50. Bioinformática: bases de datos y software de análisis

Gabriel Dorado Pérez

Departamento de Bioquímica y Biología Molecular, Campus Universitario de Rabanales, Edificio Severo Ochoa, 14071-Córdoba

RESUMEN

La cantidad de información sobre ácidos nucleicos y péptidos que han sido secuenciados es tan voluminosa, que su análisis requiere el empleo de medios informáticos. De hecho, la bioinformática es una disciplina en plena expansión. En este capítulo se describen diversas herramientas informáticas que pueden resultar de gran utilidad en biología molecular. Las secuencias de ácidos nucleicos son almacenadas en diversas bases de datos, existiendo diversas aplicaciones de software para analizarlos. Existen básicamente tres formas de acceso a estas bases de datos: conectar con un potente servidor central (localizado en cualquier parte del mundo) vía InterNet, con un sistema personal, o mediante soluciones híbridas entre ambos (IntraNet). Se describe la búsqueda de la secuencia de DNA en una base de datos, el análisis de restricción de la misma, la selección del fragmento de interés y el diseño de cebadores para su amplificación mediante PCR y para la secuenciación del mismo.

Palabras clave: algoritmo, BLAST, genoma, informática, proteína, RNA.

Abreviaturas empleadas (por orden alfabético de abreviatura). DVD: disco digital versátil; EBI: Instituto Europeo de Bioinformática; EMBL: Laboratorio Europeo de Biología Molecular; GCG: Grupo Genético de Computación; NCBI: Centro Nacional para la Información Biotecnológica; PCR: reacción en cadena de la polimerasa; SIDA: síndrome de inmunodeficiencia adquirida.

1. INTRODUCCIÓN Y OBJETIVOS

La bioinformática es una disciplina en plena expansión. En este apartado se describirán diversas herramientas informáticas que pueden resultar de gran utilidad en biología molecular.

2. BASES DE DATOS

La cantidad de información sobre ácidos nucleicos y péptidos que han sido secuenciados es tan voluminosa, que su análisis requiere el empleo de medios informáticos. Las secuencias de ácidos nucleicos son almacenadas en diversas bases de datos, siendo GenBank/EMBL la más popular. Actualmente esta base de datos (que incluye secuencias de DNA) ocupa aproximadamente 200 GB, por lo que suele distribuirse en formato DVD-ROM y en el futuro lo será en

otros formatos (HD-DVD, Blu-Ray y sistemas holográficos) y —sobre todo— vía InterNet (ver más adelante).

Se han organizado también otras bases de datos donde se recopila información sobre RNA y proteínas en general, así como específicas sobre tRNA, virus del SIDA, etc.

Existen básicamente tres formas de acceso a estas bases de datos: conectar con un potente servidor central (localizado en cualquier parte del mundo) vía InterNet, con un sistema personal, o mediante soluciones híbridas entre ambos (IntraNet).

La primera opción (InterNet) utiliza superordenadores (p.ej., Cray) o potentes servidores (p.ej., SiliconGraphics/MIPS, Sun/Sparc, HP/PA-RISC, IBM/PowerPC, Digital/Alpha, AMD o Intel) con procesamiento masivamente paralelo a los que se conectan los usuarios mediante terminales informáticas. Suele ser el método más barato, pero tiene la desventaja de que los tiempos de espera pueden llegar a ser muy grandes, debido a la saturación del sistema y a la escasa velocidad que ofrece la línea telefónica de datos en estos momentos. Se esperan mejoras sustanciales en el futuro (satélites, tecnología ATM, fibra óptica, etc).

La segunda posibilidad (el sistema personal) puede resultar algo más cara, pero su eficiencia suele ser significativamente superior. En este caso, el factor limitante es disponer de un potente microprocesador para realizar las búsquedas y analizar los resultados. Lo ideal es un sistema basado en Unix, y por tanto multiproceso (p.ej., Xserve/Xserve RAID/Xserve Cluster con MacOS X Server). Gracias a los logros tecnológicos alcanzados en la fabricación de microprocesadores, se está haciendo realidad el sueño de explotar la potencia de los superordenadores (casi) al precio de los ordenadores personales.

