Bioinformática: una oportunidad y un desafío

Bioinformatics presents both an opportunity and a challenge

Emiliano Barreto Hernández[1]

Recibido: mayo 9 de 2008 Aprobado: mayo 29 de 2008

En los últimos diez años las diferentes áreas de la biología han tenido que lidiar con nuevas metodologías provenientes del área de la computación, tales como el desarrollo de nuevos algoritmos y aplicaciones matemáticas, diseñadas especialmente para la integración y el análisis del cúmulo de datos que se han generando de la aplicación de las metodologías de alto rendimiento en la investigación biológica, desarrolladas paralelamente con el proyecto genoma humano (The Internacional Human Genome Sequencing Consortium, 2001) en áreas como la genómica, la transcriptómica, proteómica, metabolómica y otras ómicas.

Los datos biológicos crecen exponencialmente

El crecimiento de los datos biológicos, que han pasado de 606 secuencias de ADN almacenadas en 1982, a más de 82 millones hoy en día, fue impulsado por el desarrollo de la técnica Reacción en Cadena de la Polimerasa (PCR) en el año de 1986 (Mullis, 1990), y ahora por la aparición de las denominadas nuevas tecnologías de alto rendimiento en experimentación biológica.

Los vertiginosos avances en el desarrollo de la siguiente generación de las tecnologías de secuenciación han llevado a producción de equipos comerciales de altísimo rendimiento y economía, capaces de producir diariamente enormes cantidades de datos, como es el caso de la secuenciación masiva en paralelo de *Illumina (Solexa)* que con su "Genome Analyzer" puede secuenciar en una simple corrida más de un billón de bases (Hall, 2007). Las primeras aplicaciones de estas nuevas metodologías de secuenciación incluyen análisis de genómica comparativa, genómica de microorganismos, la detección de SNP de alto rendimiento, secuenciación y análisis de ARN pequeños,

1 Ph D. c, Ms C. Profesor Asociado, Director Grupo de Bioinformática, Instituto de Biotecnología, Universidad Nacional de Colombia – Sede Bogotá.

identificación de mutaciones de genes en rutas metabólicas asociadas a enfermedades, análisis transcriptomas de organismos de los que se tiene muy poca información (metagenómica de microorganismos no cultivables, por ejemplo), determinación de genes con bajas tasas de expresión en sus ambientes naturales, y la obtención de información sobre la poco estudiada variación genética a nivel de especies, poblaciones y ecosistemas.

Los datos obtenidos por las nuevas técnicas de secuenciación, además de ayudar al descubrimiento de mejores estrategias para el diagnóstico, tratamiento y prevención de enfermedades, jugarán un papel clave en hacer realidad la medicina personalizada ya que, relativamente pronto, se espera alcanzar la meta de secuenciar un genoma humano por 1000 dólares, haciendo que bases de datos como el GeneBank, EMBL y DDBJ, ahora cuenten con un nuevo conjunto de datos con más de 27 millones de secuencias adicionales provenientes de la secuenciación de genomas usando la técnica de Whole-Genome Shotgun (WGS) (Benson et ál., 2008).

La masiva cantidad de datos que se generarán en el futuro inmediato por la utilización de las nuevas tecnologías de secuenciación, unida a la que diariamente se produce por la aplicación de metodologías como los microarreglos, por medio de la cual en un solo experimento se pueden generar millones de datos, almacenados en bases de datos como ArrayExpress (EBI, 2008), o la obtención de datos provenientes de la espectrometría de masas en tándem para la detección de proteomas completos, entre muchas otras, conllevan necesariamente al reto de diseñar procesos que permitan almacenar, actualizar y poner a disposición de otros investigadores estos datos de manera permanente y confiable.

Los grupos de investigación dedicados a la bioinformática tendrán que desarrollar formas distribuidas de almacenamiento para sus bases datos a fin de hacer más fácil su almacenamiento local, ya que se espera que pasen muy rápidamente del orden de 103 millones (Gigas bites) a 106 millones (Tera bites), y aunque los costos del almacenamiento siguen disminuyendo a nivel mundial, la disponibilidad de almacenamiento que se requiere en un futuro implicará costos elevados, y por ello se hará necesario una mayor optimización de este recurso. Así mismo, se prevé que tendrán que implementarse, a través de Internet, estrategias de comunicación con distribuidores de datos como el Centro Nacional de Información de Biotecnología de los Estados Unidos (NCBI), o el Instituto Europeo de Bioinformática (EBI), cada vez más eficientes y veloces, para permitir la actualización, sincronización y consulta permanente de las bases de datos, y que funcionen adecuadamente dentro de las limitaciones de la cada vez más congestionada Internet, y el enorme volumen de datos que implican las principales bases de datos biológicos.

