قاعدة بيانات التركيب الرباعى للبروتين:

http://pqs.ebi.ac.uk/

تقارير جودة التركيب:

http://www.cmbi.kun.nl/gv/pdbreport

مصادر على الانترنت Web Resources لقواعد بيانات لعائلات معينـــة من بروتين:

Protein kinase :

http://www.sdsc.edu/kinases/

HIV proteases

http://www-fbsc.ncifcrf.gov/HIVdb/

Icosahedral viruses :

http://mmtsb.scripps.edu/viper/main.html

قاعدة بيانات علم المناعة Immunology :

IGMT : (International ImMunoGene Tics database):

http://imgt.cines.fr

وهى قاعدة بيانات متكاملة عالية الجودة، ومتخصصة فى:

Immunoglobulins (Ig)، و T-Cell receptors (TcR) ، و

Histocompatibility Complex (MHC) molecules of all vertebrate species.

KABAT : http://immuno.bme.nwu.edu./

وهى قاعدة بيانات لتتابعات البروتينات ذات الأهمية المناعية.

MHCPEP : http://wehih.wehi.edu.au/mhcpep/

وهى قاعدة بيانات لمعقد البيبتيدات الرئيسية المرتبطة والمتوافقة هيستولوجيا.

مجموعات من روابط الاتصال لقواعد بيانات لعائلات معينة من البروتين:

http://www2.ebi.ac.uk/msd/Links?family.shtml

قواعد بيانات ناتج تعبير الجين والبروتيومكس Expression and proteomics databases:

من المعرف أن الدنا يصنع الرنا ويقوم الرنا بصنع البروتين. وتحتوى قواعد بيانات الجينوم على تتابعات الدنا. كما تسجل قواعد بيانات التعبير الجينى مقاييس مستويات الرنا الرسول mRNA وذلك عبر (short terminal sequences of cDNA synthesized from mRNA) ESTs مع وصف طرز نسخ الجين. بينما تسجل قواعد بيانات البروتيومات proteomics مقاييس للبروتينات مع وصف طرز ترجمة الجين.

تقدم مقارنات طرز التعبير الجينى معلومات موثقة عن:

(١) وظيفة وآلية فعل نواتج الجين.

(٢) كيف تنسق الكائنات عملية التحكم فى العمليات الأيضية تحت ظروف مختلفة. على سبيل المثال: الخميرة تحت ظروف هوائية ولاهوائية.

(٣) الاختلافات فى تعبئة الجين أثناء مراحل مختلفة من دورة الخلية وتطور الكائن.

(٤) آليات المقاومة للمضادات الحيوية فى البكتريا وبالتالى اقتراح أماكن مستهدفة لاستحداث أدوية.

(٥) الاستجابة لتحدى طفيل ما.

(٦) الاستجابة لأنواع وجرعات مختلفة من الأدوية للوصول الى علاج فعال.

وهناك عدة قواعد بيانات ل ESTs، تحتوى مدخلات معظمها على مجالات تشير الى موقع نسيج و/ أو مكون تحت خلوى، حالة التطور، ظروف النمو، والمستوى الكمى للتعبير. حاليا تحتوى dbEST من خلال بنك الجين على ما يقرب من تسعة ملايين مدخل من ٣٤٨ نوع.

بعض مجموعات EST متخصصة فى أنسجة معينة (عضلات، أسنان) أو أنواع. كما توجد جهود لربط طرز التعبير بمعلومات أخرى حول الكائن. فعلى سبيل المثال ينسق مشروع تطور الجرذان بين بيانات تعبير الجين والتشريح التطورى.

تتيح العديد من قواعد البيانات وسائل للربط بين ESTs فى أنواع مختلفة مثل ربط التماثل فى الانسان والجرذان، أو العلاقات بين جينات أمراض الانسان وبروتينات الخميرة. وهناك مجموعات أخرى متخصصة فى نوع من البروتين مثل سيتوكينات. كما يوجد جهدا هائلا للتركيز على السرطان من حيث تكامل المعلومات عن الطفرات واعادة التنظيم للكروموسومات والتغيرات فى طرز التعبير وذلك للتعرف على التغيرات الوراثية أثناء تكوين الأورام ونموها.