La tercera solución (IntraNet) pretende aprovechar las ventajas de las dos primeras posibilidades. Se trata de un servidor local (típicamente una máquina con multiproceso simétrico y basada en Unix, para uso dentro del campus universitario) al que acceden los usuarios por medio de una red EtherNet de alta velocidad basada en fibra óptica.

3. SOFTWARE DE ANÁLISIS

3.1. Accelrys GCG (GCG/Wisconsin)

El paquete Accelrys GCG (anteriormente conocido como GCG/Wisconsin) de Accelrys (San Diego, CA, USA; <http://www.accelrys.com>) también conocido como Wisconsin es, con diferencia, el más potente y exhaustivo del mercado para el análisis de ácidos nucleicos y proteínas. Su único —y muy significativo— inconveniente es que sólo está disponible para sistemas Unix bien bajo su típica interfaz de comandos o con interfaz gráfica XWindow (Unix). Ello significa que para llegar a “dominarlo” es necesario un gran esfuerzo de aprendizaje. Incluso su versión web no es todo lo intuitiva que sería de desear. Otro inconveniente es que no permite realizar las excelentes presentaciones de datos y gráficos que pueden obtenerse con productos competidores.

Actualmente se está estudiando la posibilidad de desarrollar versiones de GCG para otros sistemas operativos más fáciles de usar, debido al significativo incremento de potencia de los procesadores usados en ordenadores personales y a la mejora de sus sistemas operativos (MacOS X sobre Intel). También se está desarrollando dicha versión “intuitiva” para ser consultada vía InterNet.

3.2. Una revolución de la información llamada Internet

La información que necesitas está ahí; en internet. A un clic del ratón de tu ordenador. Hace unos años se encontraba algo dispersa, a través de diversos protocolos de comunicación y con una interfaz de comandos y, por tanto, poco amigable (Telnet, FTP, Gopher, News, WAIS, eMail vía EAN/Pine, etc). Afortunadamente, todo ha sido unificado gracias al denominado “Web” (literalmente, “tela de araña” o “red”) o “World Wide Web” (literalmente, “tela mundial de araña” o “red de redes”), a través de los programas de ordenador Navigator y Communicator (NetscapeCommunications, Mountain View, CA, USA; en <**http://www.netscape.com**>), posteriormente copiados por Microsoft (Redmond, WA, USA; <http://www.microsoft.com>) bajo el nombre de Internet Explorer. Veamos algunos ejemplos de una fuente de recursos (Internet) que está creciendo exponencialmente:

El gobierno de los Estados Unidos mantiene el acceso gratuito (por ahora) vía Internet a los recursos informáticos del National Center for Biotechnology Information (NCBI) en <**http://www.ncbi.nlm.nih.gov>**. El NCBI incluye, entre otros servicios e informaciones, las búsquedas a la base de datos GenBank en <**http://www.ncbi.nlm.nih.gov/Web/Search/index.html**> y, en particular, búsquedas de secuencias de ácidos nucleicos, proteínas o texto a través de Entrez en <**http://www.ncbi.nlm.nih.gov/Entrez**> y búsquedas de similitud de secuencias vía BLAST en <**http://www.ncbi.nlm.nih.gov/BLAST**>. También pueden enviarse secuencias inéditas de DNA o proteínas para su inclusión en las bases de datos correspondientes (p.ej., GenBank para ácidos nucleicos) vía BanKit en <**http://www.ncbi.nlm.nih.gov/BankIt/index.html**>.

Otra herramienta de búsquedas de secuencias y texto del GenBank es el Bioccelerator en <**http://sgbcd.weizmann.ac.il>**, aunque está pensado más como un servicio comercial.