Nuevas estrategias de análisis

La bioinformática dio un giro en su enfoque a mediados de los años noventa, perfilándose como un área de investigación gracias al desarrollo de proyectos de secuenciación como el del genoma humano (The Internacional Human Genome Sequencing Consortium, 2001), y de otros organismos importantes en las áreas de la salud y la industria, lo cual propició el desarrollo de herramientas bioinformáticas que se han utilizado para realizar estudios en la organización de los genomas, el "descubrimiento" de genes, la relación entre las mutaciones y la alteración de la función bioló-

gica o la evolución de diferentes organismos, entre muchas otras, que han facilitado el desarrollo de disciplinas como la genómica, la proteómica y la filogenia. El desarrollo de herramientas bioinformáticas se está proyectando al diseño y la generación de sistemas eficientes de almacenamiento y nuevos modelos para la comparación y análisis de las distintas clases de datos biológicos, rápidos y confiables, de los resultados desde el punto de vista estadístico, como es el caso del algoritmo BLAST (Altschul et ál., 1990), tal vez el más utilizado en la actualidad para comparar en unos cuantos segundos una secuencia contra los millones de secuencias almacenadas en una base de datos como GeneBank.

La utilización de tecnologías de alto rendimiento en investigación biológica, lleva a que las actuales estrategias de análisis requieran de su adaptación o de nuevos desarrollos, como ocurre en el caso de las nuevas metodologías de secuenciación, ya que las estrategias de análisis para la obtención de las secuencias de ADN que están basadas actualmente en la química de Sanger y su asignación de bases (Base Calling), no son adecuadas. Estas nuevas metodologías se caracterizan por un cubrimiento relativamente profundo, longitudes de lectura cortas y altos porcentajes de error en las secuencias, lo que requiere de nuevas formas de asignación de las bases, de ensamblaje de las secuencias y alteración de los métodos estadísticos para la determinación de puntajes de calidad de las mismas.

Los bajos costos y la rapidez de estas nuevas metodologías de secuenciación harán que los investigadores tengan acceso a un número cada vez más grande de genomas de microorganismos, que en la actualidad se aproxima los 1000 genomas secuenciados, y se cuenta con el muestreo de más de 100 condiciones de ambientes diferentes utilizando aproximaciones metagenómicas, haciendo necesarias nuevas aproximaciones más precisas, consistentes y de alto rendimiento a la anotación y el descubrimiento de la función biológica, para que esta gran cantidad de datos de diversidad genómica sea útil a la comunidad científica. Estos procedimientos, además, implican un cambio en el diseño de los algoritmos de análisis para aprovechar en toda su dimensión la computación de alto rendimiento (Guim et ál., 2007), necesaria para el manejo y análisis de cantidades tan grades de datos como los que se producen en un solo experimento realizado con estas tecnologías de secuenciación.

Este inmenso volumen de datos que se está generando, una vez almacenado, requiere de modelos de análisis nuevos que permitan hacerlos comparables en la medida que son obtenidos utilizando un variado conjunto de técnicas y condiciones experimentales, como ocurre, por ejemplo, en el desarrollo de un sistema de clasificación y seguimiento epidemiológico de ß-lactamasas (grupo de proteínas responsable de una gran parte de la resistencia de las bacterias frente a antibióticos ß-lactámicos como las penicilinas y las cefalosporinas) como sistema de información sobre ß-lactamasas BLA-Id (BLA-Id, 2008), que implica el diseño de sistemas de anotación automático que busquen las secuencias en bases de datos como EMBL y UNIPROT, las almacenen y después las clasifiquen. Procesos que aunque parecen triviales requieren de formas novedosas de manejar los errores de clasificación presentes en las bases de datos y algoritmos para la integración de las diversas clasificaciones disponibles en el proceso.

En este contexto, otro aleccionador ejemplo es la masiva cantidad de datos obtenidos de microarreglos almacenados en bases de datos, provenientes de todas partes del mundo aplicando metodologías diversas como la de lectura en uno o dos canales según el método de etiquetado, utilizado para visualizar las secuencias hibridadas; el uso de secuencias cortas, medianas o largas en los oligos que conforman los microarreglos, o el uso de microarreglos fabricados por casas comerciales como Affymetrics o fabricados por los propios investigadores (Pham et ál., 2006). La enorme generación de datos hace que muchos grupos de investigación en bioinformática se encuentren desarrollando nuevas metodologías estadísticas de análisis que modelen algunos de los factores de variación mencionados, y permitan tener una medida estándar y confiable estadísticamente de la expresión de los genes, que posibiliten su comparación independientemente de la técnica experimental utilizada.