بالرغم من العلاقة الشديدة بين طرز النسخ وطرز الترجمة، الا أن القياسات المباشرة للمحتوى البروتينى للخلايا والأنسجة – البروتيومكس proteomics تتيح معلومات أضافية قيمة. وبسبب عدلات النسخ المختلفة لمختلف الرنا الرسول فان قياسات البروتينات تعطى مباشرة وصفا أكثر دقة لطرز تعبير الجين بالمقارنة بقياسات النسخ. ويمكن اكتشاف تحورات ما بعد الترجمة فقط عن طريق فحص البروتينات.

يتضمن تحليل البروتيوم الفصل والتعريف والتقدير الكمى لبروتينات العينة. ويعتمد ذلك على فرد البروتين بواسطة الجيل ثنائى الأبعاد

وتعريف كل مكون بواسطة مطياف الكتلة. وتقوم قواعد بيانات البروتيوم بتخزين صور الجيل وتفسيراتها على هيئة طرز بروتين. كما تظهر بعض قواعد البيانات صورا وتسمح باختيار تفاعلى للنقط. وباختيار نقطة معينة تفتح نافذة لمدخلات البيانات المتطابقة. ويوجد مدخل بيانات لكل بروتين يسجل المعلومات التالية:

- تعريف البروتين.
- الكمية.
- الوظيفة.
- آلية الفعل.
- طراز التعبير.
- مكان التواجد تحت الخلوى.
- البروتينات ذات الصلة.
- تحورات ما بعد الترجمة.
- التفاعلات مع البروتينات الأخرى.
- روابط اتصال بقواعد بيانات أخرى.

وقد ساهمت البيومعلوماتية فى انشاء وتطوير قواعد البيانات هذه، وكذلك ايجاد نظم خوارزمية لمقارنة وتحليل طرز البروتينات المحتوية عليها تلك القواعد.

قواعد بيانات المسارات الأيضية Databases of metabolic pathways:

تقوم قاعدة KEGG (Kyoto Encyclopedia of Gens and Genomes) بجمع الجينوم، ونواتج الجين ووظائفها، وعمل تكامل بين المعلومات البيوكيميائية والوراثية. وتركز KEGG على تفاعلات شبكات التجميع الجزيئى والشبكات الأيضية والمنظمة.

١) وتنظم KEGG خمسة أنواع من البيانات من خلال نظام شامل وهى:

٢) كتالوج للجينات يحتوى على جزيئات وتتابعات معينة.

٣) خرئط الجينوم. حيث يتم التكامل بين الجينات طبقا لظـــهورهم علـــى الكروموسومات.

٤) خرائط المسارات والتى تصف شبكات الأنشـــط الأيضيـــة والمنظمـــة للجزيئات. وتساهم فى امكانية ايجاد مسار ايضى حقيقى فـــى كـــائن معين عن طريق مضاهاة بروتينات هذا الكائن مع انزيمات مســـارات مرجعية.

٥) جداول أرثولوج. تستخدم بيانات تلك الجداول لربط انزيم فىكائن مـــا بالانزيمات ذات الصلة فى كائنات أخرى. ويتيح ذلك تحليل العلاقـــات بين المسارات الأيضية فى كائنات مختلفة.

تتيح KEGG امكانية أخذ مجموعة من الانزيمات من كائن واختبار مـــدى تكاملها مع مسارات أيضية معروفة. وفى حالة ظهور فراغ فى مســـارما فان ذلك قد يرجع الى غياب انزيم أو مسار أيضى غير متوقع.

مداخل الى السجلات :

تمتلك قواعد بيانات الحمض النووى وتتابعات البروتين امكانـــات هائلـــة لاسترجاع المعلومات وتحليلها من خلال عديـــد مـــن العمليـــات والتـــى تتضمن:

١. استرجاع التتابعات من قاعدة البيانات.

٢. مقارنة التتابع.

٣. ترجمة تتابعات الدنا الى تتابعات بروتين.

٤. تحليل التركيب والتنبؤ.

٥. طرز التعرف.

٦. الأشكال الجزيئية. وتستخدم تلك الأشكال فى:

- خرطنة الأجزاء التى يعتقد أن لها وظيفة ما الى اطار ثلاثى الأبعاد لبروتين.
- تقسيم ومقارنة طرز الطى فى البروتينات.
- تحليل الاختلافات بين التراكيب شديدة القرابة أو بين الأوضاع الفراغية لجزئ ما.
- دراسة تفاعل جزئ بسيط مع بروتين كمحاولة لتحديد وظيفة أو لاكتشاف دواء.
- الحصول على نموذج تفاعلى لتحسين جودة صور الأشعة السينية لمقاييس تراكيب البروتين.
- تصميم ونمذجة تراكيب جديدة.

