

El Software Libre en Bioinformática

Biol. José Mauricio M. Herrera Cuadra

mauricio_at_intelligenomes_dot_com

Resumen

La bioinformática se ha convertido en una herramienta esencial para la investigación biológica. Gracias al desarrollo tecnológico e informático, es que los biólogos pueden llevar la búsqueda del significado de la vida hasta horizontes que antes no se consideraban posibles.

Se describirán brevemente algunas de las herramientas de software libre más utilizadas en bioinformática, así como su impacto y aplicación en el mundo de la investigación biológica.

¿Qué es la Bioinformática?

El término bioinformática es una invención bastante reciente, el cual apareció en la literatura alrededor de 1991, en aquel entonces únicamente dentro del contexto de la publicación electrónica. El concepto de bioinformática es mejor descrito como la convergencia de dos revoluciones tecnológicas: el crecimiento explosivo de la biotecnología, paralelamente con el crecimiento explosivo de la tecnología de la información. Esto se ilustra mediante el hecho de que tanto el tamaño de la base de datos *GenBank* como el del poder computacional se han duplicado aproximadamente al mismo ritmo (cada 18–24 meses) durante muchos años (Figura 1).

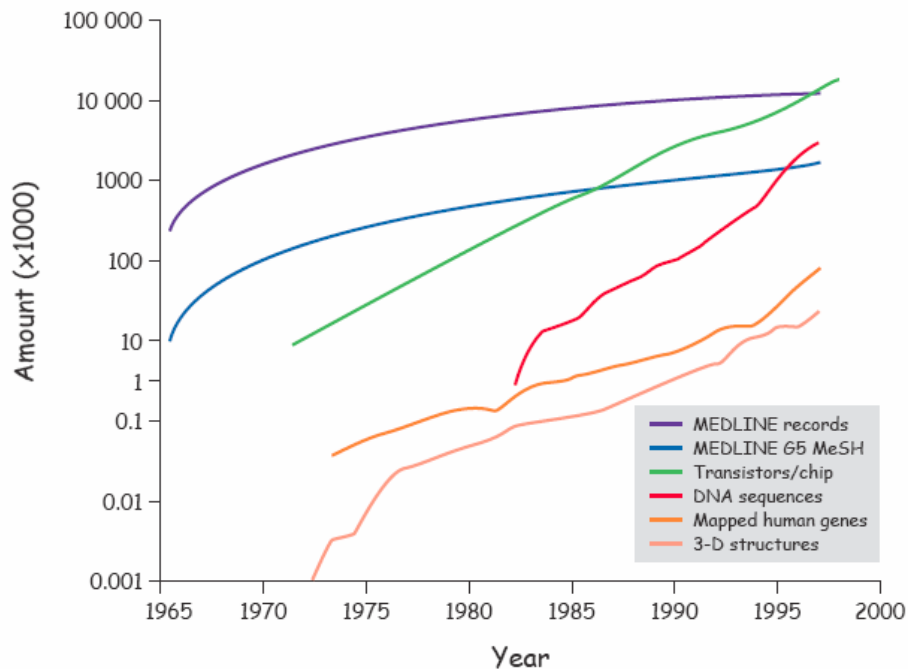


Figura 1. Crecimiento acumulado de información biomédica y poder de cómputo. *MEDLINE* (línea morada) es la base de datos bibliográfica del US National Library of Medicine (<http://www.nlm.nih.gov/Entrez/medline.html>) y actualmente contiene >10 millones de registros derivados de artículos publicados en >3900 revistas biomédicas. Los artículos categorizados

dentro del G5 Medical Subject Heading (MeSH) de biología molecular y genética (línea azul) suman cerca de 1 millón. El número total de registros de secuencias de DNA en *GenBank* (línea roja) es >2.5 millones (dato de <ftp://ncbi.nlm.nih.gov/genbank/gbrel.txt>). Nótese que ahora existen más registros de secuencias en *GenBank* que el número de publicaciones relacionadas en la literatura, esto indica un suceso importante y al mismo tiempo un aumento en nuestra falta de comprensión de las funciones de estas secuencias. Alentadoramente, las tecnologías de genómica funcional nos ayudarán a reducir este hueco. La línea que representa el número de transistores por chip (línea verde) se refiere a microprocesadores Intel™ e ilustra la Ley de Moore, la cual se refiere al ritmo de crecimiento exponencial del poder de cómputo (dato obtenido de <http://www.physics.udel.edu/wwwusers/watson/scen103/intel.html>). El número de genes humanos mapeados (línea naranja) es actualmente >30,000 (<http://www.ncbi.nlm.nih.gov/genemap>). El número de estructuras tridimensionales de proteína en el *Protein Data Bank* (pdb) (línea rosa) es actualmente ~7500 (<http://www.pdb.bnl.org>).

