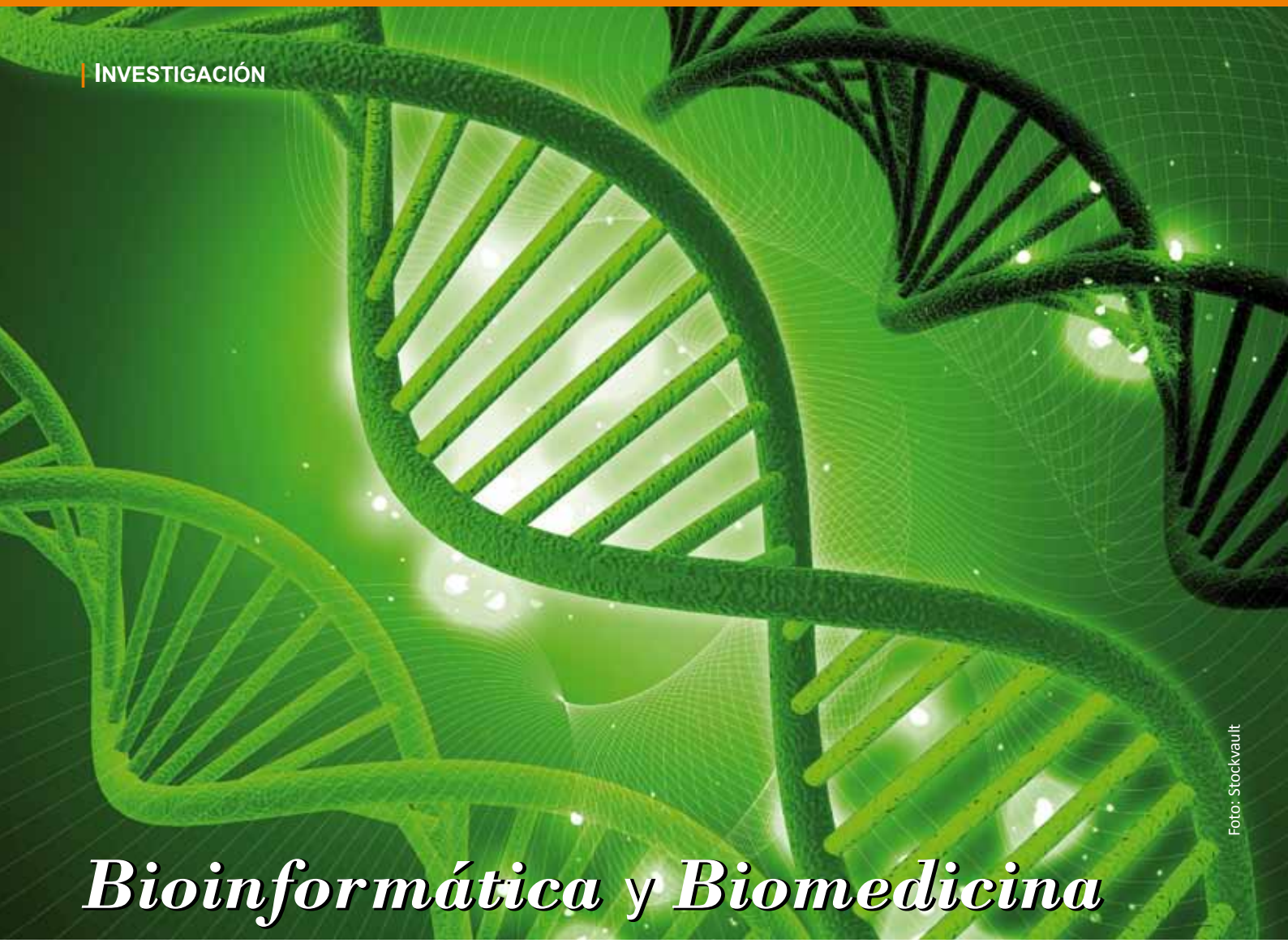


Foto: Stockvault

Bioinformática y Biomedicina

Nuevos retos de la supercomputación

Los científicos confían en la potencia de cálculo de los ordenadores para desarrollar métodos, rápidos y baratos, que en el futuro permitan a un individuo secuenciar su propio genoma. Enormes volúmenes de datos que son una nueva meta para la ciencia y ahora también para la UMA: el *Big-Data problem*.

