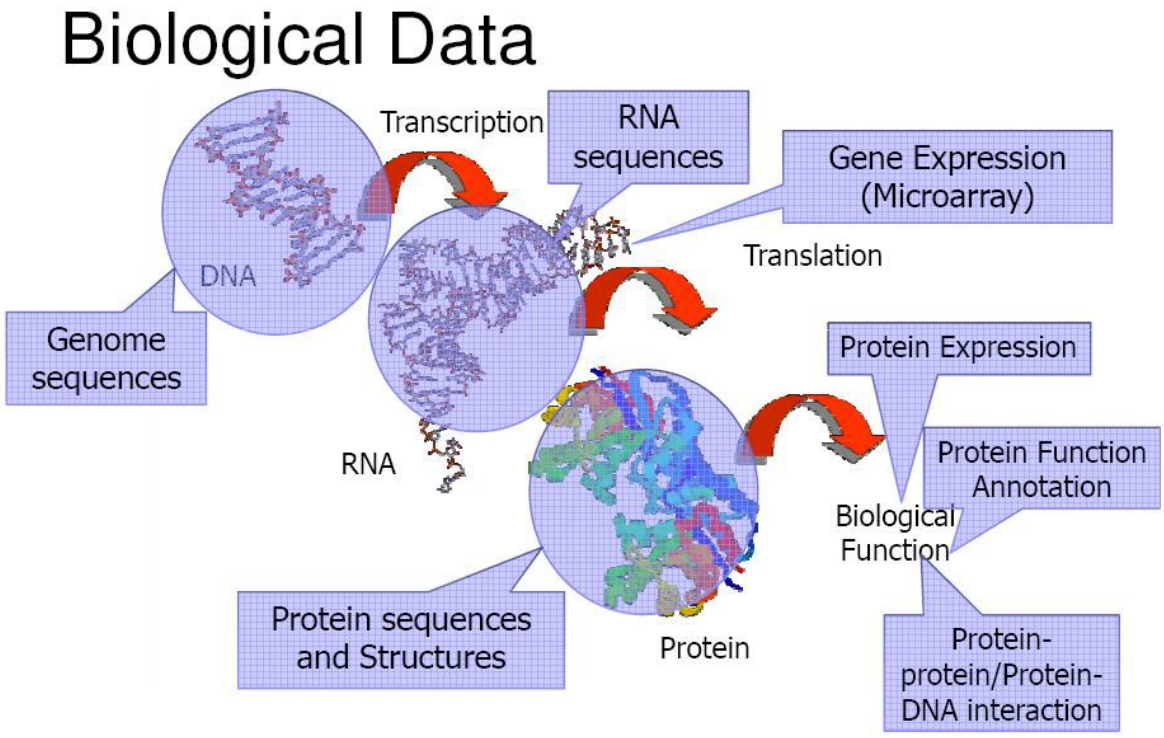
Module II

Importance of databases - Biological databases-primary sequence databases, Composite sequence databases- Secondary databases- nucleic acid sequence databases - Protein sequence data bases - structure databases, Types of databases, Data retrieval tools – Entrez



Importance of databases

- A database is any collection of related data.
- A Computerized archive used to store and organize data in such a way that information can be retrieved easily.
- A database is a collection of interrelated data store together without harmful and unnecessary redundancy (duplicate data) to serve multiple applications
- Retrieving is called by firing a query.
- Database System is an integrated collection of related files along with the detail about their definition, interpretation, manipulation and maintenance

- A database system is based on the data. Also a database system can be run or executed by using software called DBMS (Database Management System).
- A database system controls the data from unauthorized access.
- A database management system (DBMS) is a collection of programs that enables users to createand maintain a database.
- Database management systems provide severalfunctions in addition to simple file management: allow concurrency, control security, maintain data integrity, provide for backup and recovery, control redundancy, allow data independence, provide non-procedural query language, perform automatic query optimization
- relational database-a database that treats all of its data as a collection of relations

Need for databases in Biology

•Need for storing and communicating large datasets has grown.
•Need to disseminate biological information.
•Provide Organized data for analysis friendly retrieval.
•Need to make biological data available in computer-readable form.

