Bases de datos en Bioinformática

•National Center for Biotechnology Information (NCBI)

Creada en 1979 en the LANL (Los Alamos, NM). Mantenida desde 1992 NCBI (Bethesda, MD, USA)

http://www.ncbi.nlm.nih.gov/

•European Bioinformatics Institute (EBI)

Creada en 1980 en the *European Molecular Biology Laboratory* in Heidelberg. Es mantenida por el EBI- Cambridge. Desde 1994

http://www.ebi.ac.uk/embl

•GenomeNet

Reúne bases de datos diversas en Japón. Inició en 1984, en el the National Institute of Genetics (NIG) in Mishima. Mantenida por el grupo de Takashi Gojobori.

http://www.ddbj.nig.ac.jp

Inconvenientes con las Bases de Datos (DB)

Errores en las Bases de Datos

- Debido a que es enviada de forma gratuita y vía WEB no pasa ningún tipo de filtro, lo que conlleva a errores en su anotación final por omisiones, Inexactitudes e inconsistencias en algunos campos (nombres, tamaños, Fechas, etc.)
- Otra fuente de error son los intrínsecos a la secuencias de nucleótidos o aminoácidos, generados en la secuenciación, Interpretación de las electroforesis, elementos externos (partes del vector de clonaje), etc.

Redundancia

- Uno de los principales problemas en las DB lo genera la gran cantidad de redundancia que existe, debido a secuencias parciales de un solo genoma.

Bases de datos de genomas

- Se encargan de mantener y actualizar las secuencias y las anotaciones de genomas completos.
  - **Ensembl** (EBI)
  - **Genome viewer** (NCBI)
  - **Goldenpath** (UCSC)
    - Existen también recursos genómicos especializados
  - **Transfact**: sitios de unión a factores de transcripción.
  - **EST**: Expressed Sequence Tags
  - **UTRDB**: Untranslated regions
  - **SpliceSitesDB**: Pares de señales de *splicing*

Bases de datos de proteínas

- Secuencias primarias de aminoácidos
  - Sin revisión humana
  - **Trembl** (EBI)
  - **nr** (NCBI)
  - Con revisión de la anotación
  - **Swisprot** (EBI)
  - Bases de datos de proteomas
  - **Proteome analysis** (EBI)

Estructuras secundarias o dominios

Varían según la fuente de las proteínas y el análisis que se realiza sobre ellas.

- **BLOCKS:** Motivos alineados de PROSITE/PRINTS
- **PROSITE:** Expresiones regulares sobre Swiss-prot
- **PRINTS:** Conjunto de motivos que definen una familia sobre Swiss-prot/TrEMBL
- **PFAM:** Modelos de Markov sobre Swiss-prot
- **INTERPRO(EBI):** Integra la información de muchas bases de datos de dominios.
- **Domains** (Blocks, Domo, Pfam, ProDom, SBASE).

