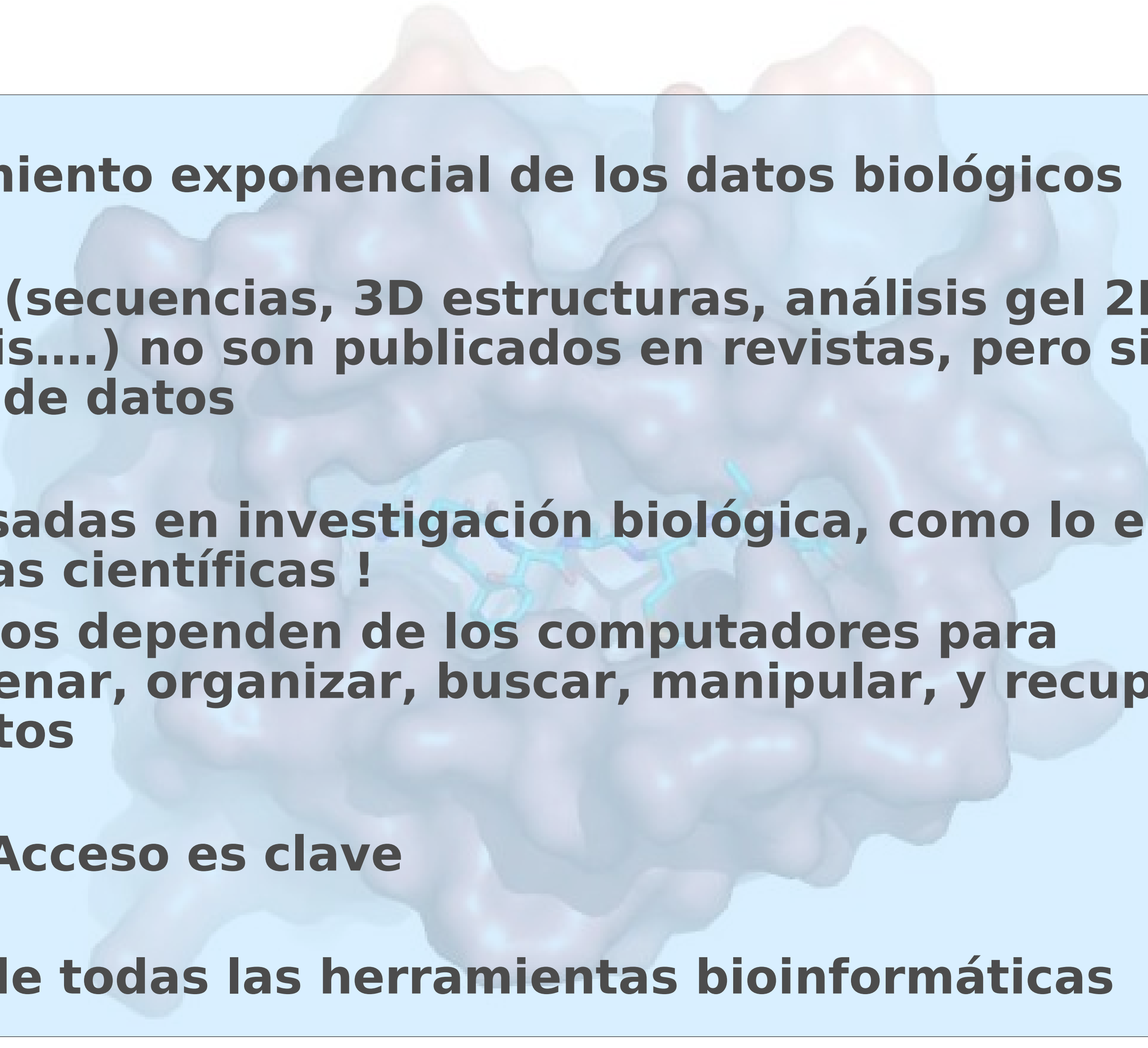
Bases de datos Biológicas

Andrés Pinzón

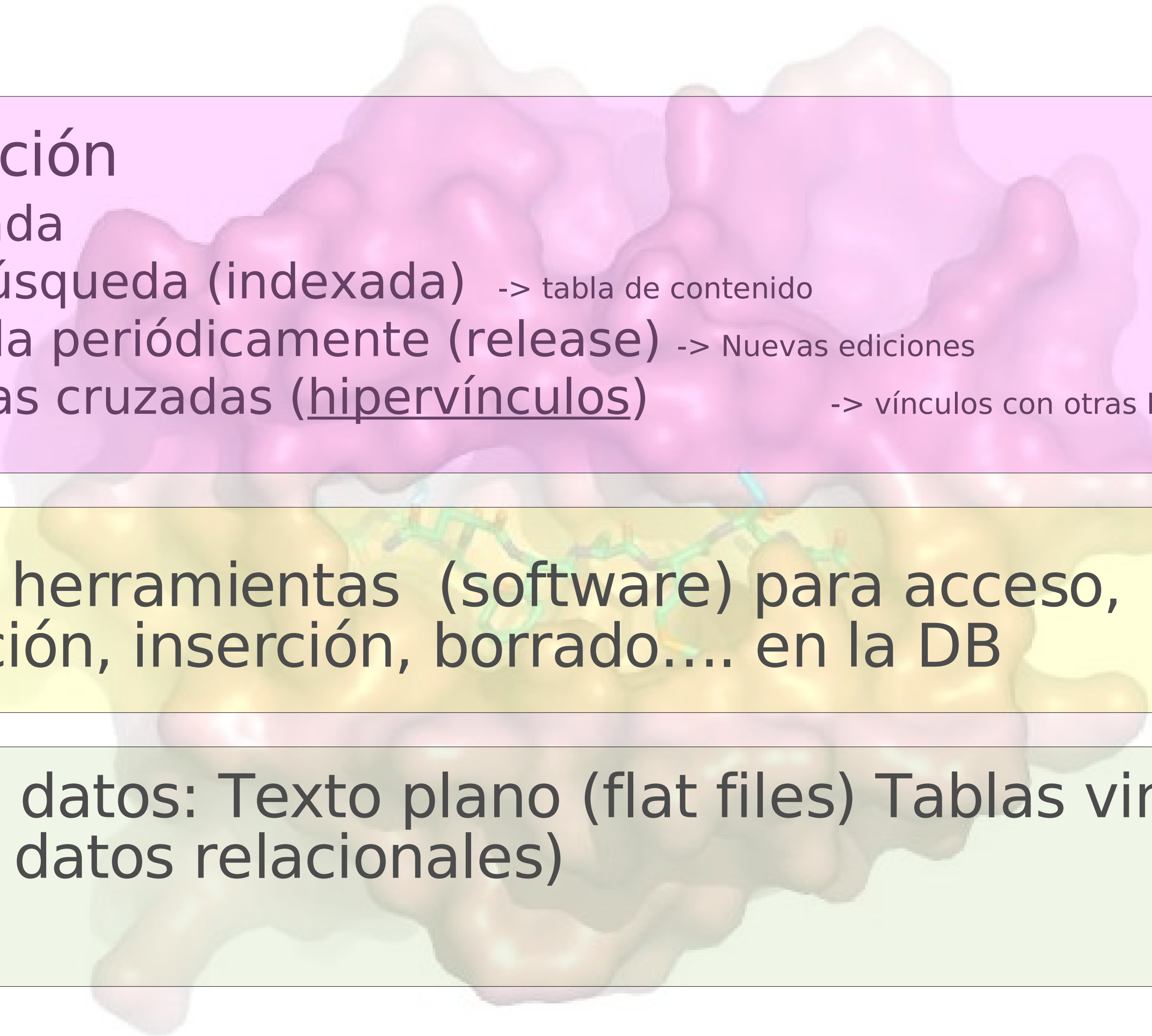
Centro de Bioinformática
Instituto de Biotecnología
Universidad Nacional de Colombia

¿Por qué las bases de datos ?



- Crecimiento exponencial de los datos biológicos
- Datos (secuencias, 3D estructuras, análisis gel 2D, MS análisis….) no son publicados en revistas, pero si en bases de datos
- Son usadas en investigación biológica, como lo eran la revistas científicas !
- Biólogos dependen de los computadores para almacenar, organizar, buscar, manipular, y recuperar los datos
- Libre Acceso es clave
- Base de todas las herramientas bioinformáticas

¿Qué es una base de datos ?



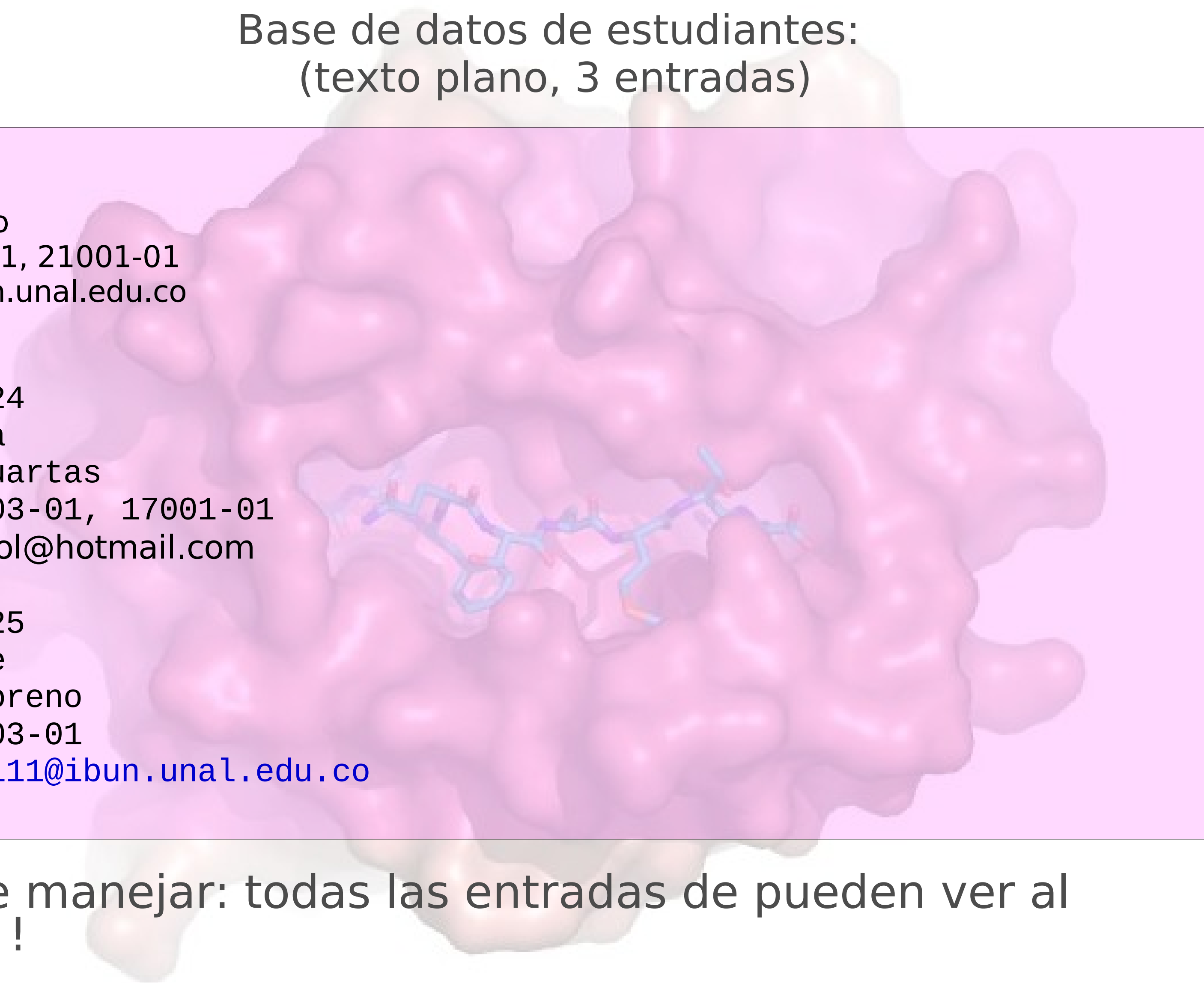
- Una colección
  - estructurada
  - De fácil búsqueda (indexada) -> tabla de contenido
  - Actualizada periódicamente (release) -> Nuevas ediciones
  - Referencias cruzadas (hipervínculos) -> vínculos con otras DB

- Incluye la herramientas (software) para acceso, actualización, inserción, borrado.... en la DB

- Almacena datos: Texto plano (flat files) Tablas vinculadas (bases de datos relacionales)

DB: Texto plano « flat file »

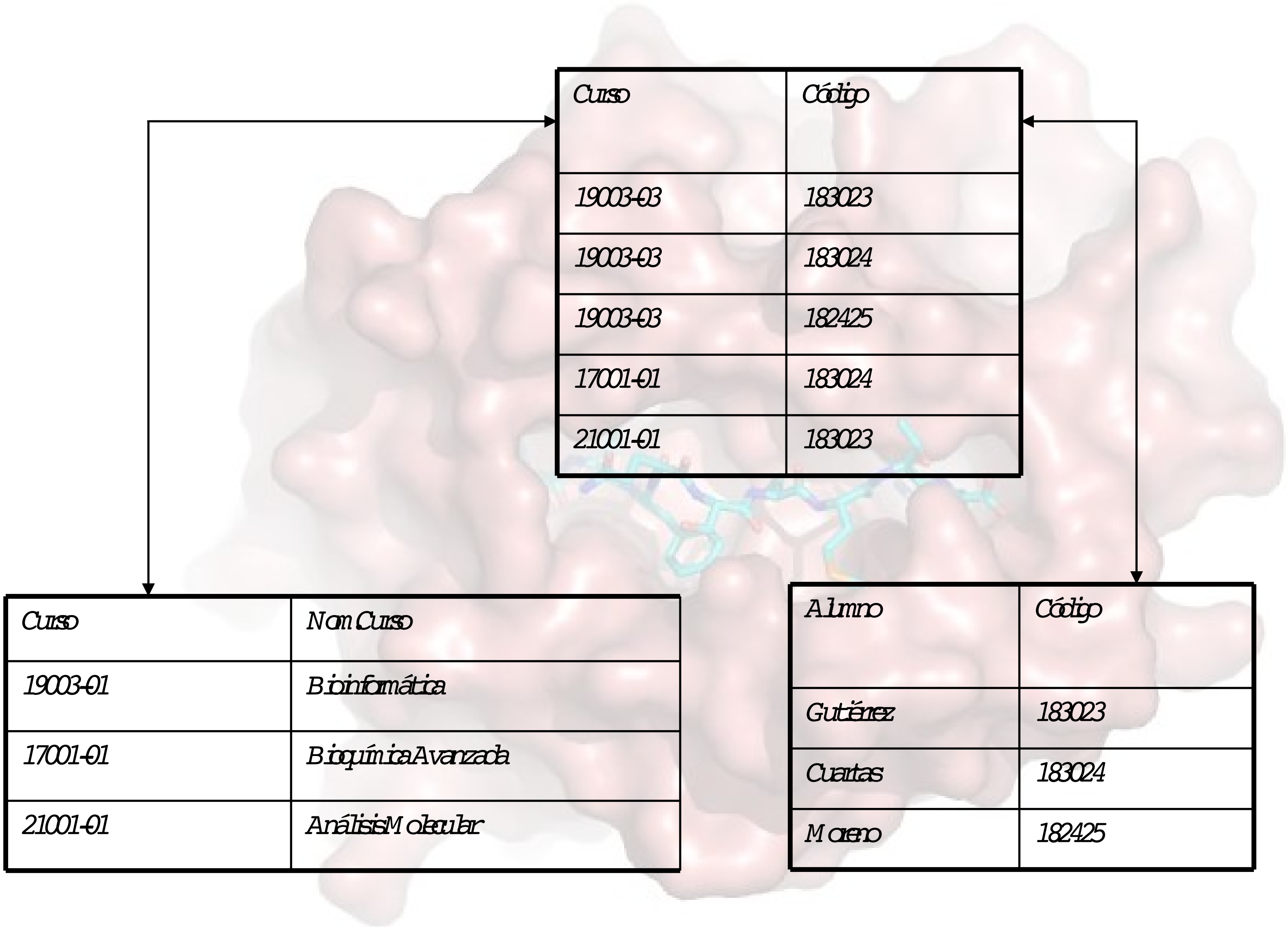
Base de datos de estudiantes:
(texto plano, 3 entradas)



```
Código: 183023
Nombre: Julián
Apellido : Pulecio
Cursos: 19003-01, 21001-01
Email: jpul@ibun.unal.edu.co
//
//
Código: 183024
Nombre: Sonia
Apellido : Cuartas
Cursos : 19003-01, 17001-01
Email: soniacol@hotmail.com
//
Código: 183025
Nombre: Jaime
Apellido : Moreno
Cursos : 19003-01
Email: pm186111@ibun.unal.edu.co
//
```

- Facil de manejar: todas las entradas de pueden ver al tiempo !

