

الفصل الأول

مقدمة

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

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

وكذلك من الخصائص الواضحة لبيانات البيومعلوماتية أنها ذات كم هائل جداً. فمثلاً يحتوى بنك البيانات لتتابع النيكلوتيد على 16×10^9 من القواعد (16 Gbp). لو استخدم الحجم التقريبى للجينوم البشرى (3,2 * 10⁹ حرفاً) كوحدة فإن ذلك يعادل (2 bugs, an apt name). وللمقارنة المرجعية فإن 1 bug يقابل عدد الحروف التى ظهرت فى اعداد نيويورك تايمز لمدة 6 سنوات. تحتوى قاعدة البيانات لتراكيب الجزيئات الكبيرة Macro molecules (الأبعاد الثلاثية للبروتينات بمتوسط طول 400 حمض أمينى) على 16000 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.

- إيجاد وربط علاقة التتابع والتركييب 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 ووصولاً 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 مرضاً وبائياً للإنسان أو الحيوان. فى هذه الحالة سوف يقوم العلماء بعزل المادة الوراثية معملياً 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 حرفاً 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 دوراً حاسماً فى التكرار والمضاهاة (١١).

ومن خلال تتابع الحمض الأمينى سوف تقوم البرامج بالحساب الآلى لتراكيب تلك البروتينات على أساس أن تتابع الأحماض الأمينية يحدد التركيب ثلاثى الأبعاد لها وبالتالي 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 وبالتالي 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 برمجة الحاسب الآلى The art of computer

programming حيث تدل الأرقام أقل من ٣٠ على مشاكل ذات حلول موجودة بالفعل أما الأرقام الأعلى تدل على موضوعات قيد البحث. وأخيرا فانه يجب أن يكون من المعلوم أن الطرق التجريبية البحتة لايجاد مواد مضادة للفيروس سوف تظل ولسنوات عديدة قادمة أكثر نجاحا من التوجهات النظرية.^٦

الدوجما: مركزيا وطرفيا Dogmas: Central & Peripheral

تعتبر مادة الوراثة دنا DNA وفي بعض الفيروسات رنا RNA هي سجل المعلومات في الكائنات (طبيعة التطور والنشاط لكل فرد). وكما هو معروف فان جزيئات دنا عبارة عن سلسلة طويلة خطية تحتوى على رسالة مكونة من أربعة حروف هجائية. والرسالة طويلة حتى فى الكائنات الدقيقة تتكون من ١٠*٦ حرفا تماما. ويتضمن تركيب الدنا على آليات للنسخ الذاتى وتكوين البروتينات. يؤدى التركيب الحلزونى المزدوج والمتماثل داخليا الى نسخ دقيق. ومع أن النسخ الدقيق ضروري لثبات الصفة الوراثية ألا أن بعض عمليات النسخ غير الدقيق أو آليات نقل مادة وراثية غريبة تكون أحيانا ضرورية لحدوث النشوء فى الكائنات اللاجنسية.

شرائط الحلزون المزدوج تكون فى وضع متوازيا وعكسيا ويعرف بالاتجاه ٣-٥ وذلك بالنسبة لأماكن حلقة الداى أوكسى ريبوز وعند ترجمتها الى بروتين يقرأ تتابع الحمض النووى فى الاتجاه ٥-٣. ويتم تنفيذ المعلومة الوراثية من خلال تخليق الرنا والبروتينات. والبروتينات هى الجزيئات المسؤولة عن غالبية التراكيب والأنشطة فى الكائنات. فالشعر والعضلات والانزيمات والمستقبلات والأجسام المضادة جميعها بروتينات وكل من الأحماض النووية والبروتينات جزيئات تتكون من سلسلة طويلة خطية. تتكون الشفرة الوراثية من ثلاثة حروف متتالية من تتابع الدنا تحدد أحماض أمينية متتالية وبامتداد تتابعات الدنا يشفر تتابعات

الأحماض الأمينية فى البروتينات. كذلك تتكون البروتينات من ٢٠٠ - ٤٠٠ حمض - أميني تتطلب من ٦٠٠ - ١٢٠٠ حرفا من رسائل الدنا لتحديدتها. تخليق جزيئات الرنا أيضا يتحكم فيه تتابعات الدنا بالرغم من أنه فى معظم الكائنات ليس كل الدنا يترجم الى بروتينات ورنـا. 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 يؤدي 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 الأميني فى البروتين التركيب ثلاثى الأبعاد له. يوجد لكل تتابع طبيعى للحمض الأميني حالة فطرية ثابتة مميزة والمتأقلمة تلقائيا تحت الظروف المناسبة. عندما يتم تسخين بروتين منقى أو يتعرض لظروف مغايرة للبيئة الفسيولوجية الطبيعية ينتج تركيب غير حلزوني 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 للجالة 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 أميني من السهل أن توصف منطقيا بواسطة الكود الوراثةى 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 - إلا أن عملية طي البروتين تحدث تلقائياً.