Estructura 3D

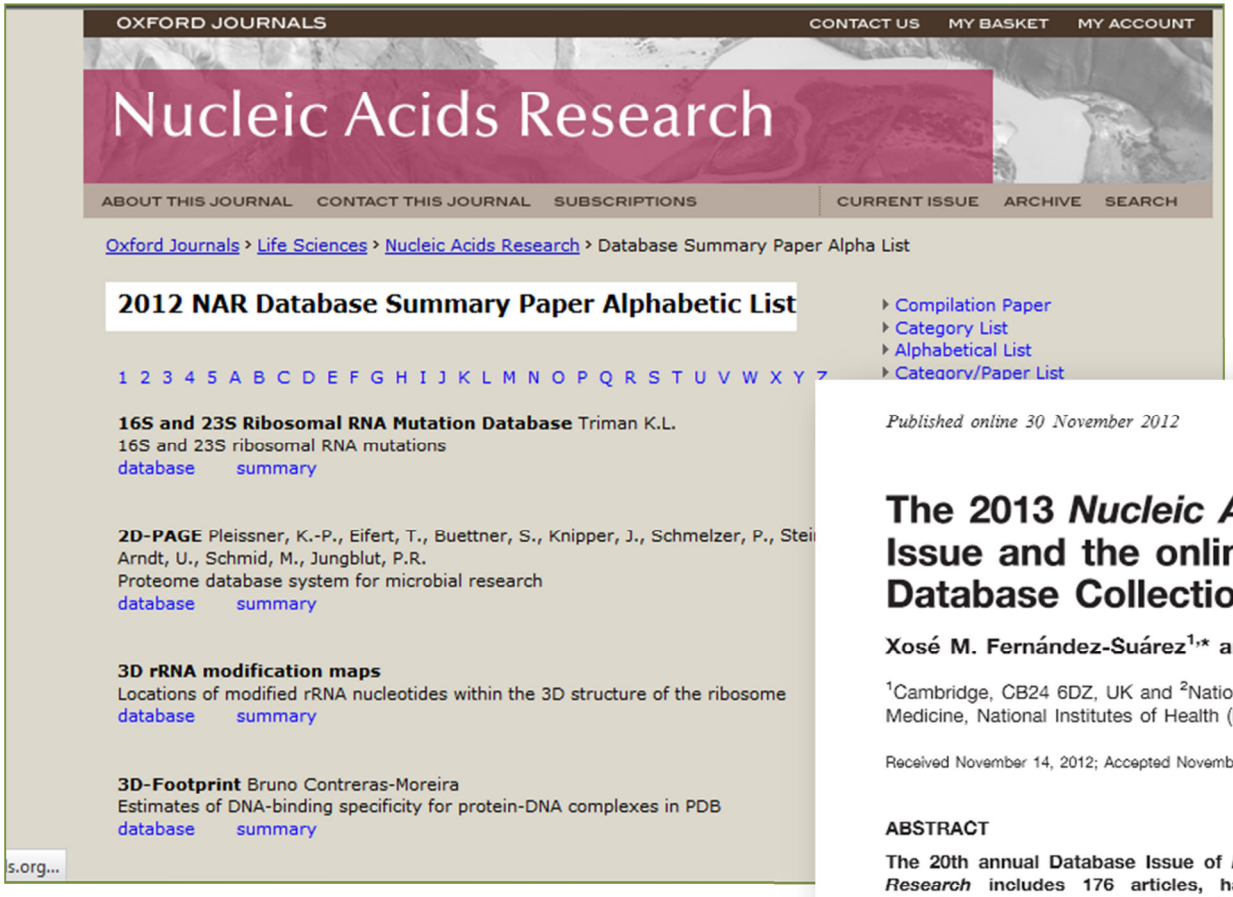
Estructuras tridimensionales de macromoléculas con las coordenadas en el espacio de cada átomo.

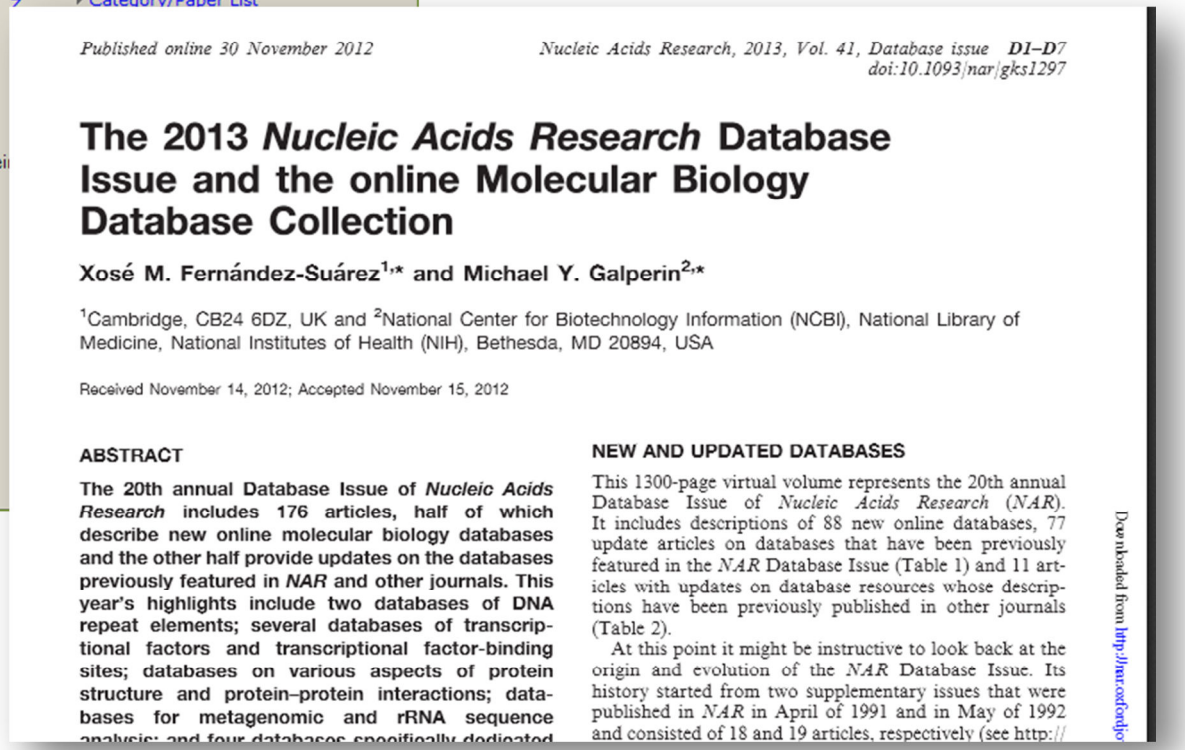
- **PDB**: Base de datos principal de estructuras tridimensionales
- **CATH**: Clasificación de **PDB** en diferentes grupos funcionales y estructurales
- **MMDB**: subset de PDB mantenido por NCBI
- **MSD**: subset de PDB mantenido por EBI
- **Dali Database:** Base de datos para comparar Estructuras 3D contra la PDB databank