> **Oswaldo Trelles** / *Doctor del Departamento de Arquitectura de Computadores*

La mejora de los dispositivos de adquisición de datos, la disponibilidad de las redes para distribuirlos y el incremento de capacidad de almacenamiento de los ordenadores, han hecho posible la adquisición y gestión de colecciones gigantescas de datos en el rango del terabyte (trillón de caracteres), o del petabyte (cuatrillones de caracteres), e incluso superiores (exa, zetta, yotta...). Ello ocurre en diversas áreas que van desde la astronomía hasta las ciencias sociales, pasando por el medioambiente, la agricultura, la medicina, la biología, los negocios, o simplemente dispersa en la web por las redes sociales. Este hecho ya se resaltaba en diversas publicaciones científicas, como en la edición especial que *Nature* lanzó en 2008 bajo el título *Big Data: welcome to the petacentre, science in the petabyte era*.

Veamos un buen ejemplo en el campo de la biología. Al empezar este nuevo milenio, secuenciar 1 Mbp (millón de pares de bases —las "letras"— de un genoma) de ADN —la molécula de la vida— costaba alrededor de 10 mil dólares. Pocos años más tarde, en 2008, secuenciar el genoma humano, con algo más de 3.200 millones de letras, ya era posible en unas seis semanas y su coste total caía hasta los 60 mil dólares. Por aquel tiempo los proyectos especulaban con alcanzar el genoma humano a unos mil dólares en los siguientes tres años. En efecto, en octubre de 2009 varias compañías apuntaron al genoma de los mil dólares con un nuevo avance tecnológico, la secuenciación con nanoporos, y ya en marzo de 2011, el coste había caído a menos de medio dólar por Mbp. Como referencia, en 1990, el Congreso de los Estados Unidos de América aceptó financiar con 3 billones de dólares

Al empezar el nuevo milenio, secuenciar 1 Mbp de ADN costaba sobre 10 mil dólares, en 2011 el coste había caído a menos de medio dólar