طرق الدخول إلى قواعد بيانات البيولوجيا الجزيئية:

ENTREZ: هى نقطة البداية لاسترجاع تتابعات وتراكيب من على الموقع التالى:

http://www.ncbi.n1m.nih.gov/ENTREZ/

وتقدم ENTREZ طرق دخول عبر الأقسام الآتية لقواعد البيانات:

- Protein
- Peptide
- Nucleotide
- Structure
- Genome
- Popset- information about populations
- OMIM- Online Mendelian Inheritance in Man

فمثلا للبحث فى قاعدة البروتين: ادخل على موقع ENTREZ، ثم اختر بروتين، ثم ادخل نوع البروتين المطلوب البحث عنه، ثم اضغط على GO وسيقوم البرنامج باظهار الاجابات.

وللبحث فى قاعدة بيانات النيكليوتيدات: اختر نيكليوتيد من على موقع ENTREZ واضغط على INDEX ثم اختر الكائن من القائمة وادخل اسم الموضوع فى صندوق البحث، ثم اضغط على AND ثم اتبع التعليمات من القائمة.

:OMIM

وهى قاعدة بيانات للجينات البشرية والعيوب الوراثية. وتعمل بتكامل تام مع ENTREZ.

نظام استرجاع التتابع The Sequence Retrieval System (SRS):

يستطيع نظام SRS البحث فى ١٤١ من قواعد بيانات البروتين، وتتابعلت النيكليوتيد، والمسارات الأيضية، والتراكيب ثلاثية الأبعاد، والوظائف، والجينوم، والأمراض.

مصدر تعريف البروتين The Protein Identification Resource (PIR):

يمكن البحث باستخدام SRS من على المواقع التالية :

فى الولايات المتحدة الأمريكية :

http://www.-nbrf.georgetown.edu/pirwww/search/textpsd.html

فى أوروبا: http://www.mips.gsf.de

: ExPASY-Expert Protein Analysis System

وهو نظام استرجاع وتحليل المعلومات لمعهد البيومعلوماتية السويسرى بالتعاون مع معهد البيومعلوماتية الأوروبى والذى ينتج قواعد بيانات تتابع البروتين SWISS-PROT وقاعدة TrEMBL التى تحتوى على تراجم تتابعات النيكلوتيد. وفتح الصفحة الرئيسية ل ExPASY

(http://www.expasy.ch) واختيار SWISS-PROT و TrEMBL يعطـــى مداخل توصل الى أدوات استرجاع المعلومات.

:ENSEMBL

وهـــو عبـــارة عـــن مصـــدر معلومـــات شـــامل للجينـــوم البشـــــرى (http://www.ensemble.org). وتهدف الى جمع وتفسير كل المعلومـــات المتاحة عن تتابعات الدنا البشرى، وروابط اتصال الـــى تتـــابع جينـــوم قياسى، واتاحة هذه المعلومات لعديد من العلماء. ويمكن للمستخدم تعريف مناطق فى التتابع عبر عدة أدوات بحث:

- BLAST ويستخدم للبحث عن تتابع أو جزء منه.
- التصفح، ويبدأ على مستوى الكروموسوم ثم يتعمق داخله.
- العلاقة بالأمراض بواسطة OMIM.
- ENSEMBL ID.
- أداة بحث عامة للنصوص.

الفصل الرابع

المحازاة والأشجار العرقية

محازاة التتابع:

فى حالة وجود اثنان من التتابعات يتم عمل الآتى:

- قياس درجة التشابه بينهما.
- تحديد تقابل للأجزاء.
- ملاحظة طرز التحفظ والاختلاف.
- علاقات النشوء.

The dotplot:

وهى صورة بسيطة تعطى رؤية عامة للتشابه بين اثنان من التتابعات. يتكون dotplot من جدول أو مصفوفة. تشير الصفوف الى أجزاء أحد التتابعات بينما تدل الأعمدة على التتابع الآخر. وتظهر بقع فى حالة التماثل، بينما لاتوجد استجابة فى حالة الاختلاف.