Bases de datos « relacionales »



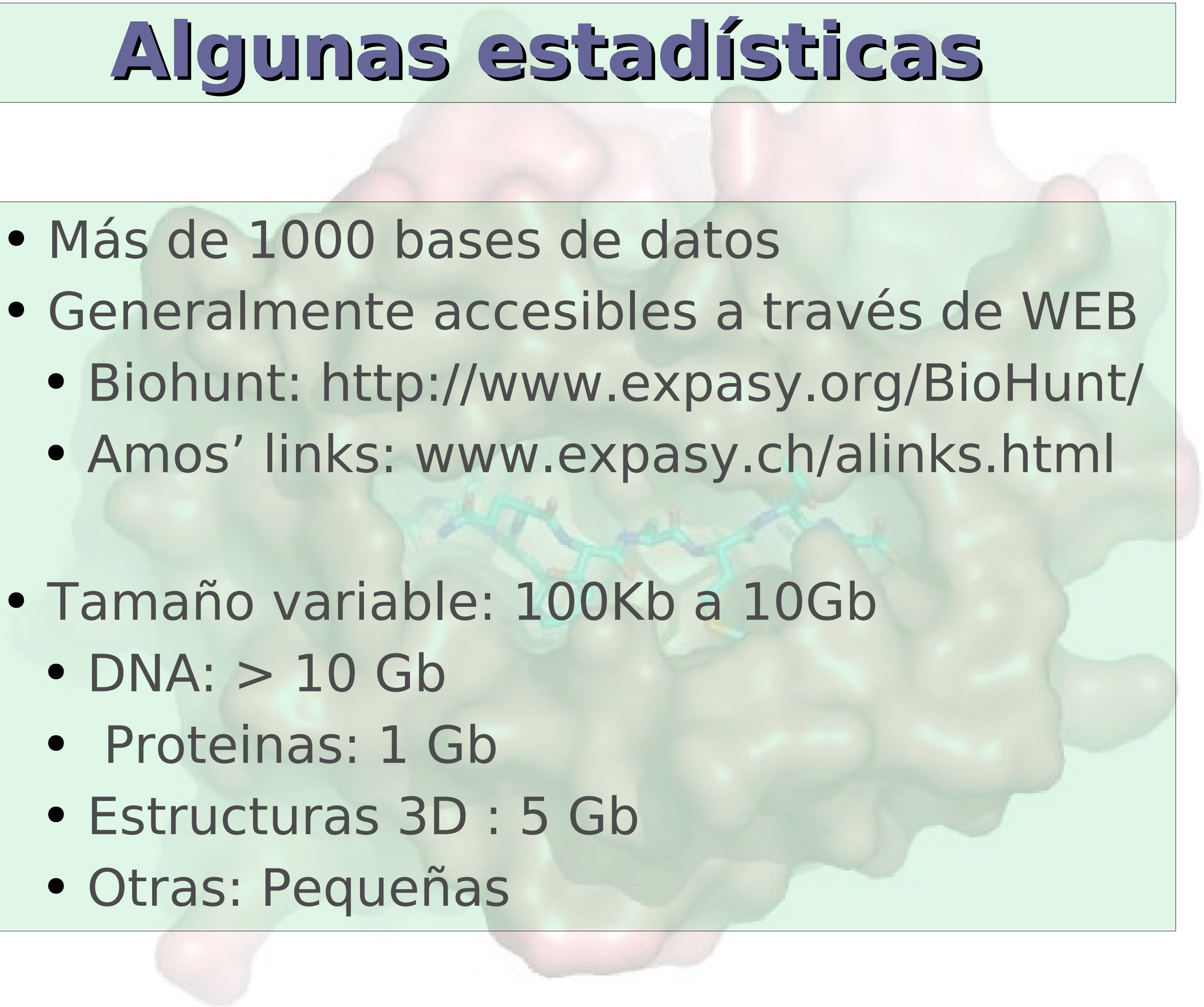
Curso	Código
19003-03	183023
19003-03	183024
19003-03	182425
17001-01	183024
21001-01	183023

Curso	Nom.Curso
19003-01	Bioinformática
17001-01	Bioquímica Avanzada
21001-01	Análisis Molecular

Alumno	Código
Gutiérrez	183023
Cuartas	183024
Moreno	182425

Fácil: manejo y selección de la salida

Algunas estadísticas

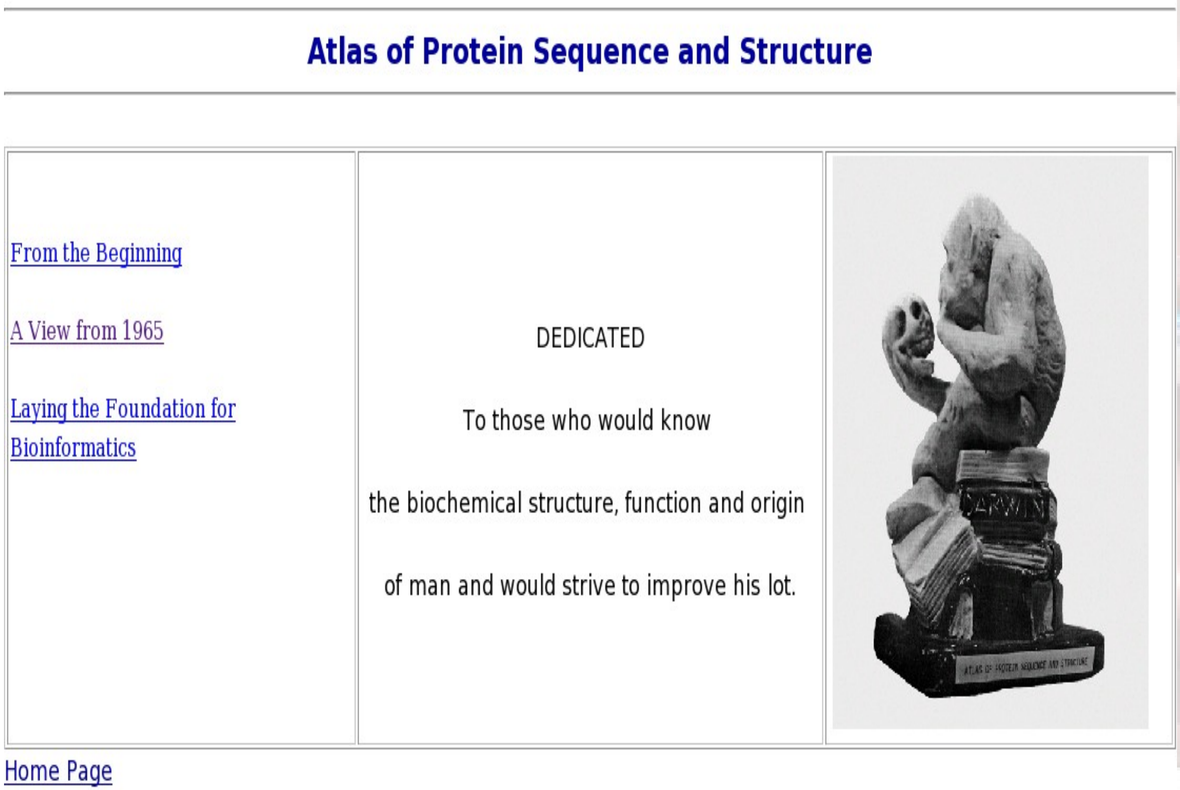


- Más de 1000 bases de datos
- Generalmente accesibles a través de WEB
  - Biohunt: http://www.expasy.org/BioHunt/
  - Amos’ links: www.expasy.ch/alinks.html

- Tamaño variable: 100Kb a 10Gb
  - DNA: > 10 Gb
  - Proteinas: 1 Gb
  - Estructuras 3D : 5 Gb
  - Otras: Pequeñas

Históricamente las bases de datos de proteínas preceden a las de DNA.

- Atlas de secuencias y estructuras protéicas (1965).
- PIR (1997).



Atlas of Protein Sequence and Structure

From the Beginning

A View from 1965

Laying the Foundation for Bioinformatics

DEDICATED

To those who would know

the biochemical structure, function and origin

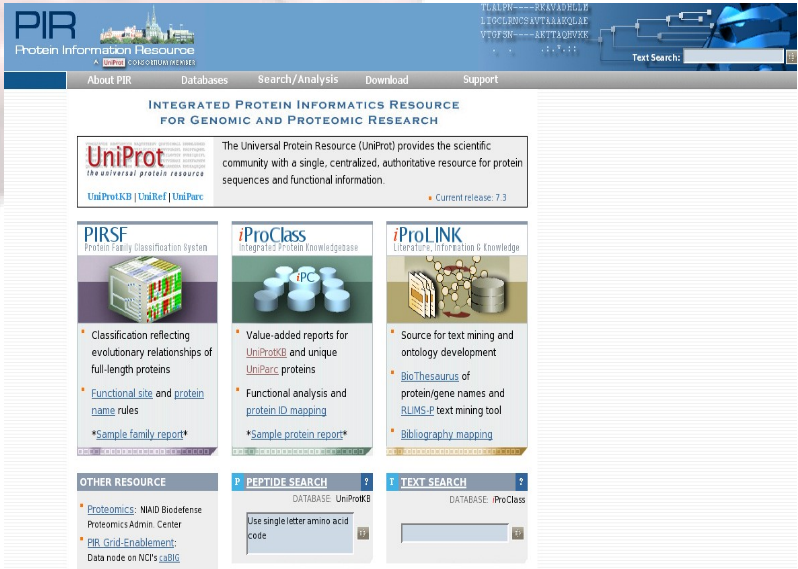
of man and would strive to improve his lot.

Home Page

http://www.dayhoff.cc/MODAtlasSummary.html

Dr. Margaret Oakley Dayhoff

(1925-1983)



INTEGRATED PROTEIN INFORMATICS RESOURCE FOR GENOMIC AND PROTEOMIC RESEARCH

UniProt — the universal protein resource

The Universal Protein Resource (UniProt) provides the scientific community with a single, centralized, authoritative resource for protein sequences and functional information.

UniProtKB | UniRef | UniParc

Current release: 7.3

PIRSF Protein Family Classification System

- Classification reflecting evolutionary relationships of full-length proteins
- Functional site and protein name rules

Sample family report

iProClass Integrated Protein Knowledgebase

- Value-added reports for UniProtKB and unique UniParc proteins
- Functional analysis and protein ID mapping

Sample protein report

iProLINK Literature, Information & Knowledge

- Source for text mining and ontology development
- BioThesaurus of protein/gene names and RLIMS-P text mining tool
- Bibliography mapping

OTHER RESOURCE

- Proteomics: NIAID Biodefense Proteomics Admin. Center
- PIR Grid-Enablement: Data node on NCI's caBIG

PEPTIDE SEARCH

DATABASE: UniProtKB

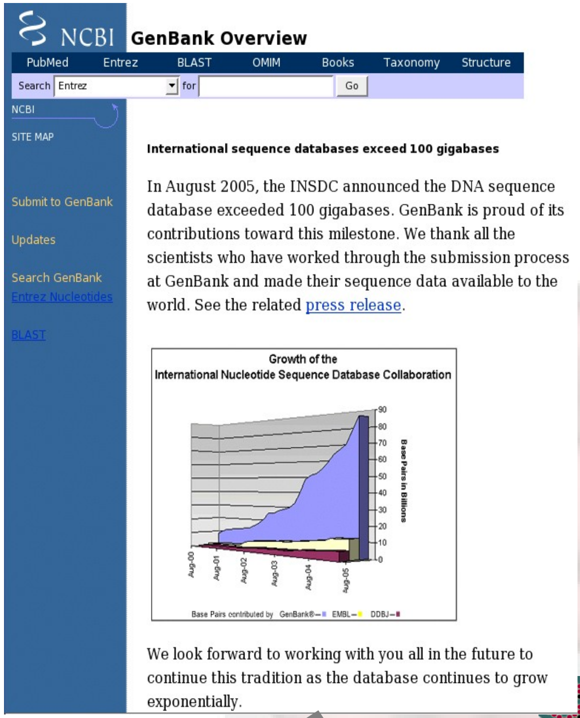
Use single letter amino acid code

TEXT SEARCH

DATABASE: iProClass

GenBank

DDBJ



International sequence databases exceed 100 gigabases

In August 2005, the INSDC announced the DNA sequence database exceeded 100 gigabases. GenBank is proud of its contributions toward this milestone. We thank all the scientists who have worked through the submission process at GenBank and made their sequence data available to the world. See the related press release.

Growth of the International Nucleotide Sequence Database Collaboration

We look forward to working with you all in the future to continue this tradition as the database continues to grow exponentially.

What is DDBJ

DDBJ (DNA Data Bank of Japan) began DNA data bank activities in earnest in 1986 at the National Institute of Genetics (NIG).
DDBJ has been functioning as the international nucleotide sequence database in collaboration with EBI/EMBL and NCBI/GenBank.
DNA sequence records the organismic evolution more directly than other biological materials and ,thus, is invaluable not only for research in life sciences, but also human welfare in general. The databases are, so to speak, a common treasure of human beings. With this in mind, we make the databases online accessible to anyone in the world.

more information...

DNA Data Bank of Japan (DDBJ)
Center for Information Biology and DNA Data Bank of Japan (CIB-DDBJ)
National Institute of Genetics (NIG)
SOKENDAI
Department of Genetics
Research Organization of Information and Systems

DDBJ uses the Super SINET computer network for its business.