تعتمد وظائف البروتينات على التركيب الثلاثي الأبعاد الفطري. على سبيل المثال يحتوى تركيب أى انزيم على تجويف على سطح الجزيء يقوم بالارتباط بجزيء صغير ليجاور أحماض أمينية حفازة. لذلك نذكر النموذج التالي:

- 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 تعميماً 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 تعقيداً هى برامج التطور خلال فترة حياة الكائن. تلك المشكلات الساحرة المتعلقة بتدفق المعلومات فى الكائن والتحكم فيها أصبحت الآن تأتى من خلال مجال البيومعلوماتية.

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

**المشاهدات وسجلات
البيانات
Observables and
Data Archives**

- بنك بيانات سجلية للمعلومات البيولوجية:
 - تتابعات الدنا والبروتين متضمنة الشرح والتفسير.
 - تراكيب الحمض النووي والبروتين وشرحها. بنك بيانات لطرز تعبير البروتين.
 - بنوك بيانات فرعية: تحتوى على المعلومات المجمعة من بنوك البيانات السجلية والناجمة 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 التفاضل 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 الموجود 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 بينك للبيانات هى مجموعة من الأدوات للإجابة على الأسئلة الآتية:

- هل يحتوى بنك البيانات على المعلومات المطلوبة ؟ (مثال: فى أى من بنوك البيانات يمكن الحصول على تتابعات الحمض الأمينى للكحول ديهيدروجينيز؟)
- كيف يمكن جمع معلومات منتقاة من بنك البيانات فى صورة مفيدة؟ (مثال: كيف يصنف قائمة لتتابعات الجلوبيين أو جدول لمصفوفة تتابعات الجلوبيين؟)
- فهارس بنوك البيانات تفيد عند التساؤل: أين يمكن أن يوجد بعض المعلومات المحددة ؟ (مثال: ما هى بنوك البيانات التى تحتوى تتابع الحمض النووى لترسين 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 بسلسلة مع برامج استدعاء المعلومات.

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

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

(٢) تركيب بروتين مفترض، أو جزء منه ، وإيجاد تراكيب مماثلة فى قاعدة البيانات. ذلك هو التعميم لسلسلة المشاكل المتشابهة للأبعاد الثلاثية.

(٣) نتابع مفترض لبروتين غير معروف تركيبه، إيجاد تراكيب فى قاعدة البيانات والتي تتبنى تراكيب ثلاثية الأبعاد مشابهة. يحدث 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 والتي ستستطيع التعرف على بروتينات متشابهة التركيب حتى ولو كان تتابع كل منهما ينحرف عن النقطة التي تقر التشابه بينهما على أساس مقارنة التتابع.

(٤) تركيب مفترض لبروتين لإيجاد تتابعات فى بنك البيانات والتي 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 والذي 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 بينوك البيانات الى بحث التفاعل داخل بيانات البنك بمعنى - كيف يتم تبادل البيانات بين بنك وآخر بدون تضحية كبيرة فى حرية كل بنك فى بناء بياناته بطرق تلائم الصفات الفردية المميزة للمادة التى تحتوىها 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 العامة (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 تحتوىها 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 مفهومها ووظيفة فردية فصلت لتلبى احتياجات مجموعات البحث الفردية أو حتى علماء منفردين.

الرعاية والتفسير والتحكم في الجودة:

Curation Annotation, and Quality Control:

تعتمد المجتمعات العلمية والطبية على جودة بنوك البيانات. ومؤشرات الجودة حتى وإن كانت لا تسمح بتصحيح الأخطاء إلا أنها قد تتجنب التوصل إلى 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 الأميني للبروتينات التي تم تقديرها بواسطة تحديد التتابع الببتيدي يتم الآن ترجمتها جميعا من تتابعات الدنا. أحد عواقب انفجار البيانات هو أن غالبية البيانات جديدة وتم الحصول عليها بتكنولوجيا حديثة والتي في معظم الحالات تؤدي عملا جيدا.

