OBTENER INFORMACIÓN EN BASES DE DATOS DE BIOLOGÍA MOLECULAR

(*OBTAINING INFORMATION FROM MOLECULAR BIOLOGY DATABASES*)

Laura de la Fuente Cuello, Teresa Carretero Vaquer, Raquel Prieto Conejero, Ana Yarte del Toro

Biblioteca Nacional de Ciencias de la Salud. Instituto de Salud Carlos III.
Crta. Majadahonda-Pozuelo. Km. 2. Madrid CP 28220 (España)

Resumen

En la actualidad cualquier proyecto de investigación de biología molecular necesita analizar los datos obtenidos procedentes de secuenciación. Gran parte de estos análisis pueden realizarse a través de Internet mediante la utilización de bases de datos de uso público. Estas bases de datos son un recurso esencial en el trabajo diario de los investigadores y profesionales del campo de la Biomedicina. El conocimiento de estas herramientas, sus aplicaciones, finalidad es de enorme utilidad.
El presente trabajo muestra un análisis de estas bases de datos organizadas por materias de manera que puedan ser de utilidad para la comunidad científica y facilite el proceso de búsqueda de la base de datos más idónea para cada proyecto particular.
El propósito del estudio no es proporcionar un listado interminable de bases de datos, sino de incluir aquellas que permiten un acceso gratuito, tengan un reconocimiento por la comunidad científica, y que permitan una forma de navegación directa sin necesidad de descargar ningún tipo de software.
El elevado número de bases de datos hace necesario la interconexión y estandarización entre las mismas. Desde esta perspectiva, uno de los sistemas más apropiados para buscar información en las bases de datos es el denominado SRS (Sequence Retrieval System). SRS es un gestor de bases de datos desarrollado específicamente para trabajar con datos biológicos.
Palabras clave (DeCS): Biología Computacional; Bases de Datos Genéticas; Bases de datos de Ácido Nucleico; Biología Molecular.

Abstract

Nowadays, all biomedical research needs to search and analyze data obtained from molecular biology processes. Most of this can be carried out through the Internet thanks to free public databases. These databases are essential resources for the daily work of researchers and professionals from the field of biomedicine.
Knowledge of these tools, their applications and purposes is very useful. These databases have been evolved to meet the changing needs of molecular biologists.
In this work an overview of these databases is organized by subject and presented in such a way as to be useful to the scientific community and to facilitate the information research process according to each project.
The main objective of this study is not to provide a never ending databases listing, but to include those that give free access, are supported by international prestigious institutions and allow easy browsing without the need for downloading any software program.

The great amount of databases makes necessary to interconnect and standardize them. From this point of view, one of the most appropriated research systems is the Sequence Retrieval System (SRS). SRS is a database administrator developed specifically to work with biological data, and enables search in over 80 databases and navigation within them.
Keywords (DeCS): *Computacional Biology; Databases, Genetic; Databases, Nucleic Acid; Molecular Biology.*

INTRODUCCIÓN

Actualmente la práctica de la biología básica genera un enorme crecimiento en el volumen y complejidad de la información disponible sobre moléculas y procesos básicos de la vida. La mayoría de los procesos de investigación en biología molecular que se llevan a cabo en los laboratorios y Centros de Investigación necesitan analizar los datos generados en los procesos de secuenciación.
Desde este punto de vista surgió la necesidad de crear bases de datos que incluyeran herramientas de análisis de los datos que contenían. En un principio estas bases de datos se concibieron como simples repositorios de información con algunos mecanismos de búsqueda, sin embargo han ido evolucionando con el tiempo y hoy en día contamos con sistemas integrados que permiten enlaces entre las distintas bases de datos existentes.

UNIVERSALIDAD DE LAS BASES DE DATOS

La aplicación de la informática y las tecnologías de la información para el procesamiento de los datos biológicos forma parte del entramado necesario para la realización de las investigaciones en biología y en el área de la biomedicina. Con el fin de conseguir una utilización universal de este enorme volumen de información, capaz de generar grandes cantidades de conocimiento, fue necesario implementar sistemas de almacenamiento en bases de datos específicas.
Desde 1980 las bases de datos del *European Molecular Biology Laboratory* (EMBL), del *National Center for Biotechnology Information* y del *DNA Databank of Japan* (DDBJ), han recopilado las secuencias nucleotídicas que se han publicado. Actualmente existe una colaboración entre todas ellas y gracias a un Proyecto de Colaboración Internacional (*Internacional Nucleotide Sequence Database Collaboration*), cada nueva entrada es intercambiada de forma automática entre las tres (Figura 1). Los avances informáticos, el desarrollo de las redes de comunicación y el rápido desarrollo de Internet han contribuido a impulsar extraordinariamente el acceso a toda esa información.