La información biológica, y los datos de secuencias de DNA en particular, se están acumulando a un ritmo excepcional. En junio del año 2000, se logró el primer ensamblaje de la secuencia completa del genoma humano. Aunque este logro pudo parecer el final de un largo camino, en realidad fue solamente el principio. Para poder explotar la riqueza de las secuencias de DNA y otros datos biológicos, surgió una nueva área que fusiona la biología con las matemáticas y las ciencias de la computación: la bioinformática.

Localizar los genes dentro de las secuencias es una tarea intensa por sí misma. Las regiones aparentemente codificantes o sin caracterizar deben de ser asignadas a una función. Subsecuentemente, deben ser comprendidas las interacciones entre los genes y sus productos a todos los niveles, no solamente en el contexto de las rutas dentro y fuera de las células, sino también en términos de la evolución de las familias de genes dentro y entre las especies. Estas preguntas pueden ser abordadas utilizando la bioinformática. La bioinformática se emplea en toda la biología, y el acceso directo a los datos por medio de Internet significa que un gran conjunto de información esta disponible, literalmente al alcance de nuestras manos.

¿Es la bioinformática el análisis del DNA y las proteínas? ¿El manejo de conjuntos de datos? ¿La utilización de computadoras para procesar grandes cantidades de datos en biología? Tal vez. Lincoln Stein –experto en bioinformática, también conocido por la creación del módulo CGI de Perl– dice que la bioinformática es: “*Biólogos utilizando computadoras, o al revés*”. El no esta exactamente seguro de lo que es, pero afirma que esta creciendo. La bioinformática es más una herramienta que una disciplina. La biología es una disciplina, una ciencia. Finalmente, la bioinformática es realmente biología también: biología de alto desempeño, biología integrativa, pero biología sin duda.

¿Por qué utilizar Software Libre en Bioinformática?

Existen varias razones por lo que es adecuada la utilización de software libre en la investigación biológica. A continuación describiré brevemente algunas de ellas:

Transparencia

Usar cualquier sistema operativo o programa libre supone utilizar una herramienta cuyo código fuente siempre está disponible. Podría argumentarse que en realidad, con ciertos programas poco importa el funcionamiento interno mientras obtengamos el resultado deseado, siendo valioso para otras actividades el tiempo invertido en entender el programa, pero utilizando software libre tenemos la garantía de que muchos programadores capacitados –que sí tienen tiempo y ganas de asegurarse que los programas funcionan como deben– han verificado su funcionamiento interno.

Siempre es bueno invertir cierto tiempo en entender cómo funcionan los programas aunque sea a un nivel superficial. Aunque al principio puede parecer difícil, a la larga es una inversión, por que nos da autonomía. En el caso del software propietario, el papel y el lápiz nos están vetados a priori. Un científico o técnico debería tener siempre la posibilidad de indagar en las herramientas que usa, hasta el nivel que él considere oportuno. La transparencia del software libre entronca perfectamente con la tradición científica de hacer públicos todos los procedimientos de investigación.

Potencia, estabilidad y flexibilidad

Los sistemas operativos libres y sus programas asociados son en general más estables y potentes. En misiones de alta exigencia de cómputo esta diferencia se hace crucial (muchos proyectos que requieren procesamiento en paralelo se llevan a cabo con software libre), pero a un nivel mucho más doméstico o de aplicaciones que no requieren muchos recursos también aumenta el rendimiento considerablemente.

Otra gran característica de los sistemas libres es su flexibilidad. En ellos se puede ajustar absolutamente todo el sistema en función de los conocimientos y necesidades del usuario.

Economía

El software libre, por lo general, es gratuito, aunque no hay que confundir nunca estos dos parámetros. La mayoría de las herramientas más utilizadas en bioinformática se pueden descargar íntegramente de la red. Por esta misma razón, hay muchos países en vías de desarrollo que están optando por la utilización de software libre en sus organismos oficiales, universidades, etc.

Independencia de plataforma

Afortunadamente, en bioinformática se han desarrollado formatos de archivo manipulables en cualquier sistema operativo, esto debido a que la gran mayoría del software utilizado en el área se encuentra disponible tanto para plataformas libres como propietarias.