What's new

- updated: Apr. 3
- mail magazine(Mar. 3)

Hot Topics

Termination of DDBJ/CIB Human Genomics Studio by DDBJ (Mar. 31) NEW

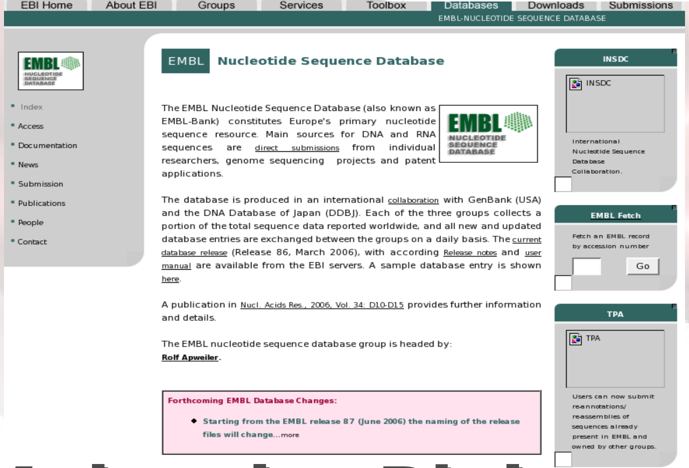
Release of new 136 thousand Human EST entries (Mar. 14)

New functions are added to "getentry" web server (Mar. 13)

- The 5th Japan-Korea-China Bioinformatics Training Course
- Release of the complete genome sequence of E. coli K-12 strain W3110(Jan. 24)

Announcement

- UniProt (Swiss-Prot) service on DDBJ (Apr. 12, 2005)
- Preliminary notice for the termination of S&W SEARCH sercice(Feb. 18, 2005)
- Accession Number Assigned by INSD(Jun. 9, 2004)



EMBL Nucleotide Sequence Database

The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications.

The database is produced in an international collaboration with GenBank (USA) and the DNA Database of Japan (DDBJ). Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis. The current database release (Release 86, March 2006), with according Release notes and user manual are available from the EBI servers. A sample database entry is shown here.

A publication in Nucl. Acids Res., 2006, Vol. 34: D10-D15 provides further information and details.

The EMBL nucleotide sequence database group is headed by:
Rolf Apweiler.

Forthcoming EMBL Database Changes:

- Starting from the EMBL release 87 (June 2006) the naming of the release files will change...more

International Nucleotide Sequence Database Collaboration.

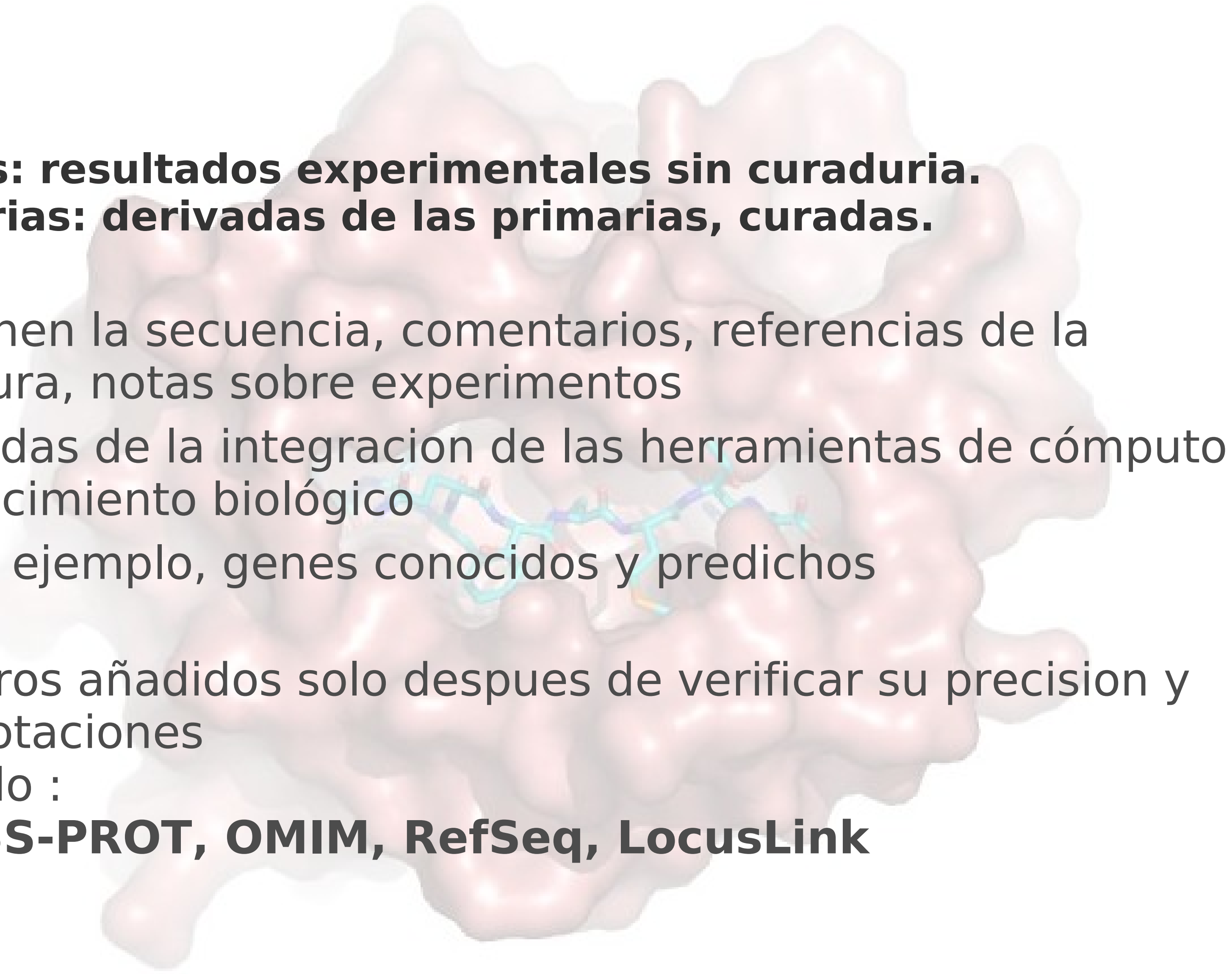
Fetch an EMBL record by accession number

Users can now submit reannotations/reassemblies of sequences already present in EMBL and owned by other groups.

European Molecular Biology (EMBL)
1982

International Collaboration of DNA sequence Databases

Bases de datos primarias y secundarias



Primarias: resultados experimentales sin curaduria.
Secundarias: derivadas de las primarias, curadas.

- contienen la secuencia, comentarios, referencias de la literatura, notas sobre experimentos
- Derivadas de la integracion de las herramientas de cómputo y conocimiento biológico
  - por ejemplo, genes conocidos y predichos

- Registros añadidos solo despues de verificar su precision y las anotaciones
- Ejemplo :
  SWISS-PROT, OMIM, RefSeq, LocusLink

EMBL-BANK

- **Release 86 on 28-FEB-2006.**
- **69,783,593 entries.**
- **126,401,347,060 nucleotides,of which 13,313,896 entries (65,362,911,476 nucleotides) are WGS (whole genome shotgun) data.**
- **The release 86 files total 61 GB compressed.**

EMBL Nucleotide Sequence Database

The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications.

The database is produced in an international collaboration with GenBank (USA) and the DNA Database of Japan (DDBJ). Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis. The current database release (Release 86, March 2006), with according Release notes and user manual are available from the EBI servers. A sample database entry is shown here.

A publication in Nucl. Acids Res., 2006, Vol. 34: D10-D15 provides further information and details.

The EMBL nucleotide sequence database group is headed by: Rolf Apweiler.

Forthcoming EMBL Database Changes:

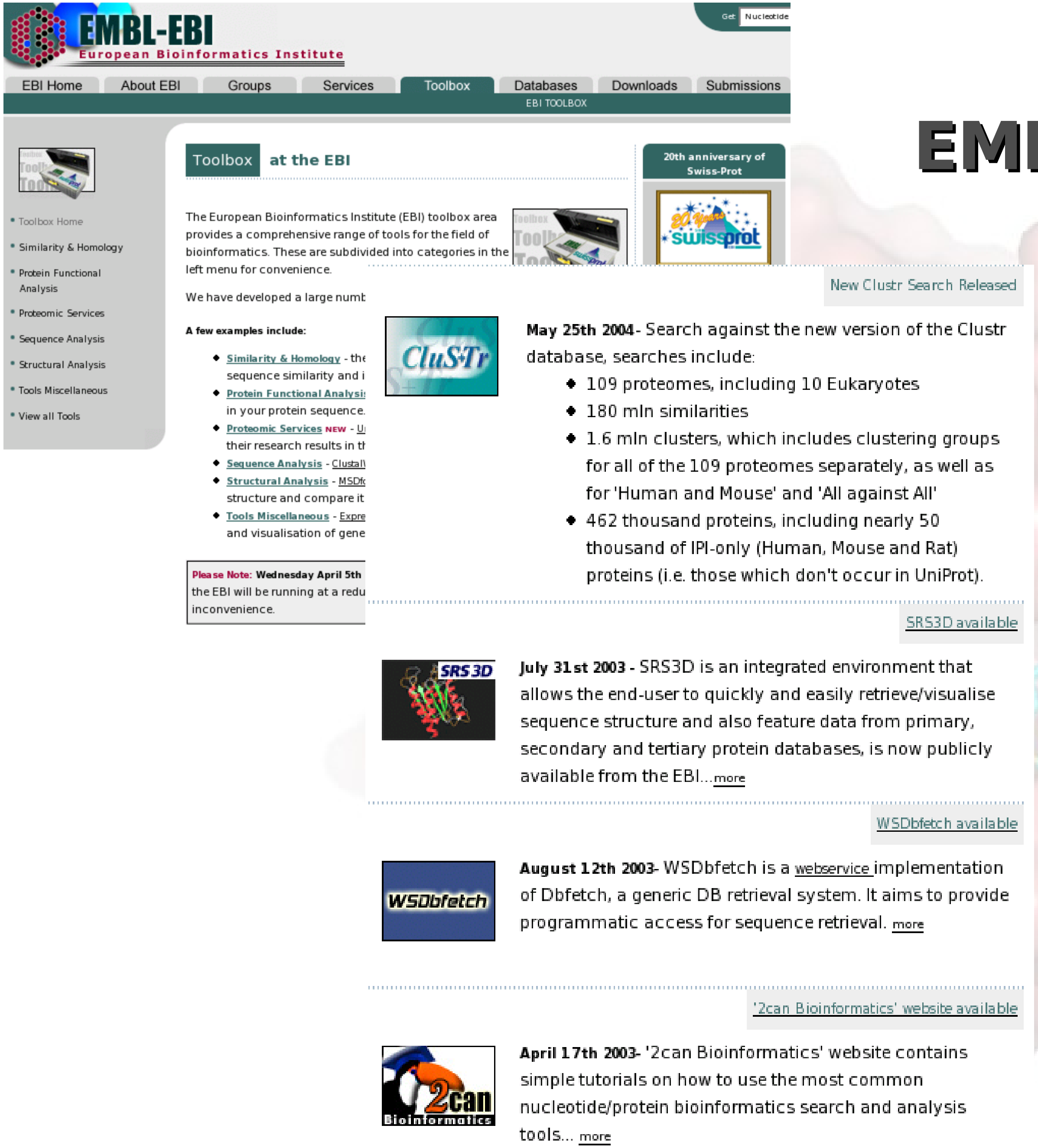
- Starting from the EMBL release 87 (June 2006) the naming of the release files will change...more

International Nucleotide Sequence Database Collaboration.

Fetch an EMBL record by accession number

Users can now submit reannotations/reassemblies of sequences already present in EMBL and owned by other groups.

EMBL-BANK (Tools)



Toolbox at the EBI

The European Bioinformatics Institute (EBI) toolbox area provides a comprehensive range of tools for the field of bioinformatics. These are subdivided into categories in the left menu for convenience.

We have developed a large numb

A few examples include:

- **Similarity & Homology** - the sequence similarity and i
- **Protein Functional Analysis** in your protein sequence.
- **Proteomic Services** NEW - U their research results in th
- **Sequence Analysis** - ClustalW
- **Structural Analysis** - MSDfo structure and compare it
- **Tools Miscellaneous** - Expre and visualisation of gene

Please Note: Wednesday April 5th the EBI will be running at a redu inconvenience.

New Clustr Search Released

May 25th 2004- Search against the new version of the Clustr database, searches include:

- 109 proteomes, including 10 Eukaryotes
- 180 mln similarities
- 1.6 mln clusters, which includes clustering groups for all of the 109 proteomes separately, as well as for 'Human and Mouse' and 'All against All'
- 462 thousand proteins, including nearly 50 thousand of IPI-only (Human, Mouse and Rat) proteins (i.e. those which don't occur in UniProt).

SRS3D available

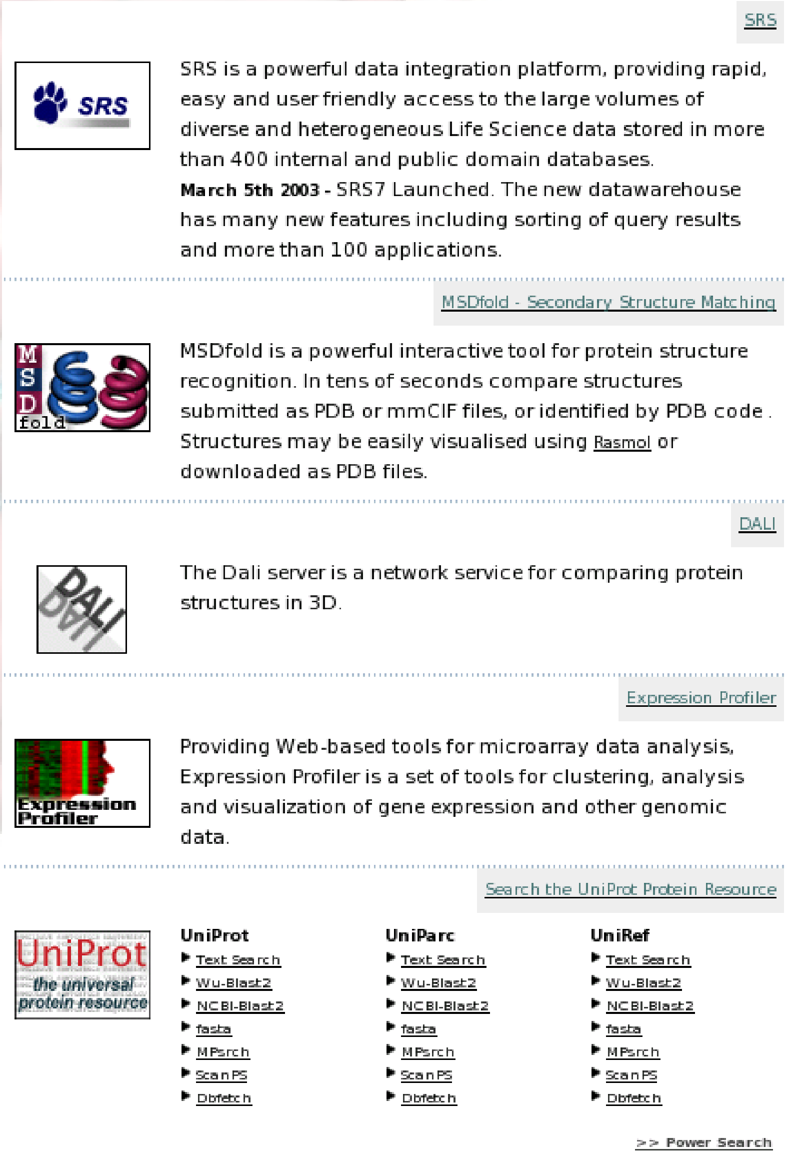
July 31st 2003 - SRS3D is an integrated environment that allows the end-user to quickly and easily retrieve/visualise sequence structure and also feature data from primary, secondary and tertiary protein databases, is now publicly available from the EBI... more

WSDbfetch available

August 12th 2003- WSDbfetch is a webservice implementation of Dbfetch, a generic DB retrieval system. It aims to provide programmatic access for sequence retrieval. more

'2can Bioinformatics' website available

April 17th 2003- '2can Bioinformatics' website contains simple tutorials on how to use the most common nucleotide/protein bioinformatics search and analysis tools... more



SRS

SRS is a powerful data integration platform, providing rapid, easy and user friendly access to the large volumes of diverse and heterogeneous Life Science data stored in more than 400 internal and public domain databases.
March 5th 2003 - SRS7 Launched. The new datawarehouse has many new features including sorting of query results and more than 100 applications.