تشتمل التفسيرات معلومات حول مصادر البيانات والطرق المستخدمة في تقديرها. كما أنها تقدم روابط مع المعلومات ذات الصلة في بنوك المعلومات الأخرى. في بنوك بيانات التتابع تتضمن التفسيرات جداول توصيف: قوائم بأجزاء التتابعات التي لها معنوية بيولوجية - على سبيل المثال مناطق من تتابع الدنا الخاصة بشفرة البروتينات. يظهر ذلك في تصميم مناسب للتعامل مع الحاسب الآلي ويكون محتوياتها مقيدة بمفردات لغوية محكمة. حتى وقت قريب ادخال تتابع نمطي 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 تجريبي للتعبير في معظم الجينات المفترضة ولا تشخيص نواتجها. يقيم الأمناء في بنوك البيانات تفسيراتهم على أساس تحليل التتابعات بواسطة برامج الحاسب الآلي.

التفسيرات هي أضعف مكون في مؤسسة الجينوم ويمكنه التفسيرات تكون ممكنة فقط بدرجة محدودة. وللحصول عليها بطريقة صحيحة تكون كثيفة

العمالة وتقسيم المصادر يكون غير كافياً. ولكن لا يمكن الاستخفاف بالتفسيرات الصحيحة. وقد علق بورك بأن تلك الأخطاء في دراسة الجينس تفسد جودة بيانات النتائج. سوف يؤدي نمو بيانات الجينوم الى تحسين جودة التفسير كما أن الطرق الاحصائية تزيد من الدقة وتسمح باعادة محسنة لتفسير المدخلات. التحسن في التفسيرات شيء جيد ولكن التلزم الحتمى - التفسير سيكون متدفقا - سيكون مزعجا. هل يجب الاطلاع دوريا على البحوث المكتملة والأخذ في الاعتبار خلاصة ما توصلت له من نتائج؟ وقد تفاقت 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 وديناميكي لمعالجة الخطأ والتفسير. سوف يقوم الأخصائيون الموزعون بدور الأمناء حيث أن العاملين بينك البيانات ليس لديهم لا الوقت ولا الخبرة للقيام بتلك المهمة. كما تسمح ديناميكية التقدم فى ميكنة تعريف وتصحيح الخطأ والتفسير باعادة التفسير لبنوك البيانات. وسوف نضطر أن نسلم بالفكرة الآمنة لبنك بيانات مستقر يتكون من مدخلات سليمة من بداية توزعها وتظل كذلك. وسوف تصبح بنوك البيانات مصدر غال واثمين للمعلومات والتى تنمو فى الحجم وتزداد نضجا ونأمل أن تكون بجودة مناسبة.

تستخدم الشبكة العنكبوتية للحصول على مواد مرجعية وأخبار وللتعامل مع قواعد البيانات فى 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.

الشبكة العنكبوتية العالمية واسعة الانتشار
The World Wide Web (www)

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 محلي. ومن الأشياء الرئيسية لبدء استخدام الشبكة بفاعلية إيجاد نقاط دخول مفيدة حيث تأخذك روابط الوصل (links) إلى المكان الذي تريده بمجرد بدء التشغيل. ومن بين أكثر المواقع أهمية تلك الخاصة ببرامج البحث والتقيب search engines والتي تفهرس الويب بأكمله وتتيح استرجاع المعلومات باستخدام كلمات رئيسية key words. ومن الممكن إدخال واحد أو أكثر من المصطلحات مثل 'phosphorylase', 'allosteric change', 'crystal structure' وسوف يظهر برنامج البحث قائمة بروابط الوصل لمواقع على الويب تحتوي على تلك المصطلحات وعندئذ يمكن التعرف على المواقع محل الاهتمام.

أثناء جلسة تصفح الويب يمكن الاحتفاظ بالمستندات التي نحتاج الرجوع إليها في جلسات تصفح قادمة وذلك بحفظ روابط الوصل الخاصة بها في ملف bookmarks أو favourites ومن ثم يمكنك في جلسات لاحقة الرجوع إلى أي موقع مباشرة دون الحاجة إلى اتباع محاولات روابط الوصل المستخدمة في أول مرة.