En los países en vías de desarrollo se opta por el uso de software libre debido a que los técnicos que se forman se convierten en una inversión a futuro. Este proceso es aplicable también para cualquier laboratorio de investigación o empresa. El ser usuarios de una plataforma abierta y libre nos permite tener un control absoluto de las herramientas que utilizamos, así como de realizar mejoras en redundancia de nuestro laboratorio y toda la comunidad científica.

Respeto a los estándares

Si utilizamos software libre tenemos la garantía de que los datos que produzcamos respetarán los estándares abiertos y públicos. Los científicos o técnicos deberían ser especialmente sensibles hacia esta problemática, ya que es precisamente la labor científico-técnica la que más escrutinio requiere, sobretodo de colegas pero también del público en general.

Software Libre destacado en Bioinformática

A continuación describiré algunas de las herramientas de software libre más empleadas en bioinformática, las cuales –desde mi personal punto de vista– han logrado un mayor impacto dentro de la investigación biológica:

Perl

Una gran parte de la biología computacional consiste de tareas frecuentes de procesamiento de textos, tales como la manipulación de cadenas, concordancia de expresiones regulares, traducción de archivos, e interconversión de formato de datos. Por consiguiente, muchos desarrolladores en la comunidad bioinformática hacen uso extenso del lenguaje de programación Perl, el cual sobresale en dichas tareas.

Perl es popular entre los biólogos debido a su carácter práctico. La información biológica en las computadoras tiende a estar organizada en archivos de texto o en bases de datos relacionales. Cualquiera de estas fuentes de datos es fácil de manejar con programas en Perl. Perl se ha convertido en una especie de fenómeno en el área, puesto que muchos biólogos lo encuentran como un lenguaje fácil de aprender que posee muchas de las herramientas que ellos necesitan: en particular su soporte para el procesamiento de textos y expresiones regulares lo hacen adecuado para tareas complejas de traducción de textos (comunes en bioinformática).

Perl ha demostrado ser un poderoso y fácil lenguaje de alto nivel para programación, desarrollo orientado a objetos, y desarrollo rápido de prototipos para software bioinformático. Los programas en Perl pueden ser vistos como modelos para bio-objetos y conceptos que puedan ser reimplementados en otros lenguajes de programación.

Perl ha madurado de un simple lenguaje de "script" a un poderoso ambiente de programación tanto para el estilo procedimental como para el orientado a objetos. Mientras que sigue siendo utilizado para crear programas simples "desechables", también se utiliza para diseñar aplicaciones complejas, modulares, bien documentadas y mantenibles. La facilidad de utilización de Perl para una variedad de tareas, tanto de alto nivel como para programación de CGI, es inigualable.

Esto sucede debido a que Perl es mucho más poderoso de lo que la gente piensa, pero su poder proviene de una manera interesante. Perl posee el inusual don de unir cosas. Esto lo hace mejor que cualquier otro lenguaje, y lo hace en muchos niveles diferentes. He aquí algunos ejemplos:

- Perl puede unir sitios Web y bases de datos a través de los módulos [CGI](#) y [DBI](#).
- El CPAN (Comprehensive Perl Archive Network) organiza todos los módulos del mundo juntos y hace fácil la búsqueda a través de ellos.
- Los módulos [Inline](#) permiten a Perl ejecutar código de otros lenguajes de programación, tal como si fuera Perl nativo. Por ejemplo, [Inline::Python](#) permite a los programadores utilizar objetos de Python tal y como si estuvieran utilizando objetos de Perl.
- En Perl existen muchos conceptos y sintaxis de diferentes lenguajes de programación. Siempre es agradable poder observar algo familiar cuando se está aprendiendo algo nuevo. Aún los programadores novatos pueden beneficiarse de esto si pueden reconocer el parecido de Perl con el Inglés hablado.

Un ejemplo sobresaliente del papel que ha jugado Perl en bioinformática, es cuando permitió a los científicos del Proyecto Genoma Humano el intercambiar datos y comparar los resultados que se estaban produciendo en 2 diferentes centros de secuenciamiento.