مصادر DOTPLOTS على الانترنت (web Resources):

يتيح برنامج E.L. Sonnhammer's Dotter للمستخدم التحكم وطرق الحساب وتغيير مظهر العرض بواسطة ضبط معايير تفاعلية والذى يجب تجهيزه فى الكمبيوتر من على الموقع التالى:

http://www.cgr.ki.se/cgr/groups/sonnhammer/Dotter.html

يقدم الموقع التالى امكانية اجراء dotplotting تفاعلى:

http://www.isrec.isb-sib.ch/java/dotlet/exonintron.html

A PERL program to draw dotplots.

```
#!/usr/bin/perl
#dotplot.pl -- reads two sequences and prints dotplot
# read input
$/ = "";
$_ = <DATA> $_ =~ s/#(.*)\n/\n/g;
$_              =~          /^(.*)\n\s*(\d+)\s+(\d+)\s*\n(.*)\n([A-
Z\n]*)\*\s*\n(.*)\n([A-Z\n]*)\*/;
$title = $1; $nwind = $2; $thresh = $3;
$seqt1 = $4; $seq1 = $5; $seqt2 = $6; $seq2 = $7;
$seq1  =~  s/\n//g;  $seq2  =~  s/\n//g;  $n   = length($seq1); $m =
length($seq2);
# postscript header
print  <<   EOF setfont scalefont 20 findfont Helvetica translate 30
setlinewidth  1.75   def load fill f closepath c newpath n rlineto r
moveto m lineto l stroke s %!PS-Adobe->

#print matrix

$dx = 500.0/$n;  $mdx = -$dx; $dy = 500.0/$m;
if  ($dy  <  $dx)  ($dx  =  $dy;}   $dy = $dx; $xmx = $n*$dx; $ymx =
$m*$dx;
print "0 510 m ($title NWIND = $nwind) show\n";
printf  "0  0  m  0  %9.2f  l  %9.2f  %9.2f  l  %9.2f  0   l c s\n",
$ymx,$xmx,$ymx,$xmx;
for ($k = $nwind - $m + 1; $k < $n - $nwind; $k++) {
    $i = $k; $j = 1; if ($k < 1) {$i = 1; $j = 2 - $k;}
    while ($i <= $n - $nwind && $j <= $m - $nwind) {
        $_    =  (substr($seq1,$i  -1,$nwind)  ^  substr($seq2,$j  -
1,$nwind));
        $mismatch = ($_ =~ s/[^\x0]//g);
        if ($mismatch < $thresh) {
            $xl = ($i - 1)*$dx; $yb = ($m - $j)*$dy;
            printf  "n %9.2f %9.2f m %9.2f 0 r 0 %9.2f r %9.2f 0 r c
f\n",
                    $xl,$yb,$dx,$dy,$mdx;
        }
        $i++; $j++;
    }
}
print "showpage\n";

__END__
ATPases lamprey / dogfish                    #TITLE
15 6                                         #WINDOW, THRESHOLD
Petromyzon marinus mitochondrion             #SEQUENCE 1
ATGACACTAGATATCTTTGACCAATTTACCTCCCCAACA
ATATTTGGGCTTCCACTAGCCTGATTAGCTATACTAGCCCCTAGCTTA
ATATTAGTTTCACAAACACCAAAATTTATCAAATCTCGTTATCACACACTA
CTTACACCCATCTTAACATCTATTGCCAAACAACTCTTTCTTCCAATAAAC
CAACAAGGGCATAAATGAGCCTTAATTTGTATAGCCTCTATAATATTTATC
TTAATAATTAATCTTTTAGGATTATTACCATATACTTATACACCAACTACC
CAATTATCAATAAACATAGGATTAGCAGTGCCACTATGACTAGCTACTGTC
CTCATTGGGTTACAAAAAAAACCAACAGAAGCCCTAGCCCACTTATTACCA
GAAGGTACCCCAGCAGCACTCATTCCCATATTAATTATCATTGAAACTATT
AGTCTTTTTATCCGACCTATCGCCCTAGGAGTCCGACTAACCGCTAATTTA
ACAGCTGGTCACTTACTTATACAACTAGTTTCTATAACAACCTTTGTAATA
ATTCCTGTCATTTCAATTTCAATTATTACCTCACTACTTCTTCTATTA
CTAACAATTCTGGAGTTAGCTGTTGCTGTAATCCAGGCATATGTATTTATT
CTACTTTTAACTCTTTATCTGCAAGAAAACGTTT*
Scyliorhinus canicula mitochondrion          #SEQUENCE 2
ATGATTATAAGCTTTTTTGATCAATTCCTAAGTCCCTCCTTTCTAGGA
ATCCCACTAATTGCCCTAGCTATTTCAATTCCATGATTAATATTTCCAACACCAAC
AATCGTTGACTTAATAATCGATTATTAACTCTTCAAGCATGATTTATTAACCGATT
TATCAACTAATACAACCCATAAATTTAGGAGGACATAAATGAGCTATCTTATTTACA
CTAATATTATTTTTAATTACCATCAATCTTCTAGGTCTCCTTCCATATACTTTTACG
ACAACTCAACTTTCTCTTAATATAGCCTTTGCCCTGCCCTTATGGCTTACAACTGTA
```

```
ATTGGTATATTTAATCAACCAACCATTGCCCTAGGGCACTTATTACCTGAAGGTACCCCA
ACCCCTTTAGTACCAGTACTAATCATTATCGAAACCATCAGTTTATTTATTCGACCATTA
GCCTTAGGAGTCCGATTAACAGCCAACTTAACAGCTGGACATCTCCTTATACAATTAATC
GCAACTGCGGCCTTTGTCCTTTTAACTATAATACCAACCGTGGCCTTACTAACCTCCCTA
GTCCTGTTCCTATTGACTATTTTAGAAGTGGCTGTAGCTATAATTCAAGCATACGTATTT
GTCCTTCTTTTAAGCTTATATCTACAAGAAAACGTATAA*
```