Parte de los esfuerzos de investigación en bioinformática estarán dedicados a la búsqueda e integración de datos de diferente índole que estarán disponibles en los bancos de datos, con el objeto de tener una mayor comprensión de las funciones biológicas. No sólo estamos frente a la necesidad de modificar los algoritmos actuales para que hagan uso de los sistemas de alto rendimiento de cómputo, sino frente al desarrollo de nuevos algoritmos que utilicen más intensivamente herramientas como los modelos ocultos de Markov, algoritmos de aprendizaje como las redes booleanas, Support Vector Machines y redes neuronales, entre otros. Esto llevará al desarrollo de aplicaciones para la anotación de alto rendimiento requeridas para responder a las necesidades de análisis de los datos que generan técnicas como la de secuenciación de alto rendimiento; al desarrollo de herramientas para la búsqueda de nuevos genes a partir de la comparación de genomas completos realizadas en supercomputadores como el Mare Nostrum con más de 6000 procesadores en paralelo, del Centro Nacional de Supercomputación de España (CNS, 2008); comparaciones que requieren de la integración de diferentes tipos de datos a los datos de secuencias, y al diseño e implementación de herramientas para la modelación y simulación de sistemas biológicos (Reyes et ál., 2007).

En el futuro próximo áreas nuevas como "System Biology" o "Network Biology" (términos en inglés que aun no tienen un consenso de cómo deben ser traducirlos al español), que se basan en la teoría de sistemas esbozada a mediados del siglo XX, estarán cada vez más presentes en las publicaciones científicas, mostrando cómo a través del descubrimiento de los principios de diseño, la identificación de componentes dentro de los sistemas biológicos, y la comprensión cuantitativa de su funcionamiento por medio de experimentos y simulación, se podrán elucidar la funciones biológicas y predecir cómo cambian frente a perturbaciones internas como las mutaciones, y externas como los fármacos, lo que podría llevar a un tratamiento más preciso y efectivo de las enfermedades. Así, aunque aún estamos lejos de entender por completo los sistemas biológicos debido a su complejidad por tratarse de sistemas naturales, y de que no es posible conocer todas las interacciones que ocurren a este nivel para modelarlas de forma integral, un esfuerzo mayor en el desarrollo de herramientas bioinformáticas para la integración de experimentación y modelación, nos acercará cada vez más a esto, permitiéndonos nuevas e interesantes aplicaciones.

La bioinformática y las perspectivas para Colombia

Considerando que en la época actual la bioinformática no sólo se restringe al análisis de datos moleculares, la integración de datos de biodiversidad constituye uno de los aspectos de la investigación y el desarrollo en los cuales los grupos de bioinformática pueden encontrar un conjunto de problemas interesantes y pertinentes para resolver, ya que aunque Colombia es uno de los países con mayor biodiversidad del mundo, enfrenta el reto de iniciar varios frentes de acción de forma sistemática y coordinada, para consolidar el conocimiento completo de dicha biodiversidad. La bioinformática ofrece las herramientas y los conceptos para sistematizar ese conocimiento.

Es importante destacar que en la actualidad cualquier proyecto a nivel mundial que busque realizar investigación en genómica, transcriptómica, proteómica, metabolómica, y cualquiera de las otras ómicas, requiere de un fuerte apoyo de la bioinformática en el desarrollo e implementación de nuevas aplicaciones particulares que permitan rentabilizar la información generada, sobre todo si se considera que, con base en las respuestas obtenidas utilizando los desarrollos bioinformáticos, se avanzará mucho más rápido y con mayor confianza en la experimentación biológica de cualquier tipo. Esto significa que no solamente se debe propiciar y financiar la investigación en el área de la bioinformática en Colombia que ha mostrado ser competitiva a nivel mundial independientemente de la experimentación biológica que se realiza en el país, sino que para hacer más competitiva la generación de conocimiento biológico debe ser integrada con igual jerarquía a las líneas de investigación, para darles la competitividad que requieren para su desarrollo.

Lo anterior se hace más apremiante si consideramos que en este momento nuestras líneas de investigación se ven enfrentadas, necesariamente, a la utilización técnicas de alto rendimiento para la generación de datos biológicos, haciéndose necesario avanzar en la implementación de una plataforma bioinformática común para el manejo sistemático de estos datos, en especial los que se obtienen de forma masiva por la aplicación de tecnologías en áreas como la genómica, transcriptómica, proteómica y metabolómica, coadyuvando a la disminución de los costos de desarrollo e implementación de herramientas.