Existe una gran variedad de aplicaciones bioinformáticas desarrolladas en Perl, algunos ejemplos se encuentran en la siguiente tabla:

Tabla 1. Aplicaciones bioinformáticas desarrolladas en Perl

Aplicación	Descripción
MULTICLUSTAL	Automatiza el proceso de escoger los parámetros de alineamiento para <i>Clustal W</i> , con objeto de generar alineamientos de alta calidad.
Oliz	Busca oligonucleótidos específicos a genes para experimentos de microarreglos.
Pise	Genera interfaces de Web a partir de programas de biología molecular.
Swissknife	Permite extraer o modificar registros dentro de archivos con formato SWISS-PROT.
Virtual PCR	Predice productos de PCR a partir de iniciadores introducidos por el usuario.
WebPHYLP	Interfaz de Web para <i>PHYLP</i> , permite realizar análisis filogenéticos a través de Internet.

Bio* Toolkits

En 1995 se formó la Open Bioinformatics Foundation (OBF) por un grupo de auto-denominados hackers de Perl, con el objeto de reunir recursos para el desarrollo de software bioinformático. Esta fundación cuenta actualmente con los siguientes proyectos: *BioPerl*, *BioPython*, *BioJava*, *BioCORBA* y *BioDAS*. Los grupos *BioRuby*, *BioLisp* y *Bioinformatics.org* comparten una visión similar y vale la pena conocerlos para obtener una perspectiva y recursos útiles, pero se encuentran oficialmente desafiliados.

Dentro de estos, *BioPerl* es el más antiguo y utilizado, y por buenas razones. Es ciertamente el más maduro, posee las características más útiles y la comunidad de desarrollo más grande. La documentación de la estructura de *BioPerl* (aproximadamente 60 niveles, ~400 módulos) es manejada en un diseño extremadamente bien logrado que hace fácil examinar la funcionalidad del proyecto. La documentación para cada módulo es bastante completa y contiene descripciones cortas del módulo, sus dependencias, sus herencias y ejemplos de código para cada método.

El proyecto *BioPerl* es un esfuerzo internacional de biólogos, bioinformáticos e ingenieros en sistemas, que ha evolucionado –a través de casi 10 años de desarrollo– en la librería de Perl más completa para el manejo y manipulación de información biológica. Los módulos de *BioPerl* han sido repetitiva y satisfactoriamente utilizados para convertir tareas complejas en pocas líneas de código. Su modelo orientado a objetos es lo suficientemente flexible como para soportar aplicaciones de nivel empresarial tales como *EnsEMBL*, pero al mismo tiempo mantiene una fácil curva de aprendizaje para programadores novatos. *BioPerl* es capaz de ejecutar análisis y procesar resultados de programas tales como *BLAST*, *Clustal W* o la suite *EMBOSS*. Su interoperabilidad con módulos escritos en Python y Java es posible a través del puente *BioCORBA*. *BioPerl* provee acceso a datos generados por *GenBank* y *SwissProt* a través de una serie de flexibles módulos de entrada/salida de datos, así como al reciente formato de almacenamiento común de datos del proyecto Open Bioinformatics Database Access.

NCBI Toolkit

Debido al creciente rol de la bioinformática en los Estados Unidos, el National Center for Biotechnology Information (NCBI) fue creado en Noviembre de 1988, con los siguientes objetivos:

1. Crear un sistema de conocimiento automatizado sobre biología molecular, bioquímica y genética.
2. Realizar investigación dentro de métodos avanzados de análisis e interpretación de datos de biología molecular.
3. Permitir a los investigadores en biotecnología y personal médico la utilización de los sistemas y métodos desarrollados.
4. Coordinar esfuerzos para recopilar información biológica a nivel internacional.

Tomando en cuenta que los datos de biología molecular provienen de una forma extremadamente heterogénea, distribuida y cambiante, la información procesada e integrada a través del NCBI debe poder ser procesada y manipulada en múltiples sistemas operativos utilizando diferentes sistemas de manejo de datos. Esto implica que los datos deben ser descritos y controlados de manera formal, de modo que cualquiera pueda comprender cuales componentes comunes se encuentran disponibles en cualquier momento, sin la necesidad de depender en alguna herramienta de software, lenguaje de programación, base de datos o arquitectura de hardware.

Debido a esto, el NCBI adoptó la utilización del Abstract Syntax Notation 1 (ASN.1) y el International Standards Organization standard (ISO 8824, 8825) para la descripción y codificación de datos de una manera legible que fuera independiente de cualquier plataforma.