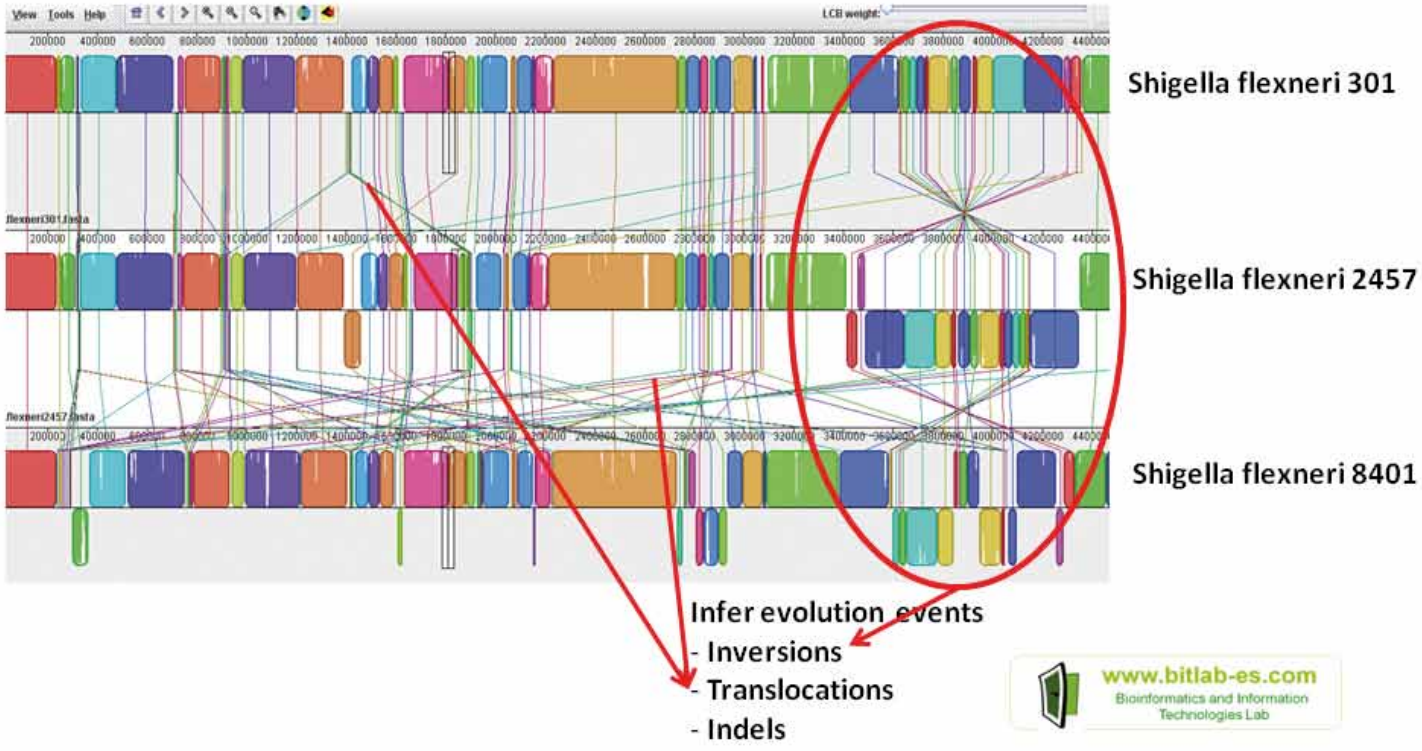
el proyecto Genoma Humano —un dólar por base—. El genoma se completó en 2003, dos años antes de lo previsto, costando "solo" 2,7 billones de dólares, lo que representa 2,7 millones de veces el coste actual.

Aún más. En febrero de 2012 se lanzó un dispositivo en miniatura del tamaño de una memoria USB, diseñado para hacer de la secuenciación de ADN una tecnología universalmente accesible por menos de 900 dólares (unos 690 euros). Pero los científicos siguen en la carrera para desarrollar nuevos métodos más rápidos y más baratos que permitan a cada individuo secuenciar su propio genoma, lo que en última instancia significaría dar inicio a la era de la medicina personalizada basada en la genética.