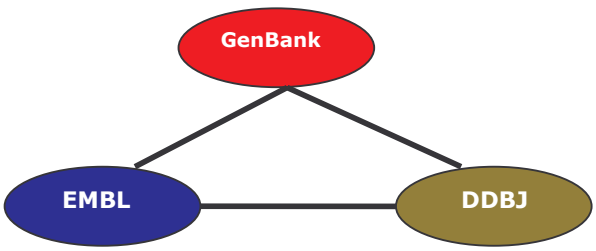

Figura 1. *Colaboración de Bases de Datos de Secuencias*

NACIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (NCBI)

El NCBI fue creado en 1988 con el fin de desarrollar sistemas de información en biología molecular. Pertenece a la *Library National of Medicine* de EEUU dentro de los Institutos Nacionales de Salud (*National Institutes of Health*, NIH). El NCBI ha propiciado la creación de bases de datos públicas y asumió el mantenimiento y la responsabilidad de la base de datos de secuencias de ADN GenBank en 1992 construida a partir de secuencias enviadas por científicos de todo el mundo y gracias al intercambio de información con bases de datos internacionales como las anteriormente mencionadas EMBL y DDBJ. Todos los recursos disponibles son gratuitos y de libre acceso a través de la página web del NCBI http://www.ncbi.nlm.nih.gov. En la Figura 2 se incluyen los principales recursos del NCBI.

RECURSOS NCBI	
DESCRIPCIÓN	**URL**
Bases de Datos	**Bibliográficas**
PubMed	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
PubMed Central	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Pmc
Online Mendelian Inheritance in Man	OMIM http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
Bookshelf	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books
Bases de Datos	**Moleculares**
Secuencias de	**Nucleótidos**
GenBank	http://www.ncbi.nlm.nih.gov/Genbank/index.html
Reference Sequences	Ref Seq http://www.ncbi.nlm.nih.gov/RefSeq/
Expressed Sequence Tags	dbEST http://www.ncbi.nlm.nih.gov/dbEST/
Genome Survey Sequences	dbGSS http://www.ncbi.nlm.nih.gov/dbGSS/
Major Histocompatibility Complex	dbMHC http://www.ncbi.nlm.nih.gov/mhc/MHC.cgi?cmd=init
Single Nucleotide Polymorphisms	dbSNP http://www.ncbi.nlm.nih.gov/SNP/index.html
Sequence Tagged Sites	dbSTS http://www.ncbi.nlm.nih.gov/dbSTS/
Third Party Annotation Database	TPA http://www.ncbi.nih.gov/Genbank/TPA.html
Trace Archive	http://www.ncbi.nlm.nih.gov/Traces/trace.cgi
Evolutionary Relatedness	PopSet http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PopSet
Vector Sequences	UniVec http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html
Whole Genome Shotgun Sequences	WGS http://www.ncbi.nlm.nih.gov/projects/WGS/WGSprojectlist.cgi
Secuencias de	**Proteínas**
Referente Sequences	RefSeq http://www.ncbi.nlm.nih.gov/RefSeq/
Conserved Domain Database	CDD http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

Estructuras	
Molecular Modeling Database	MMDB http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml
3D Domains	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Domains
PubChem BioAssay	http://pubchem.ncbi.nlm.nih.gov/
Genes	
Gene	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
UniGene	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene
Homologene	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene
Consensus CoDing Sequence	CCDS http://www.ncbi.nlm.nih.gov/CCDS/
Expresión	**genética**
Gene Expression Omnibus	GEO http://www.ncbi.nlm.nih.gov/geo/
GENSAT	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gensat
Taxonomía	
Taxonomy Browser	TaxBrowser http://www.ncbi.nlm.nih.gov/Taxonomy/
Entrez Taxonomy	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy
Datos	**Citogenéticos**
Spectral Karyotyping Multiplex Fluorescence In Situ Hybridization	SKY/M-FISH http://www.ncbi.nlm.nih.gov/sky/
Recursos	**Divulgativos**
Noticias	
NCBI News	http://www.ncbi.nlm.nih.gov/About/newsletter.html
What's New	http://www.ncbi.nlm.nih.gov/About/whatsnew.html
Estantería	
Coffee Break	http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=coffeebrk.TOC&depth=1
Genes and Disease	http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=gnd.TOC&depth=2
NCBI Handbook	http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=handbook.TOC&depth=2

Figura 2. *Principales Bases de Datos del Nacional Center for Biotechnology Information*

ENTREZ

Entrez (http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi) Sistema de búsqueda y recuperación que permite a los usuarios el acceso a secuencias, mapas genéticos, taxonomía a nivel molecular. Tiene capacidad para recuperar secuencias relacionadas, estructuras y referencias bibliográficas. Recientemente ha incorporado una nueva herramienta, *Global Query*, que permite la búsqueda simultánea en todas las bases de datos de Entrez; éstas incluyen secuencias de DNA y proteínas, taxonomía, genomas, datos de expresión génica, estructuras de proteínas, etc.

