الفصل الثاني عشر

الحسابات البيولوجية

Biologic Computing

بروفسير إريك هوفمان – اريك ريفيس

د. جافاد نازاريان

د. يتراب هاثوت

د. زيوا وانج – جوسفين شين

المركز الطبي المحلي للأطفال

Prof. Eric P. Hoffman, Erica Reeves

Dr. Javad Nazarian

Dr. Yetrib Hathout

Dr. Zuyi Wang and Josephine Chen

Children's National Medical Center

(١٢،١) مقدمة

يركز الحساب البيولوجي، كما هو معرف في هذا الفصل، على الطرق الحسابية والبايو معلوماتية لثلاثة من الجزيئات الأساسية للحياة وهي: الـ DNA، والـ RNA، والبروتين. هذه البلوكات البنائية الثلاثة تكون الخلايا، والأنسجة، والجسم البشري الحي. الـ DNA هو الشفرة التي تحتوي على مايقرب من ٣٠٠٠٠ من الجينات التي يمكن تفعيلها أو إخمادها في السياقات المختلفة والتتابعات كاستجابة للتطور والوسط المحيط. تتركز قواعد بيانات الـ DNA على كود أو شفرة خطية في البشر تتكون من ثلاثة بلايين من الأحرف (أقل من ذلك في الكائنات الأخرى). إن الطبيعة الساكنة نسبيا وقلة الأبعاد في الـ DNA تسمح له بأن يكون مرجعاً أو ملاذاً لمعظم قواعد بيانات الحساب البيولوجي الأخرى. الـ mRNA والبروتين يكونان أكثر أو أعلى أبعاداً، مع نطاق ديناميكي، وتعديلات، ونماذج للتعبير، والتفاعل مع الـ mRNA والبروتينات بحيث تملئ الأحوال العادية والمرضية. تعمل معظم جهود الحسابات البيولوجية الحالية والمستقبلية باتجاه تخزين التغيرات في الـ mRNA ونماذج البروتين كدالة في متغيرات معينة، بهدف تركيب هذه التغيرات في صورة شبكات وطرق مرور ذات صلة بهذه الحسابات. إن الهدف طويل المدى هو فهم واستنتاج أو توقع استجابات الخلايا، والأنسجة والمريض للأحوال المرضية والوسطية على الرغم من أن الأبعاد العالية جداً تجعل من ذلك هدفاً صعباً.

(١٢،٢) نظرة عامة على الطرق الجينومية

إن الأنواع الثلاثة من الجزيئات، الـ DNA، mRNA، والبروتين، كل منها له التحديات المعلوماتية الخاصة به. لقد كان الـ DNA من أكثرهم تكثيفاً في الدراسة وهو في جميع الأحوال أكثرهم وضوحاً وسهولة. يتكون الـ DNA من أربعة مكونات ممكنة فقط وهي A, G, T, C، وهذه تكون مرتبة في شفرات عالية الخطية لتكون الجينات والمناطق الجينية الجانبية. نحن هنا لن نمر على الكيمياء الخاصة بالـ DNA، ولكن من المهم أن نفهم بعض المفاهيم الأساسية القليلة حتى نستطيع أن نقدر وجود الموارد والتحديات المتبقية في حسابات المعلوماتية الحيوية.

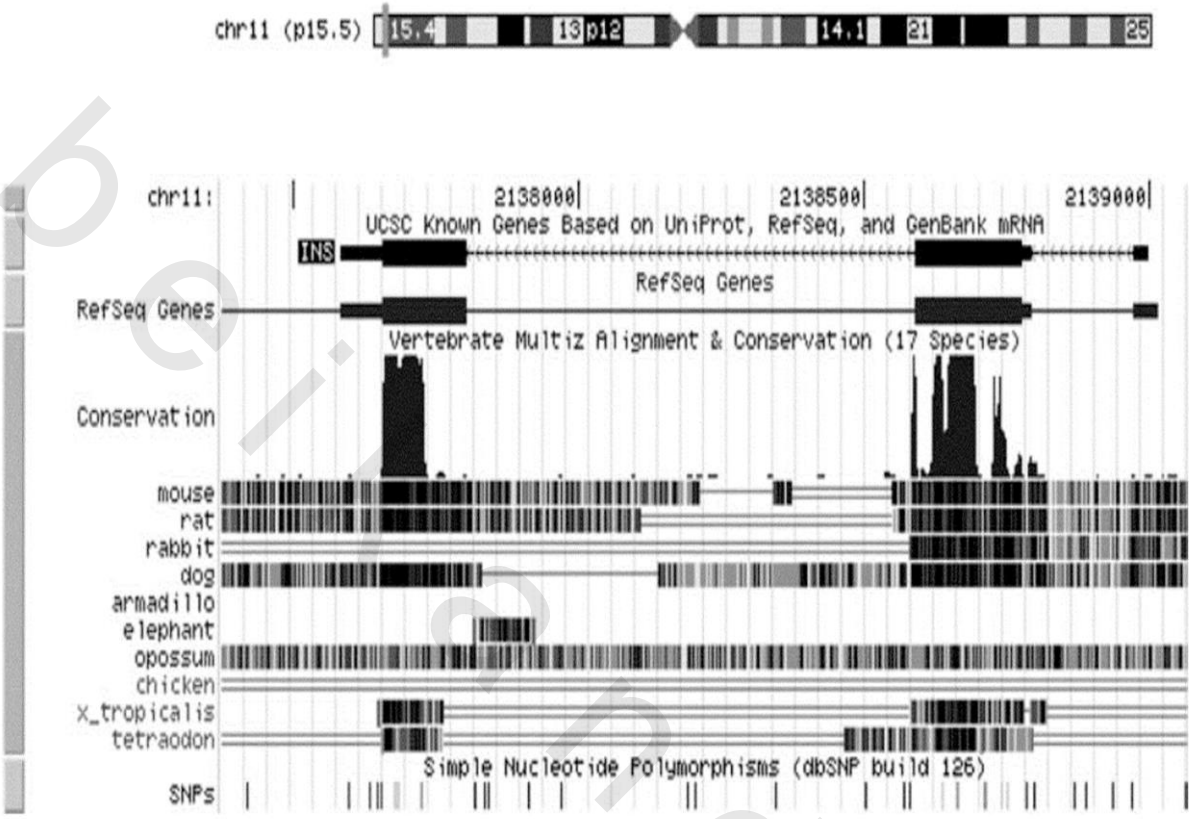
يأخذ الـ DNA الشكل الحلزوني المزدوج مع شريطين أو جديلتين مكملين تماما لبعضهما. وعلى ذلك، فليس من المهم أي الشريطين أو الجديلتين تعرضهما قاعدة البيانات لأن الشريط الآخر يمكن استنتاجه بسهولة. إن السلسلة المكونة من ثلاثة بلايين من الأحرف للـ DNA يتم ترتيبها في هيكل أو مركب فوقي (الكروموزومات، مع القسم الطرفي والقسم المركزي داخل كل كروموزوم). وعلى الرغم من ذلك فمن أجل حسابات المعلوماتية الحيوية، فقد أصبحت الكروموزومات خاصية ثانوية نسبياً. الأكثر أهمية من ذلك هي الوحدات الوظيفية لل DNA، وهي الجينات أو الوحدات النصية transcript، والتي تكون موجهة في كل من الاتجاهين خلال الـDNA.

لكي نشرح الجينات وقواعد البيانات المصاحبة لها سنستخدم واحداً من أشهر مواقع الإنترنت، وهو مستكشف الجينوم (http://www.genome.ucse.edu). لقد تم بناء هذا الموقع عن طريق أستاذ المعلوماتية الحيوية Jim Kent ومجموعته في عام ٢٠٠٠ باستخدام قاعدة البيانات MySQL في جامعة كاليفورنيا في سانتا كروز، باستخدام مجموعة من حاسبات الـLinux pentihm class التي تعمل كخادم إنترنت [1]. لقد استخدم مستكشف الجينوم الشفرة الجينية للبشر والكائنات الأخرى كمرتكز أو مرجع، والتي تشير إليها قواعد البيانات الأخرى. تقريباً نصف الاتصالات مع هذه القاعدة تتم عن طريق أعضاء هيئة التدريس في UCSC، والنصف الآخر يتم عن طريق الباحثين من أماكن أخرى بعيدة يرغبون في جعل قواعد بياناتهم شفافة وممكن الاتصال والاستعلام منها عن طريق موقع مستكشف الجينوم.

يتراوح حجم الجينات من الصغير جداً (ربما 100-300 bp) كما في الشكل رقم(١٢.١)، إلى أكبر جين معروف وهو جين الـ dystrophin الذي مقاسه ٢.٣ مليون زوج من القواعد base pairs كما في الشكل رقم(١٢.٢). وفوق كل ذلك فهناك حوالي ٢٠٠٠٠ -٢٥٠٠٠ من الوحدات النصية أو الجينات المتناثرة على الثلاثة بلايين زوج من القواع، ولكن العدد يعتمد على ما يمكن تحديده كجين وأيها لا يمكن.

سيسأل أحدهم بسرعة هذا السؤال، " كيف يمكن تحديد الجين أو الوحدة النصية transcript unit؟". إذا عملت إحدى مناطق الـ DNA كوحدة نصية، فإنها لابد أن تنتج جزيء RNA (مثلا: تم ترجمته إلى RNA). يمكن لأحدهم بعد ذلك أن يأخذ نسيجاً مثل عضلة أو المخ، وفصل كل الـ RNA، وبعد ذلك يرتب أو يتابع كل مقاطع الـ RNA. كل RNA يجب أن تكون قادمة من جزء أو مقطع من الـ DNA الموجود في مكان ما في الجينوم؛ ولذلك يمكن لأحدهم أن يرسم أو يحول مرة أخرى تتابع الـ RNA إلى تتابع جينومي في الـ DNA. لقد تم تنفيذ هذه العملية لمئات من الأنسجة والخلايا على مدار العقدين الأخيرين، مما أدى إلى قواعد بيانات كبيرة لعلامات معبرة عن التتابعات expressed sequence tags, EST، وقصاصات من تتابعات الـ RNA التي يمكن استخدامها كقواعد بيانات يمكن تحويلها مرة أخرى إلى جينومات DNA وتحديد الوحدات النصية خلال جينومات الـ DNA.

إن عملية تحويل الـ ESTs مرة ثانية إلى التتابع الخطي الجينومي DNA تعتبر حجر أساس في الحسابات البيولوجية. تعتمد هذه العملية كلية على الاعتقاد المركزي في الحسابات البيولوجية وهو أن ترتيب القواعد في الـ DNA والـ RNA يمكن توقعها بالكامل بالاعتماد على التتابع (مثلا: التحويل عن طريق تماثل التتابع). إذا تتطابق جزء من الـ RNA مع اثنين من المناطق المنفصلة في الـ DNA والتي تكون في نفس منطقة التجاور، فإنه يمكن افتراض أن هذه عبارة عن اثنتين من الإكسونات المنفصلة عن طريق إنترون. بإجراء هذه العملية تكرارياً على مئات من الأنسجة والخلايا وملايين من الـ RNA، فإن ذلك يؤدي خريطة التحويل النصي التي يمكن رؤيتها حاليا في المستكشف الجينومي.



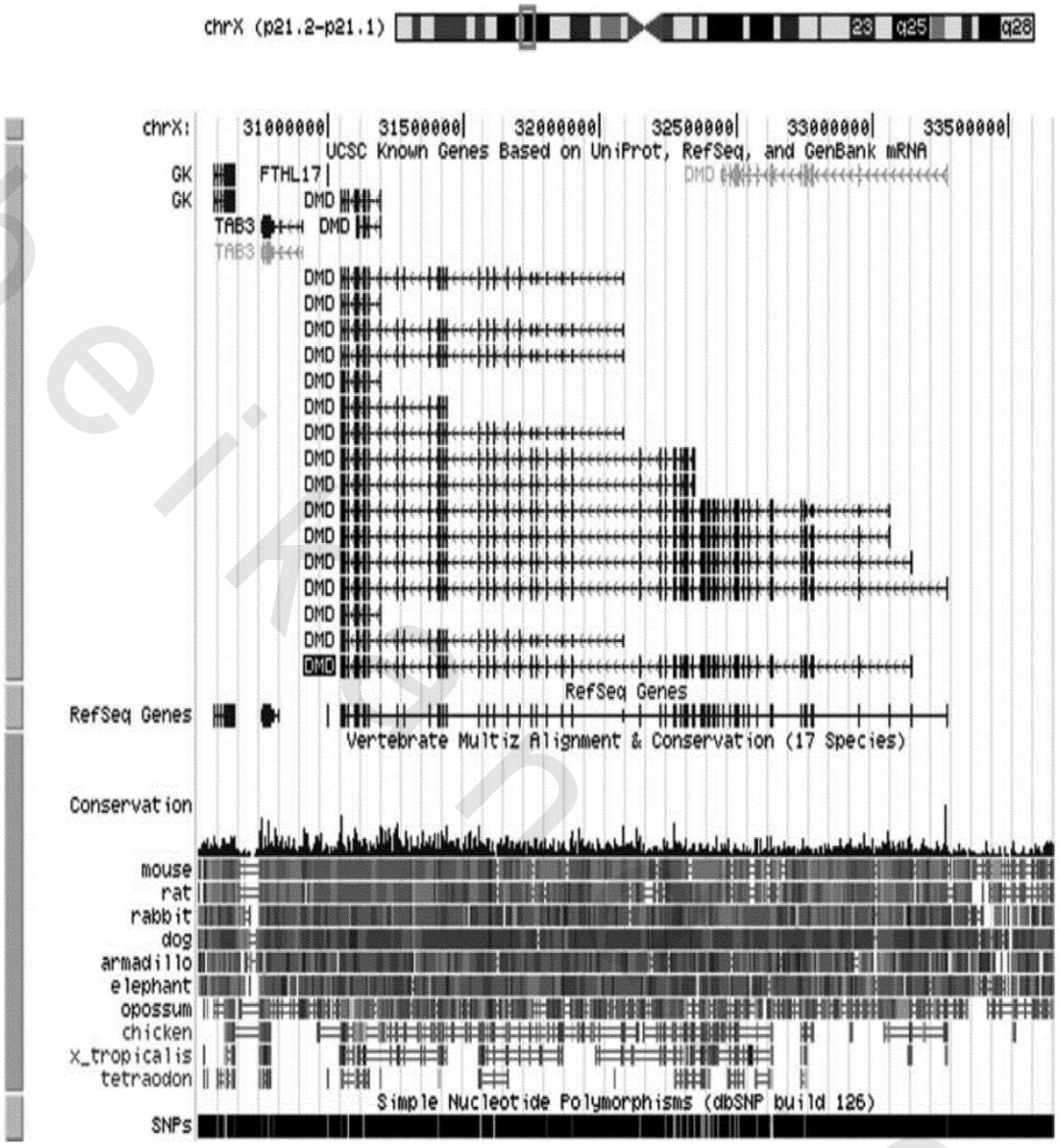
الشكل رقم (١٢،١) منظر من مستكشف الجينوم لجين الإنسيولين INS . يقع الجين على الذراع القصير للكروموزوم ١١ (11p15.5) وهو مبين في المنظر الأعلى ، مع وضع محاور أزواج القاعدة على الكروموزوم ١١ كما هو مبين (2138000 bp) . إن جين الإنسيولين يكون منسوخاً في الاتجاه من اليمين للشمال (موضح بسهم في الرسم التخطيطى لجين الإنسيولين) وله ثلاثة إكسونات (خطوط ثقيلة) . الجين الكامل يكون حوالى 1500bp (1.5kbp) . الإكسون الابتدائى الأول من اليمين يكون صغيراً وغير مشفر (خط قصير ، '5 ، منطقة غير محولة في النص) ، بينما يحتوى الإكسون الثانى على كمية صغيرة من غير المشفرة وكمية كبيرة من التتابع المشفر (شفرات الحمض الأمينى ، الشريط الأكبر) . الإكسون الأخير من الشمال يحتوى ٣/٢ تتابع شفرات الحمض الأمينى ، والباقى يكون '3 غير مشفرة (غير محولة) . يمكن إظهار تتبع الحفظ التطورى هنا ، والذي يوضح أن الإكسونات ٢ و ٣ محفوظان بدرجة عالية خلال التطور ، بينما إكسون ١ لا يكون محفوظا . المناطق عالية الحفظ تعنى أهمية وظيفية عالية للتابع في هذه المنطقة نتيجة وجود ضغط تطورى للحفاظ على هذه الجينات كما هى . الأثر الأسفل يعطى ملخص للنيوكليدات الوحيدة متعددة الأشكال single nucleotide polymorphism, SNP من خلال منطقة الجين . هذا الشكل مأخوذ من http://www.genome.ucsc.edu[1] .

