Bioinformática: ¿una revolución pendiente?

J.M. Martín-Campos

Servicio de Bioquímica. Institut de Recerca. Hospital de la Santa Creu i Sant Pau. Departament de Bioquímica i Biologia Molecular. Universitat Autònoma de Barcelona. Barcelona. España.

El pasado 25 de abril la comunidad científica celebró el 50 aniversario de la publicación, en la revista *Nature*, del artículo firmado por Watson y Crick en el que proponían una estructura para el ADN[1]. Algunos consideran esa fecha como el principio de lo que hoy conocemos como biología molecular, puesto que si bien la identificación del ADN como portador de la información genética es anterior, la determinación de su estructura fue lo que realmente permitió establecer la base molecular de la herencia (la naturaleza del gen) y comprender el mantenimiento de la información de generación en generación (modelo de replicación), abriendo el camino al estudio de su posible manipulación.

A finales de los años setenta, el desarrollo de técnicas de secuenciación de ácidos nucleicos estimuló paralelamente el desarrollo tanto de bases de datos en las que depositar la información generada, como de herramientas informáticas con las que poder procesar, analizar y manipular dicha información. Diez años más tarde, cuando empezaba a tomar forma el proyecto destinado a secuenciar el genoma humano, la base de datos de secuencias nucleotídicas, mantenida de forma colaboradora por Japón (DDBJ), Europa (EMBL) y Estados Unidos (GenBank), mantenía un crecimiento exponencial, duplicando el número de entradas aproximadamente cada 2 años. El ritmo de crecimiento persiste aún hoy día: en el período comprendido entre septiembre de 2001 y septiembre de 2002 la base de datos pasó de 12,9 millones de entradas y 13,8 gigabases a 18,3 millones de entradas y 23 gigabases[2].

Como consecuencia de este crecimiento, las consultas a la base de datos requieren progresivamente un mayor tiempo de espera. La explicación es sencilla: los proyectos de secuenciación de genomas enteros han contribuido a un desarrollo notable de las técnicas de secuenciación, lo que comporta una mayor facilidad en abordar proyectos que impliquen la resecuenciación de regiones genómicas. Por otra parte, los ordenadores cada día son más potentes, con microprocesadores más rápidos y eficientes, y sistemas de almacenamiento de información de mayor capacidad y rápido acceso. Sin embargo, las técnicas de programación y el desarrollo de aplicaciones informáticas, aunque han experimentado un crecimiento importante, éste no ha sido de la misma magnitud, por lo que se ha producido una especie de desfase entre el *hardware* y el *software*. Para explicar mejor este punto pongamos por ejemplo nuestro ordenador personal. Basta con leer periódicamente la publicidad de los distribuidores informáticos para percibir el gran aumento en la configuración básica de los ordenadores. También basta con mirar el tamaño de la carpeta de las distintas versiones de los sistemas operativos para percibir que una buena parte de esos recursos se necesitan para garantizar el normal funcionamiento del ordenador. Por poner un ejemplo, de los aproximadamente 120 megabytes que ocupaba Windows 95 se ha pasado a los aproximadamente 1,2 gigabytes de Windows XP. Cabe preguntarse si las prestaciones de la versión XP son igualmente superiores en un orden de magnitud a las de 95.

Pongamos un ejemplo más científico. La bioinformática, entendida como la disciplina científica que se ocupa del análisis cuantitativo sistemático

Correspondencia: Dr. J.M. Martín-Campos.
Servei de Bioquímica. Institut de Recerca.
Hospital de la Santa Creu i Sant Pau.
Sant Antoni M.ª Claret, 167. 08025 Barcelona. España.
Correo electrónico: jmartinca@hsp.santpau.es

Manuscrito recibido el 30 de octubre de 2003 y aceptado el 30 de octubre de 2003.

de las bases de datos con información biológica, ha experimentado un importante crecimiento en los últimos años, como demuestra la aparición de diversas revistas especializadas en el tema. Sin embargo, si se compara el número de secuencias depositadas en bases de datos y el número de publicaciones en el campo de la biología molecular y la genética, aunque ambas presentan un crecimiento exponencial, es mucho más acusado en el primer caso. En 1994 el número de entradas en GenBank igualaba el número de publicaciones, mientras que un año más tarde casi lo doblaba. En resumen, el crecimiento de la información cruda sobre los genes es mucho más rápido que el conocimiento obtenido de dichos genes.