Type of data

–nucleotide sequences

–protein sequences

–proteins sequence patterns or motifs

–macromolecular 3D structure

–gene expression data

–metabolic pathways

–proteomics data

Bioinformatics database categorized on the basis of

- Data type
- Maintainer status
- Technical design
- Data source
- Data access
- Any other parameter

Type of data:
genome database,
sequence database,

proteomic database,
structure database etc

Maintainer status:
NCBI-National center for Biotechnology Information
EBI- European Bioinformatics Institute

Technical Design:
Flat file ,XML, Relational Model, Object oriented/object relational model

Data source:
Primary database, secondary database

Data Access:
Various kinds of access status
-publicly available with no restrictions (NCBI,EBI)
- available with copyright

Others
Complete or incomplete entries in the database
Annotation- not annotated or annotated(have the analysis of data)
Curation- When annotation is established, db known as curated

Different classifications of databases….

Primary or derived databases

–Primary databases: experimental results directly into database

–Secondary databases: results of analysis of primary databases

–Aggregate of many databases

•Links to other data items

•Combination of data

•Consolidation of data

Biological databases play a central role in bioinformatics. They offer scientists the opportunity to access sequence and structure data for tens of thousands of sequences from a broad range of organisms.

Table 2 Essential aspects of primary and secondary databases.

	Primary database	**Secondary database**
Synonyms	Archival database	Curated database; knowledgebase

