EDITORIAL

Hacia una mejor comprensión del proceso de integración de los recursos bioinformáticos

Toward a better comprehension about the integration program of the bioinformatic resources

M.Sc. Carlos M. Martínez Ortiz

Grupo de Desarrollo de la Bioinformática. Centro de Cibernética Aplicada la Medicina (CECAM). E-mail: cmmo@infomed.sld.cu

El proceso de integración de los recursos bioinformáticos ha estado ocurriendo aceleradamente durante ya varias décadas, casi desde el comienzo mismo de la creación de los primeros repositorios y herramientas de análisis bioinformáticos. El propio canon fundamental de la biología molecular, es decir, uno o varios genes (nucleótidos, secuencia/estructura) →proteínas (aminoácidos, secuencia/estructura) →función biológica (reacción, proceso y componente), en el cual datos de diferente naturaleza se integran a través de relaciones de diversa índole, exigía este proceso de integración, la compresión del todo a través de sus partes relacionadas.

En su forma más simple veámoslo así: cuando hacemos investigación biomédica leemos artículos que nos informan del tema, sus problemáticas y nos permiten hacer informes al respecto. A su vez, en nuestro intento de producir datos, nos enlazamos a otros datos, referenciados en los propios artículos que leemos aunque hospedados en bases de datos diferentes. Paralelamente utilizamos (a través de interfaz web o de web-services) o descargamos (librerías de código) herramientas de análisis que nos permiten procesar y analizar estos datos y que eventualmente podemos integrar a las nuestras. Cada etapa en esta secuencia de acciones es recursiva en sí misma, deteniéndose sólo cuando nuestras hipótesis logran un grado de confirmación adecuado, produciéndose lo que asumimos como un aporte al conocimiento en la materia.

Servicios de este tipo los ofrece un modelo integrador como el del Centro Nacional para la Investigación Biotecnológica (NCBI, siglas en inglés) de los Estados Unidos. El NCBI (http://www.ncbi.nlm.nih.gov/) almacena y actualiza periódicamente información referente a secuencias genómicas en GenBank (que en sí mismo obedece a un modelo integrador centralizado, nutriéndose de dos bases de datos más, DDBJ y la del EMBL), un índice de artículos científicos referentes a biomedicina, biotecnología, bioquímica, genética y genómica en PubMed, una base de datos de enfermedades genéticas humanas en OMIM, además de otros datos biotecnológicos de relevancia en diversas bases de datos. El modelo funciona básicamente a través de protocolos de hipertexto (hiperlinks) y su interfaz de búsqueda para todas las bases de datos que indexa es el Entrez. Además contiene un gran número de herramientas para el análisis de secuencias biológicas, de las cuales el BLAST es quizá el más usado.

Al existir integración en los datos, el paisaje que se le brinda al investigador es mucho más amplio y mejor estructurado permitiéndole tomar mejores decisiones en un tiempo significativamente menor.

Esta es la vista que se ofrece desde la perspectiva del usuario, no obstante, desde la perspectiva del desarrollador, el responsable de hacer tangible dicha integración, el paisaje se torna mucho más complejo, el reto es enorme y, a pesar de los excelentes e imponentes ejemplos de integración que se observan en la actualidad, los resultados obtenidos hasta ahora distan mucho de lo que se espera en este sentido.

El incremento exponencial de los datos biológicos, complejos y heterogéneos en cuanto a formatos y tipos, hace difícil la tarea de mapear estos objetos y hacerlos accesibles de forma integrada y flexible al usuario. Se trata, entre otras cosas, de crear motores de búsqueda inteligentes que extraigan la información de repositorios en diferentes sitios, la procesen y la vuelvan a presentar en un nuevo recurso, pero esta vez integrada desde el punto de vista estructural y conceptual. Las bases de datos que hospedan esta información emplean estructuras de datos y vocabularios específicos convirtiendo el proceso de integración en una tarea nada trivial. No existe un enfoque universal en los modelos de integración de datos biológicos (hasta ahora hemos mencionado dos: centralización e hiperlinks) y las actuales metodologías que abordan este problema están en constante desarrollo.

La biología de sistemas, establecida como disciplina académica desde el año 2000, es quizá la ciencia donde este proceso de integración se vuelve imprescindible y su formalización a través de modelos de redes de interacción es necesariamente más exhaustiva y rigurosa. Se trata de formular modelos generalizados que permitan entender integralmente los sistemas biológicos y la interacción entre ellos, observándolos dentro de la dinámica de un contexto en particular. Aunque la integración a la que nos hemos referido es mucho más diversa, podemos usar esta disciplina como ejemplo arquetípico del potencial que ofrece el proceso de integración de datos biológicos, en la cual se trata de entender el comportamiento celular a través de interacciones espacio-temporales entre componentes celulares tales como genes, proteínas, metabolitos y organelos, reduciendo la dimensionalidad de los datos que se exponen para producir información valiosa del sistema sujeto a observación.

Veamos un ejemplo, desde la perspectiva de los genes, que se puede clasificar dentro del modelo de integración de conjuntos de datos. Las tecnologías de Análisis en Serie de Expresión de Genes (SAGE) y de microarreglos de ADN permiten medir simultáneamente los niveles de expresión de miles de genes en un tejido particular. El cáncer, por ejemplo, es el resultado de cambios en la secuencia de DNA.

Estos cambios se ven reflejados en los niveles de expresión de genes que directa o indirectamente son regulados por los genes mutados. Como resultado, la comparación y análisis de perfiles de expresión génica de los tejidos normales y los afectados por el tumor, van dirigidos a profundizar en el conocimiento sobre la etiología molecular de esta enfermedad.