PUBMED CENTRAL

PubMed Central (http://www.pubmedcentral.nih.gov/) Archivo digital de más de 160 revistas que proporciona el acceso a alrededor de 300.000 artículos a texto completo. El requisito para participar en PubMed Central es precisamente el acceso libre y gratuito al texto completo de los artículos, aunque en algunos casos éste se produzca con un cierto retraso a la fecha de publicación de los mismos.

BASIC LOCAL ALIGNMENT SEARCH TOOL (BLAST)

Blast (http://www.ncbi.nlm.nih.gov/blast/) Programa de búsqueda de similitudes de secuencias. Una mejora introducida en Mayo de 2005 permite identificar fácilmente diferencias entre secuencias con un alto grado de similaridad, ofreciendo un color distinto en las bases diferentes.

GENBANK

GenBank (http://www.ncbi.nlm.nih.gov/Genbank/index.html) Base de datos de secuencias génicas del NIH. En febrero de 2004 había aproximadamente 37.893.844.733 de bases en los 32.549.400 registros de secuencias. Forma parte del Proyecto *International Nucleotide Sequence Database Collaboration*, en el que se produce un intercambio diario de datos entre las organizaciones DDBJ, EMBL y GenBank.
Muchas revistas, antes de publicar un artículo en el que se describe una secuencia, demandan a los autores, como garantía, que dicha secuencia haya sido sometida a una base de datos y tenga un número de acceso.

Bases de datos de secuencias de nucleótidos

- **Ref Seq** (*Referente Sequence*) Colección de secuencias de referencia cuyo propósito es proporcionar conjuntos de secuencias de ADN, RNA y proteínas de diferentes organismos de forma integrada, exhaustiva y no redundante. Hasta julio de 2005 esta base de datos incluye 1.695.929 proteínas de 2969 organismos.

- **dbEST** (*Expressed Sequence Tags*) contiene las secuencias de gran número de organismos y otras informaciones acerca de secuencias de ADN complementario o EST (pequeños segmentos secuenciados a partir de los extremos de clones de ADN complementario).

- **dbSNP** (*Single Nucleotide Polymorphisms*) Conjunto de polimorfismos que incluye tanto sustituciones de un único nucleótido (SNPs), como pequeñas delecciones o inserciones. En humanos se cree que la mayor parte de las variaciones genéticas son debidas a SNPs.

- **dbSTS** (*Sequence Tagged Sites*) está formada por secuencias y mapas de STS que son segmentos breves de ADN (200-500 bp) compuestos de una secuencia nucleotídica única, es decir que no se repite en todo el genoma.

- **PopSet** está integrada por secuencias de ADN recopiladas con el fin de analizar las relaciones evolutivas de una población determinada.

Bases de datos de Estructuras

- **MMDB** (*Molecular Modeling Database*) es una base de datos que recoge estructuras tridimensionales de biomoléculas. Proporciona información acerca de la función biológica, evolución, mecanismos de acción y relaciones entre macromoléculas.

Bases de datos de genes

- **UniGene** Sistema experimental de partición automática de secuencias de GenBank, incluyendo EST. Actualmente incluye secuencias de más de 25 animales y por encima de 20 plantas. Se actualiza semanalmente con nuevas secuencias de EST y bimensualmente con nuevas secuencias caracterizadas.

- **HomoloGene** Procedimiento de detección automática de homólogos entre los genes anotados (aquellos que poseen una descripción de localización precisa, tamaño, función de las secuencias de nucleótidos de un genoma por comparación con otras secuencias homólogas descritas en bancos de datos).

Bases de datos de expresión génica

- **GEO** (*Gene Expresión Omnibus*) Repositorio de datos y sistema de recuperación de expresiones genéticas de alto rendimiento.

Bases de datos de Taxonomía

- **TaxBrowser** (*Taxonomy Browser*) Contiene los nombres de todos los organismos que aparecen en las bases de datos genéticas con al menos una secuencia de nucleótidos o proteínas.