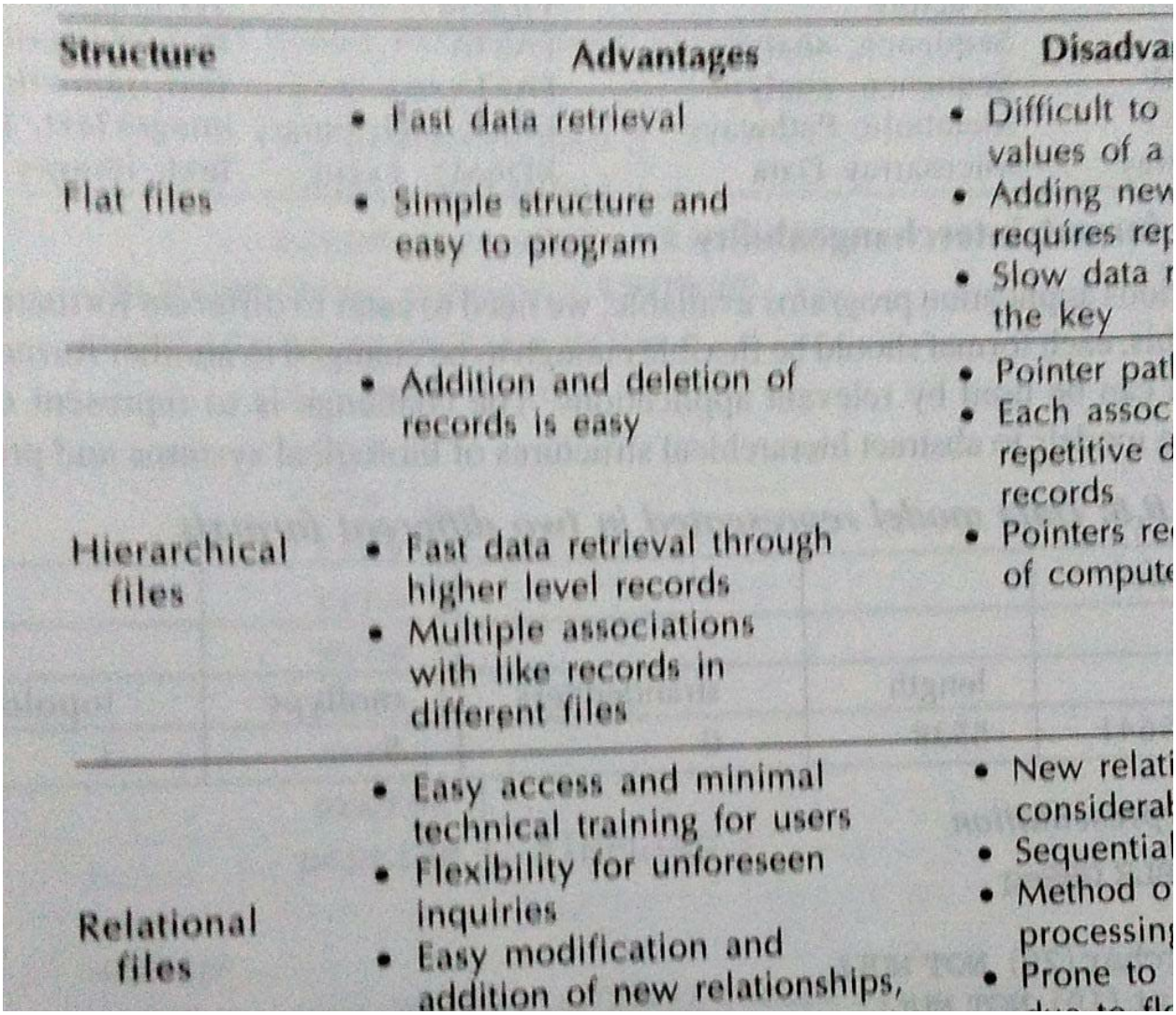
Source of data	Direct submission of experimentally-derived data from researchers	Results of analysis, literature research and interpretation, often of data in primary databases
Examples	• ENA, GenBank and DDBJ (nucleotide sequence) • ArrayExpress Archive and GEO (functional genomics data) • Protein Data Bank (PDB; coordinates of three-dimensional macromolecular structures)	• InterPro (protein families, motifs and domains) • UniProt Knowledgebase (sequence and functional information on proteins) • Ensembl (variation, function, regulation and more layered onto whole genome sequences)

The different types of databases

One may characterize the available biological databases by several different properties. Here is a list to help you think about the various properties a particular database may have.

- Type of data
  - nucleotide sequences
  - protein sequences
  - proteins sequence patterns or motifs
  - macromolecular 3D structure
  - gene expression data
  - metabolic pathways
- Data entry and quality control
  - Scientists (teams) deposit data directly
  - Appointed curators add and update data
  - Are erroneous data removed or marked?
  - Type and degree of error checking
  - Consistency, redundancy, conflicts, updates
- Primary or derived data
  - Primary databases: experimental results directly into database
  - Secondary databases: results of analysis of primary databases
  - Aggregate of many databases
    - Links to other data items
    - Combination of data
    - Consolidation of data
- Technical design
  - Flat-files
  - Relational database (SQL)

- Object-oriented database (e.g. CORBA, XML)



Structure	Advantages	Disadva
Flat files	• Fast data retrieval • Simple structure and easy to program	• Difficult to values of a • Adding nev requires rep • Slow data r the key
Hierarchical files	• Addition and deletion of records is easy • Fast data retrieval through higher level records • Multiple associations with like records in different files	• Pointer patl • Each assoc repetitive d records • Pointers re of compute
Relational files	• Easy access and minimal technical training for users • Flexibility for unforeseen inquiries • Easy modification and addition of new relationships,	• New relati consideral • Sequential • Method o processing • Prone to

- Maintainer status
  - Large, public institution (e.g. EMBL, NCBI)
  - Quasi-academic institute (e.g. Swiss Institute of Bioinformatics, TIGR)
  - Academic group or scientist
  - Commercial company

- Availability
  - Publicly available, no restrictions
  - Available, but with copyright
  - Accessible, but not downloadable
  - Academic, but not freely available
  - Proprietary, commercial; possibly free for academics.