محازاة التتابع بواسطة Dotplots:

لا يقدم **Dotplots** فقط صورة شاملة لتشابه اثنان من التتابعات ولكن أيضا يعطى مجموعة متكاملة ذات جودة نسبية للمحازاة المحتملة وذلك بالمرور على الصورة من الجانب الأيسر العلوى الى الجانب الأيمن السفلى.

محازاة التتابع المتعدد Multiple sequence alignment :

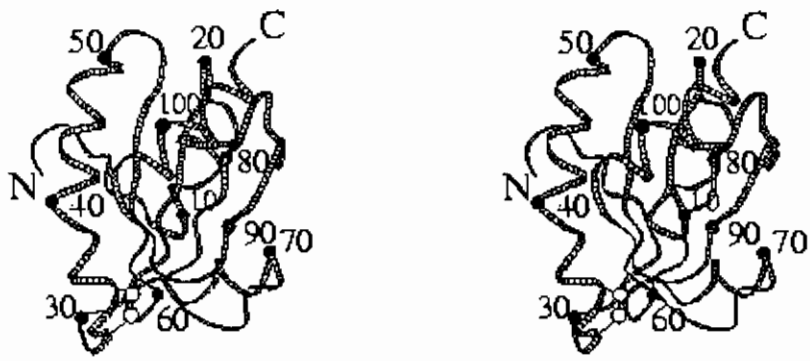
تكون نتائج الكشف عن العلاقات البعيدة بين التتابعات أكثر واقعية فى حالة تعدد التتابعات. وأيضا تعطى أدوات التنبؤ بالتركيب نتائج أفضل عندما تعتمد على محازاة تتابعات متعددة وذلك بالمقارنة بالتتابعات المنفردة. ولذلك يجب أن تحتوى محازاة التتابع المتعدد على توزيع من التتابعات قريبة وبعيدة الصلة.

الاستدلال التركيبى بواسطة محازاة التتابع المتعدد:

يمكن التعرف على الملامح التركيبية والوظيفية بواسطة المحازاة للتتابعات المتعددة من خلال موقع الخادم التالى:

http://www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi

ومثال ذلك التعرف على الملامح التركيبية لعدد ١٦ ثيوريدوكسينات (انزيمات موجودة فى كل الخلايا وتساهم فى مدى واسع من العمليات البيولوجية):



The structure of *E. coli* thioredoxin [2TRX] contains a central five-stranded β-sheet flanked on either side by α-helices. Residue numbers correspond to those in the multiple sequence alignment table. The N- and C-termini are also marked. Spheres indicate positions of the Cα atoms of every tenth residue. The reactive disulphide bridge between Cys32 and Cys35 appears between the numberals 30 and 60 .

Hiden Markov Models (HMMS) :

وهى نماذج تركيب حسابية لوصف طرز عائلات التتابعات المتماثلة. كما أنها تعتبر من الأدوات القوية للكشف عن القرابة البعيدة والتنبؤ بطرز طى البروتين.