En cuanto a la portabilidad del software y sus diferentes niveles de acceso, el NCBI desarrollo el *NCBI Toolkit*. Este toolkit se utiliza internamente para procesar y analizar datos de una variedad de fuentes, para construir y mantener bases de datos unificadas y al mismo tiempo sirve de componente para las aplicaciones de usuario final que el NCBI distribuye (*Entrez*, *Sequin*, *BLAST*, etc.). El código fuente del toolkit se encuentra disponible sin restricción para la utilización de cualquier interesado en aprovechar el trabajo realizado por el NCBI. El software funciona sobre una gran variedad de plataformas y consiste de varios niveles, que permiten a los programadores utilizar herramientas tanto de bajo como de alto nivel para la manipulación de la información.

El toolkit consiste de los siguientes componentes principales:

- **ASN.1:** un lenguaje formal para la descripción de datos, desarrollado, probado y utilizado dentro de la industria computacional, más no es un formato desarrollado por biólogos.
- **Modelo de Datos para Secuencias Biológicas:** un modelo para información biológica (especificado en ASN.1) cuyo concepto es el de una secuencia biológica como un sistema simple y lineal de coordenadas. Los mapas genéticos y físicos, piezas secuenciadas de ácidos nucleicos y proteínas, y los ensamblajes complejos de tales componentes, pueden ser considerados como especializaciones del concepto básico de secuencia de un sistema de coordenadas específico.
- **CoreLib:** un pequeño conjunto de funciones, macros y pautas en lenguaje C, que permiten la escritura de programas que se compilan y ejecutan sin cambios dentro de 14 combinaciones diferentes de hardware/sistema operativo/compilador.
- **AsnLib:** una librería de funciones escritas utilizando CoreLib, que provee métodos para la lectura y validación de especificaciones ASN.1, así como la generación de árboles para la codificación y decodificación de datos conforme a la especificación.
- **Cargadores de objeto (Object Loaders):** cada módulo de especificación ASN.1 en el modelo de datos NCBI posee un módulo "cargador de objeto". Estos son archivos `.c` y `.h` que tipifican una estructura por cada entidad definida en ASN.1 (llamada "objeto").
- **Utilidades:** un gran numero de funciones útiles escritas para la manipulación o análisis de estructuras definidas en los cargadores de objeto.
- **Acceso a Datos:** una familia de funciones que proveen acceso de alto nivel a datos de secuencias y bibliográficos (*Entrez*, *MEDLINE*). Estas funciones han sido implementadas para el acceso a los servidores de datos del NCBI y se encuentran disponibles desde 1993.
- **Vibrant:** un portable sistema de ventanas que utiliza CoreLib y que permite desarrollar aplicaciones idénticas en código fuente para Macintosh, Windows y X11.

De todas las aplicaciones bioinformáticas derivadas del *NCBI Toolkit*, la más conocida y utilizada es *BLAST* (Basic Local Alignment Search Tool), y puede encontrarse en una variedad de sitios de bioinformática a nivel mundial. La creciente demanda de búsquedas dentro de bases de datos que se encuentran en constante crecimiento, requirió de una solución para obtener resultados rápidos y

precisos. El algoritmo de *BLAST* genera resultados de alineamiento (regiones locales de similitud) contra elementos de una base de datos a partir de una secuencia de interés. Existen 2 implementaciones principales de *BLAST*: la desarrollada por el NCBI (*BLAST*), y la desarrollada por la Universidad de Washington (*WU-BLAST*).

EMBOSS y EMBASSY

EMBOSS (European Molecular Biology Open Software Suite) es una colección de utilidades de bioinformática y librerías de software, diseñada para ser utilizada de manera individual, empotrada en scripts o para el desarrollo de programas. Miembros de EMBnet y la comunidad de usuarios en general han contribuido al rango de programas incorporados en este paquete libre. La documentación para cada una de las aplicaciones puede ser encontrada en el sitio Web de *EMBOSS*. La integración de estos programas permite que la salida de una aplicación se convierta en la entrada de otra, sin la necesidad de reformatear los datos. Esta integración ha sido desarrollada más profundamente por la salida de una serie estándar de formatos de reporte. *EMBOSS* también ha sido diseñado para reconocer más de 40 formatos de datos, y puede ser utilizado para crear archivos en diferentes formatos, dependiendo de los requerimientos del usuario.