All databases use a system where an entry can be identified and traced. The entry can be called by these names

- Identifier
- GI number
- Version number
- Accession code(number)
- Nucleic acid identifier

Identifier:

It is a unique integer which identifies a particular sequence.
This number will change every time the sequence changes.
The identifier is a string of letters and digits .
Identifier serves three main purposes

1. An identifier is assigned to all sequences processed. This number provides a unique sequence identifier , which is independent of the database source.
2. When a sequence is modified a new identifier is assigned to it. However its accession number remains unchanged.
3. The identifier is stable and retrievable.

GI Number:

It stands for GenInfo identifier. It is a series of digits that are assigned consecutively by NCBI to each sequence it processes.

It is not consistent across different databases.

- The nucleotide sequence GI number is shown in the VERSION field of the database record
- The protein sequence GI number is shown in the CDS/db-xref field of a nucleotide database record and the VERSION field of a protein database record.

Version numbers:

It consists of the accession number followed by a dot and a version number. It is consistent across databases.

- The nucleotide sequence version contains two letters followed by six digits, a dot and a version number.
- The protein sequence version contains three letters followed by five digits, a dot and a version number.
  Eg: GI:6995995, VERSION: NM_000492.2

Accession number:

It is the unique identifier assigned to the entire sequence record when the record is submitted to the GenBank.

The GenBank accession number is a combination of letters and numbers which are usually in the format of one letter followed by five digits (M12345) or two letters followed by six digits(AC123456).

The accession number for a particular record does not change even if the author submits a request to change some of the information in the record.

It is often called the primary key for the entry.

Nucleic acid identifier(NI) :

It is a number assigned to each version of an entry.

A new NI number is allocated each time the sequence is modified.

Nucleotide sequence databases

Primary nucleotide sequence databases

The databases EMBL, GenBank, and DDBJ are the **three primary nucleotide sequence databases**: They include sequences submitted directly by scientists and genome sequencing group, and sequences taken from literature and patents. There is comparatively little error checking and there is a fair amount of redundancy.

The entries in the EMBL, GenBank and DDBJ databases are **synchronized** on a daily basis, and the accession numbers are managed in a consistent manner between these three centers.

The nucleotide databases have reached such large sizes that they are available in **subdivisions** that allow searches or downloads that are more limited, and hence less time-consuming. For example, GenBank has currently 17 divisions.

There are **no legal restrictions** on the use of the data in these databases. However, there are some patented sequences in the databases.

EMBL

The EMBL (European Molecular Biology Laboratory) nucleotide sequence database is maintained by the European Bioinformatics Institute (EBI) in Hinxton, Cambridge, UK. Its size is given below, in total number of bases, and total number of records. Note its speed of increase since one year.

Date	# records	# bases

30 Oct 2001	13,771,247	14,745,640,065
16 Oct 2000	9,156,113	10,333,087,560

It can be accessed and searched through the SRS system at EBI, or one can download the entire database as flat files..

SRS-Sequence Retrieval System

GenBank

The GenBank nucleotide database is maintained by the National Center for Biotechnology Information (NCBI), which is part of the National Institute of Health (NIH), a federal agency of the US government.

It can be accessed and searched through the Entrez system at NCBI, or one can download the entire database as flat files.

DDBJ

The DNA Data Bank of Japan began as a collaboration with EMBL and GenBank. It is run by the National Institute of Genetics. One can search for entries by accession number, and little else.

Other nucleotide sequence databases

The following databases contain subsets of the EMBL/GenBank databases. Some also contain more information or links than the primary ones, or have a different organization of the data to better some specific purpose. However, the nucleotide sequences themselves should always be available in the EMBL/GenBank databases. In this sense, the databases below are secondary databases.