مصادر على الانترنت Web Resources :

http://cse.ucsc.edu/research/compbio/sam.html

http://cse.ucsc.edu/research/compbio/HMM-apps/HMM-applications.html

http://www.sanger.ac.uk/Software/Pfam.

http://stash.mre-lmb.cam.ac.uk/SUPERFAMILY/

النشوء النوعى Phylogeny :

هناك العديد من الأمثلة للتطور فى البروتينات والجينوم. ويهدف هذا المجال العمل على التوصل الى العلاقات بين الأنواع والعشائر والأفراد والجينات وذلك باستخدام الخصائص الجزيئية.

أشجار القرابة الوراثية:

وتمثل الشجرة فى علم الحاسبات شكلا خاصا يتكون من نقط nodes يتـــم توصيلها ببعض عن طريق خطوط edges. والشجرة قد تكـون rooted أو unrooted. والشكل الأخير لا يظهر النسب بينما فى الشكل الأول ينحـــدر من كل نقطة أصلان ويسمى binary tree ويستخدم برنامج PERL لرســـم الشجرة كما يلى:

```
 drawtree.prl - PERL

program to draw binary trees.

#!/usr/bin/perl
#drawtree.pl -- draws binary trees (root at top)
#usage:  echo '(A((BC)D)(EF))' | drawtree.pl > output.ps

print <; chop($tree); $_ = reverse($tree); s/[()]//g;


$x = 0; $y = 0;
while ($nd = chop()) {
    print "$x $y m ($nd) stringwidth pop -0.5 mul 0 rm ($nd)
show\n";
    $xx{$nd} = $x; $x+=20; $yy{$nd} = 10;
}


while ($tree =~ s/(?([A-Z])([A-Z]))?/1/) {
    print "n $xx{$1} $yy{$1} m\n";
    ($yy{$1} > $yy{$2}) || ($yy{$1} = $yy{$2}); $yy{$1} += 20;
    print "$xx{$1} $yy{$1} l $xx{$2} $yy{$1} l $xx{$2} $yy{$2} l
s\n";
    $xx{$1} = 0.5*($xx{$1} + $xx{$2});
}print "n $xx{$tree} $yy{$tree} m 0 20 rl s showpage\n";


$rx = 2*$x + 30; $yt = 2*$yy{$tree} + 146;
print "%%BoundingBox: 40 95 $rx $yt\n";
```

مصادرعلى الانترنت لأدوات الوراثة العرقية :

http://evolution.genetics.washington.edu/phylip/software.html

الفصل الخامس

تركيب البروتين واكتشاف الدواء

يتضمن هذا الفصل من الكتاب تركيب البروتينات من حيث:

ثبات وطى البروتين وتطبيقات

Hydrophobicity:

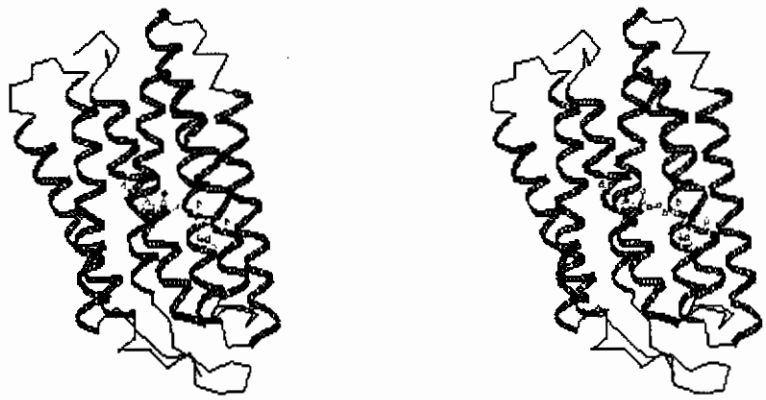
يستخدم تحليل hydrophobicity profile للتنبؤ بأماكن التفاف العناصر في التركيب الثانوى، الوحدات الظاهرة والمختفية، امتداد أجزاء الغشاء، ومواقع الأنتيجين. وفيمايلى مثال لاستخدام برنامج PERL لرسم مايسمى ب The helical wheel التى تعتمد على دراسة تتابع الوحدات من حيث التبادل مابين محب وكاره للماء وذلك فى alf-helics in globular protiens:

helwheel.pl - program to draw helical wheels.

```
#!/usr/bin/perl
#helwheel.pl -- draw helical wheel
#usage: echo DVAGHGQDILIRLFKSH | helwheel.prl > output.ps
# or    echo 20DVAGHGQDILIRLFKSH | helwheel.prl > output.ps
#       the numerical prefix sets the first residue number

# The output of this program is in PostScript (TM),
#      a general-purpose graphical language


# The next section prints a header for the PostScript file

print <;                                    #  read line of input
chop();$_  =~  s/\s//g;                        #  remove terminal
carriage return and blanks

if  ($_ =~ s/^(\d+)//)                      #  if input begins with
integer
     {$resno  =  $1;}                          #  extract it as
initial residue number
else   {$resno = 1}                          #  if not, set initial
residue number = 1

$radius  =  50;                               #  initialize values
for radius,
$x = 0; $y = -50; $theta = -90;             #  x, y and angle theta
```

```
#     print light gray spiral arc as succession of line segments, 10
per residue

$npoints = 10*(length($_) - 1);
print   "0.8 0.8 0.8 setrgbcolor\n";          #  set colour to light
gray
print "newpath\n";                           #  draw spiral arc
printf("%8.3f %8.3f moveto\n",$x,$y);
foreach  $d  (1  ..  $npoints)  {                    #  10 points per
residue
    $theta   += 10; $radius += 0.6;           #  increase radius and
theta
    $x   = $radius*cos($theta*0.01747737);    #  calculate new value
of x
    $y = $radius*sin($theta*0.01747737);    #    and y
    printf("%8.3f %8.3f lineto\n",$x,$y);
}
print "stroke\n";


#  print residues and residue numbers


$radius   = 50;                               #  reinitialize values
for radius,
$x = 0; $y = -50; $theta = -90;              #  x, y and angle theta
print ``0 setgray\n''                        #  set colour to black


foreach  (split ("",$_)) {                    #  loop over characters
from input line
    print "/$font{$_} findfont";             #  set font appropriate
    print "20 scalefont setfont\n";          #  for this amino acid
    printf("%8.3f  %8.3f  moveto\n",$x,$y);       #  move to current
point
    print  "   ($resno$_) stringwidth";       #  adjust position to
center residue
    print  " pop -0.5 mul -7 rmoveto\n";     #     identification on
point on spiral
    print  " ($resno$_) show\n";              #  print residue number
and id
    print "% $theta $resno$_\n";
    $theta  +=  100;   $radius += 6;            #  set new values of
angle, radius
    $x  =   $radius*cos($theta*0.01747737);    #  compute new values
of x
    $y = $radius*sin($theta*0.01747737);    #     and y
    $resno++;                                  #  increase residue
number
}


print  "showpage\n";                           #  postscript signals
to
print "%%BoundingBox:";                      #  print
$xl = 297.5 - 1.05*$radius;                  #   x
$xr = 297.5 + 1.05*$radius;                  #    and
$yb = 421.  - 1.05*$radius;                  #     y
$yt = 421.  + 1.05*$radius;                  #      limits
printf("%8.3f %8.3f %8.3f %8.3f\n",$xl,$xr,$yb,$yt);


print "showpage\n";
print "%%EOF\n";                             #  and wind up
```

المثال الثانى: التنبؤ بالتركيب اللولبى لبروتينات الأغشية، حيث أن العديد من بروتينات الأغشية تتركب من سبع أشكال لولبية متصلة مع بعضها بواسطة عقد كما هو الحال فى bacteriorhodopsin :



Bacteriorhodopsin membrane from the bacterium *Halobacterium salinarum* . The ligand shown in ball-and-stick representation is the chromophore, retinal .

Web Resources : Transmembrane Helix Prediction :

TMHMM (A.Krogh and E. Sonnammer
Markov Model :
http://www.cbs.dtu.dk/krogh/TMHMM/

PHDhtm (B.Rost) :
http://dodo.bioc.columbia.edu/predictprotein

Membrane protein explorer (S. White) :
http://błanco.biomol.uci.edu/mpex/

Web Resources :

matrix Alignment) :
http://www2.ebi.ac.uk/dali/

Classifications of Protein Structure :
http://www2.ebi.ac.uk/msd/Links/fold.shtml.

http://www.bioscience.org/urllists/protdb.htm
http://scop.mrc-lmb.cam.ac.uk/scop

Protein Structure Prediction and modelling :
http://cubic.bioc.columbia.edu/eva

Homology Modelling :

SWISS-MODEL (automatic homology modelling) :
http://www.expasy.ch/swissmod/SWISS-MODEL.html