Dentro de *EMBOSS* se pueden encontrar programas que cubren las siguientes áreas:

- alineamiento de secuencias
- búsqueda rápida en bases de datos utilizando patrones de secuencias
- identificación de motivos de proteínas, incluyendo análisis de dominios
- análisis de patrones de secuencias de nucleótidos
- análisis de utilización de codones para genomas pequeños
- identificación rápida de patrones de secuencias dentro de conjuntos grandes de secuencias
- herramientas de presentación para publicación

EMBOSS se encuentra bajo la licencia GPL. Las librerías se encuentran bajo la LGPL. Los programas de terceros que han sido incluidos y que poseen sus propios términos de licencia se encuentran separados bajo el agrupamiento *EMBASSY*. Esto permite a las librerías de *EMBOSS* enlazarse con otros programas y solamente se requiere que el software posea una licencia compatible con la LGPL. Sin embargo, para el usuario estos programas funcionan exactamente como si fueran aplicaciones de *EMBOSS*.

Los programas incluidos en *EMBASSY* son: *PHYLIP*, *PHYLIPNEW*, *DOMAINATRIX*, *CONSTRUCT*, *HMMER*, *EMNU*, *ESIM4*, *MEME*, *MSE*, *TOPO* y *CRYSTALBALL*.

Clustal W/X

La serie de programas *Clustal* es ampliamente utilizada para la elaboración de alineamientos múltiples y la preparación de árboles filogenéticos. Su desarrollo actual lo patrocinan el INSERM (Institut National de la Santé et de la Recherche Médicale), CNRS (Centre National de la Recherche Scientifique), el Ministère de l'Éducation Nationale de la Recherche et de la Technologie y el EMBL (European Molecular Biology Laboratory).

Los programas han pasado por varias encarnaciones, hasta que en 1997 se liberó la versión 1.7 de *Clustal W* y *Clustal X* (el cual posee una interfaz de ventanas). Podría pensarse que la gente utiliza los programas *Clustal* debido a que producen buenos alineamientos, sin embargo, una de las razones de su amplia utilización ha sido su portabilidad a diferentes sistemas operativos.

La portabilidad puede poseer un lado negativo, y por muchos años la interfaz de *Clustal* tuvo que permanecer simple. *Clustal X* provee una interfaz grafica de usuario más agradable para X11, Macintosh y Windows, manteniendo su portabilidad gracias a la utilización de las librerías Vibrant del *NCBI Toolkit* anteriormente descritas. *Clustal X* presenta los alineamientos utilizando colores para la conservación de residuos y posee una herramienta para el marcado de regiones de alineamiento pobres. Además, el usuario puede seleccionar tales regiones para realineamiento. Gracias a esto, *Clustal X* posee una mayor flexibilidad ante las estrategias existentes para la preparación de alineamientos múltiples.