UniGene

The UniGene system attempts to process the GenBank sequence data into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location.

SGD

The Saccharomyces Genome Database (SGD) is a scientific database of the molecular biology .

EBI Genomes

This web site provides access and statistics for the completed genomes, and information about ongoing projects.

Genome Biology

The Genome Biology site at NCBI contains information about the available complete genomes.

Ensembl

Ensembl is a joint project between EMBL-EBI and the Sanger Centre to develop a software system which produces and maintains automatic annotation on eukaryotic genomes.

Protein sequence databases

The two protein sequence databases SWISS-PROT and PIR are different from the nucleotide databases in that they are both **curated**. This means that groups of designated curators (scientists) prepare the entries from literature and/or contacts with external experts.

SWISS-PROT, TrEMBL

SWISS-PROT is a protein sequence database which strives to provide a **high level of annotations** (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases.

It was started in 1986 by Amos Bairoch in the Department of Medical Biochemistry at the University of Geneva. This database is generally considered one of the best protein sequence databases in terms of the quality of the annotation. Its size is given in the table below.

TrEMBL is a **computer-annotated supplement** of SWISS-PROT that contains all the translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT. The procedure that is used to produce it was developed by Rolf Apweiler. The annotation of an entry in TrEMBL has not (yet) reached the standards required for inclusion into SWISS-PROT proper. Its size is given in the table below.

	SWISS-PROT	TrEMBL

Date	Release	# entries	Release	# entries
24 Oct 2001	40.1	101,737	18.0	484,388
2 Oct 2000	39.7	88,757	14.17	300,152

SWISS-PROT and TrEMBL are developed by the SWISS-PROT groups at Swiss Institute of Bioinformatics (SIB) and at EBI. The databases can be accessed and searched through the the SRS system at ExPASy, or one can download the entire database as one single flat file..

The SWISS-PROT database has **some legal restrictions**: the entries themselves are copyrighted, but freely accessible and usable by academic researchers. Commercial companies must buy a license fee from SIB.

Trembl

- Created in 1996 as a computer annotated supplement to SWISS-PROT.
- Contains translations of all coding sequences (CDS) in EMBL.

Has 2 main sections:

1.SP-TrEMBL: contains entries that will eventually be incorporated into SWISS-PROT, but that have not yet been manually annotated.

2. REM-TrEMBL: contains sequences that are not destined to be included in SWISS-PROT, these include immunoglobulins and T-cell receptors, synthetic and patented sequences and codon translations that do not encode real proteins.Computer-annotated supplement to SWISS-PROT, as it is impossible to cope with the flow of data…TrEMBL contains all what is **not yet** in SWISS-PROT.

PIR

The Protein Information Resource (PIR) is a division of the National Biomedical Research Foundation (NBRF) in the US. It is involved in a collaboration with the Munich Information Center for Protein Sequences (MIPS) and the Japanese International Protein Sequence Database (JIPID). The PIR-PSD (Protein Sequence Database) release 70.01 (22 Oct 2000) contains 254,293 entries.

PIR grew out of Margaret Dayhoff's work in the middle of the 1960s. It strives to be **comprehensive**, well-organized, accurate, and consistently annotated. However, it is generally believed that it does not reach the level of completeness in the entry annotation as does SWISS-PROT. Although SWISS-PROT and PIR overlap extensively, there are still many sequences which can be found in only one of them.

One can search for entries or do sequence similarity searches at the PIR site. The database can also be downloaded as a set of falt files..

PIR also produces the **NRL-3D**, which is a database of sequences extracted from the three-dimensional structures in the Protein Databank (PDB)

NRL_3D database makes the sequence information in PDB available for similarity searches and retrieval and provides cross-reference information for use with the other PIR Protein Sequence Databases.