El Mapa del Transcriptoma Humano (HTM, http://bioinfo.amc.uva.nl/HTMseq) es un recurso bioinformático que fue diseñado con este propósito y la pregunta que le dio origen es un ejemplo nítido de un empeño integracionista: "¿Es posible desarrollar una herramienta que permita identificar genes candidatos en regiones cromosómicas relacionadas con la formación de neuroblastomas (u otros tipos de cáncer) partiendo de perfiles de expresión génicas?" Para responder esta pregunta, el HTM integra los datos de las posiciones de genes humanos en los cromosomas, fruto de los proyectos de secuenciación y mapeo físico del genoma humano, con los perfiles de expresión suministrados por librerías SAGE construidas como parte del Proyecto de Anatomía Genómica del Cáncer (CGAP). Aunque parece un simple mapeo de objetos, en este caso posiciones y perfiles de expresión, la aplicación es mucho más compleja y añade algoritmos de análisis de secuencia, métodos estadísticos que hacen inferencia sobre los datos y un sistema de bases de datos relacional que permite la integración con otros recursos bioinformáticos de carácter público. Entre estos recursos se encuentran las bases de datos GenMap, UniGene y RHdb. Actualmente el alcance de HTM va más allá de la identificación de genes candidatos relacionados con el cáncer, brindando una visión mucho más holística de la organización del genoma humano.

Ocurre así también en la proteómica funcional donde las tecnologías actuales han permitido representar redes de expresión de proteínas que brindan información sobre la co-regulación de estas moléculas y sus respuestas bajo condiciones específicas. Esta información no es totalmente informativa sin indagar en la función biológica de estos productos de los genes, por lo que para solucionar este problema se necesita conocer qué otro componente celular interacciona con ellas y es aquí donde surgen las redes de interacción de proteínas. La función de una proteína no es completamente entendida hasta que se conoce su papel en las rutas celulares y su interacción con otros componentes como DNA, RNA, metabolitos, lípidos y otras proteínas.

Además de los mencionados, otros esquemas de integración se han empleado para la creación de este tipo de recurso bioinformático entre los que podemos mencionar Almacenes de Datos o Warehousing (en Pathway Commons y STRING) e Integración de Vistas (en BioZon).

Pathway Commons (http://www.pathwaycommons.org) incluye reacciones bioquímicas, complejos moleculares, eventos de transporte y catálisis, e interacciones físicas donde participan proteínas, DNA, RNA, entre otros. Permite recopilar toda esta información e integrarla en un formato estándar. Contiene datos de 9 bases de datos con 1400 rutas biológicas y 687000 interacciones. STRING (http://string-db.org) por su parte integra igualmente redes de interacción de proteínas tanto experimentales como predichas y brinda grados o puntuaciones de confiabilidad para cada una de las interacciones.

BioZon (http://biozon.org), es un sistema para la unificación, gestión y análisis de datos biológicos heterogéneos. Unifica múltiples bases de datos de objetos como (DNA, proteínas, interacciones y rutas celulares). Emplea un esquema de grafo fuertemente conectado y a su vez envuelto en una ontología jerárquica de relaciones y documentos. También implementa un sofisticado algoritmo de consultas que abarca múltiples tipos de datos.

Veamos ahora otro tipo de integración, el Sistema de Anotación Distribuido (DAS), que se puede clasificar entre los modelos de integración federados. El modelo federado se refiere a múltiples bases de datos interconectadas a través de la red e integradas de forma transparente en el sistema, brindando así un único punto de entrada para la formulación de consultas de datos. Las anotaciones genómicas son información de variada naturaleza relacionadas con la función biológica, caracterizando las secuencias genómicas en posiciones específicas. En el 2001, Dowell y colaboradores se percataron de que la realización de anotaciones genómicas no podía seguirse haciendo por un grupo centralizado debido al crecimiento exponencial de estos bancos de secuencias. El DAS surgió como solución a este problema y permite que la anotación de secuencias quedara descentralizada entre múltiples anotadores (Ensembl, UniProt, InterPro, UCSC, CBS) e integrada por diversos software clientes (Ensembl, Gbrowse, Dalliance, IGB, entre otros) comunicándose con sus respectivos servidores a través del estándar XML. Hasta el momento existen más de 1000 fuentes de anotaciones genómicas (http://www.dasregistry.org), lo cual demuestra la expansión que ha sufrido el sistema para poder cubrir estas necesidades.

A modo de conclusión podemos decir que las herramientas y modelos de integración de recursos bioinformáticos, dada su capacidad de brindarnos una visión del todo y sus partes relacionadas, incluso en la dinámica de un contexto específico, nos permiten no solo confirmar hipótesis, usualmente el paso lento en el proceso de investigación, de una forma significativamente más rápida, sino, y quizá más importante y curioso aún, generar nuevas hipótesis, de mejor calidad y en una medida significativamente mayor. Por otro lado, más allá de los intereses particulares de una investigación aislada, la integración de los recursos permite compartir esta información entre laboratorios a modo de evitar la innecesaria duplicación en los experimentos. El reto actual y futuro consiste en seguir perfeccionado estos modelos de integración y como fin último lograr la integración entre los diversos recursos, comportándose como subsistemas con personalidad propia, en sistemas de mayor alcance y poder de discernimiento.

Recibido: 10 de diciembre de 2013.
Aprobado: 6 de enero de 2014.