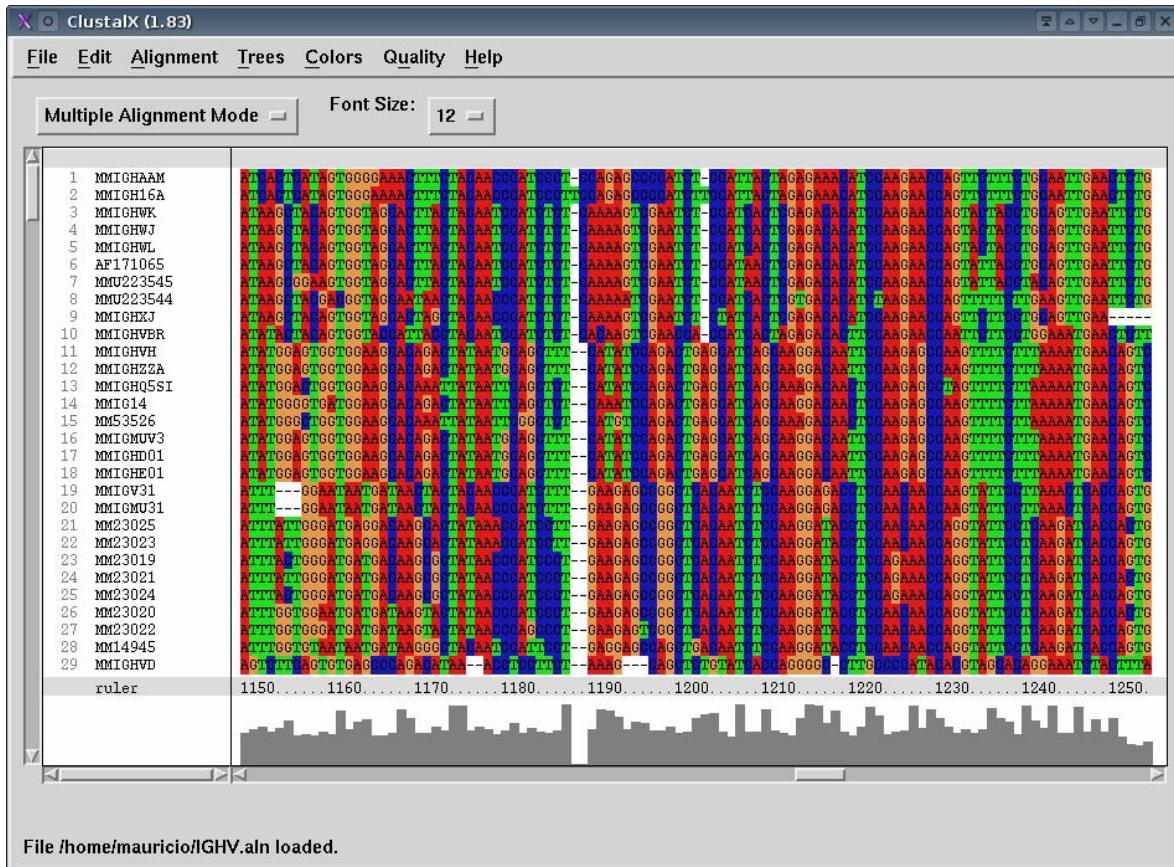


Figura 2. Pantalla de *Clustal X* mostrando un alineamiento múltiple.

Tendencias y Perspectivas

Tradicionalmente, las bases de datos biológicas consistían de pequeños registros en formato de texto plano, que se relacionaban con otros registros mediante campos conocidos como “referencias cruzadas” y permitían interconectar diversas bases de datos (*PROSITE*, *EMBL*, *GenBank*, etc.). En la actualidad, existen robustos desarrollos que permiten no solo la recopilación de registros de gran tamaño, sino la integración múltiple de dichos registros para generar resultados detallados y exhaustivos, que inclusive pueden llegar a permitir la navegación a través de genomas completos (*EnsEMBL*).

Así mismo, existe la necesidad de utilizar modelos de cómputo en paralelo y/o distribuido para la resolución de análisis que requieren de un intenso poder de procesamiento, así como del manejo de grandes volúmenes de datos. Dentro de esta tendencia existen aplicaciones para el análisis de secuencias mediante clusters (*BLAST*, *HMMER*, *Clustal W*, etc.), predicción de estructuras de proteína (*Folding@home*, *Genome@home*, etc.), predicción del clima mediante simulaciones Monte Carlo

(*Climateprediction.com*), actividad de fármacos sobre estructuras de proteína (*THINK*, *AutoDock*, *Find-a-Drug*, etc.), entre otros.

Actualmente, los campos donde el software libre no científico (Perl, XML, CORBA, SOAP, etc.) desempeña un mayor papel dentro de la bioinformática, es en la generación de interfaces de Web (*Pise*, *WebPHYLIP*), así como de Web Services (*BioMOBY*).

Cada día se recomienda más la aplicación del modelo de desarrollo de software libre dentro de los proyectos de software científico, debido a que son mucho más rentables y financiables.

Bajo esta misma perspectiva, toda la información generada a partir de la investigación biológica deberá ser de acceso público. De lo contrario, toda la información de dominio público en el futuro será restringida más por las patentes que por el código fuente propietario.