الويب ليس طريقاً ذو اتجاه واحد بمعنى أن العديد من مستندات الويب تتضمن نماذج يمكن إدخال معلومات فيها وإجراء برنامج للحصول على النتائج خلال نفس الجلسة. وتعتبر برامج البحث والتقيب خير مثال على ذلك. وتتطلب الآن العديد من العمليات الحسابية في البيومعلوماتية عبر أجهزة الخدمة servers. وفي 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 ينبغي أن يكون لكل مستند على الوب ملف في 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 في صيغة بروتوكول نقل النص المحوري hypertext transfer protocol أما www.lib.berkeley.edu فهو اسم الحاسب: المكتبة المركزية في جامعة كاليفورنيا بيركلي. وباقيّة الأجزاء تشير الى مكان واسم الملف في الحاسب.

النشر الإلكتروني Electronic publication:

هناك العديد والعديد من المواد المنشورة على صفحات الوب. وفي 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 في:

• التحليل العددي (الخوارزميات) Analysis of algorithms:

وهو الوصف الكامل والدقيق لطريقة حل المشكلة. بالنسبة لاسترجاع تتابعات مماثلة نحتاج قياس مدى تشابه تتابع المجس لكل تتابع موجود في قاعدة البيانات. ومن المحتمل أن يكون ذلك أفضل بكثير من تلك الاتجاه الساذج لفحص كل زوج من الأماكن في كل تجاور محتمل وهي طريقة حتى بدون السماح بفراغات فانها تحتاج الى وقت يتناسب مع ناتج ضرب عدد حروف مجس التتابع في عدد حروف تتابعات قاعدة البيانات. ويركز تخصص الحاسب والمعروف عامة 'stringology' على إيجاد 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 كفاء للسؤالات؟. 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 السؤالات؟.

• هندسة البرامج Software engineering:

لم يعد من الصعب بتاتا كتابة برامج بلغة الحاسبات الأصلية. يعمل المتخصصون في 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 في البيومعلوماتية تكتب بواسطة متخصصون.

البرمجة Programming:

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

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

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

كيف يمكنك تعلم PERL بدرجة كافية لاستخدامها في البيومعلوماتية ؟ تقدم العديد من المعاهد دروساً 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/>) والذي يتيح مصدرا لبرامج PERL ومكوناته المستخدمة في مجال البيومعلوماتية.

قوة PERL في تناول سلسلة الحروف جعلها تلائم عمليات تحليل التتابع في علم البيولوجي. وفيما يلي مثال لاستخدام برنامج PERL بسيط لترجمة تتابع نيكلويدية الى تتابع حمض نووي 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 التي 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. ويحتوي هذا الملف على بيانات مصاحبة (جدول ترجمة الكود الوراثي)، وعبارات تخبر الحاسب لعمل شيء معين بالمدخلات (مثل التتابع المطلوب ترجمته)، والبيانات التي يتم ادخالها (وهي تظهر بعد سطر `_END_`). كما تلخص التعليقات أقسام من البرنامج وتوصف تأثير كل عبارة.

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

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

```
...
```

```
}
```

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

ويظهر في البرنامج ثلاثة أنواع من تراكيب البيانات. سطر ادخال البيانات ويشار اليه \$ line وهو سلسلة من الحروف البسيطة والتي تجزأ الى منظومة متعددة البيانات array أو حامل لثلاثيات triplets. وتخزن المنظومة بيانات لموضوعات عديدة في ترتيب خطي. ويمكن استرجاع بيانات كل موضوع على حدي من أماكنها في المنظومة. لتسهيل التقاط الكود الثلاثي لحامض أميني فانه يتم تخزين الكود الوراثي كمنظومة ترتيب. ومنظومة الترتيب أو جدول التكرار عبارة عن تعميم لمنظومة بسيطة أو تسلسلية. اذا كانت عناصر المنظومة البسيطة مفهومة بواسطة أرقام متتابعة فان عناصر منظومة الترتيب تفهرس باستخدام سلاسل من حروف وهي في هذه الحالة تكون 64 ثلاثية. يتم معالجة الثلاثيات المدخلة طبقا لترتيب ظهورهم في تتابع النيكلويدية مع الاستعانة بعناصر جدول الكود الوراثي بترتيب تحكيمي كما يقرأ في الثلاثيات المتتالية. المنظومة البسيطة أو حامل سلاسل الحروف يكون مناسباً لمعالجة ثلاثيات متتالية بينما تلائم المنظومة التسلسلية النقاط الأحماض الأمينية المقابلة للثلاثيات.