هناك بعض الجوانب الأخرى في الـ mRNA، والتي تكون متغيرات في تطوير واستخدام الحساب البيولوجي. كما هو موضح في الشكل رقم(١٢.٢) فإن المتعهد أو المؤسس (إشارات تقوم بإدارة النهاية '5 في الجين) والإكسونات الأولى المصاحبة يمكنها أن تكون متغيرات. هناك متغير إضافي وهو الربط البديل، حيث يمكن استخدام الإكسونات المختلفة عن طريق الخلايا المختلفة عند أوقات مختلفة.كمثال على أن الوحدة النصية الوحيدة يمكنها أن تظهر عدد من المؤسسين، والربط البديل، ومواضع الإنهاء البديلة الموضحة في الشكل رقم(١٢.٣) (tropomysin 3, TPM3).

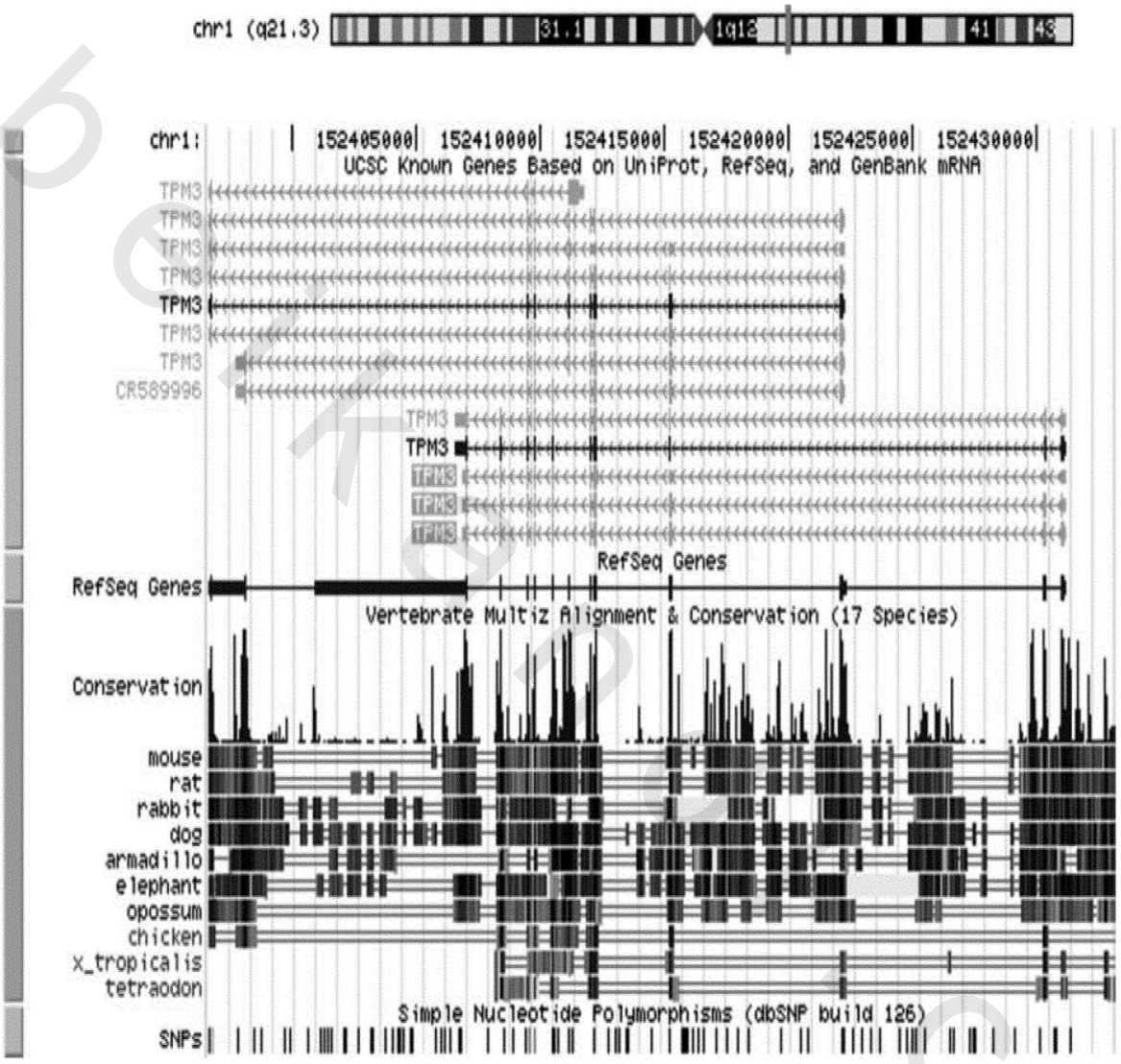
من الضروري أن نشير هنا إلى أن استخدام الوحدات النصية عن طريق الخلايا والأنسجة يعتمد على العديد من العوامل، بما في ذلك حالة التطوير أو البناء، وتلميحات الوسط، والأحوال المرضية. في الحقيقة، فإن الخلايا المتجاورة في أي نسيج من الممكن أن تفعل أو تخمد الجينات المختلفة، وحتى عند استخدام نفس الجين فإنه قد يستخدم بعض المتغيرات المختلفة لهذا الجين. ومن الأكثر أهمية، فإن النموذج الجيني والتعبير البروتيني في الخلية الواحدة من الممكن أن يغير نموذج هذا التعبير في الخلية المجاورة أو في خلية على مسافة بعيدة في الجسم. وعلى ذلك، فإنه من الممكن أن نلاحظ سريعا الثبات والإتقان في تنظيم هذه الثلاثة بلايين وحدة من جينوم الـ DNA وكيف يصل إلى هذا التعقيد المهول عند مستوى الـ RNA. على الرغم من هذا التعقيد، فإنه يبقى نسبياً سهل الاستجواب، والفهرسة، وتوليد قاعدة بيانات للـ RNA نتيجة الحساسية الرائعة والتحديد لأبحاث الحاسب المتماثلة والحلول المعتمدة على المعمل والتهجين للتتابعات المحددة. سيتم شرح ذلك مؤخراً في هذا الفصل، والذي سيشتمل أيضاً على أدوات وتحديات المعلوماتية الحيوية الحالية.

(١٢،٣) نظرة عامة على الطرق البروتينية

إن تعقيدات نماذج وشبكات الـ RNA تتضاءل بالمقارنة مع تعقيدات شبكات البروتين. إننا عادة ندرس أن حوالي ١٪ حتى ٤٪ من هذا التعقيد تكمن في تعبير الـ RNA، و ٩٦٪ حتى ٩٩٪ من هذا التعقيد تكمن في التعبير البروتيني ووظائفه. إن البروتينات تتزايد تعقيداً بصورة أسية على عدد من المستويات. أولاً: هناك ٢٠ حمضاً أمينياً قياسياً (وحدات بناء) بدلاً من أربعة في الـ DNA والـ RNA. ثانياً: بمجرد حدوث التحول البروتيني، فإن هناك قائمة طويلة من المتغيرات تؤثر في نشاط ووظيفة البروتين. يتضمن ذلك بكثافة تعديلات ما بعد التحول (الفسفرة phosphorylation، الجلوكزة glycosylation، الانشقاق البروتيني proteolytic cleavage). ثالثا: عملية الطي البروتيني تملئ النشاط، كما يحدث مع التفاعلات مع النسخ الإضافية من نفس البروتين أو البروتينات الأخرى. رابعاً: التحديد الخلوي الجانبي والتركيز الموضعي للقواعد والعوامل المساعدة من الممكن أن يؤثر بدرجة كبيرة على وظيفة البروتين.



الشكل رقم (١٢،٢) منظر للمستكشف الجينومى لجين ال dystrophin ، والذي يعتبر أكبر جين معروف حتى الآن . الموضح في هذا الشكل هو جين الديستروفين dystrophin أو ال Duchenne muscular dystrophy, DMD . يغطى هذا الجين ٢،٣ مليوناً من أزواج القواعد ويشتمل على أكثر من ٨٠ إكسوناً exon . هناك أكثر من موقع بديل لبداية هذا الجين (انظر إلى إكسون ١ البديل على يمين قائمة الأشكال المختلفة) . إن تحليلية هذه الإكسونات ، والحفظ ، وال SNPs كلها تم ضغطها تماما في هذا الشكل؛ مما يجعل هذه المعلومات صعبة التفسير . هناك وحدات نصية إضافية على يسار الشكل (GK, TAB3, FTHL17) .إن مقارنة جين الإنسولين البسيط نسبيا في الشكل رقم (١٢,١) وهذا الجين الكبير والأكثر تعقيدا توضح المدى الواسع في الوحدات النصية في الجينوم البشرى . مأخوذ من .[1] http://www.genome.ucsc.edu .



الشكل رقم (١٢،٣) الوحدة النصية المركبة لجين tropomyosin 3 . يغطى جين ال TPM3 حوالي (35 kbp) 35000bp وهو محول من اليمين لليسار . يتم استخدام على الأقل ثلاث منظمات أو مؤسسات مختلفة وأكسون واحد عن طريق أنواع مختلفة من الخلايا عند أوقات مختلفة (المحول في القمة يستخدم منظم والأكسون الأول في مركز الجين ، بينما تستخدم المحولات الأخرى الإكسونات 10-15 kbp من اليمين) . يوجد أيضاً على الأقل ثلاثة إكسونات طرفية مختلفة يتم استخدامها . التحويلات السفلى تنتهي مبكراً، ولكنها تشتمل أيضاً على إكسونات إضافية غير مشاركة مع التحويلات العليا (ربط بديل) . هذه النتائج مع التنوع في بروتين ال TPM3 يتم إنتاجها بهذه التحويلات ، وكل ذلك من من نفس الوحدات النصية الأبوية . مأخوذة من .[1] http://www.genome.ucsc.edu

بالإضافة لهذه التعقيدات المتأصلة، فإن فهم الشبكات البروتينية قد تباطأ بسبب مشكلتين تقنيتين. إنه من المفروض أنه من الصعب تنقية وترتيب أو تتبع البروتين، بالمقارن مع الطرق المتاحة الآلية لاستنساخ وترتيب الأحماض النووية (DNA, RNA). وأيضاً الحساسية الزائدة والتتبع المحدد مع التهجين المحدد المستخدمة للاستعلام عن الأحماض النووية غير متاحة بالنسبة للبروتينات.

حتى تاريخه لم يكن من الممكن إنتاج مسبار يمكنه التقاط بروتين واحد من محلول مركب كما هو ممكن مع التهجين بين المحاليل المركبة من الـ DNA أو الـ RNA.

إن نقص الطرق الحساسة والمحددة للاستعلام البروتيني في المحاليل المركبة قد تغير أخيراً مع ظهور المقاييس الطيفية spectrometers, MS للكتلة ذات الإنتاجية العالية، وطرق فصل البروتين الحساسة الكثيرة، والبروتين الجانبي باستخدام طرق التعليم المستقرة بالنظائر. هذه الطرق موصوفة بتفاصيل أكثر قليلاً في هذا الفصل. على الرغم من ذلك، فإننا سنعطي هنا وصفاً ملخصاً لأساسيات تحليل البروتينات ذات الإنتاجية العالية (proteomics).

إن فهم أسس البروتين يصبح حرجاً عند اعتبار الجوانب المختلفة للبروتين، كما أن أساسيات البروتين تكون كما يلي:

في الـ MS، مقاطع البروتين المؤين التي تتحرك خلال الفضاء تكون لها نسبة كتلة نوعية على الشحنة مقدارها m/z.

الهيكل الثانوي (تتابع الحمض الأميني) للبروتينات يمكن توقعهاعن طريق الترجمة في السيليكو للجينات (الوحدات النصية تكون معروفة ومتوقعة).

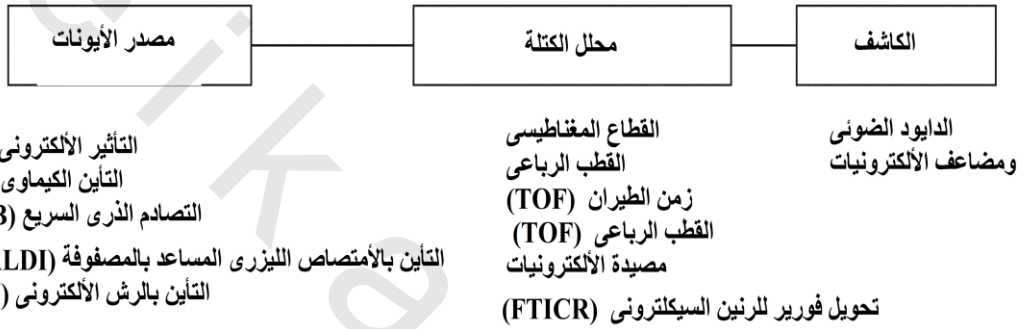
يمكن تقسيم التتابع البروتيني في السيليكو إلى نمازج أو أنماط متوقعة من الأقسام التي لها m/z.

بعد ذلك يمكن مطابقة أنماط المقاطع التي تمت ملاحظتها والتي لها m/z مع كل الأنماط النظرية ذات الـ m/z لكل البروتينات الغير معروفة أو المتوقعة؛ وبالتالي فإن البروتين ذا الاهتمام يمكن تخصيصة وتحديده.