MSDfold - Secondary Structure Matching

MSDfold is a powerful interactive tool for protein structure recognition. In tens of seconds compare structures submitted as PDB or mmCIF files, or identified by PDB code . Structures may be easily visualised using Rasmol or downloaded as PDB files.

DALI

The Dali server is a network service for comparing protein structures in 3D.

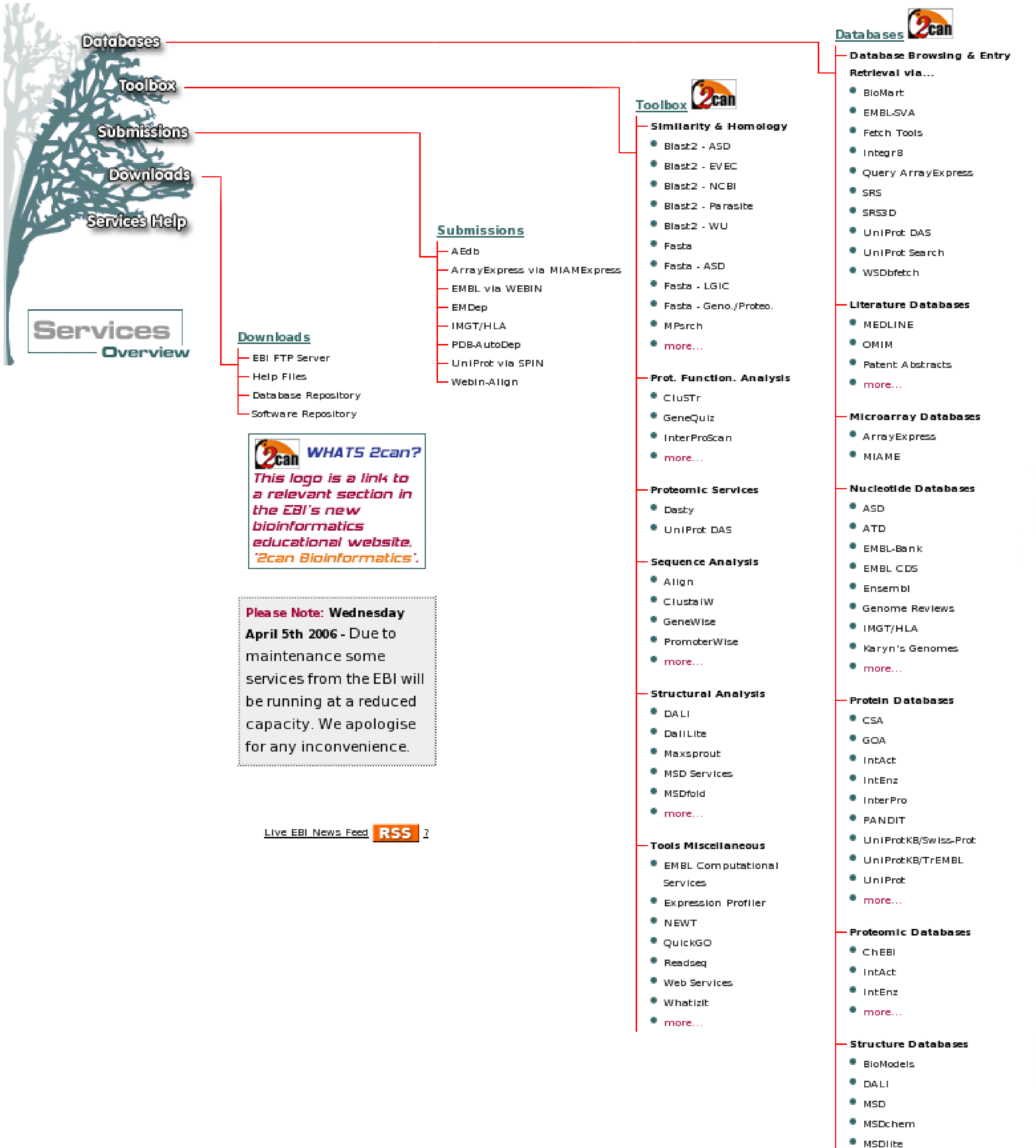
Expression Profiler

Providing Web-based tools for microarray data analysis, Expression Profiler is a set of tools for clustering, analysis and visualization of gene expression and other genomic data.

Search the UniProt Protein Resource

UniProt	UniParc	UniRef
Text Search	Text Search	Text Search
Wu-Blast2	Wu-Blast2	Wu-Blast2
NCBI-Blast2	NCBI-Blast2	NCBI-Blast2
fasta	fasta	fasta
MPsrch	MPsrch	MPsrch
ScanPS	ScanPS	ScanPS
Dbfetch	Dbfetch	Dbfetch

>> Power Search


Databases
Toolbox
Submissions
Downloads
Services Help

Services Overview

Downloads
- EBI FTP Server
- Help Files
- Database Repository
- Software Repository

WHATS 2can?
This logo is a link to a relevant section in the EBI's new bioinformatics educational website. '2can Bioinformatics'.

Please Note: **Wednesday April 5th 2006 -** Due to maintenance some services from the EBI will be running at a reduced capacity. We apologise for any inconvenience.

Live EBI News Feed RSS ?

Submissions
- AEdb
- ArrayExpress via MIAMExpress
- EMBL via WEBIN
- EMDep
- IMGT/HLA
- PDB-AutoDep
- UniProt via SPIN
- Webin-Align

Toolbox

Similarity & Homology
- Blast2 - ASD
- Blast2 - EVEC
- Blast2 - NCBI
- Blast2 - Parasite
- Blast2 - WU
- Fasta
- Fasta - ASD
- Fasta - LGIC
- Fasta - Geno./Proteo.
- MPsrch
- more...

Prot. Function. Analysis
- ClusTr
- GeneQuiz
- InterProScan
- more...

Proteomic Services
- Dasty
- UniProt DAS

Sequence Analysis
- Align
- ClustalW
- GeneWise
- PromoterWise
- more...

Structural Analysis
- DALI
- DaliLite
- Maxsprout
- MSD Services
- MSDfold
- more...

Tools Miscellaneous
- EMBL Computational Services
- Expression Profiler
- NEWT
- QuickGO
- Readseq
- Web Services
- Whatizit
- more...

Databases

Database Browsing & Entry Retrieval via...
- BioMart
- EMBL-SVA
- Fetch Tools
- Integr8
- Query ArrayExpress
- SRS
- SRS3D
- UniProt DAS
- UniProt Search
- WSDbfetch

Literature Databases
- MEDLINE
- OMIM
- Patent Abstracts
- more...

Microarray Databases
- ArrayExpress
- MIAME

Nucleotide Databases
- ASD
- ATD
- EMBL-Bank
- EMBL CDS
- Ensembl
- Genome Reviews
- IMGT/HLA
- Karyn's Genomes
- more...

Protein Databases
- CSA
- GOA
- IntAct
- IntEnz
- InterPro
- PANDIT
- UniProtKB/Swiss-Prot
- UniProtKB/TrEMBL
- UniProt
- more...

Proteomic Databases
- ChEBI
- IntAct
- IntEnz
- more...

Structure Databases
- BioModels
- DALI
- MSD
- MSDchem
- MSDlite

UNIPROT

Uniprot Knowledge Base
UniprotKB/Swissprot
UniprotKB/TrEMBL

Uniprot reference Clusters
Clusters a partir de UniprotKB y Uniparc (Uniref100, Uniref90, Uniref50).

Uniprot Archive
Todas las secuencias protéicas conocidas, no redundante.

Welcome to UniProt

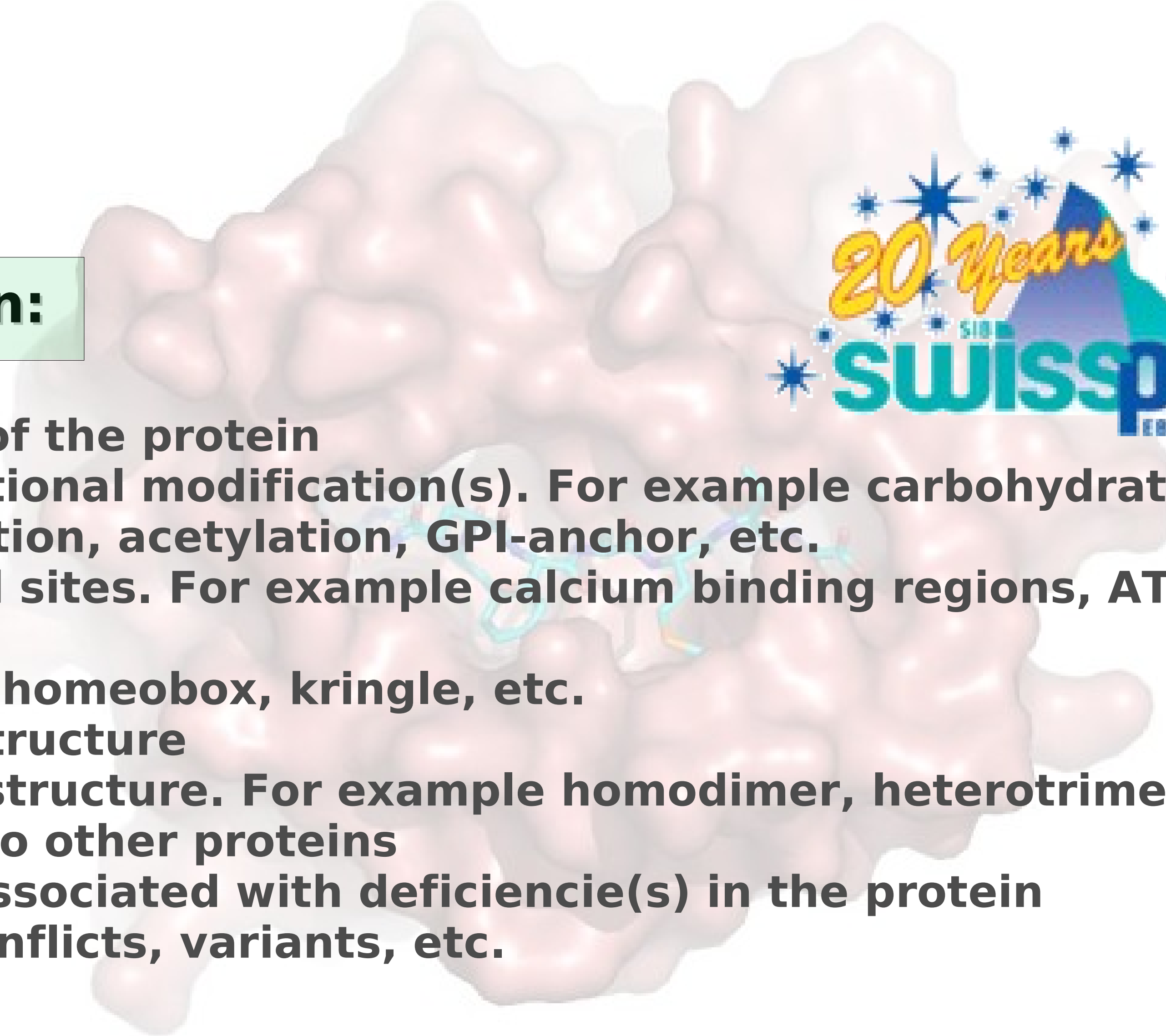
UniProt (Universal Protein Resource) is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.

UniProt is comprised of three components, each optimized for different uses. The **UniProt Knowledgebase (UniProtKB)** is the central access point for extensive curated protein information, including function, classification, and cross-reference. The **UniProt Reference Clusters (UniRef)** databases combine closely related sequences into a single record to speed searches. The **UniProt Archive (UniParc)** is a comprehensive repository, reflecting the history of all protein sequences.

The sequences and information in UniProt are accessible via **text search**, **BLAST similarity search**, and **FTP**.