مثال : برنامج بيرل لترجمة تتابع حامض نووي الى تتابع حامض أميني

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 $codon (@triplets) { # loop
over triplets # print
        print "$standardgeneticcode($codon)"; # print
out translation of each # end loop
    } # end loop
cn triplets # skip
    print "\n\n"; # skip
line on output # end loop
} # end loop
cn 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                    # to
    $prefix[$i] = "noprefixfound";   #
    "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
$1
            }
        }
    }
    $prefix[$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
```

وهناك 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

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 في تحليل العلاقات البيولوجية:

المثال الأول: استرجاع تتابع الحمض الأميني لانزيم الريبونوكليز في
بنكرياس الحصان باستخدام ExPASy server في المعهد السويسري
للبيومعلوماتية وعنوانه:

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

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

٢- ثم اضغط على مفتاح ENTER

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

```
>sp|P00674|RNP_HORSE RIBONUCLEASE PANCREATIC (EC 3.1.27.5) (RNASE
1) ...
KESPAAMKFERQHMDSGSTSSSNPTTCNQMMKRRNMTQGWCKPVNTFVHEP
LADVQAICLQKNITCKNGQSNQYQSSSMHITDCRLTSGSKYPNCAYQTS
QKERHIIIVACEGNPYVPVHFDA$VEVST'
```

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

التصنيف والتسمية البيولوجية

Biological Classification and nomenclature:

المثال الثاني: حدد باستخدام تتابعات انزيم الريبونيوكلسيز البكترياسي للحصان والحوت والكنجر النوعين الأكثر قرابة:

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

