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 ١٦X ١٠*٩ من القواعد (١٦ Gbp). لو استخدم الحجم التقريبى للجينوم البشرى (٣,٢ *١٠ ٩ حرفا) كوحدة فان ذلك يعادل (٢ buges, an apt name). وللمقارنة المرجعية فان ١ buge 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', 'losteric 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 شبكات الحاسب computer networks وكذلك بالنسبة للشبكة العنكبوتية العالمية.www.

علوم الحاسب مجال صغير ومزدهر يستهدف تعظيم الاستخدام الفعال لمكونات تكنولوجيا المعلومات. مناطق معينة من علم الحاسب تمس البيومعلوماتية مسا وثيقا ومباشرا. دعنا نتصور مشكلة بيولوجية مثل استرجاع كل التتابعات المماثلة لمجس التتابع probe sequence من قاعدة بيانات سيكون الحل الصواب هو اللجوء الى علم الحاسب فى:

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

وهو الوصف الكامل والدقيق لطريقة حل المشكلة. بالنسبة لاسترجاع 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 طريقة حتى بدون السماح بفراغات فانها تحتاج الى وقت يتناسب مع ناتج ضرب عدد حروف مجس التتابع فى عدد حروف تتابعات قاعدة البيانات. ويركز تخصص الحاسب والمعروف عامة 'stringology'على ايجاد طرقا ذات كفاءة للتعامل مع هذا النوع من المشاكل وتحليل فاعلية أدائها.

- **تراكيب البيانات واسترجاع المعلومات Data structures, and information retrieval:**

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

- **هندسة البرامج Software engineering:**

لم يعد من الصعب بتاتا كتابة برامج بلغة الحاسبات الأصلية. يعمل 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
}


$cutstring =~ s/\\n/\n/g;                    #   change signal \n to
real carriage return
print "$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 البكنرياسى 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 entry 145557 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% تطـابق كـامل فـى التتابعات المتراصة) كما أن هناك تماثل فى التركيب. وكل بروتين يمكـن 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 للتنبؤ بالتركيب Critical Assessment of Structure Prediction (CASP):

يتطلب الحكم على تقانات التنبؤ بتراكيب البروتينات الى استخدام اختبارات مصممته blind tests. لهذا الغرض صمم جاى مولت برامج CASP. العلماء المنشغلون بتقدير تركيب البروتين باستخدام طرق القياس البللورى والرنين النووى المغناطيسى مدعوون الى: (١) نشر تتابع الحمض الأمينى عدة شهور قبل التاريخ المتوقع الى اكتمال التجارب، (٢)الاحتفاظ بسرية النتائج حتى تاريخ متفق عليه. يقدم القائمون بالتنبؤ بنماذج تنبؤ يحتفظ بها حتى التاريخ المتفق عليه لاعلان النتائج التجريبية. وهنا يتم مقارنة نتائج التنبؤ والتجريب. ولقد سجلت نتائج التقييم باستخدام برامج CASP تقدما فى فاعلية التنبؤ والذى يرجع الى نمو بنوك المعلومات وأيضا بسبب التحسن فى الطرق المستخدمة. والباب الخامس من الكتاب يناقش التنبؤ بتركيب البروتين.

هندسة البروتينات Protein engineering:

كان هناك تشابه بين طبيعة عمل كلا من علماء البيولوجيا الجزيئية وعلماء الفلك من حيث القدرة على ملاحظة الأشياء دون اجراء تعديلات عليها. الا أن هذا لم يعد حقيقيا الآن. ففى المعمل يمكننا عمل تعديل فى الأحماض النووية والبروتين واحداث طفرات لمعرفة التأثير على وظائفهم. فمن الممكن أن يصبح لبروتين قديم وظيفة جديدة مثل انتاج أجسام مضادة حفازة ومحاولة تحضير بروتينات جديدة.

تشتق العديد من قواعد تركيب البروتينات من ملاحظات للبروتين الطبيعى. وليس ضروريا تطبيق تلك القواعد فى البروتينات المهندسة. تمتلك البروتينات الطبيعية صفات مرتبطة بالأسس العامة للكيمياء الطبيعية وآليات نشوء البروتين. ويجب أن تخضع البروتينات المهندسة

لقوانين الكيمياء الطبيعية وليس لقيود النشوء. وتستطيع 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 فى 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 iir, yet I am
yet I am well;
I ai n well; another
anc ther is wise, yet I am well;
yet I am well; another virtuous,
another 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.