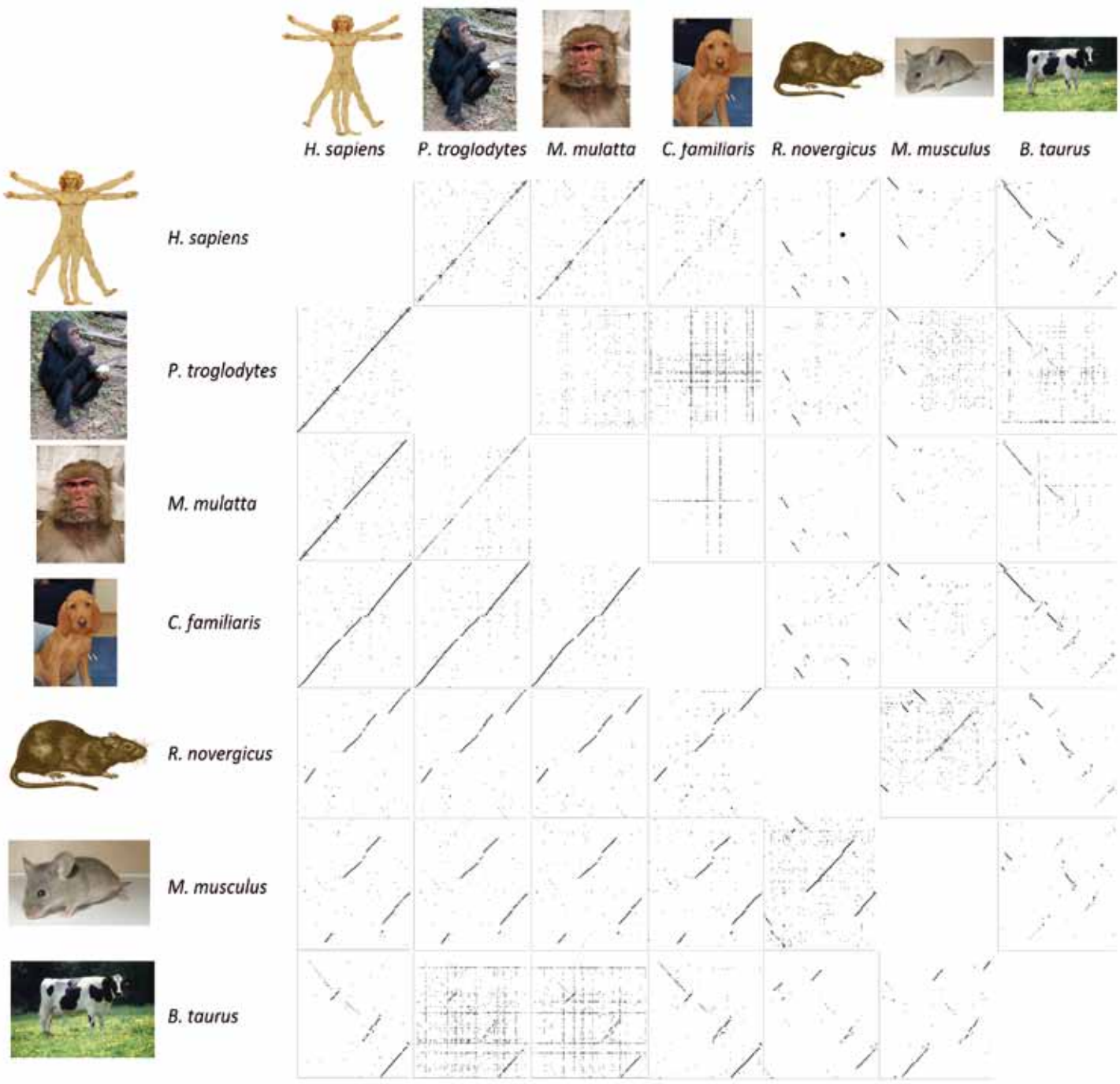
Las grandes cantidades de datos, y la seguridad de que seguirán creciendo, plantea nuevos retos para transformar esa información en conocimiento útil

Pero no solo el genoma humano está en el interés de los científicos. En términos prácticos, lo dicho es válido para todas las demás especies. Actualmente, ya hay miles de organismos en todos los reinos que han sido secuenciados y existen cientos de proyectos en marcha para seguir obteniendo estos datos. Ello pone a la genómica comparativa en el foco de interés para el desarrollo de nuevos métodos destinados a estudiar las relaciones existentes en esta avalancha de datos.

Las grandes cantidades de datos —y la seguridad de que seguirán creciendo— plantea nuevos retos para transformar esa información en conocimiento útil. El primero se refiere al manejo de estos volúmenes tan grandes, lo que se ha dado en llamar "*the Big-Data problem*". Recordemos que las redes, tan en boga e imprescindibles hoy en día, están diseñadas para mover tamaños de datos relativamente pequeños, en el orden de los pocos megabytes. Todos hemos experimentado las tremendas demoras cuando los ficheros que descargamos por la red pesan unos cientos de megas (un vídeo, por ejemplo). Podemos entonces imaginar lo tedioso que resultaría mover algunas decenas de gigas por la



Comparación de tres cepas de Shigella flexneri (enterobacteria, principal causa de mortalidad infantil en países en desarrollo). Se puede observar una inversión de una zona de su genoma (a la derecha en un círculo) de la segunda cepa respecto a la primera y cómo en la tercera cepa se ha revertido parte de esta inversión. Ademas se aprecian varias pequeñas translocaciones (reordenamientos) que afectan a su virulencia.