Referencias

- **Achard, F., Vaysseix, G. y Barillot, E.** 2001. XML, bioinformatics and data integration. *Bioinformatics*. 17, 115-125.
- **Altschul, S., Gish, W., Miller, W., Myers, E. y Lipman, D.** 1990. Basic Local Alignment Search Tool. *Journal of Molecular Biology*. 215, 403-410.
- **Birney, E., Andrews, D., Bevan, P., Caccamo, M., Cameron, G., Chen, Y., Clarke, L., Coates, G., Cox, T., Cuff, J., Curwen, V., Cutts, T., Down, T., Durbin, R., Eyas, E., Fernandez-Suarez, X., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H., Iyer, V., Kahari, A., Jekosch, K., Kasprzyk, A., Keefe, D., Keenan, S., Lehvaslaiho, H., McVicker, G., Melsopp, C., Meidl, P., Mongin, E., Pettett, R., Potter, S., Proctor, G., Rae, M., Searle, S., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Ureta-Vidal, A., Woodwark, C., Clamp, M. y Hubbard, T.** 2004. Ensembl 2004. *Nucleic Acids Research*. 32, D468-D470.
- **Boguski, M.** 1998. *Bioinformatics – a new era*. Trends Guide to Bioinformatics. Elsevier Science.
- **Cerami, E.** 2002. *Web Services for Bioinformatics*. Publicado en The O'Reilly Network. <http://www.oreillynet.com/pub/a/webservices/2002/05/14/biows.html>
- **Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T., Higgins, D. y Thompson, J.** 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*. 31, 3497–3500.
- **Egaña, M.** 2003. *Software Libre (GNU/Linux) para Biólogos*. http://www.sindominio.net/~pik/web_files/linuxbiologos.pdf
- **EMBOSS Project.** 2005. EMBOSS Homepage. <http://emboss.sourceforge.net/>
- **Gilbert, D.** 2002. Pise: Software for building bioinformatics webs. *Briefings in Bioinformatics*. 3, 405-409.
- **Hokamp, K., Shields, D., Wolfe, K. y Caffrey, D.** 2003. Wrapping up BLAST and other applications for use on UNIX clusters. *Bioinformatics*. 19, 441-442.
- **Jeanmougin, F., Thompson, J., Gouy, M., Higgins, D. y Gibson, T.** 1998. Multiple sequence alignment with Clustal X. *TIBS*. 23, 403-405.
- **Krieger, E. y Vriend, G.** 2002. Models@Home: distributed computing in bioinformatics using a screensaver based approach. *Bioinformatics*. 18, 315-318.
- **Letondal, C.** 2001. A Web interface generator for molecular biology programs in UNIX. *Bioinformatics*. 17, 73-82.
- **Lim, A. y Zhang, L.** 1999. WebPHYLIP: a web interface to PHYLIP. *Bioinformatics*. 15, 1068-1069.
- **Loewe, L.** 2002. Global computing for bioinformatics. *Briefings in Bioinformatics*. 3, 377-388.
- **Mangalam, H.** 2002. The Bio* toolkits – a brief overview. *Briefings in Bioinformatics*. 3, 296-302.

- **McMillan, R.** 2002. Natural Open Source. *Linux Magazine*. Junio 2002.
- **Mulder, N. y Apweiler, R.** 2001. Tools and resources for identifying protein families, domains and motifs. *Genome Biology*. 3, 1-8.
- **Mullan, L. y Bleasby, A.** 2002. Short EMBOSS User Guide. *Briefings in Bioinformatics*. 3, 92-94.
- **Mullan, L. y Williams, G.** 2002. BLAST and Go? *Briefings in Bioinformatics*. 3, 200-202.
- **National Center for Biotechnology Information.** 2002. Information Engineering Branch Home Page. <http://www.ncbi.nlm.nih.gov/IEB/>
- **Olson, S.** 2002. Emboss opens up sequence analysis. *Briefings in Bioinformatics*. 3, 87-91.
- **O'Reilly, T.** 2001. What's Next for Linux and Open Source? *Linux Magazine*. Octubre 2001.
- **Stajich, J., Block, D., Boulez, K., Brenner, S., Chervitz, S., Dagdigian, C., Fuellen, G., Gilbert, J., Korf, I., Lapp, H., Lehtväslaiho, H., Matsalla, C., Mungall, C., Osborne, B., Pocock, M., Schattner, P., Senger, M., Stein, L., Stupka, E., Wilkinson, M. y Birney E.** 2002. The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Research*. 12, 1611-1618.
- **Stein, L.** 1996. How Perl Saved the Human Genome Project. *The Perl Journal*. 1, 5-9.
- **Stewart, J., Mangalam, H. y Zhou, J.** 2001. Open Source Software meets gene expression. *Briefings in Bioinformatics*. 2, 319-328.
- **Story, D.** 2003. *An Understandable Definition of Bioinformatics*. Publicado en The O'Reilly Network. <http://www.oreillynet.com/cs/user/wlg/2716>
- **Thompson, J., Gibson, T., Plewniac, F., Jeanmougin, F. y Higgins, D.** 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*. 25, 4876-4882.
- **Trelles, O.** 2001. On the parallelisation of bioinformatics applications. *Briefings in Bioinformatics*. 2, 181-194.
- **Wilkinson, M. y Links, M.** 2002. BioMOBY: An open source biological web services proposal. *Briefings in Bioinformatics*. 3, 331-341.
- **Wong, L.** 2002. Technologies for integrating biological data. *Briefings in Bioinformatics*. 3, 389-404.