•**Eduardo López Viñas**
Licenciado en Biología Molecular "Severo Ochoa" (CSIC-UAM)

•**Paulino Gómez-Puertas**
Científico Titular del consejo Superior de Investigaciones Científicas y Profesor Honorario de la Universidad Autónoma de Madrid

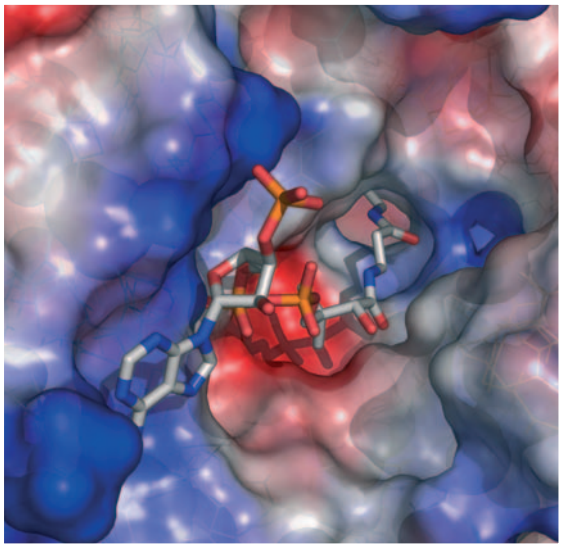
BIOINFORMÁTICA, LA INFORMACIÓN AL SERVICIO DE LA CIENCIA

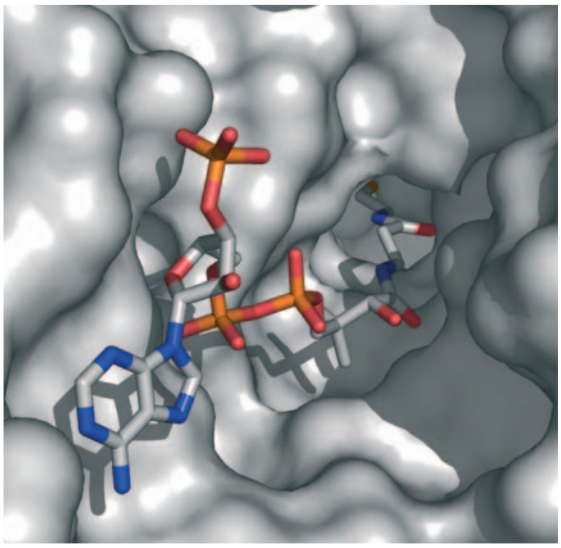
La Bioinformática es la disciplina que se encarga de estudiar el contenido y flujo de la información en sistemas y procesos biológicos. Esta disciplina, entre la informática y la biología, surgió principalmente como respuesta a las necesidades de computación y análisis de datos genéticos producidos en el estudio del Proyecto "Genoma Humano". Hoy en día la bioinformática ofrece grandes posibilidades para el avance de la medicina.

El fin último de la investigación en Bioquímica y Biología Molecular es siempre el conocimiento de los procesos que gobiernan el funcionamiento de los seres vivos, generalmente con el ánimo de utilizar esta información en provecho del ser humano y de su entorno, con nuevos avances en medicina, mejoras en el medio ambiente, o simplemente para satisfacer la curiosidad innata de saber cómo es el mundo que nos rodea y porqué se comporta como lo hace.

En los últimos años, este ansia de conocimiento ha provocado la generación de una cantidad cada vez mayor de datos genómicos, bioquímicos y funcionales que, paralelamente al auge de las tecnologías del almacenamiento y transmisión de la información, ha derivado en el nacimiento de una nueva rama de la Biología conocida en general con los nombres de "Bioinformática", "Biología Computacional" o incluso "Biología Digital".

En las primeras décadas de auge de la Biología Molecular (de los años 60 a principios de los 90), la investigación se caracterizaba por la abundancia de conocimiento detallado sobre una cantidad no muy grande de genes y proteínas de especial importancia biomédica, implicados en procesos conocidos de, señalización, división y metabolismo celular. Sin embargo, la publicación en 1995 del primer genoma completo (Haemophilus influenzae) abrió las puertas a un fenóme- ⇨





¬ **La integración de datos estructurales** y de secuencia permitirá a la Bioinformática diseñar, de forma eficiente, fármacos específicos dirigidos a centros activos de proteínas de interés biomédico. En la ilustración, una molécula de inhibidor ocupa el sitio que correspondería al sustrato en la superficie de una enzima responsable de la degradación de ácidos grasos, punto clave en futuras terapias de procesos ligados a la obesidad o la diabetes.

no diferente: desde entonces se han obtenido un total de 274 genomas completos, de los cuales 36 pertenecen a organismos eucariotas, algunos tan emblemáticos como los de levadura, arroz, maíz, ratón, chimpancé y, por supuesto, el humano, publicado en junio de 2000. Los dos últimos genomas completos disponibles son el del hongo Aspergillus fumigatus y el del pez-cebra (Danio rerio), ambos publicados en julio de 2005. En la actualidad están en proceso de secuenciación 1.222 genomas, incluyendo 494 eucariotas (fuente: "Genomes OnLine Database" www.genomesonline.org).