In-Silico Analysis of Proteins, Celebrating the 20th Anniversary of Swiss-Prot

UNIPROT-SWISSPROT

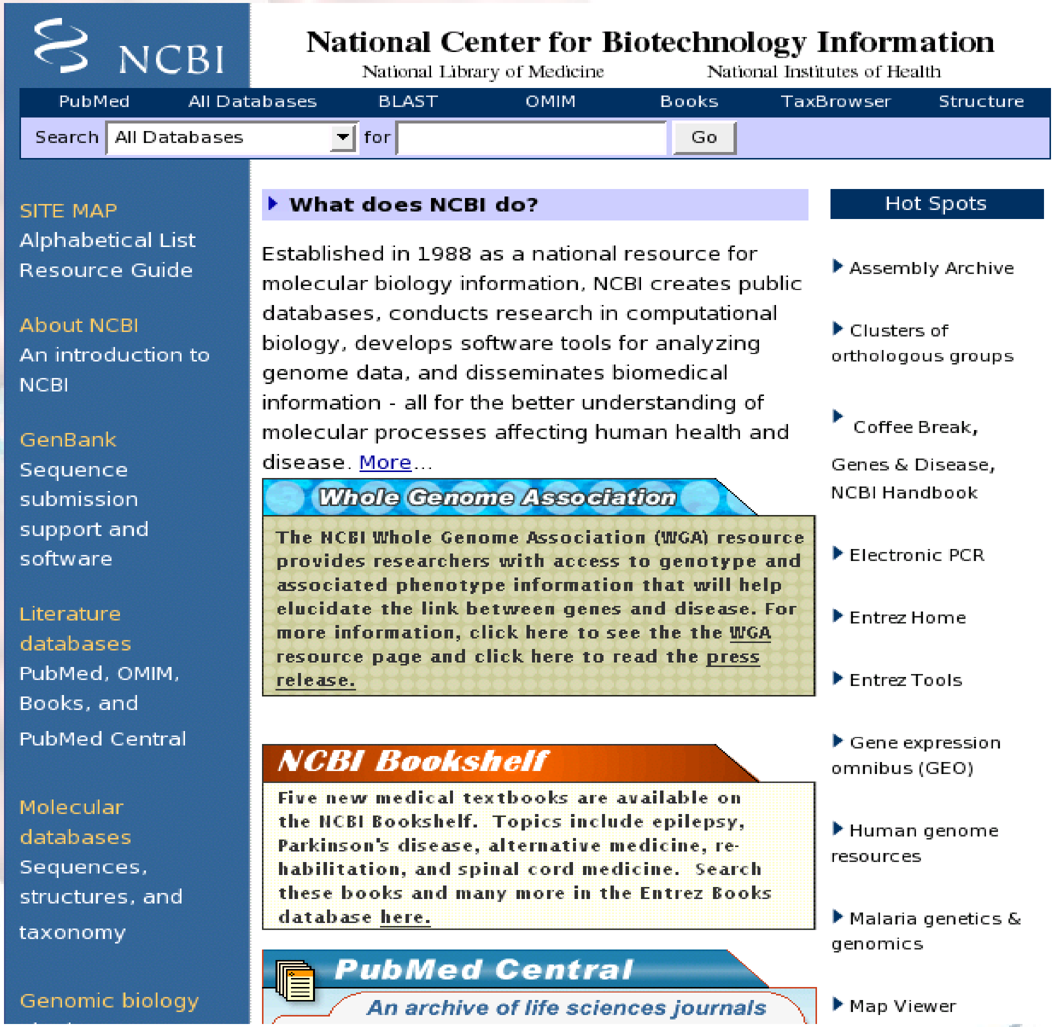


Anotación:

- **Function(s) of the protein**
- **Post-translational modification(s). For example carbohydrates,**
- **phosphorylation, acetylation, GPI-anchor, etc.**
- **Domains and sites. For example calcium binding regions, ATP-binding sites,**
- **zinc fingers, homeobox, kringle, etc.**
- **Secondary structure**
- **Quaternary structure. For example homodimer, heterotrimer, etc.**
- **Similarities to other proteins**
- **Disease(s) associated with deficiencie(s) in the protein**
- **Sequence conflicts, variants, etc.**

NCBI
National Center for Biotechnology Information

Establecido en 1988. Es un recurso nacional (USA) en información biológica molecular, el NCBI **crea bases de datos públicas**, **lleva a cabo investigaciones en biología computacional**, **desarrolla herramientas de software para el análisis de información genómica**, y **da a conocer la información biomédica** – todo esto con el propósito de ***mejorar nuestro entendimiento acerca de los procesos moleculares que afectan la salud y la enfermedad en los seres humanos.***



What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...

Whole Genome Association

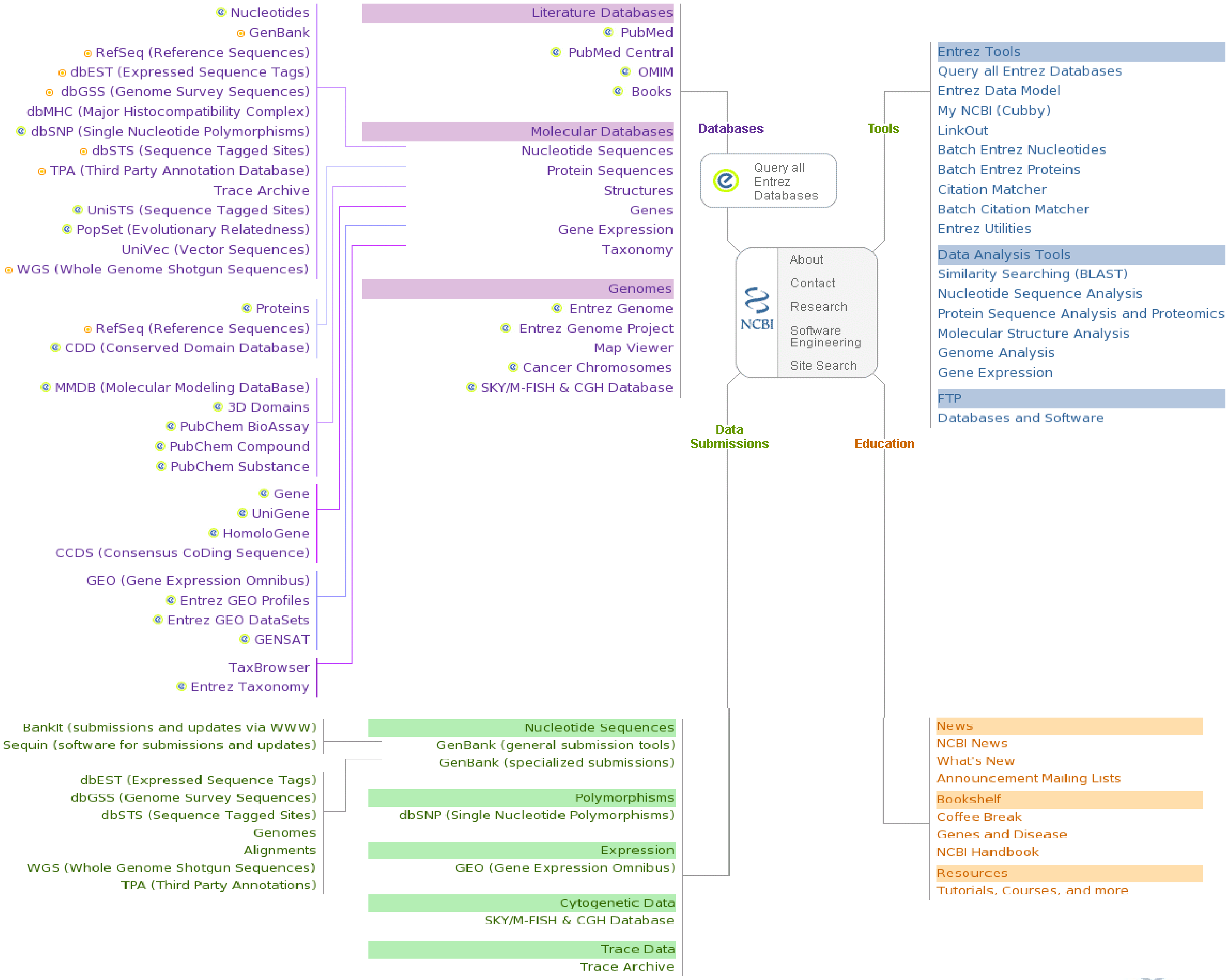
The NCBI Whole Genome Association (WGA) resource provides researchers with access to genotype and associated phenotype information that will help elucidate the link between genes and disease. For more information, click here to see the the WGA resource page and click here to read the press release.

NCBI Bookshelf

Five new medical textbooks are available on the NCBI Bookshelf. Topics include epilepsy, Parkinson's disease, alternative medicine, rehabilitation, and spinal cord medicine. Search these books and many more in the Entrez Books database here.

PubMed Central

An archive of life sciences journals



- About
- Contact
- Research
- Software Engineering
- Site Search

Databases

- Literature Databases
  - PubMed
  - PubMed Central
  - OMIM
  - Books
- Molecular Databases
  - Nucleotide Sequences
    - Nucleotides
    - GenBank
    - RefSeq (Reference Sequences)
    - dbEST (Expressed Sequence Tags)
    - dbGSS (Genome Survey Sequences)
    - dbMHC (Major Histocompatibility Complex)
    - dbSNP (Single Nucleotide Polymorphisms)
    - dbSTS (Sequence Tagged Sites)
    - TPA (Third Party Annotation Database)
    - Trace Archive
    - UniSTS (Sequence Tagged Sites)
    - PopSet (Evolutionary Relatedness)
    - UniVec (Vector Sequences)
    - WGS (Whole Genome Shotgun Sequences)
  - Protein Sequences
    - Proteins
    - RefSeq (Reference Sequences)
    - CDD (Conserved Domain Database)
  - Structures
    - MMDB (Molecular Modeling DataBase)
    - 3D Domains
    - PubChem BioAssay
    - PubChem Compound
    - PubChem Substance
  - Genes
    - Gene
    - UniGene
    - HomoloGene
    - CCDS (Consensus CoDing Sequence)
  - Gene Expression
    - GEO (Gene Expression Omnibus)
    - Entrez GEO Profiles
    - Entrez GEO DataSets
    - GENSAT
  - Taxonomy
    - TaxBrowser
    - Entrez Taxonomy
- Genomes
  - Entrez Genome
  - Entrez Genome Project
  - Map Viewer
  - Cancer Chromosomes
  - SKY/M-FISH & CGH Database

Tools

- Entrez Tools
  - Query all Entrez Databases
  - Entrez Data Model
  - My NCBI (Cubby)
  - LinkOut
  - Batch Entrez Nucleotides
  - Batch Entrez Proteins
  - Citation Matcher
  - Batch Citation Matcher
  - Entrez Utilities
- Data Analysis Tools
  - Similarity Searching (BLAST)
  - Nucleotide Sequence Analysis
  - Protein Sequence Analysis and Proteomics
  - Molecular Structure Analysis
  - Genome Analysis
  - Gene Expression
- FTP
  - Databases and Software

Data Submissions

- Nucleotide Sequences
  - GenBank (general submission tools)
    - BankIt (submissions and updates via WWW)
    - Sequin (software for submissions and updates)
  - GenBank (specialized submissions)
    - dbEST (Expressed Sequence Tags)
    - dbGSS (Genome Survey Sequences)
    - dbSTS (Sequence Tagged Sites)
    - Genomes
    - Alignments
    - WGS (Whole Genome Shotgun Sequences)
    - TPA (Third Party Annotations)
- Polymorphisms
  - dbSNP (Single Nucleotide Polymorphisms)
- Expression
  - GEO (Gene Expression Omnibus)
- Cytogenetic Data
  - SKY/M-FISH & CGH Database
- Trace Data
  - Trace Archive

Education

- News
  - NCBI News
  - What's New
  - Announcement Mailing Lists
- Bookshelf
  - Coffee Break
  - Genes and Disease
  - NCBI Handbook
- Resources
  - Tutorials, Courses, and more

NCBI- ENTREZ

Welcome to the Entrez cross-database search page

PubMed: biomedical literature citations and abstracts

PubMed Central: free, full text journal articles

Site Search: NCBI web and FTP sites

Books: online books

OMIM: online Mendelian Inheritance in Man

OMIA: online Mendelian Inheritance in Animals

Nucleotide: sequence database (GenBank)

Protein: sequence database

Genome: whole genome sequences

Structure: three-dimensional macromolecular structures

Taxonomy: organisms in GenBank

SNP: single nucleotide polymorphism

Gene: gene-centered information

HomoloGene: eukaryotic homology groups

PubChem Compound: unique small molecule chemical structures

PubChem Substance: deposited chemical substance records

Genome Project: genome project information

UniGene: gene-oriented clusters of transcript sequences

CDD: conserved protein domain database

3D Domains: domains from Entrez Structure

UniSTS: markers and mapping data

PopSet: population study data sets

GEO Profiles: expression and molecular abundance profiles

GEO DataSets: experimental sets of GEO data

Cancer Chromosomes: cytogenetic databases

PubChem BioAssay: bioactivity screens of chemical substances

GENSAT: gene expression atlas of mouse central nervous system

Probe: sequence-specific reagents

Journals: detailed information *about* the journals indexed in PubMed and other Entrez databases

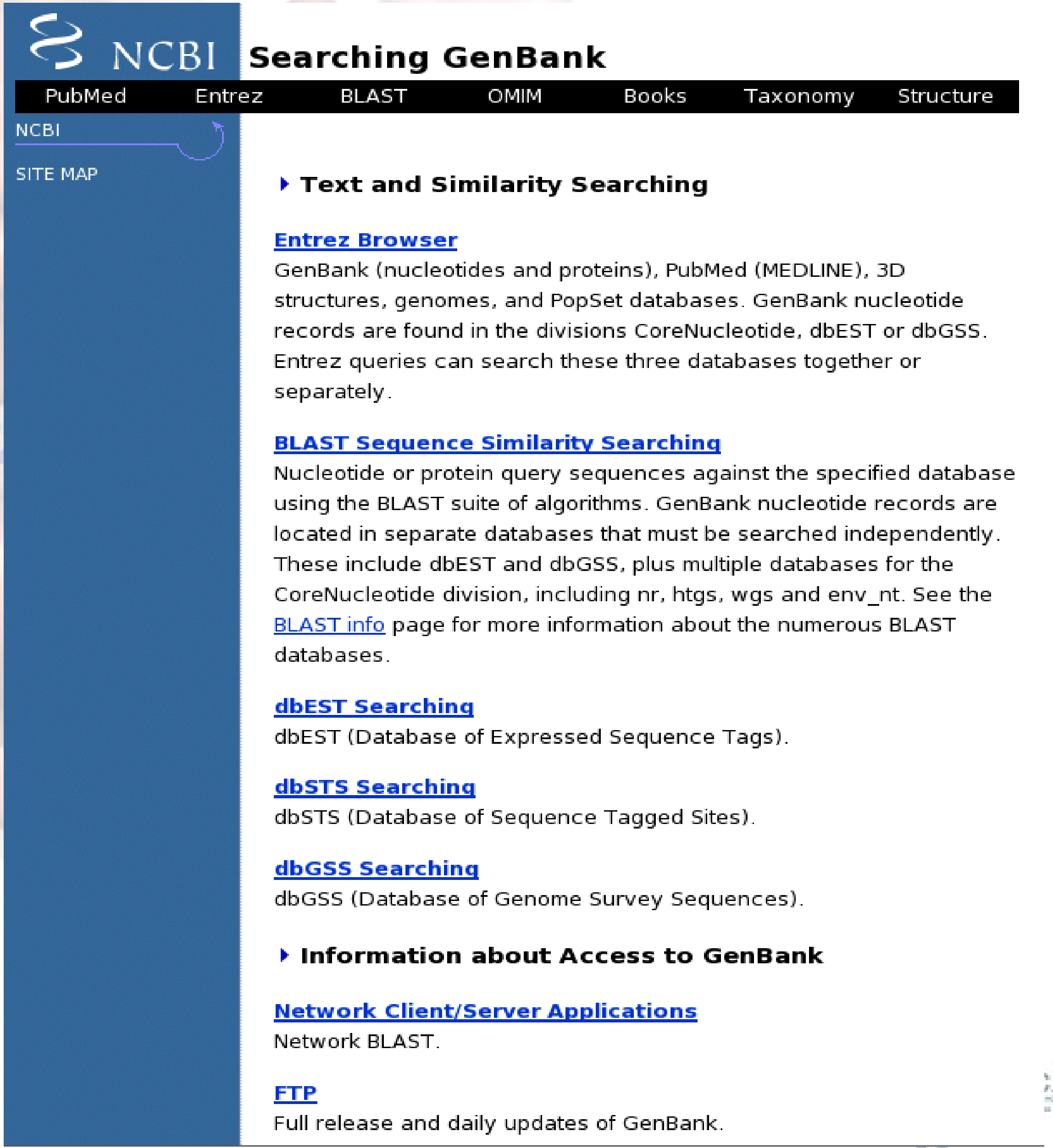
NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections

MeSH: detailed information about NLM's controlled vocabulary

NCBI- GENBANK

Base de datos pública de todas las secuencias DNA disponibles

59,750,386,305 bases en **54,584,635** entradas de secuencias en la división tradicional de GenBank y **63,183,065,091** bases en **12,465,546** entradas en la división WGS (Febrero de 2006).