لكي نبدأ مع الأساسيات الفيزيقية والكيميائية وراء استخدام الـ MS، فإن الـ MS يكون له ثلاثة مكونات أساسية: مصدر الأيون، محلل كتلة، والكاشف كما في الشكل رقم(١٢.٤). إن المصدر الأيوني ينقل شحنات إلى مقاطع البروتين (ببتايد peptides) مما يتسبب في دفعها خلال الفضاء عن طريق قوي كهرومغناطيسية خلال الفضاء. هناك عدد من الطرق المستخدمة في التأين، ولكن الاثنين الأكثر استخداماً هما التأين بمساعدة الامتزاز الليزري matrix-assisted laser desorption ionization, MALDI وطريقة التأين بالبخ الكهربي electrospray ionization, ESI كما في الشكل رقم(١٢.٤).

الفرق الأساسي بينهما هو أن عينات تحليل الـ MALDI كجوامد جافة تم تعجيلها بالامتصاص لليزر الفوق بنفسجي على مسبار، بينما طريقة الـ ESI تكون طريقة معتمدة على محلول يتم رشه خلال أنبوبة شعرية ضيقة من الصلب الغير قابل للصدأ.

يتكون المقياس الطيفى للكتلة مما يلى :



الشكل رقم (١٢،٤) نظرة عامة على مكونات مقياس طيف الكتلة (سكترومیتر) . البنود أو المكونات المكتوبة بالخط الأحمر هى مصادر الأيونات الأكثر استخداماً وطرق تحليل الكتلة . للحصول على منظر تفصيلى لهذا الشكل يمكنك زيارة الموقع التالي : http://books.elsevier.com/companions/9780123735836.

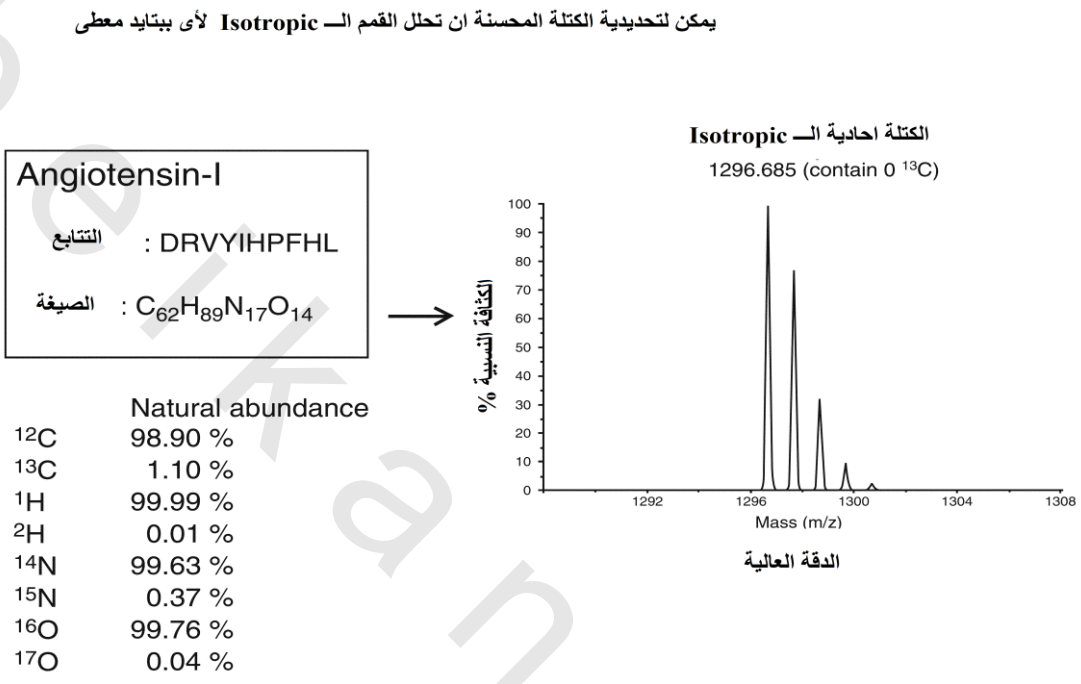
المركبة الثانية في الـ MS هي محلل الكتلة الموضح في الشكل رقم(١٢.٤). كل هذه عبارة عن وحدات فراغية يمكنها أن تتعامل مع البولي ببتايد polypeptide المشحون بطريقة تكون فيها النسبة m/z تتعلق مباشرة بالزمن الذي يتم عنده الكشف عن طريق الكاشف. من أشهر ثلاثة محللات كتلة، زمن الطيران time of flight, TOF حيث تتعلق النسبة m/z بمسار الطيران الخطي من مصدر الأيون حتى الكاشف. مصايد الأيونات والأقطاب الرباعية يكون لها تتابع من القوي الكهرومغناطيسية التي تعدل من التردد الأيوني بطريقة غير خطية. في النهاية يكون الكاشف حيثما يتم تحويل الجزيئات إلى إشارات كهربية عن طريق الدايود الضوئي والمضاعف الإلكتروني.

لابد من شحن البيبتايد (أو تأيينه) عن طريق المصدر الأيوني حتى يمكن تحليله عن طريق محلل الكتلة والكشف عنه عن طريق الكاشف، ولكن هناك عدداً من الحالات المشحونة الممكنة للببتايد خاصة (+١ {وحيدة الشحنة} ، +٢ {مزدوجة الشحنة} ، ... إلخ) وبالذات عند استخدام طريقة الـ ESI كطريقة للتأين. يجب معرفة حالة شحن الأيون حتى يمكن تحديد قيمة النسبة m/z، وبالتالي القيمة الدقيقة للكتلة الجزيئية، ويمكن حساب ذلك عن

طريق وجود الـ isotopes المستقرة والموجودة طبيعيا، وبالتحديد الـ isotopes المستقرة لذرة الكربون كما في الشكل رقم(١٢.٥). الكربون بالتحديد له وزن جزيئي ١٢ (مثلا 12C). على الرغم من ذلك، فإن الـ isotope المستقر (غير المشع) 13C، مع نيترون إضافي، يوجد في جميع الببتايد الطبيعية عند مستوى ١٪ تقريبا كما في الشكل رقم(١٢.٥). إن هذا يؤدي إلى الكشف عن كل الببتايدات في الـ MS موضحا تتابع من القمم بدلا من قمة isotope وحيدة كنتيجة لاستبدال الـ ١٪ بدلا من الـ 13C (والمستويات المنخفضة لل isotopes المستقرة الأخرى). إن كتلة الـ isotope الأحادي (لا يوجد 13C في الببتايد) يمكن رؤيتها ك MW 1296.685 في الشكل رقم(١٢.٥)، وإضافة 13C واحدة تؤدي إلى قمة إضافية على اليمين المباشر، ومزاحة بمقدار نيترون واحد (وحدة كتلة واحدة). إذا كان فرق الكتلة بين قمة الـ isotope الأحادي (القمة عند أقصى اليسار عند m/z 1297.685/1-1296.685/1=1)، والقمة مع واحد 13C (القمة المجاورة مباشرة لل m/z 1297.685) تساوي واحد، فإن حالة الشحنة تكون (z=1) (1296). وعلى الرغم من ذلك، فإن الببتايد إذا كان مزدوج الشحنة، فإن الوزن الجزيئي الظاهري الذي تم الكشف عنه في الـ MS يتغير نتيجة أن z=2 (مثلا 1296.685/2). في هذه الحالة، يصبح فرق الكتلة بين القمم الـ isotopic يساوي 0.5 وحدات الكتلة (1297.685/2-1296.685/2=0.5). يمكن لبرمجيات المعلوماتية الحيوية أن تحدد بسرعة الحالة الشحنية للببتايد ببساطة عن طريق النظر على المسافة بين قمم الـ isotopic، فإذا كانت المسافة تساوي ١، فإن الببتايد يكون أحادي الشحنة، بينما إذا كانت الشحنة بين القمم تساوي 0.5 فإن ذلك يعني شحنة مزدوجة. إن التحديدية المحسنة للكتلة لل MS الحالي تساعد في الكشف الدقيق عن تغيرات الـ isotope المستقر، والتي كانت خطوة أساسية في تحديد حالة الشحن والمساعدة في حسابات الـ m/z.

إن الأساسيات المتبقية للبروتين تعتمد على الكشف عن أنماط البولي ببتايد polypeptide في الـ MS باستخدام بيانات التتابعات الجينومية (الجينات، والوحدات النصية) والتطابق مع النماذج المرئية في بيانات الـ MS. توجد ثلاثة أنواع من قواعد بيانات البوليببتايد المستخدمة عادة. من أكثر قواعد البيانات استخداماً القواعد SwissPort، وهي عبارة عن قاعدة بيانات لتتابع بروتيني معالج والتي تحقق مستوى عالي من التعليقات، ويأتي بعدها قاعدة بيانات المركز القومي للتكنولوجيا الحيوية والمعلوماتية national center for biotechnology and information, NCBI، والتي تحتوي على تتابعات من البروتينات غير المتكررة والأحماض الأمينية، وقاعدة بيانات فهرس البروتين الدولي، والتي تحتوي على، وترتب عدداً كبيراً من قواعد البيانات الحقيقية. تقريباً تكون البروتينات الكاملة الطول كبيرة جداً (وزن جزيئي عالٍ) على التحليل خلال معظم الـ MS. ولذلك؛ فإن قاعدة البيانات هذه لا تستخدم مباشرة للتطابق مع بيانات الـ MS ولكنها تستخدم بدلا من ذلك كنقطة بداية للكشف في السيليكو عن مقاطع البولي ببتايد من البروتين الأصلي الكامل الطول.

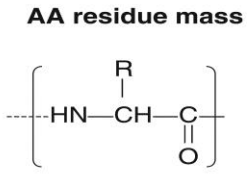
طيف الكتلة المكتشفة يستخدم الكتلة الجزيئية لكل حمض أميني كما في الشكل رقم(١٢.٦) وبعد ذلك يحسب الوزن الجزيئي المتوقع لمجموع الأحماض الأمينية للمقاطع البروتينية (الببتايد، أو البولي ببتايد).



الشكل رقم (١٢،٥) توزيع ال isotopic لببتايد تم تحليله عن طريق مقياس كتلة طيفي ذو تحديدية عالية . توزيع ال isotopic والكثافات تكون نتيجة الوفرة الطبيعية لل isotopes المختلفة من ال C, H, N, and O .

هناك اثنين من نماذج التجزئة التي يتم تجميعها فيما بعد في قواعد البيانات: بصمات البروتين (مثلا البيبتايدات التريبسينية tryptic peptides) وتجزئة الطيف العشوائي (طيف الـ MS/MS). بالنسبة لنموج بصمة البروتين، فإنه من أشهر البروتينات استخداماً هو التربسين trypsin وهو الإنزيم الذي يشق سلاسل البولي ببتايد مباشرة بعد متبقيات اليساين lysine (K) والأرجينين arginine R. تستخدم قواعد بيانات بصمة البيبتايد تتابع البروتين المبدئي، وتهضم البروتين بالتربسين في السيليكو، وتولد قاعدة بيانات لبصمة الببتايد المتوقعة، وبعد ذلك تطابق هذا مع بصمة الببتايد الملاحظة على الـ MS. إذا تم تنقية أي بروتين وبعد ذلك تم هضمه بالتربسين ثم تم الكشف عنه على الـ MS، فإن بصمة الببتايد وحدها تكون كافية لتحقيق تحديد للبروتين لا لبس فيه (تحديد للأصل). تتطلب بصمات الببتايد فقط تحديداً للبتايدات السليمة المنبثقة من البروتين الأصلي (MS).

Amino acid	3-Letter code	1-Letter code	Monoisotopic mass
Glycine	Gly	G	57.021
Alanine	Ala	A	71.037
Serine	Ser	S	87.032
Proline	Pro	P	97.053
Valine	Val	V	99.068
Threonine	Thr	T	101.048
Cysteine	Cys	C	103.001
Leucine	Leu	L	113.084
Isoleucine	Ile	I	113.084
Asparagine	Asn	N	114.043
Aspartic acid	Asp	D	115.027
Glutamine	Gln	Q	128.059
Lysine	Lys	K	128.095
Glutaminic acid	Glu	E	129.043
Methionine	Met	M	131.040
Histidine	His	H	137.059
Phenylalanine	Phe	F	147.068
Arginine	Arg	R	156.101
Tyrosine	Tyr	Y	163.063
Tryptophan	Trp	W	186.079



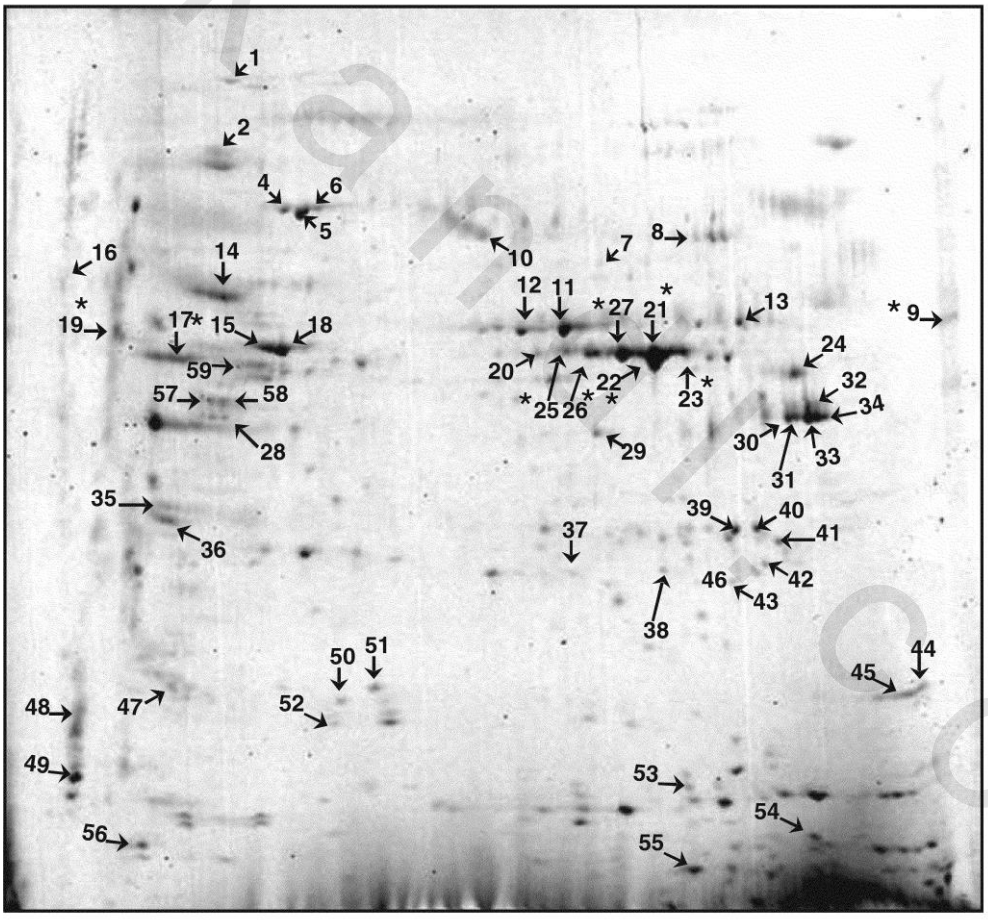
الشكل رقم (١٢،٦) الأحماض الأمينية والوزن الجزيئى المقابل لكل منها (الكتلة المتبقية) . تم أيضاً توضيح هيكل سلسلة البولي ببتايد مع وصلات الأمايد amide linkages .