El panorama, apenas diez años después de la publicación del primer genoma, ha cambiado radicalmente: en lugar de una gran cantidad de conocimiento sobre unos pocos genes de interés, encontramos encima de la mesa una enorme montaña de datos (el número de secuencias conocidas, depositadas en la base de datos pública de GenBank, es de casi 45 millones) de los

En 1995 se publicó el primer genoma. Desde entonces se han descifrado un total de 274 genomas completos, entre ellos, el humano

que de apenas un pequeño porcentaje tenemos información acerca de su posible estructura, función, localización o mecanismo en el que se encuentra implicado. Y es en este momento, como disciplina con capacidad de integración de información, cuando la Bioinformática juega un papel central en la generación de los enlaces que permiten correlacionar la información disponible para extraer los patrones subyacentes y avanzar así en el conocimiento del funcionamiento de los organismos vivos.

La fuente de la que se nutre la Bioinformática la constituyen una serie de bases de datos de acceso público donde se acumula, continuamente, toda la información disponible. Éstas se han especializado según la naturaleza de los datos almacenados, siendo las principales las que recogen secuencias de genes, genomas y proteínas (GenBank: www.ncbi.nlm.nih.gov; EMBL: www.ebi.ac.uk; DDBJ: www.ddbj.nig.ac.jp; Uniprot: www.ebi.uniprot.org), estructuras tridimensionales de macromoléculas (Protein Data Bank: www.rcsb.org/pdb), datos de expresión génica (Array Express: www.ebi.ac.uk/arrayexpress; Stanford Microarray Database: genome-www5.stanford.edu), ontologías (GeneOntology: www.geneontology.org) y literatura científica (PubMed: www.ncbi.nlm.nih.gov/Literature).

El primer reto al que se enfrenta la Biología Computacional es, precisamente, el mantenimiento y regulación de acceso a estas bases de datos, cuyo volumen ⇨

Eduardo López Viñas / Paulino Gómez-Puertas
Licenciado en Biología Molecular "Severo Ochoa" (CSIC-UAM)
Científico Titular del consejo Superior de Investigaciones Científicas y Profesor Honorario de la UAM

¬ **El Centro de Biología** Molecular "Severo Ochoa" es uno de los más activos de nuestro país

crece exponencialmente al tiempo que lo hace el número de usuarios que solicitan su utilización o descarga. A esto se unen algunos problemas históricos que arrastran algunas de ellas desde su nacimiento, como que cada una de las 30.000 entradas del Protein Data Bank sea un archivo independiente en formato de texto o, aún más complicado, que toda la base de datos de GenBank o de Uniprot sea un sólo archivo de texto de un tamaño cada vez más inmanejable. Aunque, por suerte, la memoria de trabajo de los ordenadores y la capacidad de transporte de las redes ha crecido también en los últimos años, no está claro si en un futuro la estructura de estas bases de datos no necesitará de una reforma radical para permitir que sigan siendo útiles.

Es, sin embargo, esta enorme cantidad de datos disponibles y la esperanza de encontrar los nexos de unión entre ellos lo que empuja a los especialistas en Bioinformática a plantearse una serie de retos ambiciosos que parecen encontrarse al alcance de la mano, algunos de los cuales se enumeran a continuación:

· **Modelado virtual de la estructura tridimensional de proteínas y complejos proteicos.** En la década de los 90 fue este campo de la Bioinformática uno de los que registraron un mayor crecimiento. En unos pocos años se pasó de sistemas capaces de apenas sugerir características unidimensionales como la capacidad de los aminoácidos de plegarse en láminas beta y hélices alfa a obtener modelos tridimensionales completos de gran calidad para una cantidad que se aproxima al 30% de todas las entradas de la base de datos Uniprot (con 2 millones de secuencias de proteína anotadas). Los avances se fueron haciendo patentes desde las primeras ediciones de la competición internacional CASP (Critical Assessment of Techniques for Protein Structure Prediction: www.predictioncenter.org) y su variante para sistemas completamente automáticos (CAFASP). Hay, sin embargo, en este campo un tema pendiente que aún no termina de avanzar a la velocidad deseada: la obtención, utilizando únicamente el conocimiento de las fuerzas físicas que gobiernan las interacciones entre los átomos que componen una cadena polipeptídica desplegada de la trayectoria de plegamiento que permita alcanzar una forma final energéticamente estable y similar a la que posee la proteína in vivo. Pese a que los esfuerzos dedicados son cada año mayores, entre los que destacan la actividad del ordenador "BlueGene" de IBM o la inauguración reciente del "Mare Nostrum" en Barcelona, los resultados, aunque esperanzadores, son aún escasos. En todo caso, merece la pena seguir en el empeño: si se encontrase la manera correcta de plegar una proteína completa in silico, se habría dado con la Piedra Rosetta que permitiría el paso de gigante de la simulación efectiva de muchos de los procesos que tienen lugar dentro de la célula y que están guiados por fuerzas de interacción entre aminoácidos muy similares.

· **Simulación in silico de los mecanismos de interacción entre macromoléculas.** Un problema en parte similar al anterior, pero en el que entran en juego factores como la definición de superficies de interacción y sus características fisicoquímicas. Los sistemas actuales de predicción de interacciones entre macromoléculas, aunque han avanzado considerablemente en los últimos 4-6 años como lo demuestran los cada vez mejores resultados de la competición internacional CAPRI (Critical Assessment of Prediction of Interactions: www.capri.ebi.ac.uk), no han llegado aún a la capacidad de los sistemas de predicción de plegamiento, sobre todo cuando se utilizan modelos tridimensionales obtenidos a su vez mediante ⇨

El auge de las tecnologías de almacenamiento y transmisión de información, ha derivado en el nacimiento de una nueva rama de la Biología conocida como Bioinformática

modelado informático. En coordinación con técnicas recientes de obtención de volúmenes tridimensionales de complejos macromoleculares mediante microscopia electrónica de alta resolución, en los últimos tiempos están apareciendo sistemas que permiten acoplar en estos volúmenes los modelos atómicos de las moléculas que lo constituyen. Aún es pronto para asegurarlo, pero probablemente sea éste uno de los puntos más cercanos de conexión entre el mundo macromolecular (microscópico) con el atómico a desarrollar en los próximos años.

· Integración de datos de secuencia y estructura tridimensional. Mejor suerte que los anteriores corren los sistemas de simulación de interacciones de fármacos con los centros activos de determinadas proteínas de interés biomédico. Aunque lejos aún de poder sustituir efectivamente a las técnicas tradicionales de muestreo in vitro, la integración de análisis de datos de secuencia aminoacídica y de su conservación a lo largo de la evolución en familias de proteínas junto con el modelado tridimensional de su estructura, están permitiendo conocer cada vez mejor los mecanismos responsables de la especificidad funcional de las proteínas y de cómo utilizarla para diseñar fármacos "a la carta" en un futuro cada vez más cercano.

· Análisis de la expresión génica. El uso de microarrays para obtener información de la relación entre determinados estados celulares, tisulares e incluso sistémicos con los niveles de expresión específicos de determinados genes o grupos de genes es, de entre todas las técnicas bioinformáticas, la que probablemente se utilice de forma más generalizada en los laboratorios de experimentación en Biología Molecular. El avance en los métodos estadísticos que permiten una correcta interpretación de los datos y su integración con información funcional y de secuencia hace de esta técnica una de las más poderosas en el estudio simultáneo de grandes cantidades de información experimental.

La Bioinformática, como disciplina independiente, posee la capacidad de generar conocimiento nuevo además de procesar eficientemente el ya existente

· Análisis de genomas y secuencias de genes y proteínas. Entre los sistemas estrella de la Bioinformática figuran aquellos capaces de predecir, a partir de la maraña de bases nucleotídicas que componen la salida de los sistemas automáticos de secuenciación genómica, la localización de los genes y secuencias reguladoras. Una vez localizados, el siguiente paso es la asignación funcional utilizando para ello métodos que integran el análisis de la composición física del DNA con la comparación de secuencias a nivel tanto de familias de proteínas como de genomas completos. La asignación de función por comparación con secuencias similares, idealmente homólogas es, al tiempo, un éxito de la Bioinformática y uno de sus mayores riesgos. Un éxito porque permite extrapolar la función de una proteína en un organismo del que apenas si es necesario dispone de poco más que su secuencia de DNA (en ocasiones apenas si se conoce el organismo en sí, no habiéndose trabajado nunca con él en el laboratorio como ocurre con determinadas bacterias y arqueas presentes en ecosistemas extremos). Y un riesgo porque la cantidad de proteínas de las que su función sólo se conoce por este método supera ya a aquellas de las que se ha obtenido experimentalmente y las probabilidades de que, por proximidad de secuencia y de forma transitiva, determinadas funciones se asignen equivocadamente a terceros que a su vez pueden ser el origen de nuevos errores en anotación, aumentan cada vez más. Aunque se utilizan métodos para prevenir estos problemas, es éste sin duda un riesgo latente que habrá que cuidar en el futuro.