Searching GenBank

Text and Similarity Searching

Entrez Browser
GenBank (nucleotides and proteins), PubMed (MEDLINE), 3D structures, genomes, and PopSet databases. GenBank nucleotide records are found in the divisions CoreNucleotide, dbEST or dbGSS. Entrez queries can search these three databases together or separately.

BLAST Sequence Similarity Searching
Nucleotide or protein query sequences against the specified database using the BLAST suite of algorithms. GenBank nucleotide records are located in separate databases that must be searched independently. These include dbEST and dbGSS, plus multiple databases for the CoreNucleotide division, including nr, htgs, wgs and env_nt. See the BLAST info page for more information about the numerous BLAST databases.

dbEST Searching
dbEST (Database of Expressed Sequence Tags).

dbSTS Searching
dbSTS (Database of Sequence Tagged Sites).

dbGSS Searching
dbGSS (Database of Genome Survey Sequences).

Information about Access to GenBank

Network Client/Server Applications
Network BLAST.

FTP
Full release and daily updates of GenBank.

Protein Data Bank (PDB)

Base de datos de estructuras tridimensionales de proteínas

- **Cristalografia de Rayos X.**
- **41136 estructuras (Enero 16 de 2007)**
- **Archivo de coordenadas.**

An Information Portal to Biological Macromolecular Structures

As of Tuesday Nov 01, 2005 there are 33367 Structures | PDB Statistics

Welcome to the RCSB PDB

The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

The RCSB is a member of the wwPDB whose mission is to ensure that the PDB archive remains an international resource with uniform data.

This beta site offers tools for browsing, searching, and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

A **narrated tutorial** illustrates how to search, navigate, browse, generate reports and visualize structures using this new site. [This requires the Macromedia Flash player download.]

Comments? betafeedback@rcsb.org

Molecule of the Month: Acetylcholine Receptor

Nerve cells need to be able to send messages to each other quickly and clearly. One way that nerve cells communicate with their neighbors is by sending a burst of small neurotransmitter molecules. These molecules diffuse to the neighboring cell and bind to special receptor proteins in the cell surface. These receptors then open, allowing ions to flow inside.

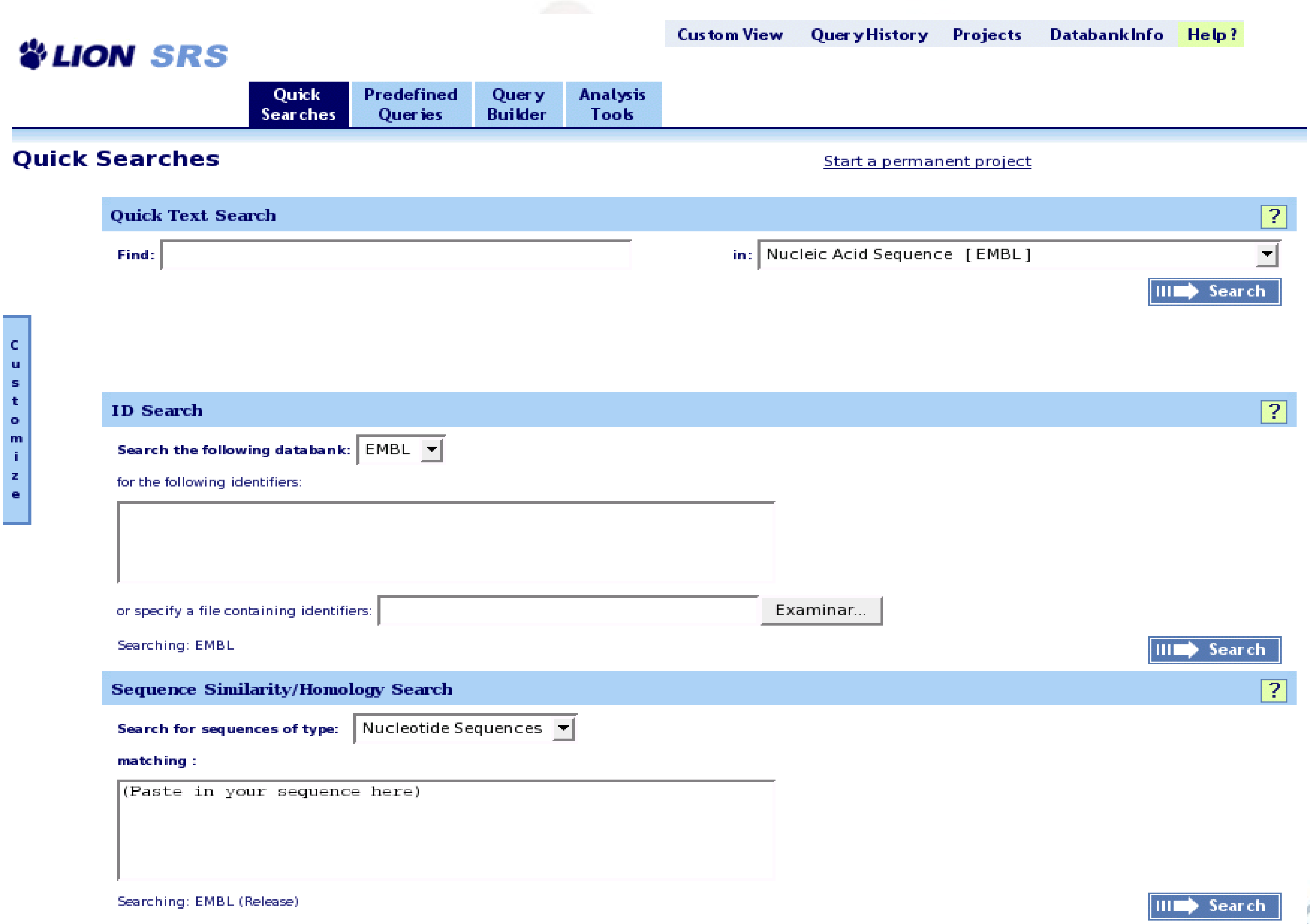
01-Nov-2005

Newsletter Published

The New Beta Site and Interviews with William L. Duax (International Union of Crystallography and Hauptman-Woodward Medical Research Institute) and Miriam Rossi (Vassar College). The Fall 2005 issue of the RCSB PDB Newsletter has been published in HTML and PDF formats. Highlights include a look at the new beta site that will become the main RCSB PDB site on January 1, 2006;

http://www.rcsb.org/pdb/

Sequence Retrieval System



Quick Searches

Quick Text Search

Find:

in: Nucleic Acid Sequence [EMBL]

ID Search

Search the following databank: EMBL

for the following identifiers:

or specify a file containing identifiers:

Searching: EMBL

Sequence Similarity/Homology Search

Search for sequences of type: Nucleotide Sequences

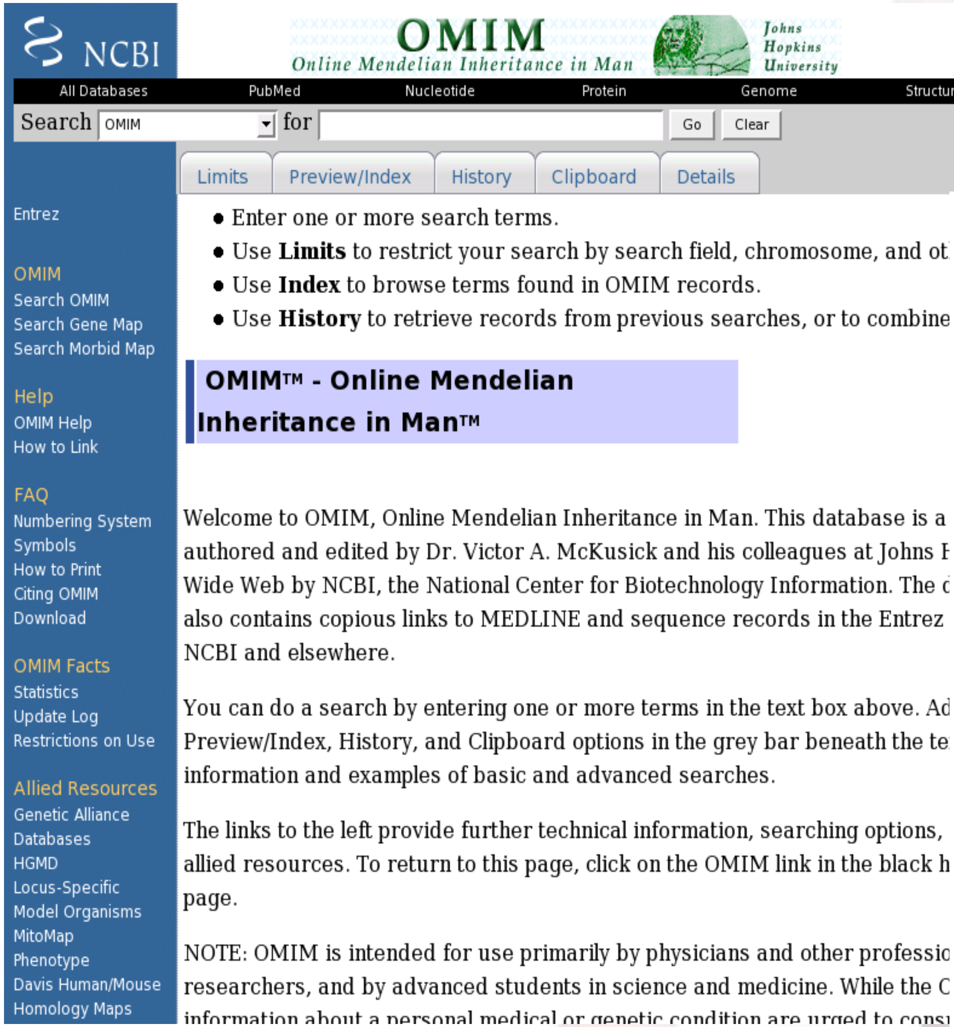
matching :

(Paste in your sequence here)

Searching: EMBL (Release)

http://srs.ibun.unal.edu.co:8080/srs81/

Más bases de datos I



- Enter one or more search terms.
- Use **Limits** to restrict your search by search field, chromosome, and ot
- Use **Index** to browse terms found in OMIM records.
- Use **History** to retrieve records from previous searches, or to combine

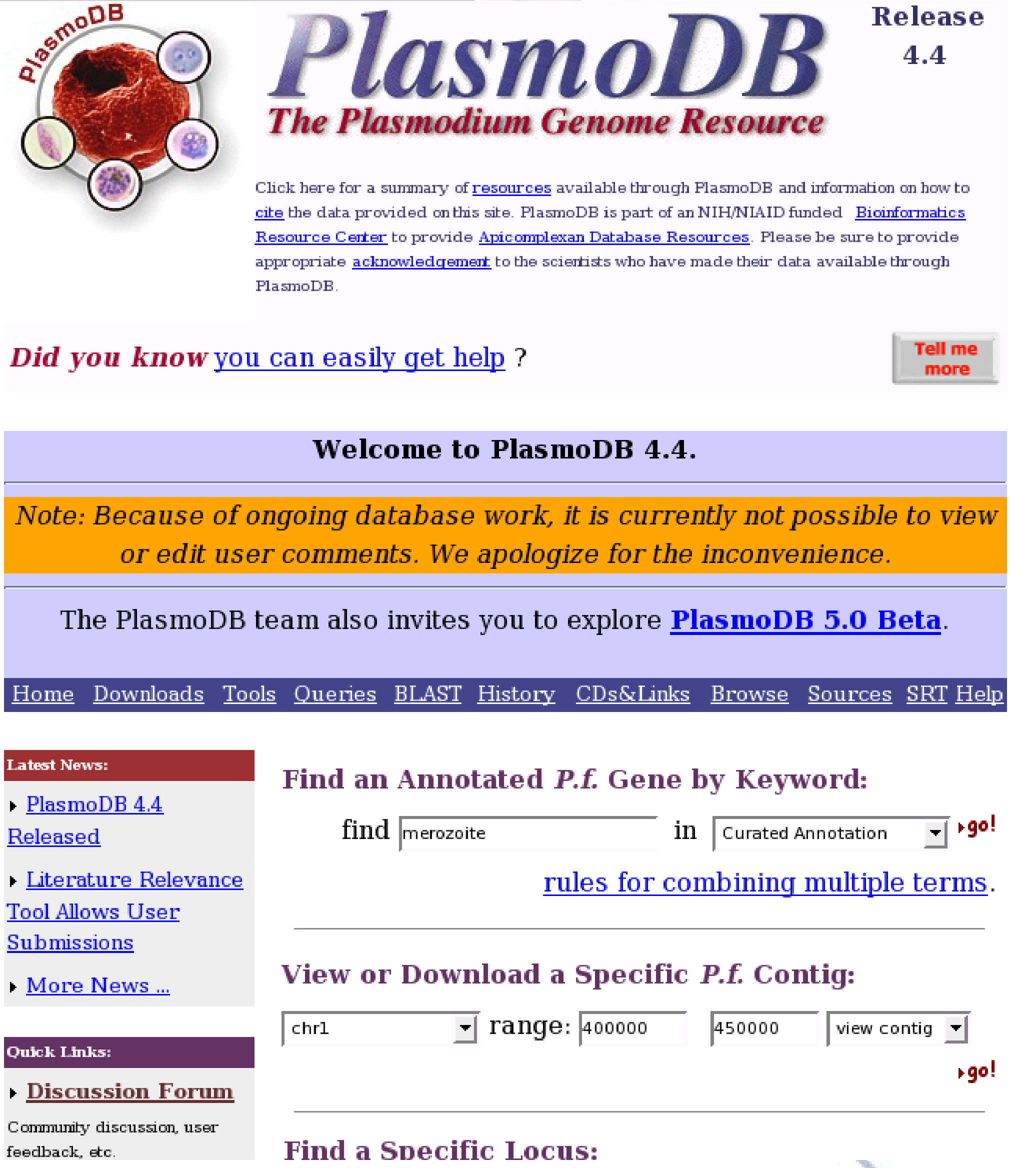
OMIM™ - Online Mendelian Inheritance in Man™

Welcome to OMIM, Online Mendelian Inheritance in Man. This database is a
authored and edited by Dr. Victor A. McKusick and his colleagues at Johns H
Wide Web by NCBI, the National Center for Biotechnology Information. The d
also contains copious links to MEDLINE and sequence records in the Entrez
NCBI and elsewhere.

You can do a search by entering one or more terms in the text box above. Ad
Preview/Index, History, and Clipboard options in the grey bar beneath the te
information and examples of basic and advanced searches.

The links to the left provide further technical information, searching options,
allied resources. To return to this page, click on the OMIM link in the black h
page.

NOTE: OMIM is intended for use primarily by physicians and other professio
researchers, and by advanced students in science and medicine. While the O
information about a personal medical or genetic condition are urged to cons



Click here for a summary of resources available through PlasmoDB and information on how to cite the data provided on this site. PlasmoDB is part of an NIH/NIAID funded Bioinformatics Resource Center to provide Apicomplexan Database Resources. Please be sure to provide appropriate acknowledgement to the scientists who have made their data available through PlasmoDB.