Disponer de genomas completos complica el panorama aún más. El análisis de genomas enteros o de porciones importantes, lo que se conoce como genómica, supone un cambio cuantitativo que requiere el desarrollo de aplicaciones informáticas complejas y mucho tiempo de computación. Ahora bien, es un cambio cuantitativo que ha llevado a un cambio cualitativo, puesto que estudiar muchos genes a la vez no requiere suposiciones *a priori* sobre qué genes son interesantes en un proceso concreto, sino que simplemente se estudian todos, de forma que se elimina el sesgo. Como resultado, de los estudios clásicos basados en hipótesis (estudiamos una batería de genes que *presuntamente* están implicados) pasamos a estudios que generan hipótesis (a partir de ahora estudiaremos un grupo de genes que son los que *realmente* presentan alteraciones).

Como consecuencia de la tendencia actual a trabajar con una mayor cantidad de información de forma simultánea, la demanda de mejoras en el *hardware* y el *software* no es exclusiva de los especialistas (teóricos, estadísticos y programadores), sino que se ha extendido al resto de investigadores que hasta ahora se limitaban, en muchos casos, a paquetes estadísticos clásicos y programas de análisis de datos relativamente sencillos. Sin embargo, las necesidades no se limitan a programas para el análisis, sino también a programas destinados al manejo preanalítico de las muestras, tendentes a conseguir una cierta automatización de los procesos de laboratorio y una mejor organización de la información asociada a las muestras (elaboración de genealogías, manejo de bibliografía, bases de datos).

El artículo de Coltell et al[3] publicado en este mismo número cabe enmarcarlo en este último grupo. Como bien señalan los autores, a menudo es preciso intercambiar información entre aparatos de laboratorio gobernados por programas distintos o enlazar el fichero de resultados de un aparato con un programa de análisis. El problema viene cuando el formato no es el mismo, lo que sucede con mucha frecuencia. La conversión manual de dichos archivos, además de ser en muchos casos larga y tediosa, es una fuente de introducción de errores a menudo difíciles de detectar *a posteriori*. Por tanto, el desarrollo de aplicaciones que automaticen dicho proceso resulta de gran interés, y si además están basados en programas de uso común, como es Microsoft® Excel™, tienen la ventaja añadida de una pronta familiarización por parte del usuario.

Desde hace años, para algunos tipos de análisis como son, por citar 2 ejemplos, el de ligamiento[4] o el de secuencias, tanto nucleotídicas como proteínicas[5], existen formatos de fichero más o menos estándar y que son usados como entrada de datos por un gran número de programas. Con ello, se consigue eliminar la tarea de rescribir los datos crudos cada vez que se necesita reanalizar las muestras con un programa nuevo y facilitar el intercambio de información entre programas. De la misma forma, podría desarrollarse un formato común a todas las máquinas que usen, por ejemplo, el sistema de placas de pocillos. La estructura de dicho formato podría ser modular, a semejanza del formato de secuencias NEXUS[5], con un primer módulo destinado a la asignación de las muestras a los distintos pocillos, y módulos posteriores que incorporen el resultado del procesamiento de las muestras (genotipificación, secuenciación, cuantificación, detección) y que sean reconocidos por los programas de análisis, que a su vez añadirían módulos de resultados finales, reconocidos por programas de presentación (tomen nota las casas comerciales).

La era de la genómica y la proteómica ha revolucionado el panorama científico. Es de esperar que la revolución acabe por alcanzar también al trabajo diario en el laboratorio, provocando una mayor automatización de los procesos. Permanezcan atentos a su pantalla de ordenador.

Bibliografía

1. Watson JD, Crick FHC. Molecular structure of nucleic acids: a structure of deoxyribose nucleic acid. Nature 1953;171:737-8.
2. Stoesser G, Baker W, Van den Broek A, García-Pastor M, Kanz C, Kulikova T, et al. The EMBL Nucleotide Sequence Database: major new developments. Nucleic Acids Res 2003;31:17-22.
3. Coltell O, Corella D, Tai ES, Guillén M, Chalmeta R, Ordovás JM. PLATEX: una herramienta bioinformática para la conversión de datos en el estudio genético de la arteriosclerosis. Clin Invest Arterioscl 2004;16:43-52.
4. Terwilliger JD, Ott J. Handbook of human genetic linkage. Baltimore: Johns Hopkins University Press, 1994.
5. Maddison DR, Swofford DL, Maddison WP. Nexus: an extensible file format for systematic information. Syst Biol 1997;46:590-621.