· Análisis de la literatura científica. Uno de los proyectos más novedosos, por lo alejado de los conceptos biológicos clásicos, es el análisis de texto científico. Estos sistemas de minería de datos utilizan parámetros de frecuencia de aparición de determinadas palabras (nombres de genes, proteínas y funciones) o combinaciones de palabras, así como su relación mediante determinados verbos o conjunciones para extraer información específica de función o de relaciones entre macromoléculas a partir de resúmenes o textos completos de artículos científicos. Verdadera arqueología del conocimiento guardado en las bases de datos de literatura (en la base de datos de PubMed hay, en la actualidad, 14 millones de entradas), estos métodos permiten combinar el conocimiento previo con el actual para encontrar relaciones que pasarían desapercibidas de otro modo ante la imposibilidad de consultar personalmente tal cantidad de información. ⇨

Eduardo López Viñas / Paulino Gómez-Puertas
Licenciado en Biología Molecular "Severo Ochoa" (CSIC-UAM)
Científico Titular del consejo Superior de Investigaciones Científicas y Profesor Honorario de la UAM

La Bioinformática, ordena y regula la basta información que existe sobre genes y proteínas de toda naturaleza

Y, por último, el gran reto de la Biología Molecular en las próximas décadas, en el que la Bioinformática puede jugar un papel protagonista, es dar un paso más allá de la biología de lo pequeño para encontrar grandes patrones comunes que permitan conocer el funcionamiento completo de organismos y ecosistemas complejos, en lo que se ha venido a llamar la "Biología de Sistemas".

La Bioinformática, como disciplina independiente posee ya la capacidad de generar conocimiento nuevo además de procesar eficientemente el ya existente. Precisamente, por ello, tiene ahora por delante la obligación, no necesariamente fácil de asumir, de compartir realmente este conocimiento con la amplia comunidad de científicos experimentales, expertos en los diferentes campos de la Biología Molecular. Ésta interacción no siempre es de la naturaleza deseada. A modo de ejemplo, no es infrecuente encontrar especialistas en un grupo de genes o proteínas que, al consultar las bases de datos donde se recoge su secuencia y función, experimentan una cierta frustración al comprobar que la información no es lo precisa que desearían o que ésta se ha generado de forma automática sin tener en cuenta pequeños detalles específicos que resultan ser clave en la función real de ese grupo concreto de genes. La interacción estrecha con los grupos experimentales expertos en cada tema permitiría conocer los puntos en los que la Bioinformática no genera las respuestas adecuadas, el primer paso para poder corregirlos. Sería esta retroalimentación la que posibilitaría un nuevo avance de la Biología Computacional.

Para hacer frente a este reto, la mejor herramienta con la que se cuenta es la formación de un número elevado de especialistas en Bioinformática y Biología Computacional, con formaciones diferentes (en ciencias biosanitarias, informáticos, matemáticos, químicos, físicos) capaces de integrarse en equipos multidisciplinares con capacidad de interacción con los expertos de cada una de las áreas concretas. Por desgracia, en nuestro país y con alguna honrosa excepción, tal oferta formativa es muy escasa, reduciéndose en muchos casos a cursos de pocos días de duración para licenciados que no permiten ofrecer una panorámica adecuada de los fundamentos de la disciplina. Esto sigue provocando que la mayoría de los licenciados en Biología Molecular apenas si hayan tenido contacto durante la carrera con las herramientas básicas de acceso y manejo de lo que más tarde será parte fundamental de su trabajo: las bases de datos donde se almacena el conocimiento de secuencia, estructura y función de las macromoléculas biológicas. Pero el panorama puede cambiar, y ya algunas universidades públicas y privadas empiezan a ofrecer como asignatura, en sus estudios de segundo ciclo, materias relacionadas con la Bioinformática, además de los cursos de doctorado y "master" de especialización para posgraduados, en lo que quizá sea, esperemos, la señal de que en esta materia no perderemos el tren de la investigación puntera por falta de base formativa. ■