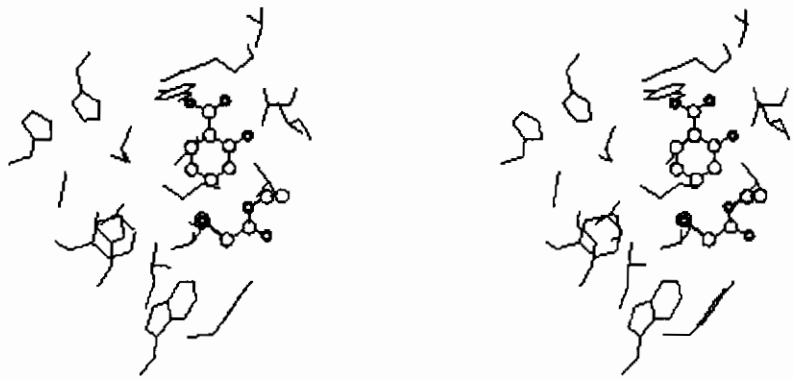
MODBASE, a database of comparative models of protein from complete genomes:
http://guitar.rockefeller.edu/modbase/

اكتشاف وتطوير الدواء:

تم استعراض المواصفات التى يجب توافرها فى المركب الكيميائى المؤهلة لاستخدامه كدواء. وكذلك مراحل انتاج دواء جديد. واستخدام الكمبيوتر فى حصر مركبات عديدة للوصول الى مركب واحد مؤهل، ثم اجراء تحسين للمركب المرشح من خلال تحديد العلاقة الكمية بين التركيب والفعالية QSAR والتى تتيح طرق للتنبؤ بالنشاط الدوائى لمجموعة من المركبات بواسطة العلاقة بين الملامح الجزيئية والنشاط الدوائى. وكذلك الاستعانة بمعلومات عن الجينوم.

تصميم الدواء بمساعدة الكمبيوتر:

مثال: تصميم مثبطات prostaglandin cyclo-oxygnase 2. ومن المعروف أن مركب الأسبرين ومركبات أخرى من NSAID تثبط اثنان من prostaglandin cyclo-oxygnase قريبى الصلة تسمى COX-1 و COX-2 :



The binding site in COX-1 for an aspirin analog, 2- bromoacetoxybenzoic acid . The ligand has reacted with the protein, transferring the bromoacetyl group to the sidechain of serine 530 . The protein is shown in skeletalrepresentation . The asprin analog is shown in ball-and-stick representation .

وللمزيد من المعلومات يمكن الرجوع الى الموقع الالكترونى للكتاب:

http://www.oup.com/uk/lesk/bioinf

الخاتمة

استقراء الحالة الراهنة لتطبيقات البيومعلوماتية تمكننا من استشراف أهمية مجال البيومعلوماتية فى المستقبل. فمن الواضح أن عمليات جمع البيانات ستستمر فى الزيادة وبسرعة. كذلك التوجه نحو زيادة كفائة وقدرة أجهزة الكمبيوتر من حيث التخزين، والتوزيع، وتحليل النتائج. وسوف يؤدى ابتكار طرق خوارزمية محسنة الى تحليل وتفسير المعلومات المتاحة لنا وتحويلها من مجرد صور بيانات الى معرفة ثم حكمة.

وسوف نصل الى الكتلة الحرجة عندما تصبح معرفتنا بالتتابعات والتراكيب أكثر قربا من الاكتمال، بمعنى جمع مجموعات غير منحازة من البيانات المتاحة عن أشكال الحياة المعاصرة. وسوف نتعرف على ذلك عندما يؤدى التعمق العشوائى فى جينوم ما أو معرفة تركيب بروتين ما الى تغيير شيئا موجودا بالفعل بدلا من الكشف عن شيئا جديدا. وبعد كل ذلك، تتكون الطبيعة من نظام غير محدود الاحتمالات ولاكن مع اختيارات لانهائية.

ستصبح التطبيقات أكثر وضوحا، ونضجا وستمر بسرعة من مراحل البحث الى الممارسات الصناعية والاكلينيكية القياسية. ونقل بعض المعلومات البيولوجية عالية المستوى مثل برامج التطور أثناء فترة حياة الأفراد، وأنشطة العقل البشرى ودخولها فى عمليات معالجة البيانات، سوف يمكننا من الوصف الكمى والتحليل على مستوى الجزيئات وتفاعلها.

رقم الإيداع: ٢٠٠٦/٢١٩١
ISBN: 977-281-297-5

مطابع الدار الهندسية/القاهرة
تليفون/فاكس : (٢٠٢) ٥٤٠٢٥٩٨