- **Bases de Datos Especializada**

- **HOVERGEN** (Homologous Vertebrate Genes Database) for vertebrates: Based on GenBank CDS.
- **HOBACGEN** (Homologous Bacterial Genes Database) for prokaryotes and yeast: Based on SWISS-PROT/TrEMBL.
- **HOBACGEN**-CG for completely sequenced genomes: Based on SWISS-PROT/TrEMBL
- **COG** (Clusters of Orthologous Groups), also for complete genomes: Based on GenBank CDS.
- **NuReBase** (Nuclear Receptors Database) for mammalian nuclear receptors: Based on EMBL CDS.
- **RTKdb** (Tyrosine Kinase Receptors), **PlasmoDB**, MaizeGDB, **Mouse Genome Informatics**, **CAMERA**, **WormBase**, SGD, **Stanford MicroArray Database,** CryptoDB**,** **KEGG PATHWAY Database**, etc.

http://www.oxfordjournals.org/nar/database/a/





Published online 30 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D1–D7***
doi:10.1093/nar/gks1297

The 2013 *Nucleic Acids Research* Database Issue and the online Molecular Biology Database Collection

Xosé M. Fernández-Suárez[1,*] and Michael Y. Galperin[2,*]

[1]Cambridge, CB24 6DZ, UK and [2]National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health (NIH), Bethesda, MD 20894, USA