إن الرحلان الكهربي الثنائي الأبعاد للبروتينات يسمح بتحديدية مقدارها ١٠٠ -١٠٠٠ من البروتينات المنفردة (بما في ذلك الحالات المعدلة بعد التحول) كما في الشكل رقم(١٢.٧)، حيث المحور X هو الشحنة (نقطة التركيز الكهربي) والمحور Y يمثل الوزن الجزيئي. هذه النقاط تم تنقيتها إلى الدرجة أن النقطة المرئية من المحتمل أن تكون بروتين وحيد. إن استئصال هذه النقط من الجيل الثنائي الأبعاد، والهضم بالتربسين، ثم الكشف بخريطة البيبتايد يسمح أحيانا بتحديد البروتين في كل نقطة.

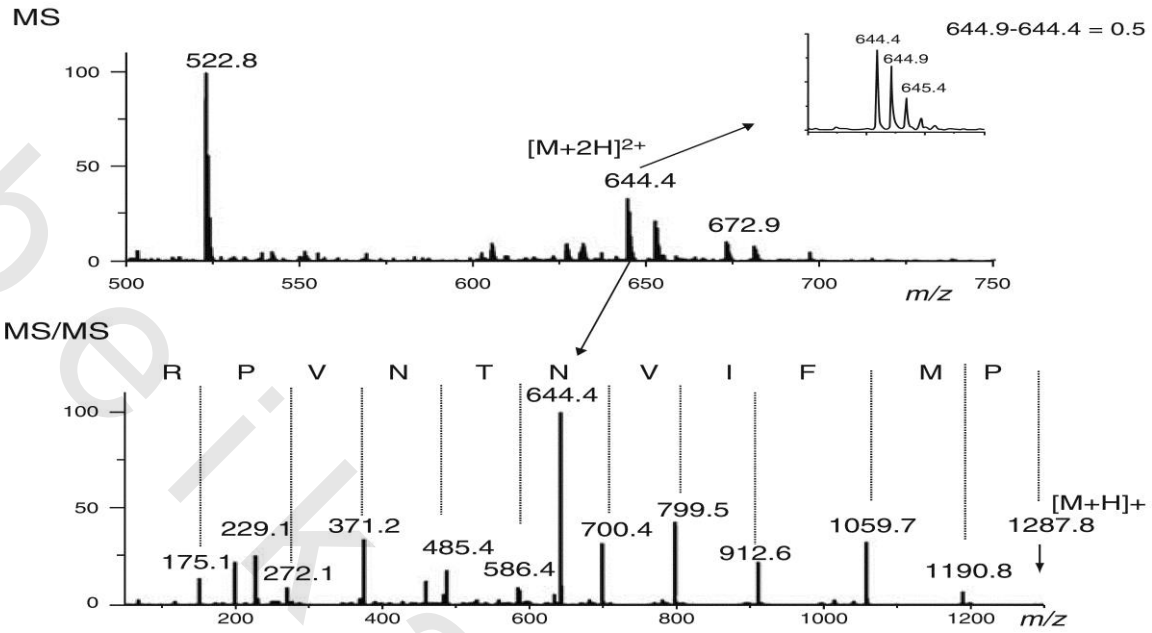
يمكن الحصول على ثقة إضافية لتحديد البروتين عن طريق طيف الـ MS/MS. في الـ MS/MS، يتم محاصرة بيبتايدات معينة (عزلها في المجال الكهرومغناطيسي في الـ MS)، وبعد ذلك يتم تقسيمها عن طريق ليزر أو عن طريق تصادم غازي متعادل. في الشكل رقم(١٢.٨) تم محاصرة قمة البتايد 644.4 ثم تقسيمها، وتم الحصول على طيف الـ MS/MS كتتابع من الأيونات المجزئة التي تسمح بتحديد تتابع الحمض الأميني المتوقع.

يتميز الجيلاتين الثنائي الأبعاد بتحديد البروتين بدرجة عالية من الثقة. كل من خرائط البتايد وتتابع الـ MS/MS للبتايدات المنفردة يجب أن تعود مرة ثانية لنفس البروتين الأصلي في قواعد البيانات. عيوب الجيلاتين الثنائي الأبعاد هي الحساسية المنخفضة نسبياً، والوزن الجزيئي المحدود/في مدى الـ PI، والمتطلبات لكميات كبيرة نسبياً من البروتينات (≥200µg).

هناك طريقة أخرى أكثر حساسية يطلق عليها طريقة الطلق الناري shotgun، حيث يتم هضم محاليل أكثر تعقيدا من البروتين باستخدام التربسين كخليط وبعد ذلك تمر على محلول معتمد على الرذاذ الكهربي باستخدام المخطط اللوني المتعدد الأبعاد المرتبط مع الـ MS. تفقد طريقة الطلق الناري القدرة على توليد بصمات للببتايد وتعتمد بدرجة كبيرة على طيف تتابع الـ MS/MS كما في الشكل رقم(١٢.٩). حيث إن الخليط المركب من الببتايد يأتي من العديد من البروتينات الأصلية، فإنه تكون من الصعب إن لم يكن من المستحيل إعادة بناء جزء من بصمة الببتايد من الخليط العالي التعقيد. أيضاً، فإنه مع طريقة الطلق الناري، فإنه يكون هناك تغطية أقل من كل بروتين أصلي، مع كشف للقليل من الببتايدات لكل بروتين أصلي. إن الميزة الأساسية لطريقة الطلق الناري هي الإنتاجية العالية والغطاء البروتيني المتزايد.



الشكل رقم (١٢،٧) التحليل الكهربي للجيلاتين الثنائي الأبعاد (جيلاتينات ثنائية الأبعاد) . الشكل يوضح مثالا على البروتينات السيتوبلازمية المحلولة والمعزولة من العضو الكهربي لسمكة التوربيدو ، وهو نسيج مخصوص قادر على توليد ٢٠٠ فولت من التيار المستمر خلال الماء لصعق فريسته ، مأخوذ من Nazarian et al. [2].

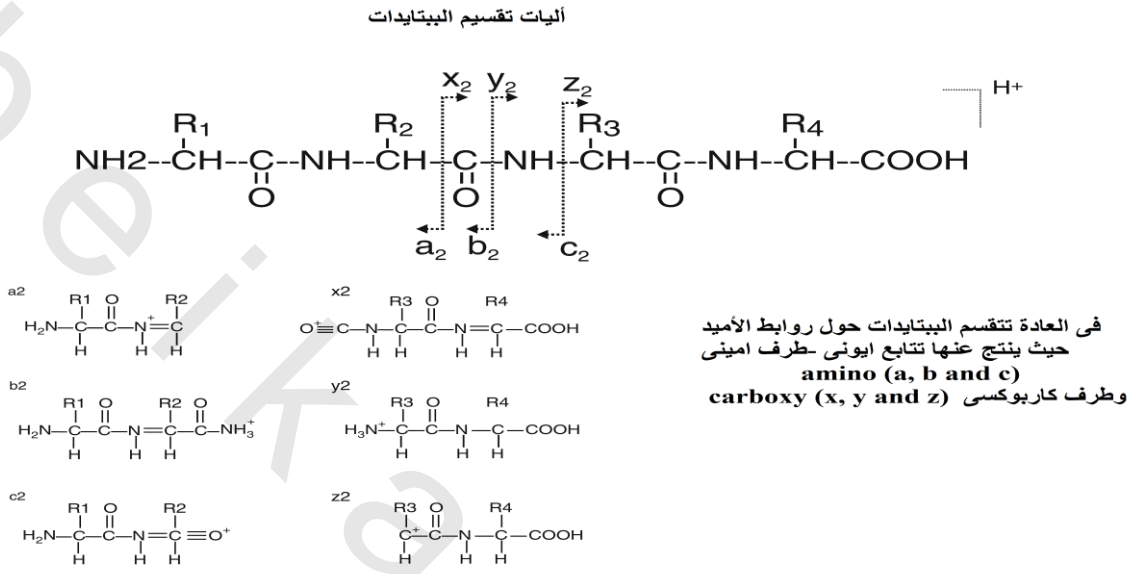


الشكل رقم (١٢،٨) محاصرة لقمة ال (MS) 644.4 ، التقسيم والكشف عن طيف ال MS/MS . لقد وجد أن ببتايد ال 644.4 يكون مضاعف الشحنة كنتيجة للمسافة المتقاربة (0.5 وحدات الكتلة) من المتغيرات الأيسوتوبية isotopic (أعلى اليمين نافذة الريح العاصفة)؛ وبالتالي التسمية بال [M+2M]2+ . الوزن الأيسوتوبي الأحادى الحقيقي لهذا الببتايد هو 1287.8 . تسمح التحديدية لأيونات ال b وال y كما في شكل (١٢–٩) باستنتاج تتابع الحمض الأيوني للببتايد.

لقد أصبحت الأهمية الإحصائية لتحديد البروتين الأصلي أكثر تحدياً مع طرق الطلق الناري MS/MS. مثالياً أو قياسياً، فإن الببتايد الأحادي المكتشف والمجزأ يعتبر غير كافٍ للتحديد القوي للبروتين الأصلي. إن التحديدات المتعددة لل MS/MS للببتايدات تعود مرة أخرى لنفس الأصل وتكون مطلوبة قبل الختام الموثوق بأن البروتين الأصلي كان موجوداً في المحلول الأصلي تحت التحليل.

تطور حديث نسبياً في البروتين وهو القدرة على مقارنة البروتينات في عينتين بطريقة كمية وإنتاجية عالية (تنميط البروتين). يشتمل ذلك على التعليم المميز للببتايدات من عينة إلى عينة أخرى، والعينات غير المعلمة. هناك العديد من طرق التعليم أو التوصيف المتاحة ولكن من أشهرها استخداماً هي طريقة الوصف بالتمثيل الغذائي مع أيسوتوب الأحماض الأمينية المستقرة. إن نمو الخلايا في الـ 13C أرجينين arginine المعلم واليسين lysine ينتج عنه أن كل ببتايد سيكون أكبر من نفس الببتايد في الخلايا غير المعلمة كما في الشكل رقم(١٢،١٠). إن خلط المحاليل المعلمة وغير المعلمة سينتج عنه مضاعفة لكل مركبة ببتايد في الخليط. كل واحدة من القمتين يمكن محاصرتها وتعريضها ل

MS/MS ؛ وبالتالي يسمح بتحديد البيبتايد مع تكميم نسبي للبروتين الأصلي في المحلولين الأصليين كما في الشكل رقم(١٢،١٠).

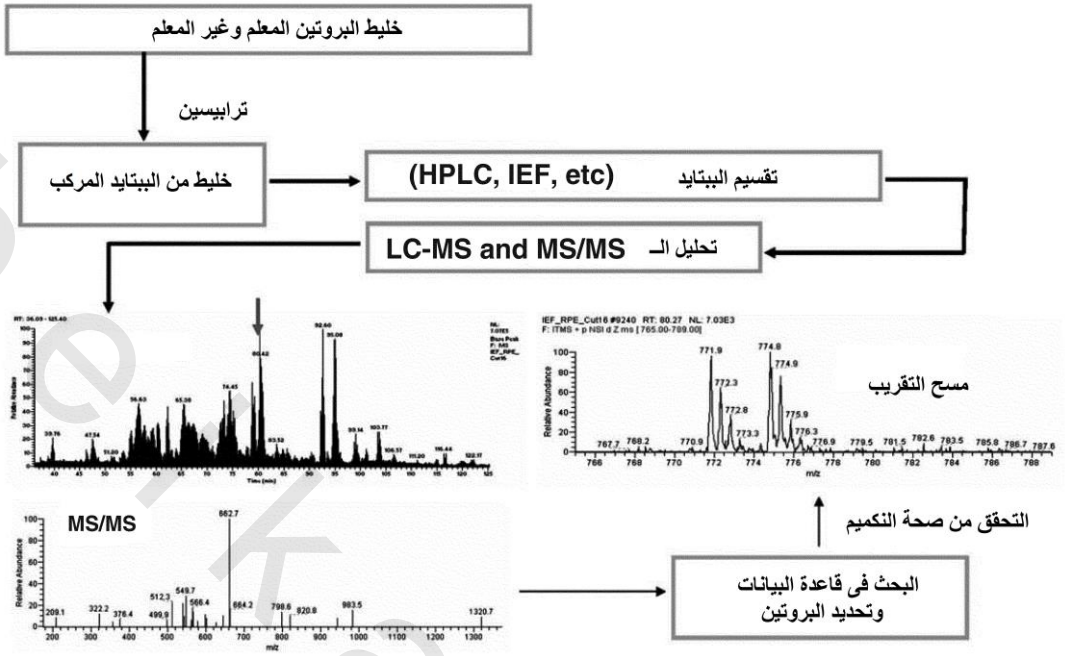


الشكل رقم (١٢،٩) آليات تقسيم البتايد في طيف ال MS/MS في العادة تنقسم الببتايدات حول روابط الأميد amide حيث ينتج عنها تتابع أيونى طرف أمينى amino (a, b, and c) وطرف كاربوكسى carboxy (y, x, and z)

(١٢،٤) المعلوماتية الحيوية والبنية التحتية للمعلومات

يمكن إنشاء المعلوماتية الحيوية والبنية التحتية للمعلومات محلياً أو موضعياً (مواقع في معامل معينة أو جامعات) لتحليل البيانات خلال معمل أو مجموعة، أو يمكن إجراؤها عن طريق مستودعات مركزية للبيانات مثل الـ NCBI. الأنظمة داخل المعامل أو داخل الجامعة يطلق عليها نظام إدارة المعلومات المعملية laboratory information management system, LIMS. يمكن لل LIMS أن يستضيف طرقاً مفصلة إلى حد ما لتتبع وتخزين المعلومات عن تصميم التجربة، وطرقها، واكتساب البيانات، ودراسة وتفسير البيانات. النظم المتقدمة من الـ LIMS .

تكون لها بوابة على الإنترنت إما للتعاملات العامة (الجمهور) على مجموعات جانبية من البيانات وإما بوابة لتصدير مجموعات مختارة من البيانات على مخازن معلوماتية عامة. كمثال على نظام متقدم لل LIMS الذي تم إنشاؤه لمصفوفات البيانات المتناهية الصغر microarray سيتم شرحه هنا.