Comparación del cromosoma sexual X de mamíferos superiores (humano, chimpancé, macaco, perro, rata, ratón y vaca). Las líneas diagonales indican zonas comunes entre las dos especies comparadas en cada caso. Como se puede comprobar el ser humano se parece mucho a los chimpancés y macacos (todos ellos simios). Pese a la distancia evolutiva existente entre las especies, el cromosoma X es una estructura poco variable, como se ve en la alta similitud con el cromosoma de perro, y algo menos con el de los roedores (rata y ratón) y la vaca. / Fotos: Wikimedia Commons.

web, pero incluso a nivel local, moverlos de un servidor a un punto de procesamiento resulta costoso.

Siendo el movimiento de datos importante, la parte más interesante es su procesamiento. Los sistemas actuales han sido diseñados para analizar pocos datos (unas pocas miles de proteínas); datos pequeños (un gen puede representar unos pocos miles de bases y una proteína unos cuantos cientos de aminoácidos), pero los datos ahora han crecido (a nivel de genoma) y el análisis se ha vuelto complejo. Por tanto, el tiempo de computación se ha disparado.

Hace pocos años estábamos convencidos de que con las tasas de crecimiento, capacidad y velocidad de nuestros ordenadores tendríamos suficiente y así la mayor parte —en términos prácticos, la totalidad— de las aplicaciones en bioinformática —la rama de la ciencia que provee soluciones para procesar los datos biológicos— no están preparadas para este cambio de escenario. Suelen cargar los datos en memoria principal para acelerar los cálculos y no hacen uso de la disponibilidad de varios procesadores para distribuir la carga de trabajo. Esto no es

sorprendente, ya que se estima que solo el uno por ciento de los programadores a nivel mundial ha recibido algún entrenamiento formal en computación de alto rendimiento, y precisamente, dar el salto desde la computación secuencial a la paralela no resulta trivial.

El siguiente reto parece ser aún más atractivo. "*Thinking big*" es, hoy en día, la forma natural de plantear proyectos en ciencias de la vida. Este "pensar en grande" significa incluir varias caras o visiones del sistema biológico en estudio, combinando desde los tradicionales datos de secuencia, hasta la observación del comportamiento dinámico de los organismos en su medioambiente, o el estudio de los organismos *in vivo* en el caso de los metagenomas; recopilando en cantidades ingentes datos fisiológicos, fenotípicos, de niveles de expresión de genes y proteínas en las células, datos clínicos…

La UMA coordina un proyecto europeo sobre computación avanzada en grandes conjuntos de datos clínicos y genéticos

Por tanto, se requiere el uso de métodos analíticos caros en el consumo de recursos computacionales, ya que deben trabajar con niveles de ruido importantes —arrastrados por la tecnología de adquisición— los valores incompletos, la variabilidad biológica de los organismos, etc. Necesitamos entonces modelos más complejos e inexistentes hoy en día para el procesamiento de estos datos.

La Universidad de Málaga (UMA) coordina el proyecto europeo 'High Performance, Cloud and Symbolic Computing in Big-Data Problems Applied to Mathematical Modelling of Comparative Genomics' sobre computación avanzada en grandes conjuntos de datos clínicos y genéticos. Este proyecto busca soluciones para gestionar grandes volúmenes de datos mediante el uso de técnicas de computación de alto rendimiento, paralela y en la nube. Los desarrollos propuestos se validarán con aplicaciones en genómica comparativa y en biomedicina. Un objetivo para el que aúnan esfuerzos grupos y empresas europeas que trabajan en distintas disciplinas, como la computación, la matemática, la estadística, la biología o la medicina. La frenética actividad de los proyectos que trabajan en estas áreas promueve como efecto colateral el desarrollo de nuevo *software* y el desarrollo de nuevas formas de analizar los resultados, lo que se traduce en la mencionada colaboración entre diversos grupos que aportarán cada uno parte de la solución.