Bases de datos de genomas

- **Entrez Genome** Proporciona el acceso a datos de genomas de especies cuya secuenciación y mapeo está concluida o en proceso de finalización. Actualmente en Entrez Genome encontramos más de 180 genomas microbianos completos, 1600 genomas víricos y por encima de 550 secuencias de orgánulos eucarióticos.

- **MapViewer** Muestra montajes genómicos, marcadores genéticos y físicos y los resultados de anotación y otros análisis utilizando para ello un conjunto de mapas alineados.

- **Cancer Chromosomes** Formada por tres bases de datos SKY (*Spectral Karyotyping*), M-FISH (*Multiplex Fluorescente In Situ Hybridization*) y CGH (*Comparative Genomic Hybridization*), proporciona una plataforma pública para que los investigadores compartan y comparen sus datos citogenéticas. Contienen datos

sobre aberraciones cromosómicas y aberraciones cromosómicas recurrentes en cáncer, mediante un menú nos permite seleccionar una enfermedad o diagnóstico que puede combinarse con diferentes especificaciones para una localización específica en el cromosoma.

Recursos divulgativos

- **Coffee Break** Colección de artículos cortos sobre descubrimientos en biología. Cada informe incorpora tutoriales interactivos que muestran como se utilizan los recursos en bioinformática como parte del proceso de investigación.

- **Genes and Diseases** Recopilación de artículos que tratan sobre genes y las patologías que pueden producir. Los artículos se organizan por las partes del cuerpo a los que pueden afectar. Actualmente están recogidos más de 80 desórdenes genéticos.

EUROPEAN BIOINFORMATICS INSTITUTE (EBI)

En 1982, el *European Molecular Biology Laboratory* (EMBL) estableció la primera base de datos de secuencias de nucleótidos del mundo. Posteriormente el EBI se constituyó en el nodo europeo encargado de coleccionar, coordinar y diseminar los datos de biología molecular. Su filosofía se basa en seis principios básicos:

1. **Accesibilidad** Todos sus recursos son libres y gratuitos sin restricciones.

2. **Compatibilidad** Desarrollo de estándares bioinformáticas para la compartir recursos.

3. **Exhaustividad** Datos fiables y actualizados.

4. **Portabilidad** Todos los datos del EBI pueden descargarse desde su página web.

5. **Navegabilidad** Favorecer el trabajo de los usuarios mediante enlaces entre datos y bases de datos.

6. **Calidad** Todas las secuencias anotadas se someten a rigurosos procesos de control de calidad.

Todos los recursos disponibles son gratuitos y de libre acceso a través de la página web del EBI http://www.ebi.ac.uk/. En la Figura 3 están incluidos las principales bases de datos del EBI.

RECURSOS EBI	
DESCRIPCIÓN	URL
BASES DE DATOS MOLECULARES	
Secuencias de Nucleótidos	
Alternative Splicing Database	ASD http://www.ebi.ac.uk/asd/
Alternate Transcript diversity Database	ATD http://www.ebi.ac.uk/atd/
EMBL Nucleotide sequence database	EMBL-Bank http://www.ebi.ac.uk/embl/index.html
Ensembl	http://www.ebi.ac.uk/ensembl/index.html
Genome Reviews	http://www.ebi.ac.uk/GenomeReviews/
IMGT/HLA	http://www.ebi.ac.uk/imgt/hla/
Secuencias de Proteínas	
Catalytic Site Atlas	CSA http://www.ebi.ac.uk/thornton-srv/databases/CSA/
GOA	http://www.ebi.ac.uk/GOA/index.html
InterPro	http://www.ebi.ac.uk/interpro/index.html
Protein and associated Nucleotide domains with Inferred Trees	PANDIT http://www.ebi.ac.uk/goldman-srv/pandit/
UniProtKB/Swiss-Prot	http://www.ebi.ac.uk/swissprot/
UniProtKB/TrEMBL	http://www.ebi.ac.uk/trembl/
UniProt	http://www.ebi.ac.uk/uniprot/
Secuencias de Proteomas	
Chemical Entities of Biological Interest	ChEBI http://www.ebi.ac.uk/chebi/
IntAct	http://www.ebi.ac.uk/intact/index.jsp
IntEnz	http://www.ebi.ac.uk/intenz/index.html
Estructuras	
DALI	http://www.ebi.ac.uk/dali/index.html
Macromolecular Structure Database	MSD http://www.ebi.ac.uk/msd/
RESID	http://www.ebi.ac.uk/RESID/index.html
SISTEMAS DE BÚSQUEDA	
BioMart	http://www.ebi.ac.uk/biomart/index.html
Integr8	http://www.ebi.ac.uk/integr8/EBI-Integr8
Sequence Retrieval System	SRS http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+srsq2+-noSession
Sequence Retrieval System 3D	SRS 3D http://srs3d.ebi.ac.uk/