الشكل رقم (١٢،١٠) التعليم أو التوصيف التفريقي للببتايدات في عينتين حيويتين يسمح بالتكميم النسبي للبروتين الأصلي . القمة القاعدية تبين أن الببتايدات تم الكشف عنها في مسح بمقياس طيف الكتلة MS . القمة الواحدة في مسحة قمة القاعدة (السهم الأحمر) يتم مدها بعد ذلك إلى تحديدية أعلى في مسح التقريب ، حيث القمتان تمثلان نفس الببتايد من العينتين الحيويتين (الببتايد المعلم وغير المعلم) . المتغيرات الأيسوتوبية لكل ببتايد تكون واضحة في مسح التقريب . الببتايدات الأحادية يتم محاصرتها بعد ذلك وتقسيمها لتحقيق مسح تقسيمي تعبيري لتتابع الحمض الأميني (MS/MS) كما في الشكل رقم (١٢,٩) . لنظرة أكثر شمولا عن هذا الشكل يمكنك زيارة الموقع المصاحب : http://books.elsevier.com/companions/9780123735836

إن قواعد بيانات الـ DNA ووحدات الـ DNA متعددة الأشكال DNA and single nucleotide polymorphism, SNP تكون صريحة ومباشرة نسبيا، بمعلومية الخطية (ثنائية الأبعاد) للبيانات الجينومية. نحن لن نشرح ذلك في هذا الجزء من النص، ولكننا نحيل القارئ إلى مصادر عامة متميزة مثل مستكشف الجينوم genome browser (www.genome.ucsc.edu) ومشروع الـ HapMap (www.hapmap.org) كأمثلة على ذلك.

سنركز في هذا الجزء على جوانب التحدي في المعلوماتية الحيوية والبنية التحتية المعلوماتية للتنميط الـ mRNA. لقد ظل تنميط الـ mRNA باستخدام إما مصفوفات الـ Affymetrix المتناهية الصغر (www.affymetrix.com) أو الـ cDNA الأكثر تحديداً وإما مصفوفات الـ oligonucleotide المتناهية الصغر في انتشار عملي لمدة حوالي ١٠ سنوات، مع عشرات الآلاف من نماذج المصفوفات المتناهية الصغر في النطاق الجماهيري أو العام. إن الاستخدام المكثف للمصفوفات المتناهية الصغر لتنميط الـ mRNA قد أدت إلى ثراء نسبي في المنشورات عن ضبط الجودة وخطوات

التشغيل القياسية، واستنتاج الإشارة، والمعلوماتية الحيوية، والتحليل الإحصائي (أحسن مجموعة عمل تدريبية لتحليل الأورام tumor analysis best practice working group 2004).

هناك مخازن بيانات عامة ممتازة لبيانات المصفوفات المتناهية الصغر لتنميط الـ mRNA مثلاً (ArrayExpress www.ebe.ac.uk/arrayexoress، والتعبير الجيني المتعدد المسار gene expression omnibus, GEO، http://www.ncbi.nlm.nih.gov/geo/) [3, 4]. على الرغم من مقدرتها على تخزين عدد كبير من المشاريع والكثير من البيانات التجريبية، إلا أنها تكون محدودة أيضاً. مع العديد من المشاركين، يكون في الغالب من الصعب التأكد من عملية التجميع الدقيق للبيانات التعريفية، وكيفية الوصول، والتشكيل المناسب للبيانات. أيضاً، فإن قواعد البيانات هذه تقبل العديد من ساحات العمل التجريبي المختلفة (مثلاً، طرق مختلفة لإجراء التنميط التعبيري: مصفوفات الـ cDNA، والـ Affymetrix arrays، والـ SAGE)، وأصبح من الصعب إن لم يكن مستحيلاً مقارنة التجارب خلال ساحات العمل. إحدى الطرق لتحقيق بعض الوسائل لمقارنة مجالات البيانات خلال التجارب وساحات العمل التجريبية كانت تطوير أقل معلوماتية عن معايير تجارب المصفوفات المتناهية الصغر [5. 6]. إن هذا لم يضع معياراً لساحات العمل التجريبي أو البيانات ولكنه يحاول توفير مجالات بيانات معينة يمكن تحويلها خلال مجموعات البيانات المختلفة وقواعد البيانات.

في الغالب تكون الـ LIMS الموضعية قادرة على التطوير الأكثر في المعايير التجريبية ولذلك فإنها توفر قدرة أكبر في طرق استعلام وتحليل البيانات. اثنتان من القواعد الأشهر استخداماً هما قاعدة بيانات استانفورد Stanford لمصفوفات البيانات المتناهية الصغر ([7]، http://genome-www5.stanford.edu/) ومصدر تنميط التعبير العام public expression profiling resource, PEPR في مركز طب الأطفال القومي في واشنطن D.C ([8,9]، http://pepr.cnmcresearch.org).

بالنسبة لل PEPR سنركز على تحسين ثلاثة جوانب لاكتساب البيانات وتحليل البيانات العامة:

تحسين عملية تجميع البيانات المحتملة بالكامل.

تحسين واجهة برمجة التطبيقات application programming interface, API بحيث تحول آليا بيانات الصور المعينة الخام للمصفوفات المتناهية الصغر (ملفات ال.cel) إلى تتابع من الإشارات الملخصة لكل مجموعة مسبار باستخدام خمسة خوارزميات لمجموعات المسبار.

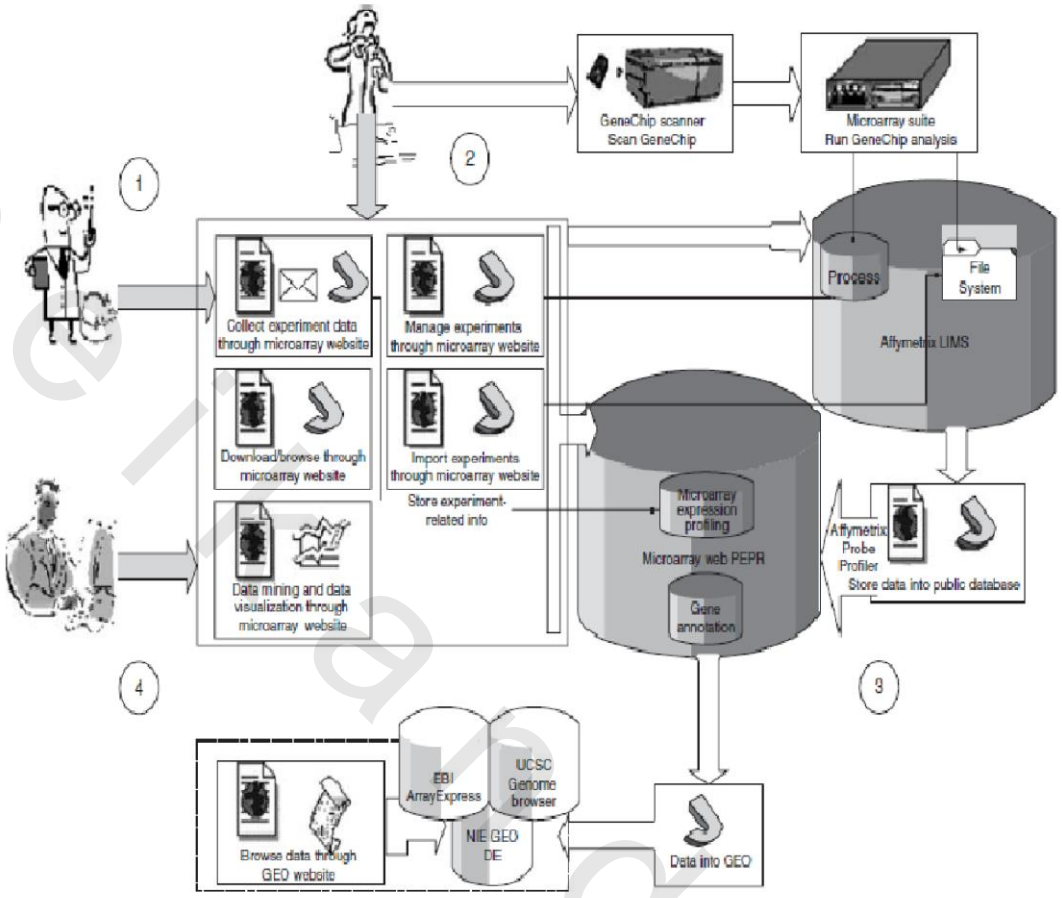
تحسين الواجهات العامة المبسطة للاستعلامات الديناميكية على الويب.

بالنسبة لعملية تجميع البيانات المحتملة، فقد صممنا وأنشأنا عملية حيث يبدأ منشئ مشروع المصفوفة المتناهية الصغر في التفاعل مع الـ PEPR مباشرة بعد المفهوم الأولى للتجربة. إن هذا يسمح بتنفيذ الاختبارات الآلية

والتوازنات على مجالات البيانات وعملية القبول عند الخطوات المختلفة في توليد البيانات وإطلاق سراحها للعامة. الشكل رقم(١٢.١١) يبين رسماً تخطيطياً لتصميم الـ PEPR.

لأي تطوير معلوماتي حيوي وحساب بيولوجي، يكون من الضروري معرفة أشياء أو مفاهيم بالنسبة لدقة هذه البيانات. بالنسبة لتجارب مصفوفات الـ mRNA، فإن هناك جدلاً معتبراً بالنسبة لكيفية التفسير المناسب لإشارات التهجين على المصفوفات واستنتاج إشارة مطبعة لكل نص لل mRNA على الشريحة خلال المشروع. بالنسبة لمصفوفات الـ Affymetrix (www.affymetrix.com)، فإن كل نص يتم استعلامه عن طريق مسبار أو مجس مع ١١ تطابقاً تاماً و 25mers مغطاة ضد آخر 500bp في النص و ١١ تطابقاً مزدوجاً 25mers تخدم كتحكم ممكن لارتباط غير محدد لمجس التطابق التام؛ لذلك فإن هناك على الأقل ٢٢ إشارة تهجين يمكن اعتبارها لكل نص، وطرق تفسير الإشارة من الضوضاء وطرق التطبيع خلال وبين المصفوفات المتناهية الصغر تعتبر مساحة بحثية نشيطة جداً (وحامية الجدل). إن المعلوماتية الحيوية والطرق الإحصائية المستخدمة لنقل ملف صورة محول إلى إشارات نصية لكل جين تسمى خوارزميات مجموعة المجس، وهناك العشرات من الطرق المتاحة في هذا الموضوع (انظر إلى Seo and Hoffman [10] للمراجعة).

إن تصميم الـ PEPR يفعل الدوال البحثية في البيانات الفوقية الغنية (بمعنى، البحث عن طريق تصميم نوع التجربة أو النموذج الحيواني العمر أو الجنس)، بما في ذلك نظام واجهة الويب لإدخال البيانات للحصول على معلومات التجربة مستقبلياً (وعن بعد مثلما يكون أحد الباحثين في السويد يبدأ في إدخال البيانات الفوقية ويصمم البيانات قبل البدأ في مشروع التنميط في واشنطن). على العكس من أي حزمة تنميط مستخدمة حالياً، فإن واجهة الويب لإدخال البيانات وعملية التقديم تعرض مرونة عالية للحصول على البيانات المطلوبة للتجربة (مثلاً، إضافة تصميم لنوع تجربة) للتحليل والعرض. إن ذلك يحقق آلية لفرض التناسق لبيانات الدخل والتحقق منها ويتخلص من الجداول الملحقة ومجموعات العمليات لترشيح البيانات. إن تناسق البيانات يزيد من المقدرة البحثية والمقدرة على العرض. يستخدم الـ PEPR أيضاً الـ GEO المنفذ حديثا أو الـ APIs المحدثة لتقديم تجارب جديدة أو تعديل تجارب سبق نشر بياناتها. يستخدم الـ PEPR تصميما معدلا للـ API للتحويل الآلي إلى كل المشاريع لخوارزميات مجموعة المجسات الخمسة (MAS5.0, dCHIP PMonly, dCHIP diff, ProbeProfiler PCA, RMA). إن هذا الـ API المرتبط مع خواص المواجهة مع المستخدم تسمح لمستخدم الويب العام أن يرى بسرعة تأثير خوارزم مجموعة المجسات على تفسير البيانات. أخيراً، فإن الـ PEPR يحقق تصديراً لمجموعات البيانات في غير الزمن الحقيقي التي تسمح للباحث بتنزيل/استيراد تتابع من مجموعات البيانات الكبيرة بينما يقوم بالتجول خلال الموقع، كما أن توليد ملفات البيانات.chp,.dat, and.cel يتم تنفيذها بعيدا عن أوقات الازدحام.



الشكل رقم (١٢،١١) تصميم هيكلي لموارد تعريف التعبير العام PEPR . من http://pepr.cnmcresearch.ofg.

بالنسبة لهيكل تصميم عملية الـ PEPR وتنفيذه، فإن الـ PEPR عبارة عن تطبيقات مشروعات جافا ثلاثية الطبقات تتكون من طبقة ويب، وطبقة وسطى، وطبقة النهاية الخلفية. تشتمل طبقة الويب على خادم ويب، وخادم تطبيقات Tomcat، ومكونات مختلفة للويب التي تحقق وظائف المقدمة مثل التجول، واستكشاف البيانات، والبحث عن البيانات، وتقديم المشروع، ونشر المشروع، وأدوات الاستفسار عن الجين، وملاحظات المستخدم. معظم مكونات مواجهة الويب تتقابل بشفافية مع قواعد البيانات في النهاية الخلفية. طبقات المواجهة هذه تسمح للمستخدمين ببدء التعامل مع الطبقة الوسطى للتطبيقات.

تتكامل الطبقة الوسطى مع العديد من الخدمات الخارجية. لبعض هذه الخدمات تم شراء إصدارات جديدة لبرمجيات كانت موجودة، وللبعض الآخر تمت كتابة عقود مخصوصة لل PEPR (Popchart, Lucene, Affymetrix SDK andCorimbia Probe Profiler SDK).

لقد تم تصميم هذا التطبيق للتعامل مع العمليات المستهلكة للوقت مثل استخلاص البيانات بالـ Affmetrix وتنزيل البيانات في غير الزمن الحقيقي بينما يسمح للمستخدم بالتجول في الموقع بدون انتظار لإتمام العملية. تحتاج تطبيقات الطبقة الوسطى إلى موارد حسابية مكثفة، وتنزيل بيانات في غير الزمن الحقيقي، وفهرسة بيانات للبحث عن كلمات مفتاحية، وتقديم بيانات NCBI GEO، واستخلاص وتحويل بيانات Affymetrix، وخليط تعريف المجسات لخوارزميات توليد البيانات.

معظم العمليات في هذه الطبقة لا تتطلب استجابة متزامنة من النهاية الأمامية لـ PEPR. بالإضافة لتطبيقات خواص نقرات الويب والانتظار العادية، فإن الـ PEPR يسمح للمستخدمين من تقديم طلبات بدون انتظار لاستكمال العملية حينما يكون من المؤكد أن هذه العملية سيتم استكمالها. لتحقيق هذه العملية غير المتزامنة بطريقة يعتمد عليها فإنه يتم تقديم خادم طابور مفتوح Open JMS في تنفيذ الـ PEPR، ويعمل ذلك على تحسين تطبيقات الـ PEPR الوظيفية. لقد تم تصميم الـ JMS ليقوم بتداول الرسائل بين مكونات الويب. عندما يقدم المستخدم طلبا بتنزيل مجموعة كبيرة من البيانات في الـ PEPR، فإن مكون ويب في خادم التطبيق Tomcat سيقوم بتغليف طلب المستخدم في صورة رسالة ثم يضع هذه الرسالة في طابور الـ JMS. إن طابور الـ JMS يكون هو المسئول عن استقبال وتوزيع الرسالة كموزع أو موجه خاص ينظر في عنوان الرسالة وإرسالها إلى الجهة المناسبة (بمعنى، عملية تنزيل البيانات في غير الزمن الحقيقي في المخطط). إن عملية تنزيل البيانات في غير الزمن الحقيقي تقوم بعد ذلك بالتوزيع والتعامل مع طلب التنزيل. إنها تستمر في بحث وضغط البيانات المطلوبة وبعد ذلك تقوم بإرسال ملاحظة عن وصلة الويب URL إلى المستخدم. في أثناء هذه العملية، فإن المستخدم ليس عليه أن ينتظر عملية ضغط الملفات المطولة أن تتم، فإن طابور الـ JMS سيجعل عملية تنزيل المجموعة ممكنا.