El proyecto que ahora coordinamos es un buen ejemplo de colaboración interdisciplinar, que ofrecerá como resultado aplicaciones capaces de lidiar con el análisis e interpretación de genomas completos y la interrelación de varios de ellos, poniendo al alcance de cualquiera con una conexión a internet grandes recursos e infraestructuras computaciones a las que hasta ahora solo podía acceder personal con un entrenamiento muy específico. En nuestro caso, la empresa de supercomputación RISC y la UMA proporcionarán la capacidad de cálculo —en la nube y en supercomputadores— y el desarrollo de software. La Universidad Johannes Kepler de Linz (Austria), y el grupo Baobab de la Universidad de Lyon (Francia) desarrollarán nuevos modelos para la comparación de genomas, distancias evolutivas entre organismos, correlaciones entre las variaciones genéticas de los pacientes con la posible respuesta a determinados tratamientos.

Por su parte, el Centro de Supercomputación de Leibniz (Alemania) diseñará interfaces, no solo para los dispositivos tradicionales, sino también para acceso con dispositivos móviles, como los teléfonos inteligentes y las tabletas. Dispositivos que serán validados por los usuarios finales, en concreto, por el equipo clínico del Hospital Carlos Haya de Málaga. Asimismo, la empresa Integromics ayudará a crear interfaces de aplicaciones y *apps* móviles profesionales. Todo ello sazonado con el uso de nuevas tecnologías para crear círculos abiertos de discusión, como las redes sociales a través de los bien conocidos Facebook, Twitter, etc… Compromiso y esfuerzo por informar al público de los avances que le pueden beneficiar y que casan a su vez con la meta de una ciencia cada vez más abierta y accesible. ●

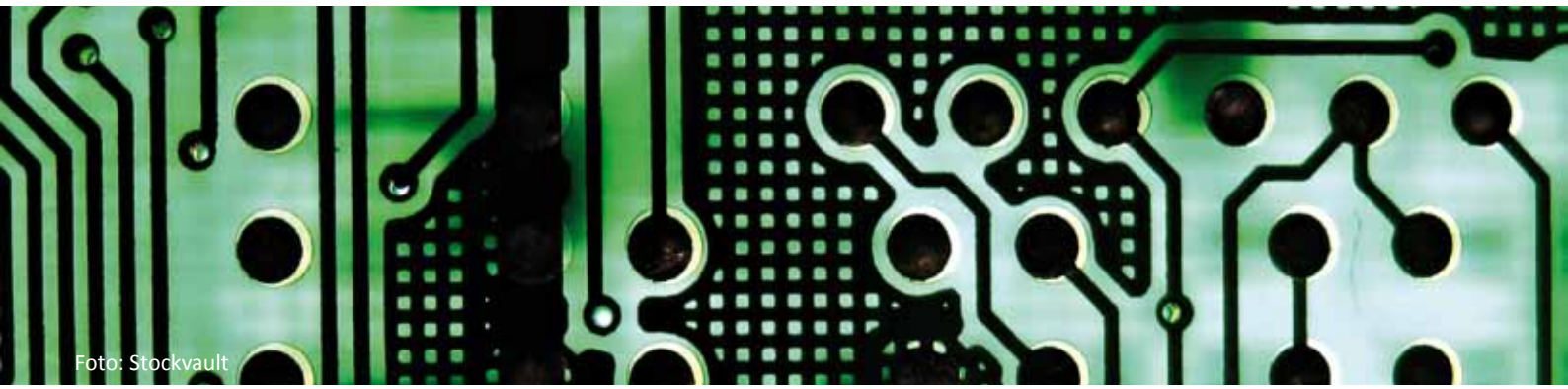


Foto: Stockvault