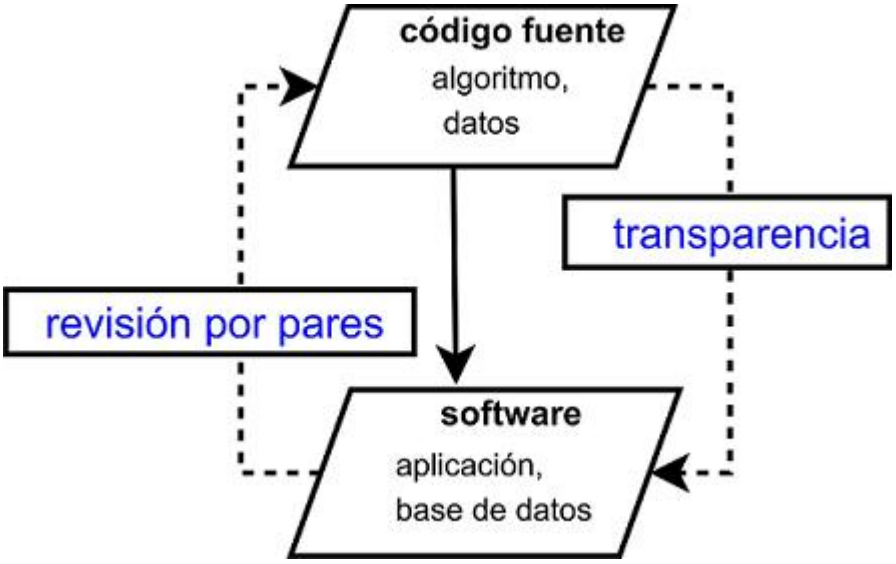
03-09-2009

La filosofía open source en la bioinformática

Desde hace una década el fenómeno open source o código abierto ha explotado y sus efectos se pueden ver en todas partes y también en la ciencia. Este modelo de software se centra en los beneficios prácticos de compartir el código, más allá de cuestiones morales y filosóficas. Así, en la ciencia, los programas open source permiten que investigadores de todo el mundo pongan a prueba e incluso mejoren soluciones a problemas científicos, como si fueran experimentos a escala global.

El motor de esta revolución cultural y económica ha sido en gran medida el sistema operativo Linux, que puede servir de ejemplo para explicar en qué consiste realmente todo esto. Antes de GNU/Linux, el desarrollo tecnológico de la informática había sido moldeado en exclusiva por las compañías con más influencia y cuota de mercado. Este modelo se basa en licencias que los usuarios, nosotros, pagamos para instalar y utilizar programas en nuestros ordenadores. Bajo este modelo, las compañías tienen un control total sobre los programas que distribuyen, y el papel de los usuarios es pasivo. Desde la irrupción de Linux, la comunidad open source, que incluye a pequeñas y grandes empresas, y a muchos programadores voluntarios, tiene cada vez más peso en el diseño de los programas informáticos que usamos cotidianamente. En este nuevo escenario los usuarios asumen un papel más activo en el desarrollo de software, ya que al tener acceso a su código fuente pueden leer, modificar (y generalmente redistribuir en las mismas condiciones) el software, haciendo que evolucione. De esta forma, los propios usuarios pueden adaptar un programa a sus necesidades, e incluso corregir errores a mayor velocidad que en el modelo de software convencional o cerrado, dando como resultado la producción de software robusto. El mejor exponente de este fenómeno es quizás SourceForge, una central de desarrollo de software que gestiona y distribuye multitud de proyectos de software libre y actúa como un repositorio donde cada día programadores de todo el mundo actualizan miles de líneas de código fuente.

Los programas open source son por tanto más transparentes, ya que es posible entender exactamente cómo funcionan. Por supuesto esta nueva filosofía se adapta mejor a la realidad de la ciencia, puesto que facilita que investigadores de diferentes laboratorios trabajen sobre el mismo código. Esta filosofía lleva a su extremo el proceso de revisión por pares (peer review) que es una de las bases del avance de la ciencia. Cuando un científico quiere publicar sus resultados debe convencer a varios revisores anónimos, normalmente otros investigadores (pares) que trabajan en áreas cercanas, de la validez de su trabajo. De la misma manera, la metodología open source lleva implicada una revisión por pares a gran escala ya que el software puede llegar a ser desarrollado en simultáneo por toda una comunidad de programadores que ponen a prueba el código aportado por otros colegas que probablemente no conozcan.