تتكون طبقة النهاية الخلفية من اثنتين من قواعد البيانات: قاعدة بيانات PEPR DB وقاعدة بيانات الـ Affymetrix LIMS DB. تقوم قاعدة البيانات PEPR DB بتخزين كل أنواع البيانات التعريفية أو الفوقية للمشاريع والتجارب وحدها مع قيمة تحليلية مصاحبة لأغراض التنقيب في البيانات في الزمن الحقيقي. تقوم قاعدة بيانات الـ Affymetrix LIMS DB بتخزين كل تعبيرات الـ Affymetrix التعريفية للبيانات الطبيعية ومعلومات عملية التقطيع.

العدد الكلي للمصفوفات المتناهية الصغر الشائعة حاليا في الـ LIMS الداخلي يساوي ٧٠٠٠ حيث الكثير منها يكون من العينات البشرية. إن موارد الـ PEPR العامة تكون مليئة بعدد ٢٨٢٧ من مصفوفات الـ Affymetrix وتوضح حوالي ٦٠٠٠ تنزيل تعريفي في الشهر من الـ PEPR. هناك طلب آلي لأنبوب تم تطويره مع الـ NCBI GEO والـ PEPR وهي المشارك #1 لحسابات الـ GEO ل ١٢٪ من كل ملفات الـ Affymetrix للفقاريات.

(١٢،٥) التنقيب عن البيانات وقواعد البيانات الحيوية الكبيرة الحجم

هناك تزايد كبير في تنوع مصادر البيانات في النطاق العام وتزايد في الأدوات المرنة التي بها يمكن التنقيب في قواعد البيانات الحيوية الكبيرة الحجم. مجموعات البيانات المتاحة للعامة قد أتاحت حلقة وصل بين علماء البيولوجيا المسئولون عن توفير البيانات الكبيرة الحجم وعلماء الحاسبات أو الإحصاء الذين يريدون الاتصال بهذه البيانات بغرض تطوير طرق لتحليل هذه البيانات.

من واقع خبرتنا، يوجد هناك خطئان شائعان يمكنهما أن يمنعا من تطور الحسابات البيولوجية. أولها: أن علماء الحاسبات والإحصائيين يكون لديهم ميلا لافتراض أن البيانات المتوفرة عن طريق علماء البيولوجيا تكون جيده على قدر إمكانهم (يعتمد عليها، ودقيقة، ونسبة إشارة لضوضاء جيدة). وهذا في العموم يكون غير حقيقياً، حيث إن علماء البيولوجيا يميلون أن يهملوا إخبار العلماء كمياً بذلك. ثانياً: أن علماء البيلوجيا يميلون لافتراض أن تحليل البيانات الذي يتم عن طريق علماء الحاسبات والإحصائيين قد حول البيانات إلى شيء يعتمد عليه، ودقيق، وله نسبة إشارة لضوضاء عالية، وهذا يعتبر أيضاً غير حقيقي حيث إنه يعاكس النظرية الكلاسيكية التي تقول " أدخل قمامة، تخرج قمامة".

إن مشكلة أدخل قمامة، تخرج قمامة تكون نادرة نسبياً مع التنقيب عن بيانات الـ DNA، أساساً بسبب البساطة النسبية للمشاكل (الأبعاد المنخفضة) والطبيعة الناضجة والتي يعتمد عليها لقواعد البيانات الجينومية. إن المشكلة تزداد وضوحاً مع زيادة التعقيد وعندما تصبح الطرق أقل قياسية ومتانة واعتمادية. من الأمثلة الجيدة على ذلك موضوع خوارزمات مجموعة المجسات لمصفوفات الـ Affymetrix الذي قدم سابقاً. إن عالم الحاسبات يقوم أساساً بتوفير إشارة لكل نص لكل مصفوفة في المشروع. هذه الإشارة تكون تحليلاً للعديد من الإشارات المركبة (٢٢ oligonucleotidv) مع التطبيع المكثف المقدم لجعل النص مقارناً بين المصفوفات. كل خوارزم لمجموعة مجسات يقوم بعمل العديد من الافتراضات المتعلقة بالتطبيع، والعقوبات من التهجين لإشارة المجس الغير مطابقة، وتقييم الخلفية [10-12]. حتى الآن فإن عالم الحاسبات نادرا ما يتم إخباره بهذه الافتراضات، والتأثير على تفسير البيانات التالية لا يتم تقييمه في الغالب.

لقد تم دراسة مشكلة خوارزم مجموعة المسبار في وضع افتراض لنسبة الإشارة للضوضاء لمشاريع مختلفة للمصفوفات [10, 13] والأسئلة المتعلقة بحسابات القدرة لأنواع وأنسجة معينة [14]. يختار معظم البيولوجيين خوارزم مجموعة مجسات واحد ومعين (مثلا: MAS5.0, PLIER, dCHIP, RMA) لتحويل مشاريع المصفوفات الخاصة بهم إلى مجموعة إشارات. إن اختيار خوارزم مجموعة المجس يعتمد على الاعتقاد بأن الواحد الذي يتم اختياره يهمل تنفيذ الأخرى التي لم يتم اختيارها، وذلك اعتماداً على ما نشر أو على خبراتهم. على العكس من

ذلك، فقد وجدنا أن مشروعات مختلفة للمصفوفات تتطلب استخدام خوارزميات مجموعة مجسات مختلفة لأن كل واحد من خوارزميات مجموعة المجسات يتأثر تفاضلياً بالمصادر غير المحكومة للضوضاء البيولوجية والتقنية (المتغيرات الحائرة). بعض خوارزميات مجموعة المجسات تكون محكمة أو منيعة نسبيا ضد الضوضاء المكثفة ولكنها بعد ذلك تكون قليلة الحساسية نسبياً. بعض خوارزميات مجموعة المجسات الأخرى تكون حساسة بشكل رائع، ولكن ذلك ينتج عنه نسبة عالية من الأخطاء الموجبة إذا كانت المتغيرات الحائرة (الضوضاء) عالية [14 ,10]. ونحن نقترح أن تكون طريقة توليد الإشارة (اختيار خوارزيم مجموعة المجسات) مفصلة أو مصممة لكل مشروع من مشاريع المصفوفات على حده.

من الواضح أن اختيار خوارزيم مجموعة بيانات معين يكون خطوة أساسية مبكرة في التفسير والتحليل المناسب للبيانات. إن خوارزميات مجموعة المجسات يكون لها تأثير عميق على البيانات المولدة اعتماداً على الافتراضات الضمنية لكل خوارزيم. في الحقيقة، إذا أخذ أحدهم مشروعاً محدداً للمصفوفات واستعلم عن التطابق أو الانسجام بين خوارزيمين، فإنه سيجد فقط ١٠٪ -٣٠٪ انسجاماً في الفروق التعبيرية الإحصائية المهمة. بصرف النظر عن هذا التوافق القليل، فإن العالم البيولوجي يختار خوارزيم مجموعة مجسات واحد، معتقدا أنه هو الأفضل، مع المعرفة القليلة أو المنعدمة عن الافتراضات الكامنة أو أي تقدير أو تخمين لكفاءة هذا الاختيار. بعد ذلك يأخذ عالم الحاسبات أو الإحصائي هذه الإشارات الناتجة من العالم البيولوجي. إنهم في العادة لا يهتمون بخوارزيم مجموعة المجسات المستخدم أو الافتراضات الكامنة وراء ذلك والتي تم الأخذ بها في توليد بيانات الإشارة المركبة. هذه الخطوة تنفذ عادة عن طريق العالم البيولوجي، ويتم توفير الإشارات الناتجة لعالم الحاسبات الذي يعتقد أن هذه تكون أرقاما قوية ودقيقة ويعتمد عليها.

إن خوارزميات مجموعة المجسات تعتبر خطوة واحدة على الطريق من عالم البيولوجيا إلى عالم الحاسبات، ولكنها تكون مثالية في الوادي الضيق العميق نسبيا الموجود بين النظامين. حيث أن البيانات البيولوجية تصبح كبيرة بصورة أسية وأكثر تعقيداً، فإنها تصبح أكثر أهمية لعبور هذا الوادي الضيق، حيث يكون البيولوجيون أكثر حرصاً من الطرق الحسابية وعلماء الحاسب أكثر حرصا من المتغيرات البيولوجية والتقنية التي تؤثر على البيانات المعطاة لهمم.

إحدى الطرق السهلة نسبياً لعبور هذه المشكلة هي بأن يصر الباحثين على الحاسب أن يتم إعطاؤهم بيانات خام بدلاً من بيانات معالجة. بالنسبة لمصفوفات الـ Affymetrix، فإن ملفات البيانات الخام تكون هي بيانات الصورة من المصفوفة (ملفات.dat) وحسابات كثافة التهجين غير المعالجة لكل واحدة من المليون خلية (oligonucleotides) على المصفوفة (ملفات.CEL). هذه الملفات تكون كبيرة نسبياً على جداول الملخص المتولدة عن طريق خوارزميات

مجموعة المجسات. هذه الأنواع للملفات كانت متاحة على الـ PEPR منذ عدة سنوات وهي متاحة الآن أيضاً للعديد من المشروعات في كل من NCBI GEO والـ ArrayExpress. إذا كان متخصص الحسابات البيولوجية سيتعامل مع البيانات الخام، فإنه من الممكن عمل اختبار نسبي لخوارزميات مجموعة المجسات (أو خوارزميات متعددة) ويمكن تفسير البيانات الناتجة بحساسية أكثر.

لقد بدأت فقط قواعد البيانات الكبيرة الحجم للبروتينات أن تكون مطورة، وهذا يجعل كل من البيولوجيا وتجميع البيانات أكثر تعقيداً. بينما تكون تجارب الـ RNA يتم تنفيذها عن طريق طحن كل العينة إلى تجمع من النصوص، فإن التجارب البروتينية تدرس غالبا حجيرات جوانب خلوية مختلفة (السيتوبلازم، والغشاء، والنواة)، مضيفاً أبعاداً أخرى لعملية اكتساب البيانات، والتخزين، والتفسير. إن تجارب مصفوفات الـ RNA يكون لها في العادة: عينة واحدة = مصفوفة واحدة (صورة)، بينما غالباً ما تشتمل التجارب البروتينية على الالتفاف العكسي لمحاليل البروتين المركب إلى أمدية من الكهربية والأوزان الجزيئية. أيضاً، فإن عملية اكتساب البيانات البروتينية في الـ MS تكون عملية خطية ضمنياً بالنسبة للزمن (على العكس من توليد البيانات العالي التوازي في التهجين على المصفوفات). إحدى التجارب المنفذة أخيراً في مركزنا اشتملت على اختبار استجابة الخلايا الأصلية لأحد الأدوية باستخدام التعريفات البروتينية (التعليم الأيسوتوبي المستقر، انظر ما سبق). لقد اشتملت هذه التجربة على متوالية زمنية، باستخدام مرضى وتحكمات من الذين كانوا قد سبق أو لم يسبق لهم أن أخذوا الدواء، مع تحديد التحليل ليكون مقصوراً على حجيرات جانب خلوية (الإندوبلازم الشبكي). لقد تطلبت التجربة إنتاجية عالية من الـ MS (مصيدة الأيونات بالرذاذ الكهربي) لكي تنفذ على مدار ٢٤ ساعة في اليوم، لمدة سبعة أيام في الأسبوع مع كل من بيانات الـ MS والـ MS/MS (التقسيم). إن كمية البيانات المجمعة تصل مئات الجيجا بايت لهذا التجربة الوحيدة. كيف سيكون ذلك موضوعاً في صورة قاعدة بيانات وكيف سيتم توفيره للاستخدام العام ؟ إن الخلافات بين خوارزميات مجموعة المجسات مع مصفوفات الـ RNA ستضع مرحلة من النقاش الأكثر تعقيداً والمتعلق بقواعد بيانات البروتين وطرق تحليل البيانات. للمرة الثانية، فإنه من الأفضل أن يحصل علماء الحاسبات على بيانات الـ MS الخام، ولكن تبقى لوجستيات قواعد البيانات والاتصال العام بها وتعقيد البيانات من التحديات الكبيرة.

(١٢،٦) طرق البيولوجيا المدفوعة بالأحداث، والمدفوعة بالزمن، والمحاكاة المهجنة

إن تتابع الـ DNA يكون ثابتاً نسبياً ويتغير في ظروف معينة فقط. مثلاً، فإن طفرة الجين والتعدد الشكلي تتغير من الحالة الأرضية (تتابع خطي عادي) ولكنها يمكن اعتبارها أحداثا منفصلة (غير متعلقة ببعضها) ومن السهل توصيفها، ووضعها في قاعدة بيانات، وتفسيرها. يمكن تعديل الـ DNA عن طريق الأستلة acetylation (تفعيل وقتي

للتلوين)، والمثيلة methylation (تعطيل أو عدم تفعيل أكثر استدامة)، والأستلة والمثيلة لجين معين في الغالب يكون لها المركبتان الديناميكية والاستاتيكية. فمثلاً، واحد كرموسوم X في كل خلية أنثى يتم تعطيله من خلال الانتشار الواسع للمثيلة methylation في التطور المبكر، وهذا يعتبر تغيراً استاتيكياً. على الجانب الآخر، فإن استجابة أي خلية لأي تحدٍّ في الوسط المحيط ينتج عنه تغيرات أستلة acetylation لجينات معينة لتفعل أو تعطل عملية النسخ (الدفع بالشيء، أو الدفع بالزمن). طرق الاتصال بحالة المثيلة methylation و الأستلة acetylation للجينات على مدى اتساع الجينوم بدأت في الظهور. بمجرد أن تصبح هذه الطرق ناضجة بما فيه الكفاية، فإن الـ DNA سيأخذ مركبة ديناميكية ستزيد من تعقيد قاعدة البيانات بدرجة كبيرة.

كما شرحنا مسبقاً، فإن تعريفات الـ mRNA وتجارب البروتين بهما تعقيد ضمني يجعل تفسيرهما تحدياً كبيراً. هناك طريقتان مفتاحيتان يمكنهما المساعدة في تفسير هذه البيانات العالية الأبعاد: المتواليات الزمنية وشبكات المعرفة.