Received November 14, 2012; Accepted November 15, 2012

ABSTRACT

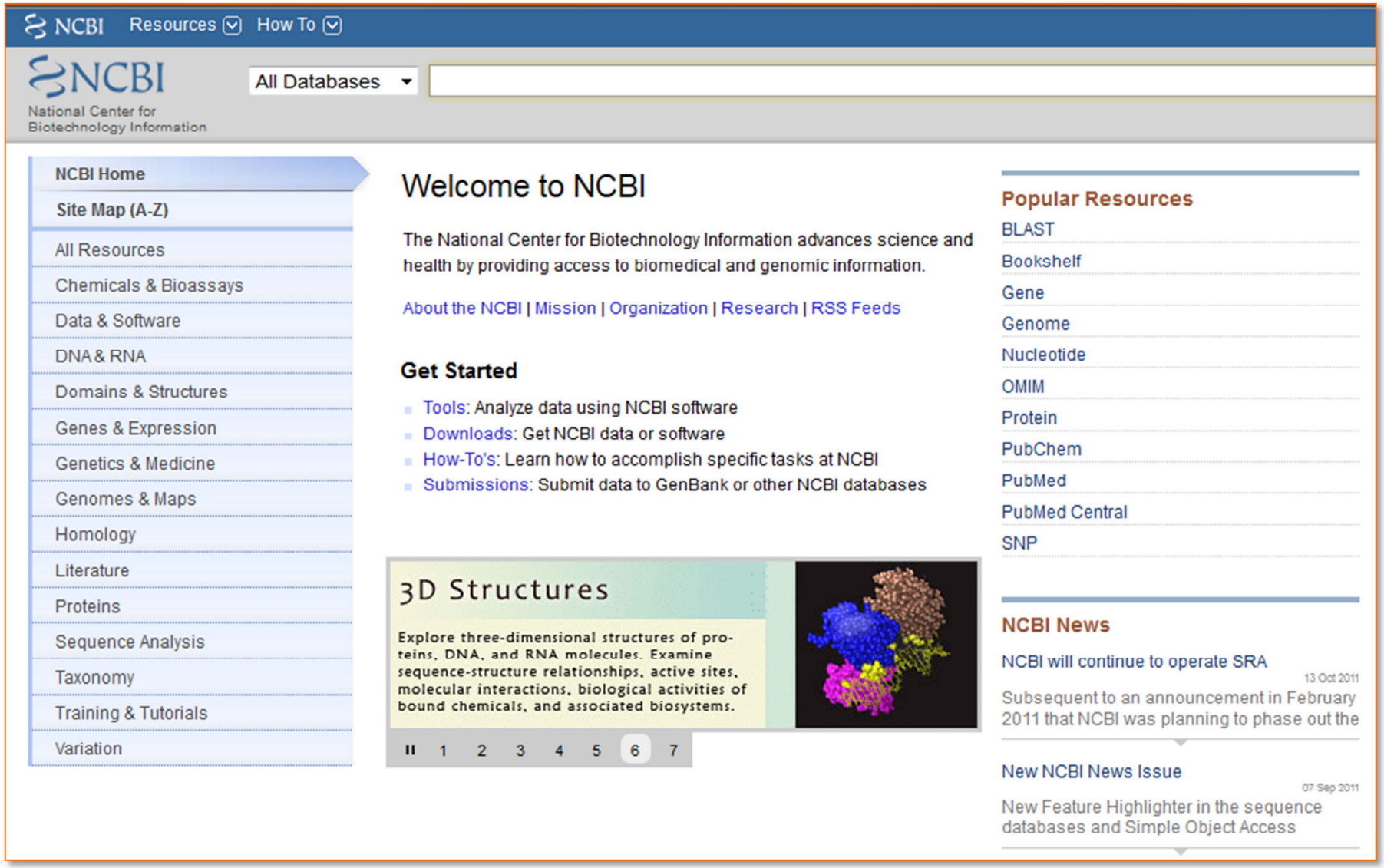
The 20th annual Database Issue of *Nucleic Acids Research* includes 176 articles, half of which describe new online molecular biology databases and the other half provide updates on the databases previously featured in *NAR* and other journals. This year's highlights include two databases of DNA repeat elements; several databases of transcriptional factors and transcriptional factor-binding sites; databases on various aspects of protein structure and protein–protein interactions; databases for metagenomic and rRNA sequence analysis; and four databases specifically dedicated

NEW AND UPDATED DATABASES

This 1300-page virtual volume represents the 20th annual Database Issue of *Nucleic Acids Research* (*NAR*). It includes descriptions of 88 new online databases, 77 update articles on databases that have been previously featured in the *NAR* Database Issue (Table 1) and 11 articles with updates on database resources whose descriptions have been previously published in other journals (Table 2).

At this point it might be instructive to look back at the origin and evolution of the *NAR* Database Issue. Its history started from two supplementary issues that were published in *NAR* in April of 1991 and in May of 1992 and consisted of 18 and 19 articles, respectively (see http://

National Center for Biotechnology Information (NCBI)



European Bioinformatics Institute (EMBL-EBI)

EMBL-EBI | Services | Research | Training | Industry | About us

The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

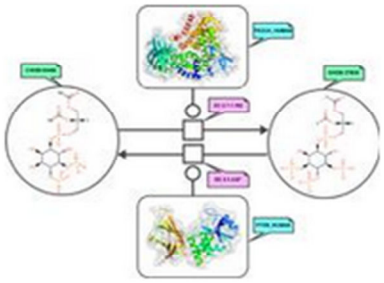
EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.