```
>RNP_HORSE
KESPAKMFQRQHMDSGSTSSSNPTYCNQMMKRNMTCGWCKPVNTFVHEP
LADVQAICLQKNITCKNGQSNQYQSSSMHITDCRLTSGSKYPNCAYQTS
QKERHIIIVACEGPNPVVPHFDASVEVST
>RNP_BALAC
RESPAMKFFQRQHMDSGNSPGNNPNYCNQMMRRKMTQGRCKPVNTFVHES
LEDVKAQVCSQKNVLCCKNGRTNCTYESNSTMHITDCRQTGSSKYPNCAYKTS
QKEKHIIVACEGPNPVVPHFDNSV
>RNP_MACRU
ETPAEKFFQRQHMDTEHSTASSSNYCNLMMKARDMTSGRCKPLNTFIHEPK
SVVDAVCHQENVTCCKNGRTNCTYKNSRLSITNCRQTGASKYPNCQYETSN
LNKQIIIVACEGQYVPHFDAYV
```

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 في 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. ففي حالة تحديد تتابع جين جديد أو التعرف من خلال الجينوم البشري على جين مسئول عن مرض معين يكون هناك رغبة لمعرفة وجود جينات مماثلة في أنواع أخرى. والطريقة المثلى لتحقيق ذلك يجب أن تكون حساسة بحيث تتعرف على كل العلاقات واختيارية بأن تكون تلك العلاقات حقيقية.

وتشتمل طرق البحث في قواعد البيانات المتناوب بين الحساسية والاختيارية. هل تستطيع الطريقة إيجاد كل أو معظم التتابعات الموجودة بالفعل أم أنها تفقد أجزاء كبيرة؟، وعلى النقيض كم من التتابعات 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)

وهو برنامج 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 الإنسان وذبابة الدروسفيل.

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>.

التنبؤ بتركيب البروتينات وهندستها: Protein structure and engineering

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

١- التنبؤ بالتركيب الثانوي Secondary structure prediction:

تحديد ماهي أجزاء التتابع التي تكون الشكل اللولبي والأخرى التي تكون الشرائط.

٢- تمييز الطي Fold recognition:

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

٣- نماذج التماثل Homology modeling:

نفترض أن هناك بروتين معروف تتابع الحمض الأميني له وغير معروف التركيب ولكنه متماثل مع واحد أو أكثر من البروتينات المعروفة 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 في حالة ما اذا كان في اثنين من البروتينات قريبة الصلة تطابق ٥٠% في التتابع عند عمل المحازاة يكون هناك تماثل ٩٠% في تركيبهما.

فيما يلي التتابعات المصفوفة والتراكيب المتطابقة لاثنتين من البروتينات قريبة الصلة وهما ليسوزيم الأبيض في بيض الدجاج وألفا لاكتوالبومين

في البابون. والتتابعات فيهما شديدة التقارب (٣٧% تطابق كامل في التتابعات المتراسة) كما أن هناك تماثل في التركيب. وكل بروتين يمكن استخدامه كنموذج جيد للبروتين الآخر:

Chicken lysozyme :

KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNT
DGS

Baboon ? - lactalbumin :

KQFTKCELSQONLY - DIDGYGRIALPELICTMFHTSGYDTQAIVEND - ES

Chicken lysozyme :

TDYGILQINSRWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVS

Baboon ? - lactalbumin :

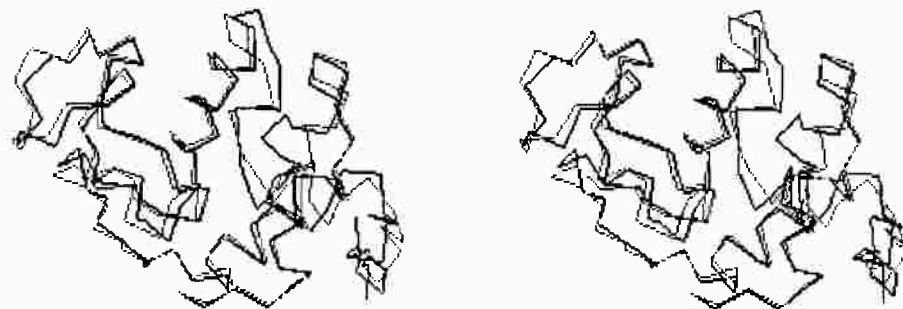
TEYGLFQISNALWCKSSQSPQSRNICDITCDKFLDDDDITDDIMCAKKILD

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 بـ بروتينات البروتينات الى استخدام اختبارات مصممة blind tests. لهذا الغرض صمم جاي مولت برامج CASP. العلماء المنشغلون بتقدير تركيب البروتين باستخدام طرق القياس البللورى والرنين النووى المغناطيسى مدعوون الى: (١) نشر نتائج الحمض الأميني عدة شهور قبل التاريخ المتوقع الى اكمال 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. ففي المعمل يمكننا عمل تعديل فى الأحماض النووية والبروتين واحدات 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. وتسطيع هندسة البروتينات أن تستكشف آفاق ومجالات جديدة.

التطبيقات الاكلينيكية:

هناك اجماع فى الرأى على أن معرفة التتابعات فى الجينوم البشرى وجينوم العديد من الكائنات الأخرى سوف يودى 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. يودى 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 فإن 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 خطيرة بالاضافة الى التكلفة الباهظة.

٣- تعريف الأماكن المستهدفة للدواء:

المكان المستهدف قد يكون بروتين يتم تعديله 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 وحوالي الربع انزيمات والربع المتبقى هرمونات وحوالي ٧% تعمل على أهداف غير معلومة.

أدى التزايد في ظاهرة مقاومة البكتريا للمضادات الحيوية الى خلق مأساة بخصوص علاج الأمراض. والحاجة الماسة لإيجاد أدوية جديدة مرهونة بقاعدة البيانات اللازمة لانتاجها. ونتائج دراسة الجينوم قد تؤدي الى اقتراح أماكن مستهدفة. الاختلافات في 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 في التعرف على تعبير بروتينات مختلفة تستخدم كأهداف لأدوية مضادة للسرطان.

٤- العلاج بالجينات:

وهو إمكانية استبدال جين - أو على الأقل مد الجسم بنتاجه - محل جين غائب أو مصاب بخلل. كما يتضمن العلاج بالجينات العمل على خفض النشاط الزائد لجين ما، وهناك بعض الحالات المرضية في الإنسان والتي أظهرت نتائج مشجعة بالعلاج الجيني.

ومن التوجهات التي تستخدم لإيقاف تعبير الجين هو ما يطلق عليه "antisense therapy"، وهو عبارة عن إنتاج شريط قصير من الدنا أو الرنا الذي يرتبط بأسلوب تتابع متخصص بمنطقة الجين. وهذا الارتباط يتداخل على التوالي مع عمليات النسخ والانتقال. لقد أظهر

هذا النوع من العلاج كفاءة ضد بعض الأمراض مثل cytomegalovirus; and Crohn disease. كما أنه ذهب مباشرة من تتابع المستهدف الى اختصار مراحل عديدة من عمليات تصميم الدواء.

المستقبل The future:

سوف يشهد القرن الحالى تطورا مذهلا فى انتاج وتوزيع وسائل العلاج والعناية بصحة البشر وتتلاشى الحواجز بين قلاع البحث والممارسة الاكلينيكية. ويصبح بإمكان قارئ هذا الكتاب أن يتوصل الى علاج لمرض قاتل. كما يمكن اكتشاف عقاقير ووسائل من شأنها منع حدوث 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/>

تمارين 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

he 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 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;
we I; 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.