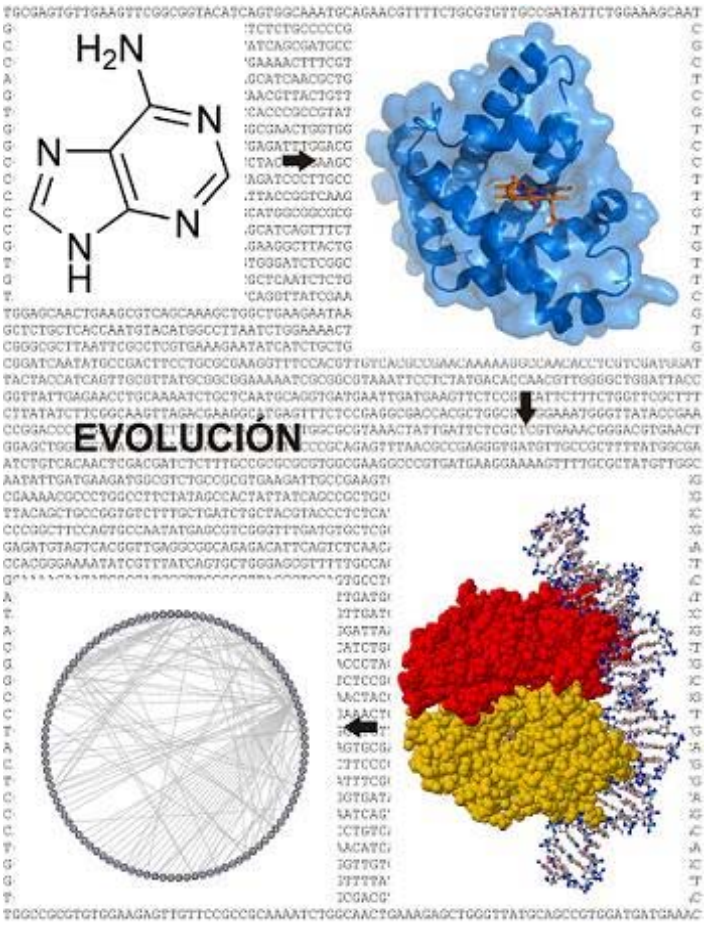
***En la imagen**, la filosofía de desarrollo open source.*

Multidisciplinariedad de la Bioinformática

Esta disciplina científica, también llamada biología computacional, atrae a investigadores de muchas áreas diferentes que se interesan por problemas de la biología que pueden ser abordados computacionalmente. En la arena bioinformática conviven biólogos, bioquímicos, físicos, matemáticos e ingenieros, y la escala de los problemas que tratan incluye desde pequeñas moléculas a sistemas complejos donde muchas moléculas u organismos coexisten. Sin embargo, de entre estas cuestiones podríamos destacar la genómica, que estudia los genomas de microorganismos, plantas y animales, que ha motivado quizás las aportaciones más importantes de nuestra disciplina.

En este contexto el tipo de datos que se manejan son secuencias (de ácidos nucléicos y de proteínas), y la operación fundamental que se realiza con ellos es la comparación. Precisamente para esto sirve el software más emblemático del área, BLAST, desarrollado en el National Center

for Biotechnology Information (Maryland, EEUU), que mantiene en el dominio público su código fuente.



***En la imagen**, escala de los problemas estudiados por la bioinformática, de las moléculas sencillas como la adenina, componente del ADN, a las redes moleculares, con la historia evolutiva como escenario.*

Bioinformática open source

La herramienta fundamental para un desarrollador de software es el lenguaje de programación en el que va a implementar un proyecto. En el contexto de la bioinformática se utilizan diferentes lenguajes, que en general son a su vez productos open source, como Perl, Python, C/C++, Java y otros lenguajes más especializados como R. Además, un recurso muy valioso para los programadores son las bibliotecas y librerías, que permiten reutilizar código para las tareas más estándar de un proyecto, ganando tiempo que se puede dedicar a programar las partes más complejas. Un ejemplo de este tipo de recursos sería el archivo CPAN, que contiene software libre escrito en Perl por más de 7500 autores, y ayuda a evitar reinventar la rueda en muchas cuestiones que ya han sido resueltas y probadas por la comunidad. En el ámbito de la biología computacional disponemos de bibliotecas especializadas como Bioconductor o Bioperl, con módulos y funciones para tareas habituales como el alineamiento de secuencias, la construcción de árboles filogenéticos o la anotación de genomas.