It appears that the PIR web site, and possibly also the underlying database, has improved considerably since one year ago. This means that if one is interested in protein sequences, there is now even more reason to check out PIR; SWISS-PROT is not the only game in town

protein structure database

In biology, a **protein structure database** is a database that is modeled around the various experimentally determined protein structures.

The aim of most protein structure databases is to organize and annotate the protein structures, providing the biological community access to the experimental data in a useful way.

Data included in protein structure databases often includes three-dimensional coordinates as well as experimental information, such as unit cell dimensions and angles for x-ray crystallography determined structures.

Though most instances, in this case either proteins or a specific structure determinations of a protein, also contain sequence information and some databases even provide means for performing sequence based queries, the primary attribute of a structure database is structural information, whereas sequence databases focus on sequence information, and contain no structural information for the majority of entries.

Protein structure databases are critical for many efforts in computational biology such as structure based drug design, both in developing the computational methods used and in providing a large experimental dataset used by some methods to provide insights about the function of a protein.

The Protein Data Bank (PDB) was established in 1971 as the central archive of all experimentally determined protein structure data. Today the PDB is maintained by an international consortia collectively known as the Worldwide Protein Data Bank (wwPDB). The mission of the wwPDB is to maintain a single archive of macromolecular structural data that is freely and publicly available to the global community

List of other protein structure databases

Because the PDB releases data into the public domain, the data has been used in various other protein structure databases.

Examples of protein structure databases include (in alphabetical order);

Database of Macromolecular Movements

describes the motions that occur in proteins and other macromolecules, particularly using moves

Dynameomics

a data warehouse of molecular dynamics simulations and analyses of proteins representing all known protein fold families

JenaLib

the Jena Library of Biological Macromolecules is aimed at a better dissemination of information on three-dimensional biopolymer structures with an emphasis on visualization and analysis.

ModBase

a database of three-dimensional protein models calculated by comparative modeling

OCA

a browser-database for protein structure/function - The OCA integrates information from KEGG, OMIM, PDBselect, Pfam, PubMed, SCOP, SwissProt, and others.

OPM

provides spatial positions of protein three-dimensional structures with respect to the lipid bilayer.

PDB Lite

derived from OCA, PDB Lite was provided to make it as easy as possible to find and view a macromolecule within the PDB

PDBsum

provides an overview macromolecular structures in the PDB, giving schematic diagrams of the molecules in each structure and of the interactions between them

PDBTM

the Protein Data Bank of Transmembrane Proteins — a selection of the PDB.

PDBWiki

a community annotated knowledge base of biological molecular structures

ProtCID

The Protein Common Interface Database (ProtCID) is a database of similar protein–protein interfaces in crystal structures of homologous proteins.

Protein

the NIH protein database, a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and Third Party Annotation, as well as records from SwissProt, PIR, PRF, and PDB

Proteopedia

the collaborative, 3D encyclopedia of proteins and other molecules. A wiki that contains a page for every entry in the PDB (>100,000 pages), with a Jmol view that highlights functional sites and ligands. Offers an easy-to-use scene-authoring tool so you don't have to learn Jmol script language to create customized molecular scenes. Custom scenes are easily attached to "green links" in descriptive text that display those scenes in Jmol.

ProteinLounge

a protein databases that includes visuals of protein structure. Also, includes protein pathways and gene sequences including other tools.

SCOP

the Structural Classification of Proteins a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known.

SWISS-MODEL Repository

a database of annotated protein models calculated by homology modeling

TOPSAN

the Open Protein Structure Annotation Network — a wiki designed to collect, share and distribute information about protein three-dimensional structures.

Data retrieval Tools

- Once the amount of biological relevant data is increasing so rapidly, knowing how to access and search this information is essential. There are three data retrieval systems of particular relevance to molecular biologist: Sequence Retrieval System (SRS), Entrez, DBGET. These systems allow text searching of multiple molecular biology database and provide links to relevant information for entries that match the search criteria.
- The three systems differ in the databases they search and the links they have to other information.