Figura 3. *Principales Bases de Datos del European Bioinformatics Information*

EUROPEAN MOLECULAR BIOLOGY LABORATORY (EMBL)

EMBL-Bank (http://www.ebi.ac.uk/embl/index.html) fue la primera base de datos que se estableció sobre secuencias de ADN y ARN. Forma parte de la colaboración internacional con GenBank y DDBJ, la actualización de la base de datos se produce mediante el intercambio de información entre las tres bases de datos.
Proporciona acceso a secuencias de genoma completas o parcialmente finalizadas.

UNIPROTKB/SWISS-PROT

UNIPROTKB/SWISS-PROT (http://www.ebi.ac.uk/swissprot/) Constituida en 1986, comprende un conjunto de secuencias de proteínas. Ofrece un alto

nivel de anotación, mínima redundancia y alto grado de integración con otras bases de datos. La versión actual (agosto 2005) contiene 188.752 entradas que suponen 68.301.856 aminoácidos.

GENE ONTOLOGY (GO)

GO (http://www.ebi.ac.uk/GO/index.html) es un consorcio internacional de colaboración que tiene como objetivo ofrecer un vocabulario controlado para la descripción de las funciones moleculares, procesos biológicos y componentes celulares que están asociados a productos génicos.

BIOMART

BioMart (http://www.ebi.ac.uk/biomart/index.html) Permite a los usuarios consultar diferentes bases de datos simultáneamente y complejas. Puede utilizarse para consultas de genomas, anotaciones realizar consultas, proteomas, estructuras macromoleculares y SNPs.

SEQUENCE RETRIEVAL SYSTEM (SRS)

SRS (http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+srsq2+-noSession) es un sistema integrado que permite realizar búsquedas en más de 400 bases de datos y navegar entre ellas. Permite la búsqueda por palabras clave basándose para ello en el uso de operadores lógicos y en enlaces entre las distintas bases de datos. Es uno de los mejores sistemas para la búsqueda de información en las bases de datos de secuencias.
No se trata de un catálogo de bases de datos, sino de un interfaz único de consulta y presentación de resultados facilitando así el acceso a los diversos recursos disponibles.
Las bases de datos a las que se puede acceder incluyen bibliografía, secuencias, estructuras de proteínas, mapas genéticos y cromosómicos, mutaciones, etc.
SRS ofrece dos formas de trabajo: a través de un interfaz gráfico de usuario, que resulta sencillo de utilizar para cualquier investigador interesado, y otro basado en un conjunto de órdenes que resulta más apropiado para especialistas.

DNA DATA BANK OF JAPAN (DDBJ)

DDBJ (http://www.ddbj.nig.ac.jp/) Banco de datos japonés de secuencias de ADN. Participa en el Proyecto *International Nucleotide Sequence Database Collaboration* junto con GenBank y EMBL. Elaborada en 1988, contiene diversos tipos de búsquedas, información, noticias relacionadas, y otros recursos de interés sobre datos de secuencias.

CONCLUSIONES

El enorme desarrollo que han tenido las bases de datos en biología molecular ha transcurrido de forma paralela a la gran acumulación de información generada por los científicos.
Los avances informáticos, y el rápido desarrollo de Internet han contribuido a impulsar extraordinariamente el acceso a toda esa información.

La principal característica de estas bases de datos es la universalidad y el acceso libre y gratuito.
La existencia de un número tan elevado de bases de datos hace necesaria la existencia de sistemas integrados que permitan realizar búsquedas simultáneas en varias de ellas.

BIBLIOGRAFÍA

1. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DV. GenBank. Nucleic Acids Res. 2005; 33: D34-D38.

2. Brooksbank C, Cameron G, Thornton J. The European Bioinformatics Institute's data resources: towards systems biology. Nucleic Acids Res. 2005; 33: D46-D53.

3. Galperin MY. The Molecular Biology Database Collection: 2005 update. Nucleic Acids Res. 2005; 33: D5-D24.

4. Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, et al. The EMBL Nucleotide Sequence Database. Nucleic Acids Res. 2005; 33: D29-D33.

5. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, et al. Database Resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2005; 33: D39-D45.

6. Zdobnov EM, López R, Apweiler R, Etzold T. The EBI SRS server--new features. 2002; 18: 1149-1150.