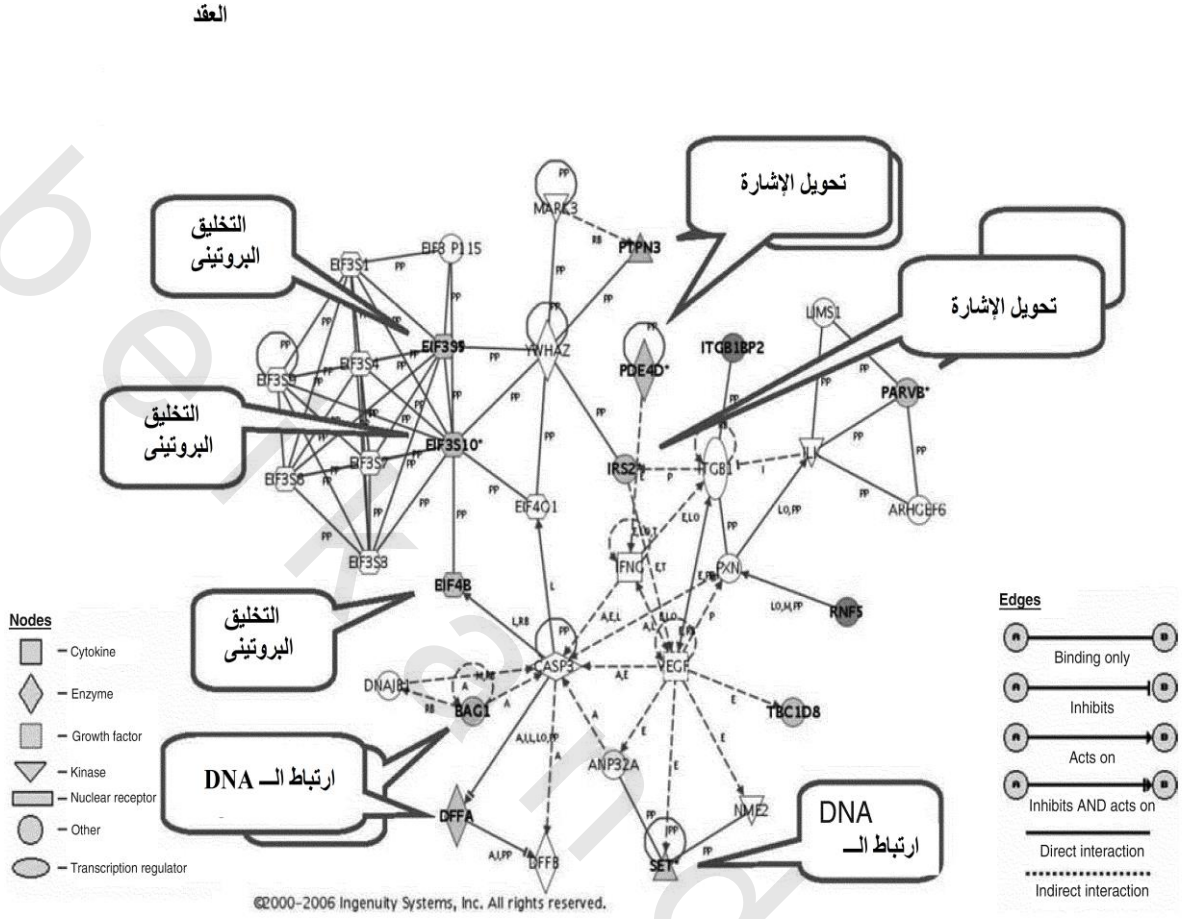
شبكات المعرفة هي قواعد بيانات الحاسبات المصحوبة بواجهات تجمع المعرفة البيولوجية الموجودة مبدئياً في صورة أدوات يمكنها أن تقبل مجموعات بيانات جينومية (مثل: تعبيرات معرفة الـ mRNA) وتساعد في تفسير البيانات في سياق المعرفة الموجودة مبدئياً. هناك العديد من أنواع شبكات المعرفة [15]. واحد من هذه المجموعات هو قاعدة بيانات جين الأورام، حيث يتم تصنيف الجينات والبروتينات المكودة إلى مجموعات محددة بالكيمياء الحيوية أو بالتتابع. طريقة أخرى هي تشفير المنشورات الموجودة إلى طرق مرور ذات صلة بيولوجية وشبكات وبعد ذلك نقارن البيانات الجينومية مع هذه الشبكات. مرة أخرى، هناك العديد من الموارد، ولكن أشهر اثنين استخداما هما مسارات التحليل البارعة Ingenuity Pathways Ananlysis (www.inginuity.com) و GenMapp(www.genmapp.org). مثلا: تخيل أن أحد البيولوجيين اختبر استجابة بعض خلايا السرطان لدواء مضاد للسرطان باستخدام تعبيرات الـ mRNA (المصفوفات متناهية الصغر). لقد تم تحميل الفروق النسخية الإحصائية (الجينات المستجيبة للدواء) في الـ Ingenuity، وبعد ذلك قامت الحزمة البرمجية بمقارنة قائمة الجينات المختلفة التعبير مع قواعد بيانات الجينات المتفاعلة والبروتينات. لقد تم حساب شبكات البروتين التي أوضحت أعلى نسبة تغيير في مجموعة بيانات دواء السرطان وتم ترتيبها حسب الأهمية الإحصائية. في هذه الحالة، وجد أن هناك شبكة بروتين تم وصفها مسبقاً لبرمجة موت الخلايا تحتوي ١٢٣ بروتين، مع النسخ الحثي ل ٨٥ من هذه البروتينات عن طريق الدواء المضاد للسرطان. إن الاحتمالية الإحصائية بالقيمة ١٢٣/٨٥ التي يمكن الكشف عنها بالصدفة تعتبر صغيرة جدا، ولذلك فإن شبكة "موت الخلية المبرمج" قد عادت للمستخدم على أنها "شبكة متغيرة عالية الأهمية". ولذلك؛ فالخلاصة هي أن الدواء المضاد للسرطان قد أثر على برمجة موت الخلايا في خلايا السرطان التي درست.

مثال على مقارنة أخذ العينات الحية (الخزعة) من عضلات أشخاص نحاف وسمان موضحة في الشكل رقم(١٢.١٢). في هذا التحليل، تم إدخال فروق التعبيرات باستخدام تشبع أهمية على برمجية الـ Ingenuity، وتم تحديد شبكة توضح أن نسبة عالية من النسخ قد تغيرت بدرجة كبيرة عن طريق حالة البدانة كما في الشكل رقم(١٢.١٢). هذه البروتينات الموضحة بالرموز الحمراء قد تم تحسن تنظيمها عن طريق البدانة وهذه التي بالرموز الخضراء قد تم إهمال تنظيمها.

برمجيات تحليل الشبكات تكون عالية الفائدة وطريقة عملية لتكثيف البيانات الجينومية إلى مسارات بيولوجية ذات صلة تحتاج إلى الدراسة المستمرة. إن الحساسية والتحديدية لهذه الحزم يجب اعتبارها مؤقتة عند أفضل وضع لها. تعتمد المسارات والشبكات بدرجة كبيرة على الخلايا، والأنسجة، والأعضاء تحت الدراسة، وهذه الحزم تفترض على العموم أن هناك تفاعلاً موصوفاً في المنشورات بين اثنين من البروتينات موصوفين في عين الفأر يكونان ذواتي صلة مع خلايا السرطان. بالإضافة لذلك، فإنه يجب افتراضها وسيلة غير حساسة نسبياً، حيث إنه نسبة ضئيلة جداً من الحقيقة البيولوجية بدلالة الشبكات والمسارات تكون معروفة حالياً وتم نشرها (ربما ١٪).

الوسيلة الثانية للمساعدة في التفسير هي المتواليات الزمنية. سلوك الجين أو البروتين كدالة في الزمن يسمح بالتقييم البيولوجي الظاهري الذي يمكن أن يساعد في تقليل الخطأ الموجب. معظم تجارب المصفوفات والبروتينات تكون صوراً لحظية عند نقطة واحدة في الزمن (مثلاً، الأشياء المتأثرة مع غير المتأثرة، انظر الشكل رقم(١٢.١٢) كمثال جيد). إنه من غير الممكن تحديد أي سبب/تأثير للاستجابات في تضخم البيانات في الشكل رقم(١٢.١٢). على الرغم من ذلك، فإنه إذا تم أخذ تتابع من الخزعات العضلية كدالة في الزمن (ربما بعد عملية تحويل معدٍّ أو نظام غذائي صارم)، فإن بعض الاستجابات النسخية المنظورة في الشكل رقم(١٢.١٢) يمكن تخصيصها لنقط زمنية سابقة عن الأخرى مؤدية إلى نماذج من الأسباب والتأثيرات.

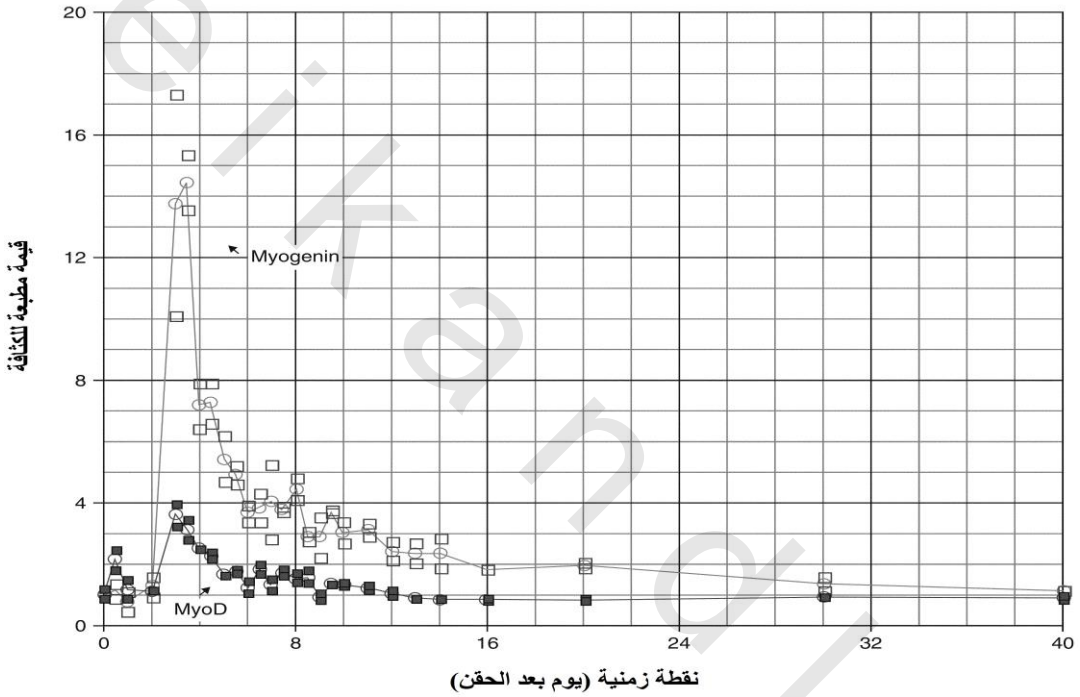
لعرض مثال على بيانات المتواليات الزمنية، فقد قمنا بوصف ٢٧ نقطة زمنية لمتوالية عضلية متكررة، بينما تم استدعاء التلف العضلي في نموذج الفأر وتم أخذ عينات عضلية كدالة في الزمن أثناء التعافي من التلف العضلي كما في الشكل رقم(١٢.١٣). يمكننا أن نرى أن معاملين للنسخ وهما الـ myogenin والـ MyoD كل منهما يؤثر بقوة على النسخ المسبب على مدار ٣ أيام أثناء التعافي. يوضح الفحص عن قرب للمتوالية الزمنية أن قمة الـ myogenin كانت نصف يوم بعد قمة الـ MyoD: وهذا يمكننا من بناء الفرض بأن الـ MyoD يتسبب في الـ myogenin (بمعنى أن الـ MyoD هو المنبع لل myogenin)؛ ولذلك يمكن إيجاد علاقة السبب والمسبب بين هذين النوعين من البروتين.



الشكل رقم (١٢،١٢) تحليل شبكة ال Ingenuity لل mRNA النصى المنظم تفاضليا عن طريق حالة البدانة . لنظرة أكثر تفصيلا لهذا الشكل عليك زيارة الموقع المصاحب : http://books.elsevier.com/companions/9780123735836

الاتجاهات المستقبلية في الحسابات البيولوجية ستركز على ربط البيانات من العديد من مجاميع البيانات ويمكن أن تحتوي على كل من صور وقتية (صور مقطعية) مسقطة وبيانات متوالية زمنية ، يمكن عملها على فصائل مختلفة. لقد تم أخيراً نشر مثالا على طريقة المساقط المتعددة [17, 18]. لقد كانت الخطوة الأولى هي عمل دراسة لصورة وقتية مقطعية، حيث تم أخذ ١٢٥ عينة حية من عضلة مريض من ١٢ مجموعة مرضية خضعت لتعريف الـ mRNA. التحليلات المعلوماتية الحيوية لهذه التعريفات تؤدي إلى شجرة علاقية للاضطرابات المختلفة كما في الشكل رقم(١٢،١٤). المرض تحت الاهتمام كانEmery Dreifuss muscular dystrophy, EDMD ، حيث يتعرض المرضى لطفرات في مكونات الغلاف النووي على الرغم من أن الفسيولوجيا المرضية لهذا الاضطراب ليست جيدة الفهم (صندوق أحمر في الشكل رقم (١٢،٤).

بعد ذلك تم أخذ جينات تشخيصية من نقطة الـ 6-EDMD في الشجرة وتم الاستعلام عنها في الـ الاضمحلال/التجديد في المتوالية الزمنية المكتملة في نموذج الفأر كما في الشكل رقم(١٢.١٣) وتم تحديد علاقة السبب والمسبب في المسار المنتظم الذي يشتمل على العديد من التعبيرات المختلفة للنصوص المحددة لل EDMD كما في الشكل رقم(١٢.١٥) [16]. لقد حدد ذلك نموذجاً مرضياً (فشل للأشياء أثناء التجديد العضلي) تم اختباره فيما بعد والتحقق منه في نموذج الفأر لـ EDMD [17].