Otros recursos informáticos de biología molecular disponibles en internet pueden encontrarse en las páginas de la revista de métodos BioTechniques en <**http://www.biotechniques.com/biosrc.html**> e Internet On Ramp en <**http://www.tulane.edu/~dmsander/biotechniquessites.html**>. También son interesantes el Laboratorio Europeo de Biología Molecular (EMBL) en Alemania <**http://www.embl-heidelberg.de>** y su sede británica, denominada European Bioinformatics Institute (EBI) <**http://www.ebi.ac.uk/ebi_home.html**>. También es interesante y muy completa la denominada página web de biología molecular de Pedro en <**http://www.public.iastate.edu/~pedro/research_tools.html**>.

Merece especial atención BioSci (Bionet Usenet News) en <**http://www.bio.net>** y, particularmente, el grupo de métodos de biología molecular en <**bionet.molbio.methds-reagnts**>, de secuenciación automática

en <**bionet.genome.autosequencing**> y de software en <**bionet.software.www**>. Toda la correspondencia de BioSciNews está archivada en <**http://www.bio.net/hypermail>**. Por ejemplo, los archivos de los grupos de News anteriores (métodos, secuenciación y software), pueden consultarse en las siguientes tres direcciones, respectivamente: <**http://www.bio.net/hypermail/METHDS-REAGNTS**>, <**http://www.bio.net/hypermail/AUTOMATED-SEQUENCING**> y <**http://www.bio.net/hypermail/BIO-SOFTWARE**>. Todos los archivos de News pueden consultarse en DejaNews <**http://www.dejanews.com**>.

3.3. LaserGene

Actualmente, el paquete integrado personal que mejor explota las posibilidades ofrecidas por la informática (facilidad de uso y potencia) es LaserGene (DNAStar, Madison, WI, USA; <**http://www.dnastar.com**>). Éste consta de:

—30 DVD-ROMs correspondientes a secuencias de ácidos nucleicos (28), secuencias de proteínas (1) y datos cristalográficos obtenidos por difracción de rayos X (1). La mayor parte de las secuencias de DNA corresponden a las denominadas “Etiquetas de Secuencias Expresadas” (“Expressed Sequence Tags”; EST), correspondientes a fragmentos de mRNAs (cDNAs), disponibles en <**http://www.ncbi.nlm.nih.gov/dbEST/index.html**>. No obstante, ya han sido secuenciados algunos genomas completos (p.ej., el de *Escherichia coli*, arabidopsis, arroz, ratón, humano, etc), y otros muchos están secuenciándose; como puede consultarse en <**http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html**>.

—Software de biología molecular para el análisis y manipulación de secuencias de ácidos nucleicos y proteínas. Son ocho módulos: EditSeq, GeneQuest, SeqBuilder, MegAlign, PrimerSelect, Protean, SeqMan II y XRayViewer.

Para facilitar su uso, el paquete dispone también de la aplicación LaserGene Navigator; una pantalla gráfica donde los diferentes módulos del conjunto aparecen en forma de botones que pueden ser activados por el usuario. La última versión de LaserGene aprovecha las posibilidades de internet, de forma que permite una interacción transparente con el web del NCBI. Así, permite realizar comparaciones de secuencias mediante el algoritmo “BLAST” y recuperar secuencias obtenidas con la aplicación “Entrez”.

Activando GeneMan se accede a las diferentes bases de datos del CD-ROM. Desde GeneMan pueden realizarse búsquedas tanto de texto (p. ej., “homo”) como de secuencias (p. ej., “ACGTACGTACGTACGT”). El resultado de la búsqueda puede grabarse para su posterior análisis. Típicamente, la secuencia de interés obtenida con GeneMan se abre después desde EditSeq, que permite su edición o modificación, según diversos criterios (p. ej., para eliminar las regiones que no interesan, obtener la cadena complementaria, traducirla a aminoácidos, etc). También puede realizarse un análisis de restricción de la secuencia de DNA mediante SeqBuilder, así como una comparación con otras secuencias empleando MegAlign (comparación de múltiples secuencias). Por otro lado, PrimerSelect permite diseñar oligos para