Sequence Retrieval System (SRS) .

SRS is a homogeneous interface to over 80 biological databases that had been developed at the European Bioinformatics Institute (EBI) at Hinxton, UK

- It includes databases of sequences, metabolic pathways, transcription factors, application results (like BLAST, SSEARCH, FASTA), protein 3-D structures, genomes, mappings, mutations, and locus specific mutations.
- The web page listing all the databases contains a link to a description page about the database including the date on which it was last updated. You select one or more of the databases to search before entering your query. After getting results you choose an alignment algorithm (like CLUSTALW, PHYLIP) enter parameters, and run it.
- The SRS is highly recommended for use.

Entrez

- Entrez is a molecular biology database and retrieval system. Developed by

the National Center for Biotechnology information (NCBI). It is entry point for exploring distinct but integrated databases.

- Of the three text-based database systems, Entrez is the easiest to use,but also more limited information to search.

DBGET

- DBGET is an integrated database retrieval system, developed at the university of Tokyo Provided access to 20 databases, one at a time. Having more limited options, the DBGET is less recommended than the two others.

Entrez

The Entrez logo

The **Entrez** Global Query Cross-Database Search System is a federated search engine, or web portal that allows users to search many discrete health sciencesdatabases at the National Center for Biotechnology Information (NCBI) website.

The NCBI is a part of the National Library of Medicine (NLM), which is itself a department of the National Institutes of Health (NIH), which in turn is a part of the United States Department of Health and Human Services. The name "Entrez" (a greeting meaning "Come in!" in French) was chosen to reflect the spirit of welcoming the public to search the content available from the NLM.

Entrez Global Query is an integrated search and retrieval system that provides access to all databases simultaneously with a single query string and user interface. Entrez can efficiently retrieve related sequences, structures, and references. The Entrez system can provide views of gene and protein sequences and chromosome maps. Some textbooks are also available online through the Entrez system.

The Entrez front page provides, by default, access to the global query. All databases indexed by Entrez can be searched via a single query string, supporting boolean operators and search term tags to limit parts of the search statement to particular fields. This returns a unified results page, that shows the number of hits for the search in each of the databases, which are also links to actual search results for that particular database.

Entrez also provides a similar interface for searching each particular database and for refining search results. The Limits feature allows the user to narrow a search a web forms interface. The History feature gives a numbered list of recently performed queries. Results of previous queries can be referred to by number and combined via boolean operators. Search results can be saved temporarily in a Clipboard. Users with a MyNCBI account can save queries indefinitely and also choose to have updates with new search results e-mailed for saved queries of most databases. It is widely used in the field of biotechnology as a reference tool for students and professionals alike

Databases

Entrez searches the following databases:

- PubMed: biomedical literature citations and abstracts, including Medline - articles from (mainly medical) journals, often including abstracts. Links to PubMed Central and other full-text resources are provided for articles from the 1990s.
- PubMed Central: free, full-text journal articles
- Site Search: NCBI web and FTP web sites
- Books: online books
- Online Mendelian Inheritance in Man (OMIM)
- *Nucleotide*: sequence database (GenBank)
- *Protein*: sequence database
- *Genome*: whole genome sequences and mapping
- *Structure*: three-dimensional macromolecular structures
- *Taxonomy*: organisms in GenBank Taxonomy

- *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
- *PopSet*: population study data sets (epidemiology)
- *GEO Profiles*: expression and molecular abundance profiles
- *GEO DataSets*: experimental sets of GEO data
- *Sequence read archive*: high-throughput sequencing data
- *Cancer Chromosomes*: cytogenetic databases
- *PubChem BioAssay*: bioactivity screens of chemical substances
- *Probe*: sequence-specific reagents
- *NLM Catalog*: NLM bibliographic data for over 1.2 million journals, books, audiovisuals, computer software, electronic resources, and other materials resident in LocatorPlus (updated every weekday).