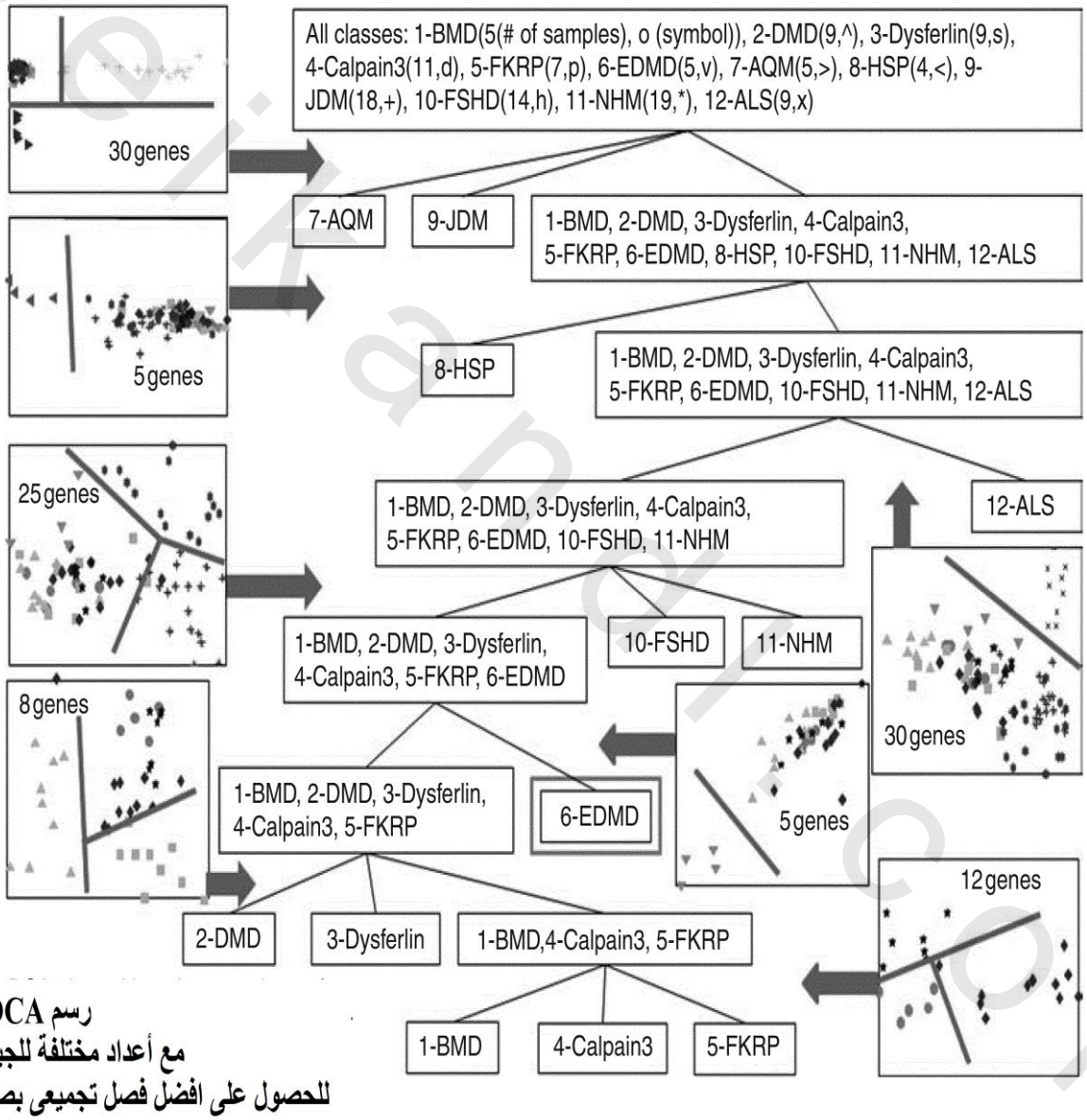


الشكل رقم (١٢،١٣) بيانات المتوالية الزمنية في التجديد العضلي في الجسم الحي (in vivo) . هذه البيانات من Zhao et al 2002 [18], and Zhao et al 2003 [19] وهي منشورة في ال PERP (http://pepr.cnmcresearch.org) .

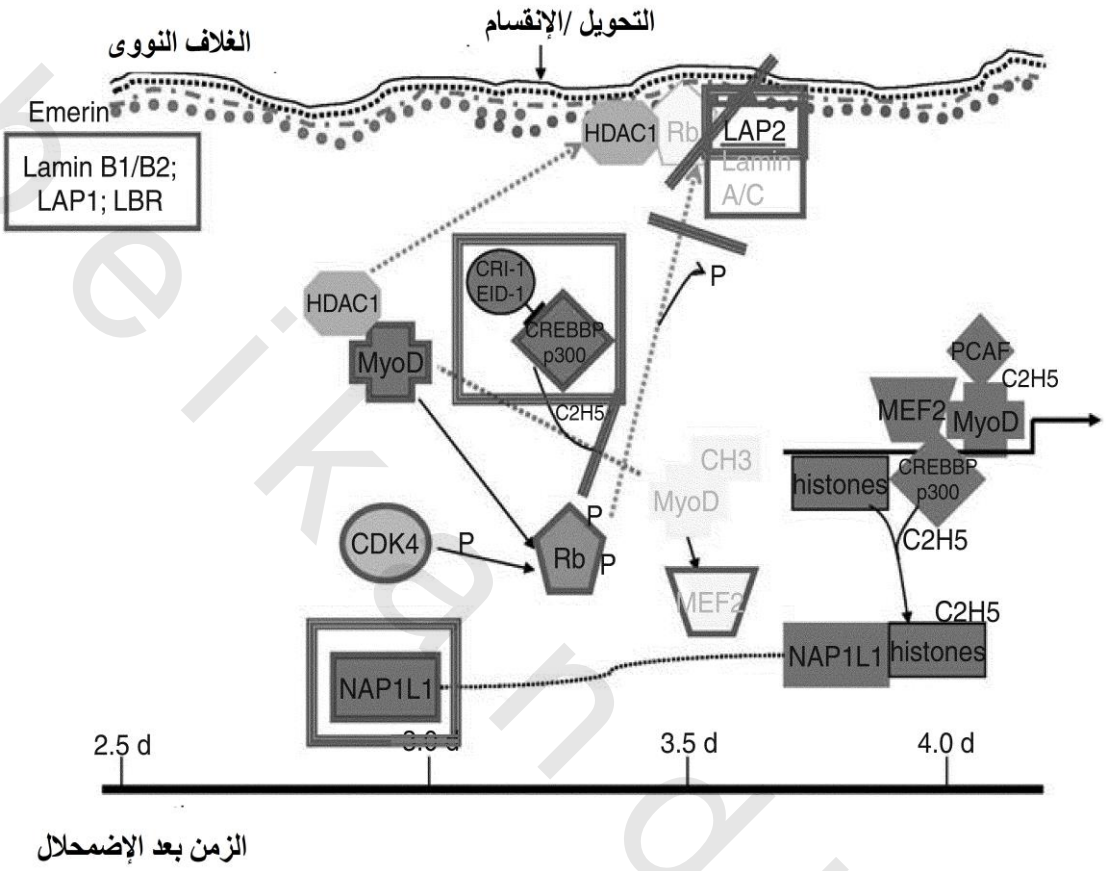
إن المعلوماتية الحيوية والحساب البيولوجي المصاحب لـ DNA أصبحت ناضجة بما فيه الكفاية. تسمح طرق التهجين بالتحديد الحساس والمحدد والتكميم لكل الجينومات في المصفوفات المتناهية الصغر الوحيدة التي تحتوي على الملايين من الخواص (oligonucleotide عند عناوين محددة في المصفوفة). لقد أوجد استخدام التتابع الخطي لـ DNA الجينومي في البشر والعديد من الأحياء الأخرى مئات من قواعد البيانات المصاحبة مثل، التعدد الشكلي لـ DNA، والتحويل لـ mRNA والـ EST (وحدات نصية أو جينات)، وحفظ التطور أو النشوء، وقواعد أخرى.

تعتبر الحسابات البيولوجية والمعلوماتية الحيوية لـ mRNA أكثر تحدياً نسبياً، حيث العديد من المتغيرات الجديدة تم تقديمها مثل التأثيرات من الوسط المحيط، والزمن، والمكان (في النسيج، وفي الجسم)، والربط البديل، وأشياء

أخرى. هناك أيضاً ساحات عمل مختلفة تجريبية أخرى مع كميات مختلفة من القياسات المتكررة والاكتساب المتين للبيانات والتسجيل الضمني لكل منها. إن هذا يجعل من الصعب بمكان تحديد معيار أو قاعدة يمكن عن طريقها مقارنة كل التجارب. لقد ظهرت المتواليات الزمنية الكثيفة التي بدأت تحدد مسارات السبب/التأثير، وذلك على الأقل مع شبكات التنسيخ المنظمة.



الشكل رقم (١٢،١٤) اختيار الجين التشخيصي في ١٢ مجموعة بيانات لعينات عضلية حية من مرضى بالضمور العضلي (انظر Bakay et al [16]) . لتفاصيل أكثر عن هذا الشكل عليك زيارة الموقع المصاحب :
http://books.elsevier.com/companions/9780123735836



الشكل رقم (١٢،١٥) نموذج للفسيولوجيا المرضية الجزيئية لنوع الضمور العضلي يشتمل على طفرات للغلاف النووي . المحور X (أسفل) يبين بيانات متوالية زمنية من تجديد الفأر (انظر شكل (١٢–١٣)) . باقي الشكل يبين الحث الزمني للمسارات النسخية أثناء التجديد العضلي على مدار زمن التحويل من الخلايا النشطة الانقسام إلى المايوتيوبس المتفاضلة بعد التقسيم (التحويل الانقسام/بعد الانقسام) . البروتينات الموجودة في المربع باللون الأحمر هى التي تم تنظيمها تفاضليا في مرضى ال EDMD ، بينما المظللة بالشرط الصليبية الحمراء تبين البلوكات الجهدية في هذا المسار الجزيئي نتيجة الطفرات في الغلاف النووي . معدلة من Bakay et al [16] . للحصول على تفاصيل أكثر عن هذا الشكل عليك زيارة الموقع المصاحب :

http://books.elsevier.com/companions/9780123735836

(١٢،٧) الملخص

إن البروتينات تكون أكثر تعقيداً من نماذج الـ mRNA بعدة درجات، مع التعديلات بعد التحويل، والتحديد الخلوي، والمشاركات المرتبطة، كل ذلك يملي أو يحدد النشاط البروتيني والوظيفي. إن البروتينات ذات الإنتاجية

العالية تتقدم نحو القدم مع حلول الـ MS ذات التحديدية العالية وقواعد بيانات المطابقة الطيفية المصاحبة. إن التعريفات البروتينية باستخدام المحاليل المعلمة تفاضلياً للببتايدات قد وصلت إلى الاستخدام الواسع الانتشار، بينما طرق المعلوماتية الحيوية والحسابات البيولوجية تبدأ فقط في التطوير.

تشتمل التحديات المستقبلية في الحسابات البيولوجية على تحديد المسارات المحددة خلوياً ونسيجياً والشبكات واستجابة الشبكات للتحديات الفسيولوجية والوسطية. سيتم التركيز على تكامل مجموعات البيانات وقواعد البيانات لـ DNA، والـ mRNA، والبروتينات، مع المحاولات لحشد الدعم لإنشاء الشبكات بينما يتم تحديد الشبكات من خلال دمج النماذج الحسابية والتحقيقات التجريبية.

(١٢،٨) تمارين

١ - ما مدى سهولة دراسة الـ DNA بالمقارنة مع الـ RNA وبالمقارنة مع البروتينات ؟

٢ - ما هي العوامل المؤثرة على استخدام الوحدات النصية عن طريق الخلايا ؟

٣ - لماذا يكون من الصعب دراسة البروتينات ؟

٤ - ما هي أساسيات البروتينات ؟

٥ - ما هي مميزات وعيوب الجيلاتين الثنائي الأبعاد والطلقة النارية ؟

٦ - ما هما المفهومان الخاطئان الشائعان اللذان يمكنهما تعطيل التطوير في الحسابات البيولوجية ؟

٧ - ما هو الموضوع الأساسي في خوارزم Affy لمجموعة المجسات مع اعتبار نسبة الإشارة/الضوضاء؟ كيف يستطيع أي شخص عبور هذه المشكلة؟

٨ - صف شبكات المعرفة ؟ وسمِّ اثنتين من هذه الشبكات الشائعة الاستخدام.

٩ - أذكر بعض التحديات المستقبلية للحساب البيولوجي ؟

١٠ -ما هو نظام الـ LIMS، وما هي استخداماته ؟

(١٢،٩) المراجع

1. J. Kent et al. Genome Browser (http://www.genome.ucsc.edu), 2000.
2. J. Nazarian, Y. Hathout, and E. P. Hoffman. The proteome survey of an electricity-generating organ (Torpedo californica electric organ). Proteomics. 7(4):617–627, 2007.
3. H. Parkinson et al. Array Express—A public database of microarray experiments and gene expression profiles. Nucleic Acids Res. 35(Database issue):D747–750, 2007.
4. T. Barrett et al. NCBI GEO: Mining tens of millions of expression profiles Database and tools update. Nucleic Acids Res. 35 (Database issue):D760–765, 2007.
5. D. W. Galbraith. The daunting process of MIAME. Nature. 444:31, 2006.
6. T. F. Rayner et al. A simple spreadsheet-based, MIAME supportive format for microarray data: MAGE-TAB. BMC Bioinformatics. 7:489, 2006.
7. J. Demeter et al. The Stanford microarray database: Implementation of new analysis tools and open source release of software. Nucleic Acids Res. 35(Database issue):D766–770, 2007.

8. J. Chen et al. The PEPR Gene Chip data warehouse, and implementation of a dynamic time series query tool (SGQT) with graphical interface. Nucleic Acids Res.32(Database issue):D578–581, 2004.
9. R. R. Almon et al. In vivo multi-tissue corticosteroid microarray time series available online at Public Expression Profile Resource (PEPR). Pharmacogenomics. 4(6):791–799, 2003.
10. J. Seo and E. P. Hoffman. Probe-set algorithms: Is there a rational best bet? BMC Bioinformatics. 7:395–410, 2006b.
11. B. Carvalho et al. Exploration, normalization, and genotype calls of high density oligonucleotide SNP array data. Biostatistics. Dec 22, 2006 [Epub ahead of print].
12. R. A. Irizarry, Z. Wu, and H. A. Jaffee. Comparison of affymetrix gene chip expression measures. Bioinformatics. 22:789–794, 2006.
13. J. Seo et al. Interactively optimizing signal-to-noise ratios in expression profiling: project-specific algorithm selection and detection p-value weighting in Affymetrix microarrays. Bioinformatics. 20(16):2534–2544, 2004.
14. J. Seo, H. Gordish-Dressman, E. P. Hoffman. An interactive power analysis tool for microarray hypothesis testing and generation. Bioinformatics. 22:808–814, 2006a.
15. Z. Fang et al. Knowledge guided analysis of microarray data. J. Biomed. Informatics 39: 401–411, 2006.
16. M. Bakay et al. Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb–MyoD pathways in muscle regeneration. Brain 129(Pt 4):996–1013, 2006.
17. G. Melcon et al. Loss of emerin at the nuclear envelope disrupts the Rb1/E2F and MyoD pathways during muscle regeneration. Hum. Mol. Genet. 15:637–651, 2006.
18. P. Zhao et al. Slug is a novel downstream target of MyoD:Temporal profiling in muscle regeneration. J. Biol. Chem. 277:20091–20101, 2002.
19. P. Zhao et al. In vivo filtering of in vitro expression data reveals MyoD targets. C. R. Biol. 326:1049–1065, 2003.

Bibliography

H. Parkinson. Tumor Analysis Best Practices Working Group. Expression prowling—Best practices for data generation and interpretation in clinical trials. Nat. Rev. Genet. 5:229–237, 2004.