hibridación, secuenciación o amplificación mediante la reacción en cadena de la polimerasa (PCR). Lo ideal es usarlo conjuntamente con Oligo (ver más abajo). Por su parte, SeqMan II permite manejar proyectos de secuenciación, realizando el solapamiento necesario de las secuencias introducidas para generar la secuencia consenso correspondiente ("contig"). No obstante, ha sido superado con creces por Sequencher (ver más abajo). Las secuencias de péptidos pueden ser analizadas mediante el módulo Protean. Por su parte, GeneQuest permite analizar exhaustiva y elegantemente las secuencias (“ORFs”, “Motifs”, sitios de poliadenilación, etc). Finalmente, XRayViewer, con la ayuda de unas gafas estereoscópicas tipo LCD acopladas al Macintosh, permite visualizar en tres dimensiones las moléculas de la base de datos que han sido cristalizadas (o sea, de las que se dispone de información sobre distancias atómicas obtenidas por difracción de rayos X).

3.4. Sequencher

Sequencher (GeneCodes, Ann Arbor, MI, USA; <**http://www.genecodes.com**>) es, probablemente, la mejor aplicación para la gestión de proyectos de secuenciación. De hecho, es el que se empleó rutinariamente en el Proyecto del Genoma Humano y otros proyectos de secuenciación satélites. Además de potente es muy fácil de usar.

3.5. Oligo

Oligo (National Biosciences, Plymouth, MN, USA; actualmente distribuido por sus desarrolladores: Molecular Biology Insights, Plymouth, MN, USA <**http://www.mbinsights.com**>) es muy valioso para el diseño de sondas de hibridación o cebadores para secuenciación o PCR. El programa aplica precisos algoritmos termodinámicos, con lo cual consigue una mayor exactitud que otros programas competidores. No obstante, la versión actual no permite buscar oligos en un rango prefijado de tamaños, por lo que utilizaremos también PrimerSelect para este menester, optimizando posteriormente el diseño de los cebadores elegidos con Oligo.

3.6. GelBase/BlotPro

GelBase/BlotPro (UVProducts, Cambridge, UK; <**http://www.uvp.com**>) es una combinación de hardware y software para la captura, digitalización, análisis e impresión de imágenes correspondientes a electroforesis de ácidos nucleicos o proteínas. También sirve para analizar membranas de hibridación ("Southern", "Northern", "Western"), etc. Actualmente ha sido superado por productos competidores —y significativamente más caros— como IntelligentQuantifier y AdvancedQuantifier (ver más abajo).

3.7. IntelligentQuantifier y AdvancedQuantifier

IntelligentQuantifier o IQ y AdvancedQuantifier o AQ (BioImage; actualmente denominada Genomic Solutions, Ann Arbor, MI, USA; <**http://www.bioimage.com**>) son, con diferencia, los paquete más potente del mercado, capaces de analizar prácticamente cualquier tipo de imagen generada en un laboratorio de biología molecular. IQ es de propósito general; AQ, además, compara imágenes 1D.

3.8. GeneConstruction/Search Kit y GeneInspector

El GeneConstruction/Search Kit o GCK (Textco, West Lebanon, NH, USA; <**http://www.textco.com**>) permite realizar dibujos precisos y profesionales de proyectos de ingeniería genética y clonación, funcionando también como una base de datos de los mismos. Contiene también otros módulos como la posibilidad de simular análisis de restricción en electroforesis en gel de fragmentos de DNA. El GCK es, sin duda, una de las herramientas más útiles para el diseño gráfico de secuencias y su administración.

Por su parte, GeneInspector es un paquete de análisis de secuencias y cuaderno de laboratorio electrónico e interactivo. Es una herramienta única en su clase; un nuevo paradigma que merece la pena explorar.

Existen también otras muchas aplicaciones específicas que pueden resultar interesantes para algunos aspectos de la biología molecular, pero que se escapan al objetivo de esta breve introducción.