Did you know you can easily get help ?

Welcome to PlasmoDB 4.4.

Note: Because of ongoing database work, it is currently not possible to view or edit user comments. We apologize for the inconvenience.

The PlasmoDB team also invites you to explore **PlasmoDB 5.0 Beta**.

- PlasmoDB 4.4 Released
- Literature Relevance Tool Allows User Submissions
- More News ...

- **Discussion Forum**

Community discussion, user feedback, etc.

Find an Annotated *P.f.* Gene by Keyword:

find: merozoite — in: Curated Annotation

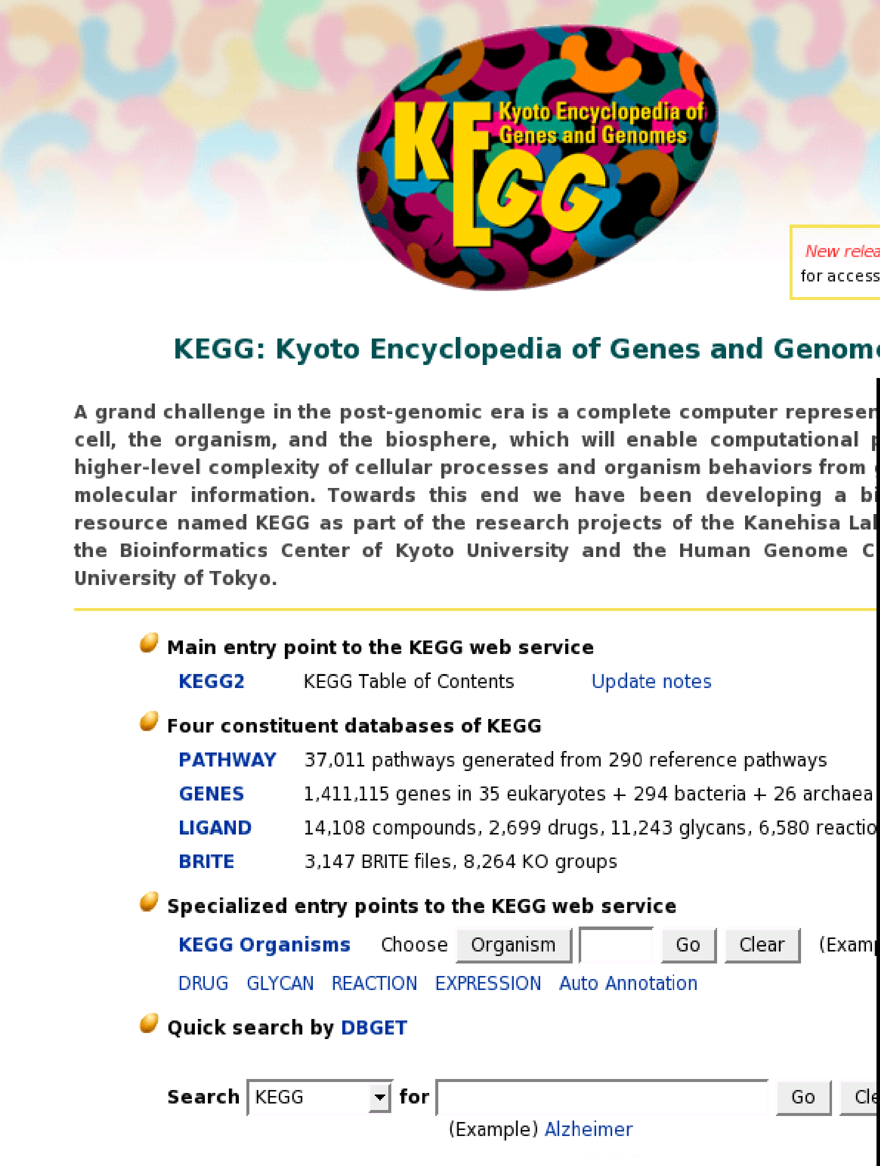
rules for combining multiple terms.

View or Download a Specific *P.f.* Contig:

chr1 — range: 400000 — 450000 — view contig

Find a Specific Locus:

Más bases de datos II



KEGG: Kyoto Encyclopedia of Genes and Genomes

A grand challenge in the post-genomic era is a complete computer represen cell, the organism, and the biosphere, which will enable computational p higher-level complexity of cellular processes and organism behaviors from molecular information. Towards this end we have been developing a bi resource named KEGG as part of the research projects of the Kanehisa Lal the Bioinformatics Center of Kyoto University and the Human Genome C University of Tokyo.

- **Main entry point to the KEGG web service**

KEGG2	KEGG Table of Contents	Update notes

- **Four constituent databases of KEGG**

PATHWAY	37,011 pathways generated from 290 reference pathways
GENES	1,411,115 genes in 35 eukaryotes + 294 bacteria + 26 archaea
LIGAND	14,108 compounds, 2,699 drugs, 11,243 glycans, 6,580 reactio
BRITE	3,147 BRITE files, 8,264 KO groups

- **Specialized entry points to the KEGG web service**

KEGG Organisms Choose Organism ☐ Go Clear (Examp

DRUG GLYCAN REACTION EXPRESSION Auto Annotation

- **Quick search by DBGET**

Search KEGG ▾ **for** ☐ Go Cle

(Example) Alzheimer

Human Disease to Drosophila Gene Database

Version 2.1

Last update: Fri, 31 Mar 2006

Credits

Homophila utilizes the sequence information of human disease genes from the NCBI OMIM (Online Mendelian Inheritance in Man) database in order to determine if sequence homologs of these genes exist in the current *Drosophila* sequence database (FlyBase). Sequences are compared using NCBI's BLAST program. The database is updated weekly and can be searched by human disease, gene name, OMIM number, title, subtitle and/or allelic variant descriptions.

This analysis is discussed in detail in the following publications:

Reiter LT, Potocki L, Chien S, Gribskov M, Bier E., "A Systematic Analysis of Human Disease-Associated Gene Sequences In Drosophila melanogaster," *Genome Research* 11:1114-1125, Cold Spring Harbor Labratory Press.

Chien S, Reiter LT, Bier E, Gribskov M., "Homophila: human disease gene cognates in Drosophila", Nucleic Acids Research, 2002, Vol. 30, No. 1 149-151

Tab delimited flatfiles are now available for download. See table below for links.

Current statistics

OMIM entries having allelic variants	1844
OMIM entries having protein reference sequences	1840
H. sapiens reference sequences	2727
Protein sequences matched against *D. melanogaster*	Entire BLAST dataset E-value < 10^{-10} E-value < 10^{-50} E-value < 10^{-100} E-value < 10^{-200}

Search for: ☐ Human text ▾

☑ Diseases ☑ Titles ☑ Subtitles ☑ Allelic variants

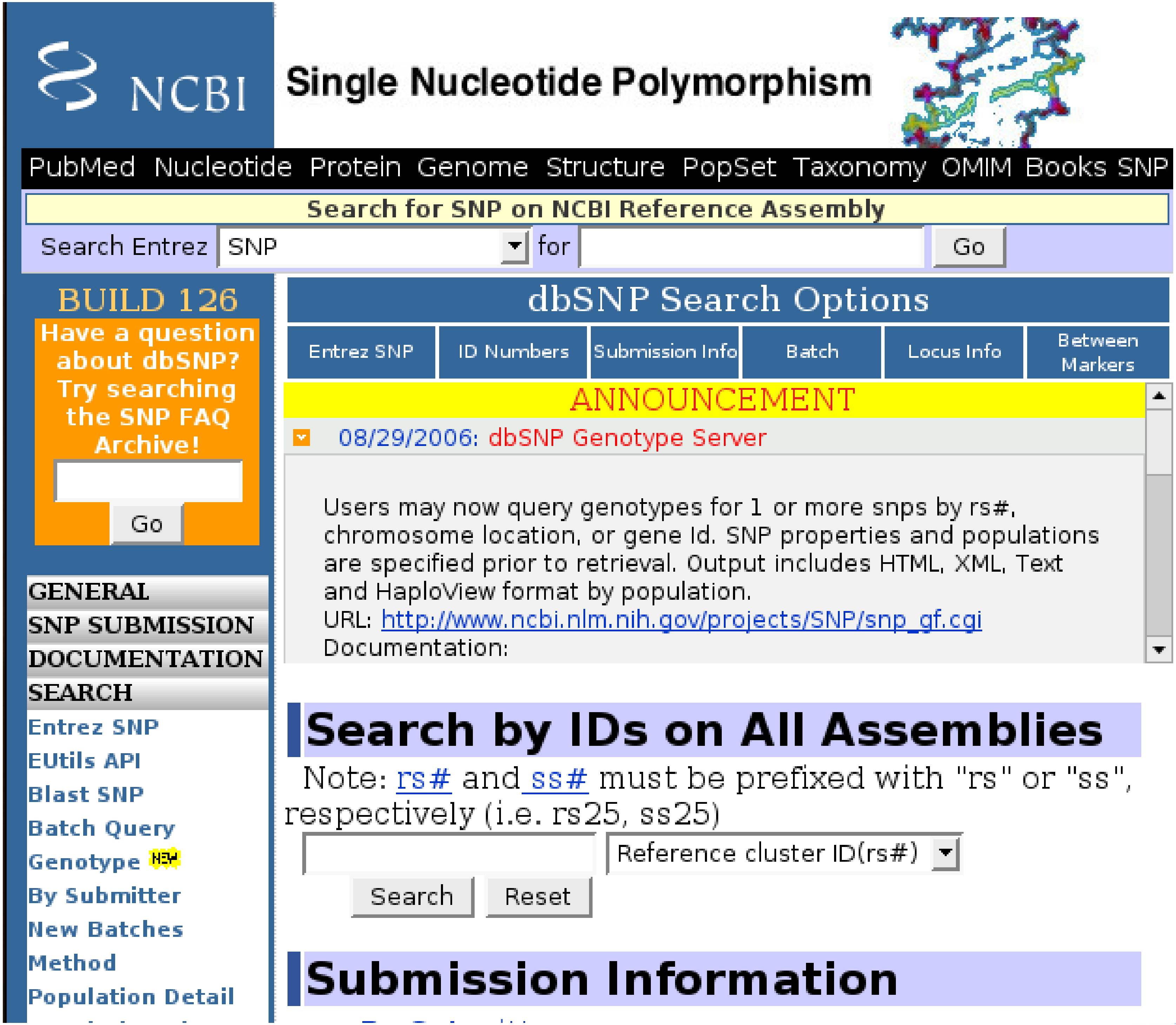
E-value cutoff: None ▾

Show best 5 matches ▾

Más bases de datos III

Bases de datos que no pueden ser clasificadas.

ReBase (enzimas restricción),
RepBase (repeticiones)
TRANSFAC (factores de transcripción),
CarbBank, GlycoSuiteDB (Azucares ligados),
Proteína-proteína interacción db (DIR, ProNet, Interact),
Proteasa db (MEROPS), patentes en biotecnología db, etc.;



ANNOUNCEMENT

08/29/2006: dbSNP Genotype Server

Users may now query genotypes for 1 or more snps by rs#, chromosome location, or gene Id. SNP properties and populations are specified prior to retrieval. Output includes HTML, XML, Text and HaploView format by population.
URL: http://www.ncbi.nlm.nih.gov/projects/SNP/snp_gf.cgi
Documentation:

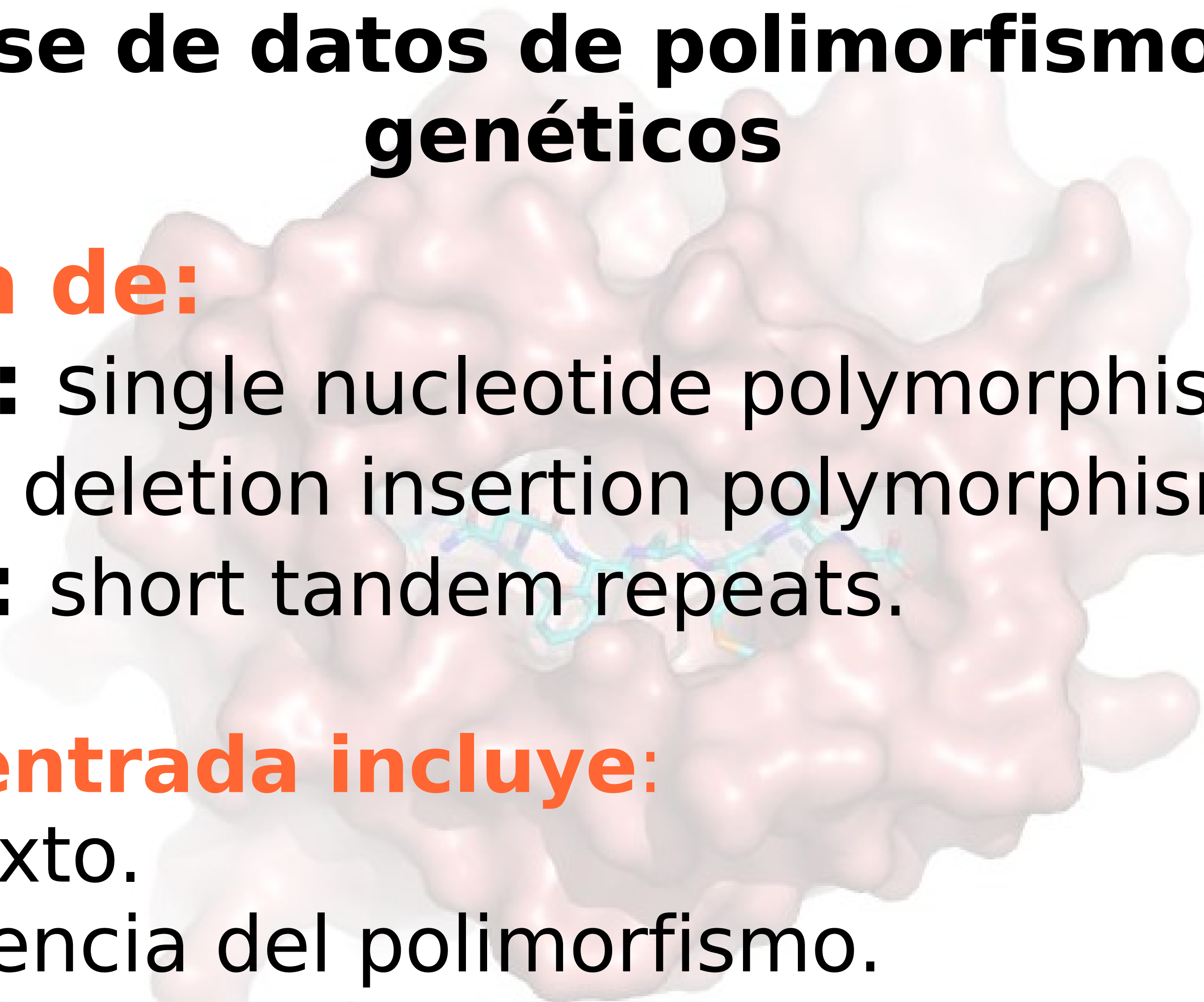
Search by IDs on All Assemblies

Note: rs# and ss# must be prefixed with "rs" or "ss", respectively (i.e. rs25, ss25)