Esta plataforma bioinformática implica implementar un sistema computacional de alto rendimiento, unido a una red de expertos en bioinformática trabajando coordinadamente en la detección de los problemas biológicos por resolver. Para tal efecto se requiere de la implementación de un centro especializado en bioinformática de alto rendimiento de cálculo y comunicación, de la formación de personal especializado en esta área, y de la adecuación o el desarrollo de las herramientas bioinformáticas específicas. Esta plataforma computacional deberá estar basada en la implementación de los desarrollos, productos y servicios seleccionados a partir de las necesidades de los diferentes grupos de investigación, de tal forma que se mantenga un intercambio permanente de necesidades y servicios con los grupos que se encargan de la generación de los datos biológicos.

Este centro podría tener un carácter virtual como ocurre con Instituto Suizo de Bioinformática, que como centro nacional coordina el desarrollo de los proyectos

de bioinformática en Suiza reuniendo a los expertos en el tema para provechar su experiencia en investigación y desarrollo en el área, y maximizar la infraestructura requerida, pero al mismo tiempo manteniéndolos vinculados con sus instituciones de origen (universidades e institutos de investigación) con el objeto de que siempre exista participación de éstos dentro de las líneas de investigación propias de cada institución.

En nuestro caso se podría comenzar por la coordinación del recurso humano y técnico que, aunque limitado, actualmente existe en Colombia. Algunos grupos y centros de la Universidad del Valle, la Universidad Nacional de Colombia con grupos como el de Bioinformática del Instituto de Biotecnología encargado de la operación del Nodo Colombiano de la Red Europea de Biología Molecular (EMBnet), y Laboratorio de Investigación en Sistemas Inteligentes (LISI) de la Facultad de Ingeniería, Cenicafé, Universidad de los Andes, la FIDIC y la Universidad Javeriana, entre otros, podrían combinar sus capacidades individuales para el beneficio de la comunidad académica que requiere de servicios, asesoría y capacitación en bioinformática.

El futuro de la bioinformática en Colombia se erigirá sobre el esfuerzo que realicen los grupos de investigación de nuestro país, de manera conjunta, para transformar la información que se genera en el área biológica en conocimiento y desarrollo tecnológico.

Agradecimientos

A los miembros del Grupo de Bioinformática del Instituto de Biotecnología de la Universidad Nacional de Colombia, Sede Bogotá, y en especial a la profesora María Teresa Reguero R., por sus aportes y comentarios a la presente nota.

Referencias bibliográficas

Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. 1990. Basic local alignment search tool. J. Mol. Biol. 215: 403-410.

Benson, D. A.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Wheeler, D. L. 2008. GenBank. Nucleic Acids Res. 36 (Database issue): D25-30.

BLA.id – Sistema de información sobre ß-lactamasas. 2008. Disponible en: http://bioinf.ibun.unal.edu.co/BLA.id [Fecha de consulta: 10/05/2008, fecha actualización: 10/05/2008].

CNS - Centro Nacional de Supercomputación de España. 2008. Disponible en: www.bsc.es. [Fecha de consulta: 10/05/2008, fecha actualización: 10/05/2008].

EBI - European Bioinformatics Institute. 2008. Disponible en: www.ebi.ac.uk [Fecha de consulta: 10/05/2008, fecha actualización: 10/05/2008].

Guim, F.; Rodero, I.; Corbalan, J.; Labarta, J.; Oleksiak, A.; Nabrzyski, J. 2007. Uniform Job Monitoring in the HPC-Europa Project: Data Model, API and Services. International Journal of Web and Grid Services 3 (3).

Hall, N. 2007. Advanced sequencing technologies and their wider impact in microbiology. J Exp Biol. 210 (9): 1518-1525.

Mullis, K. B. 1990. Target Amplification for DNA analysis by the polymerase chain reaction. Ann. Biol. Clin. 48 (8): 579-82.

Pham, T. D.; Wells, C.; Crane, D. 2006. Analysis of Microarray Gene Expression Data. Current Bioinformatics 1: 37-53.

Reyes, S.; Muñoz-Caro, C.; Niño, A.; Badia, R. M.; Cela, J. M. 2007. Performance of computation-intensive, parameter sweep applications on Internet-based Grids of computers. The mapping of molecular potential energy hypersurfaces. Concurrency and Computation: Practice and Experience 19 (4).

The International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409: 860-921.