Dado el volumen de datos que se manejan actualmente en la genómica es necesario a menudo utilizar infraestructuras de cómputo complejas, como clusters de cálculo u ordenadores con muchos procesadores, que facilitan el cálculo en paralelo. Gracias al sistema operativo de código abierto Rocks, basado en Linux, cualquier aplicación de la bioinformática puede ejecutarse en un cluster, permitiéndonos realizar tareas a gran escala. Por otro lado, para sacar el máximo rendimiento a ordenadores multiprocesador es normalmente necesario adaptar el código de la aplicación que nos interese para partir el trabajo en dos ó más tareas que son resueltas independientemente en cada procesador. Esto es precisamente lo que hicieron los desarrolladores de mpiBLAST, tomando el código fuente de BLAST, para hacer el algoritmo de alineamiento original varias órdenes de magnitud más rápido repartiendo el trabajo en hasta cientos de procesadores. Por supuesto, el código fuente de mpiBLAST también es de dominio público.

Otro recurso importante en la bioinformática son las bases de datos, que ponen a disposición de la comunidad grandes colecciones de datos como secuencias, estructuras moleculares o experimentos de chips de ADN, que capturan el estado de expresión de todos los genes de un genoma en paralelo. A diferencia del software, que con pocas excepciones es libre para usuarios académicos, los investigadores normalmente sólo comparten sus datos más recientes, por medios de repositorios como los del Instituto Europeo de Bioinformática, cuando ya han publicado los artículos en que los describen.

***En la imagen**, Cluster CAESARAUGUSTA del BIFI, en Zaragoza.*

Filosofía open access en investigación en bioinformática

Como en cualquier área de la ciencia, los bioinformáticos publicamos regularmente los resultados de nuestro trabajo con el fin de que nuestros colegas los conozcan y así contribuir al crecimiento de nuestra disciplina. Hay diferentes formas de diseminar el trabajo científico, unas más coloquiales como las conferencias, y otras más formales como los artículos, normalmente

en inglés, en revistas científicas especializadas. También en este aspecto la bioinformática está probando la vía open source, que en este caso es el modelo de acceso abierto (open access) a la literatura científica. Las revistas más importantes del área, como PLoS Computational Biology, Bioinformatics, o BMC Bioinformatics, soportan este modelo, bajo el cual los autores de un artículo asumen los costes de publicación y por tanto sus lectores tienen libre acceso a él desde cualquier parte del mundo, normalmente en formato HTML o PDF, a través de Internet.

Made in Aragón

Con maestros como Alfonso Valencia, Roderic Guigó o Joaquín Dopazo, la bioinformática ha madurado en España en los últimos 15 años. A nivel autonómico quizás el mayor impulso ha sido la creación del Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), que cuenta con una infraestructura suficiente para poner a trabajar juntos en territorio aragonés a expertos de diferentes áreas en problemas, entre otros ámbitos, de la biología computacional.

Bajo este paraguas se han desarrollado los primeros proyectos locales que han producido software bioinformático abierto. Por ejemplo, el servidor ProtSA, desarrollado en la Universidad de Zaragoza, permite estimar la accesibilidad al solvente de proteínas en su estado desplegado (artículo open access aquí, en inglés). Otro ejemplo es la herramienta de diseño de marcadores moleculares primers4clades, creado en la Estación Experimental de Aula Dei (EEAD) en colaboración con la Universidad Nacional Autónoma de México (descrita en profundidad aquí, en inglés). Finalmente, en la EEAD está en desarrollo avanzado la base de datos 3D-footprint, que estima la especificidad de proteínas que se unen al ADN a partir de descripciones atómicas de su estructura, y complementa nuestro conocimiento de procesos biológicos tan fundamentales como la regulación genética (más detalles aquí, en inglés).

Definiciones:

- **Código fuente**. Es el conjunto de instrucciones, escritas en uno o más lenguajes de programación, que constituyen lo que llamamos un programa o software informático.
- **Algoritmo.** Es una lista bien definida, ordenada y finita de operaciones que permite hallar la solución a un problema,según definición de Wikipedia
- **Anotación.** Es el proceso de recopilación de información de bases de datos de secuencias y de la literatura para asignar posibles funciones moleculares a genes, proteínas y demás componentes de un genoma.

Autor: Bruno Contreras Moreira, Investigador de la Fundación ARAID, que desarrolla su trabajo en el Laboratorio de Biología Computacional, Estación Experimental de Aula Dei perteneciente a la delegación del CSIC en Aragón

http://www.aragoninvestiga.org/

http://www.aragoninvestiga.org/La-filosofia-open-source-en-la-bioinformatica/