Submission Information

dbSNP

Base de datos de polimorfismos genéticos

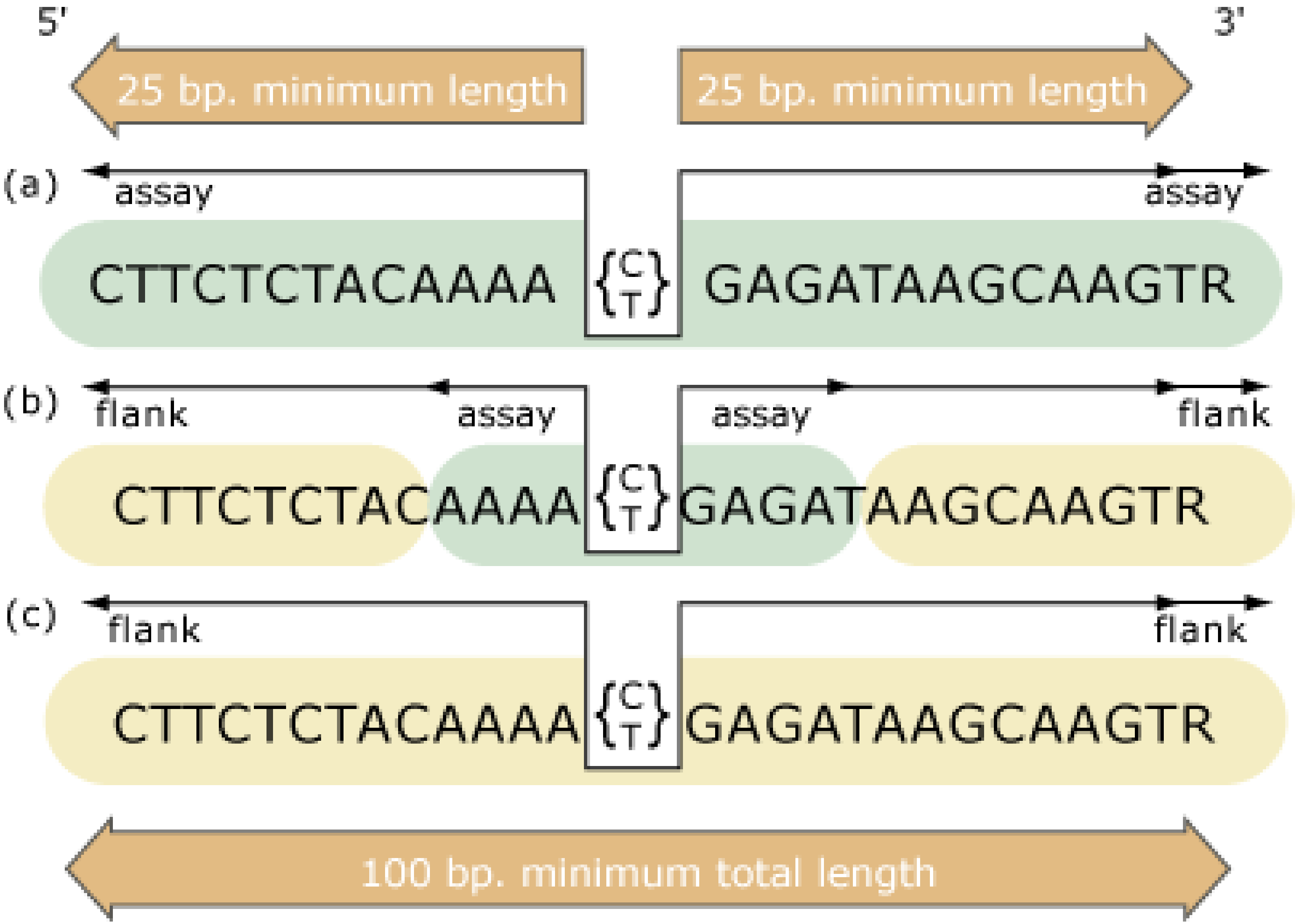


Consta de:

- **SNPs:** single nucleotide polymorphism
- **DIPs:** deletion insertion polymorphisms.
- **SSRs:** short tandem repeats.

Cada entrada incluye:

- Contexto.
- Frecuencia del polimorfismo.
- Métodos experimentales para su obtención.



5'

3'

25 bp. minimum length

25 bp. minimum length

(a) assay assay

CTTCTCTACAAAA {C T} GAGATAAGCAAGTR

(b) flank assay assay flank

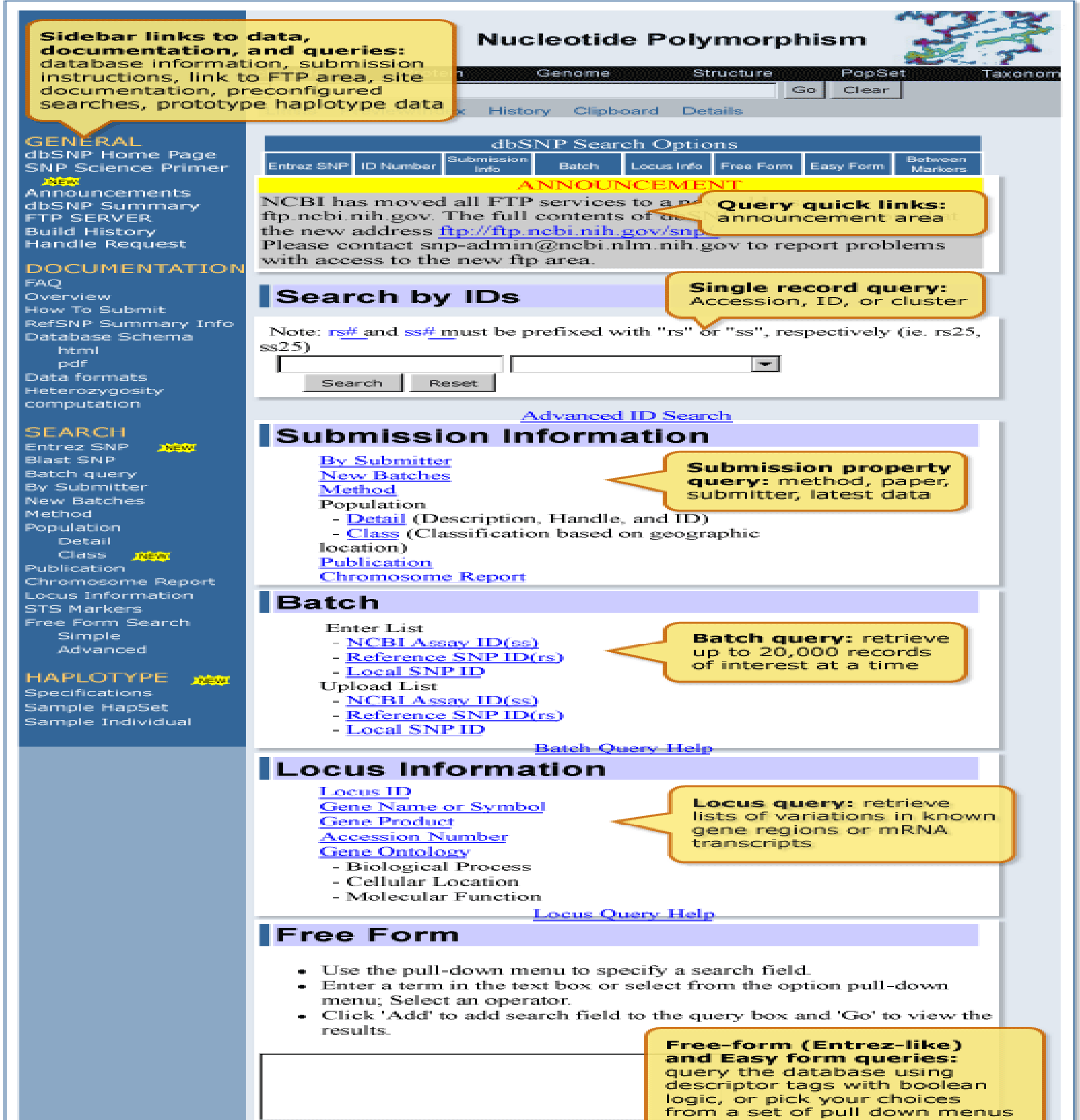
CTTCTCTACAAAA {C T} GAGATAAGCAAGTR

(c) flank flank

CTTCTCTACAAAA {C T} GAGATAAGCAAGTR

100 bp. minimum total length

Centro de BIOINFORMÁTICA



Sidebar links to data, documentation, and queries: database information, submission instructions, link to FTP area, site documentation, preconfigured searches, prototype haplotype data

Nucleotide Polymorphism

GENERAL
dbSNP Home Page
SNP Science Primer
NEW
Announcements
dbSNP Summary
FTP SERVER
Build History
Handle Request

DOCUMENTATION
FAQ
Overview
How To Submit
RefSNP Summary Info
Database Schema
html
pdf
Data formats
Heterozygosity computation

SEARCH
Entrez SNP NEW
Blast SNP
Batch query
By Submitter
New Batches
Method
Population
Detail
Class NEW
Publication
Chromosome Report
Locus Information
STS Markers
Free Form Search
Simple
Advanced

HAPLOTYPE NEW
Specifications
Sample HapSet
Sample Individual

dbSNP Search Options

ANNOUNCEMENT

NCBI has moved all FTP services to a n
ftp.ncbi.nih.gov. The full contents of d
the new address ftp://ftp.ncbi.nih.gov/snp
Please contact snp-admin@ncbi.nlm.nih.gov to report problems with access to the new ftp area.

Query quick links: announcement area

Search by IDs

Single record query: Accession, ID, or cluster

Note: rs# and ss# must be prefixed with "rs" or "ss", respectively (ie. rs25, ss25)

Advanced ID Search

Submission Information

By Submitter
New Batches
Method
Population
- Detail (Description, Handle, and ID)
- Class (Classification based on geographic location)

Publication
Chromosome Report

Submission property query: method, paper, submitter, latest data

Batch

Enter List
- NCBI Assay ID(ss)
- Reference SNP ID(rs)
- Local SNP ID

Upload List
- NCBI Assay ID(ss)
- Reference SNP ID(rs)
- Local SNP ID

Batch Query Help

Batch query: retrieve up to 20,000 records of interest at a time

Locus Information

Locus ID
Gene Name or Symbol
Gene Product
Accession Number
Gene Ontology
- Biological Process
- Cellular Location
- Molecular Function

Locus Query Help

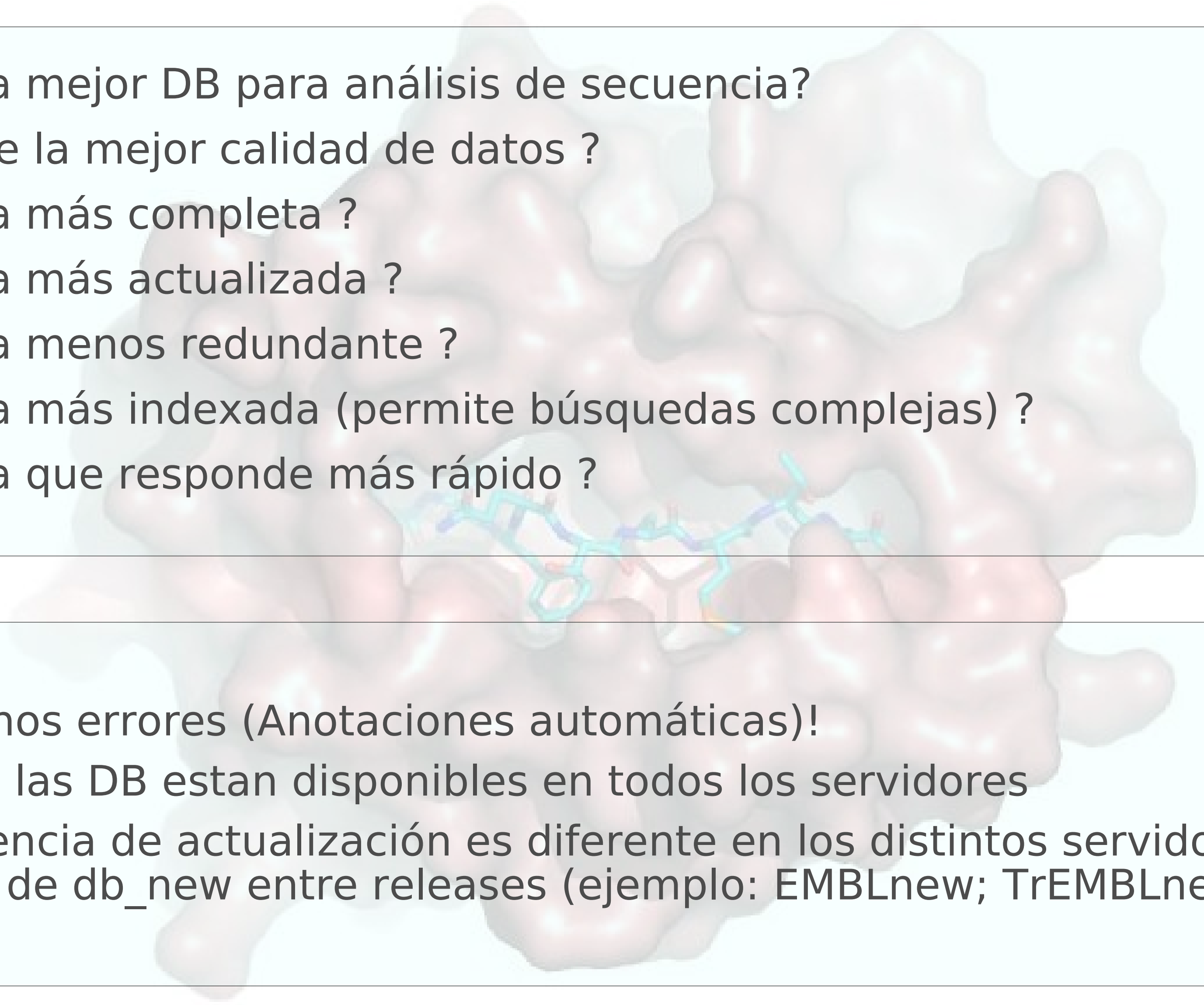
Locus query: retrieve lists of variations in known gene regions or mRNA transcripts

Free Form

- Use the pull-down menu to specify a search field.
- Enter a term in the text box or select from the option pull-down menu; Select an operator.
- Click 'Add' to add search field to the query box and 'Go' to view the results.

Free-form (Entrez-like) and Easy form queries: query the database using descriptor tags with boolean logic, or pick your choices from a set of pull down menus

Para tener en cuenta...



- Cual es la mejor DB para análisis de secuencia?
- Cual tiene la mejor calidad de datos ?
- Cual es la más completa ?
- Cual es la más actualizada ?
- Cual es la menos redundante ?
- Cual es la más indexada (permite búsquedas complejas) ?
- Cual es la que responde más rápido ?

- DB: Muchos errores (Anotaciones automáticas)!
- No todas las DB estan disponibles en todos los servidores
- La frecuencia de actualización es diferente en los distintos servidores; creación de db_new entre releases (ejemplo: EMBLnew; TrEMBLnew....)