Explore the EBI:

Search

Examples: blast, keratin, bfl1...

Press releases



Mapping metabolism

Something different

Data storage in DNA becomes a reality

Popular

- Services
- Research
- Training
- News
- Jobs
- Visit us
- EMBL
- Contacts

Events

The Bioinformatics Workshop - NIMR, London
Apr 22 2013 -Apr 24 2013
Registration deadline: Apr 12 2013

diXa: Microarray Analysis using R and Bioconductor
May 14 2013 -May 16 2013
Registration deadline: Apr 21 2013

The Bioinformatics Workshop - University College London
May 28 2013 -May 30 2013
Registration deadline: Apr 26 2013

Joint EMBL-EBI/Wellcome Trust Summer School in Bioinformatics

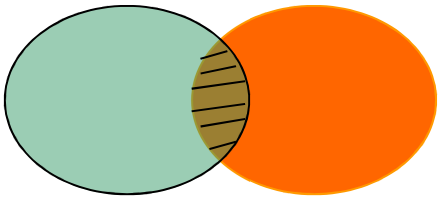
- **GI** number (**Genbank Identifier**) son una serie consecutiva de dígitos que son asignados a cada secuencia grabada y procesada por el NCBI. No tiene relación alguna con el número de acceso de la secuencia.

  - Es mostrado en nucleótidos al lado del campo **VERSION**.
  - En porteinas el GI es mostrado en el campo **CDS/db_xref** y en el campo **VERSION**
  - Cambia con cada actualización del registro correspondiente a la secuencia

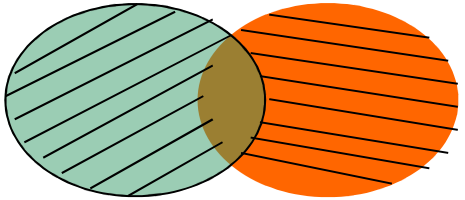
- **Accession Number**: Es un único identificador de la secuencia ingresada, puede tener una o dos letras al inicio y luego de 5 a 6 números (U12345 o NT_123456).
-
  - Algunos pueden ser más largos, es una Clave Secundaria en la DB y no cambia a pesar que el autor modifique la información.

Información: **http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html#AccessionB**

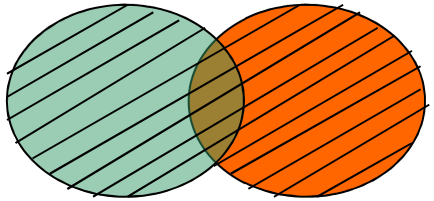
Las DB permiten utilizar operadores lógicos booleanos que establecen la relación entre los términos de búsqueda. Tomado del álgebra (George Boole) permiten combinar los términos de búsqueda de acuerdo con nuestras necesidades. Otros [ ] ó " "



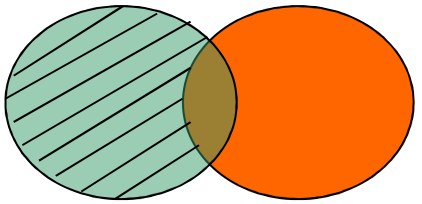
- **AND Intersección (Human AND HK):** selecciona solamente los registros en los que aparece simultáneamente los conceptos **Human y hk**



- **XOR Exclusión (Human XOR hk):** selecciona sólo los registros donde está **Human o hk** pero no **Human y hk** simultáneamente



- **OR Suma o unión (Humas OR hk):** selecciona todos los registros en los que aparece tanto A como B o ambos a la vez



- **NOT Resta o negación (Human NOT hk):** selecciona sólo los registros en los que se encuentre el término **Human** sin estar acompañado del término **hk**.

Seq

- Giardia intestinalis syntaxin 16 gene, complete cds AF404743
- Plasmodium falciparum 3D7 chromosome 11, complete sequence. NC_004315
- NM_119057 Arabidopsis thaliana HXK1 (HEXOKINASE 1); ATP binding
- AF288471 Xenopus laevis hexokinase I mRNA, partial cds. Sequence tagged site (STS)

  NM_001025935 Caenorhabditis elegans hypothetical protein (hexokinase) mRNA