4. BÚSQUEDA DE LA SECUENCIA DE DNA EN LA BASE DE DATOS, ANÁLISIS DE RESTRICCIÓN DE LA MISMA, SELECCIÓN DEL FRAGMENTO DE INTERÉS Y DISEÑO DE CEBADORES PARA SU AMPLIFICACIÓN MEDIANTE PCR Y PARA LA SECUENCIACIÓN DEL MISMO

Este protocolo corresponde al paquete informático LaserGene para Mac OS, así como a la aplicación “Entrez” del NCBI en <**http://www.ncbi.nlm.nih.gov**>:

a).-Buscar en la base de datos GenBank/EMBL las secuencias disponibles del organismo cuyo DNA se desea amplificar (*Salmonella typhimurium*) mediante el módulo GeneMan (LaserGene) o Entrez (NCBI).

b).-Elegir el operón de la arabinosa (*araBAD*). Exportar su secuencia al disco duro para su posterior análisis.

c).-Editar la secuencia mediante EditSeq.

d).-Realizar un análisis de restricción de la secuencia mediante MapDraw.

e).-Seleccionar la región del operón que se desea amplificar mediante EditSeq. Grabarla en disco.

f).-Diseñar los oligos externos y anidados necesarios para la amplificación mediante PCR (reacción en cadena de la polimerasa) y para la secuenciación del DNA elegido. Para ello emplear primero la aplicación PrimerSelect y después optimizar la pareja elegida con Oligo.

g).-Comentar las ventajas e inconvenientes de cada una de dichas aplicaciones.

h).- Imprimir los resultados obtenidos en cada una de las etapas.

Nota: un diseño eficiente de oligos para secuenciación y —sobre todo— para amplificación mediante PCR, requiere considerar diversos parámetros termodinámicos, como son la posible formación de estructuras secundarias (horquillas 3') y de dímeros entre los cebadores. Otros factores importantes a tener en cuenta son la energía libre de hibridación del pentámero 3´ de cada cebador, así como las temperaturas de fusión de los cebadores (Tm, pero calculada con el algoritmo del vecino más próximo), su contenido en GC/AT, su longitud y la del DNA que amplificarían. Recientemente, los diversos parámetros que contribuyen a la eficiencia de un cebador han sido agrupados en un algoritmo denominado Eficiencia de Hibridación (PE; "Priming Efficiency") en la aplicación Oligo.

5. BIBLIOGRAFÍA COMENTADA

+ Baxevanis AD, Davison BD, Page RDM, Petsko GA, Stein LD, Stormo GD, Leonard SA (eds) (2005): "Current Protocols in Bioinformatics". New York: Greene & John Wiley (New York). Manual de protocolos. "La nueva «Biblia» del Bioinformático" actualizada trimestralmente. Clasificación: PROTOCOLOS.

+ Campbell AM, Heyer LJ (2002): "Discovering Genomics, Proteomics & Bioinformatics". CSHL Press (New York). Excelente tratado de bioinformática.

+ DNAStar, Inc. (2005): "LaserGene. Biocomputing Software for the Macintosh and PowerMacintosh". Versiones para los diferentes módulos actualizadas bimestralmente. Manual del paquete LaserGene para la búsqueda, edición y análisis de secuencias de ácidos nucleicos y proteínas. El mejor paquete de biología molecular. Clasificación: SOFTWARE MANUAL.

+ Mount DW (2004): "Bioinformatics. Sequence and Genome Analysis". CSHL Press (New York). Excelente tratado de bioinformática.

+ Rychlik W (1998): "Oligo. Primer Analysis Software for the Apple Macintosh Computers". Versión 5.0. Manual de la aplicación Oligo para diseño de cebadores para PCR, secuenciación e hibridación. Muy útil por sus algoritmos termodinámicos. Clasificación: SOFTWARE MANUAL.

Rychlik W, Rhoads RE (1989): A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. Nucleic Acids Res 17:8543–8551. Discusión sobre los diferentes parámetros que afectan al diseño de oligos para hibridación, secuenciación y PCR. Clasificación: SOFTWARE ESPECÍFICO.

Nota: las referencias fundamentales para la preparación de la sesión se indican con el símbolo "+".

AGRADECIMIENTOS

Proyecto PAFPU 'FORMAPROFE' ('UCO-N-031') de Formación del Profesorado Universitario, Junta de Andalucía.