Lecture 2:
Biological Databases

NCBI DataBases

Database

- A **database** is an organized collection of data .
- The data is typically organized to model relevant aspects of reality (for example, the availability of rooms in hotels), in a way that supports processes requiring this information (for example, finding a hotel with vacancies).

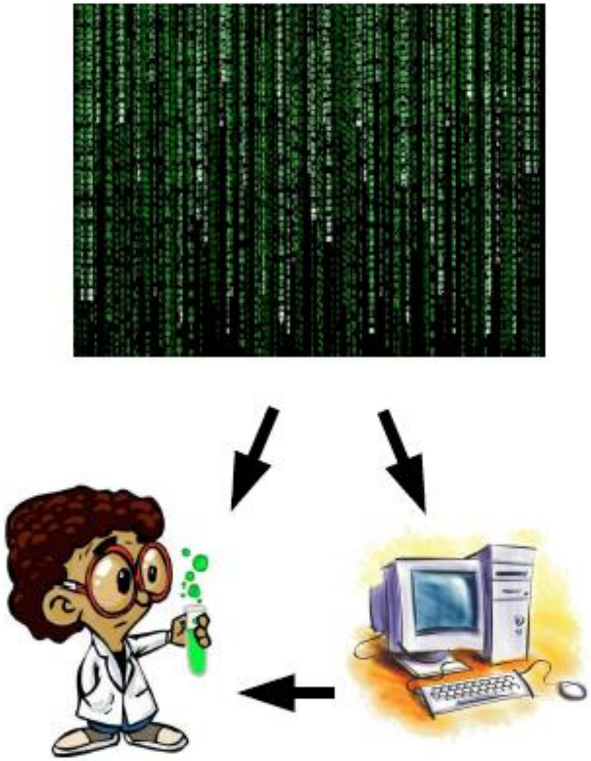
Biological databases

- **Make biological data available** ...

  1. … **to scientists**.

  2. … **in computer-readable form**.

     - Analysis (computer based)
     - Handle and share large volumes of data
     - Interface for computer based systems (Algorithms, Web interfaces)



- Store data
  - Defined formats
  - Automated storage and retrieval of experimental data
- Link knowledge with external resources

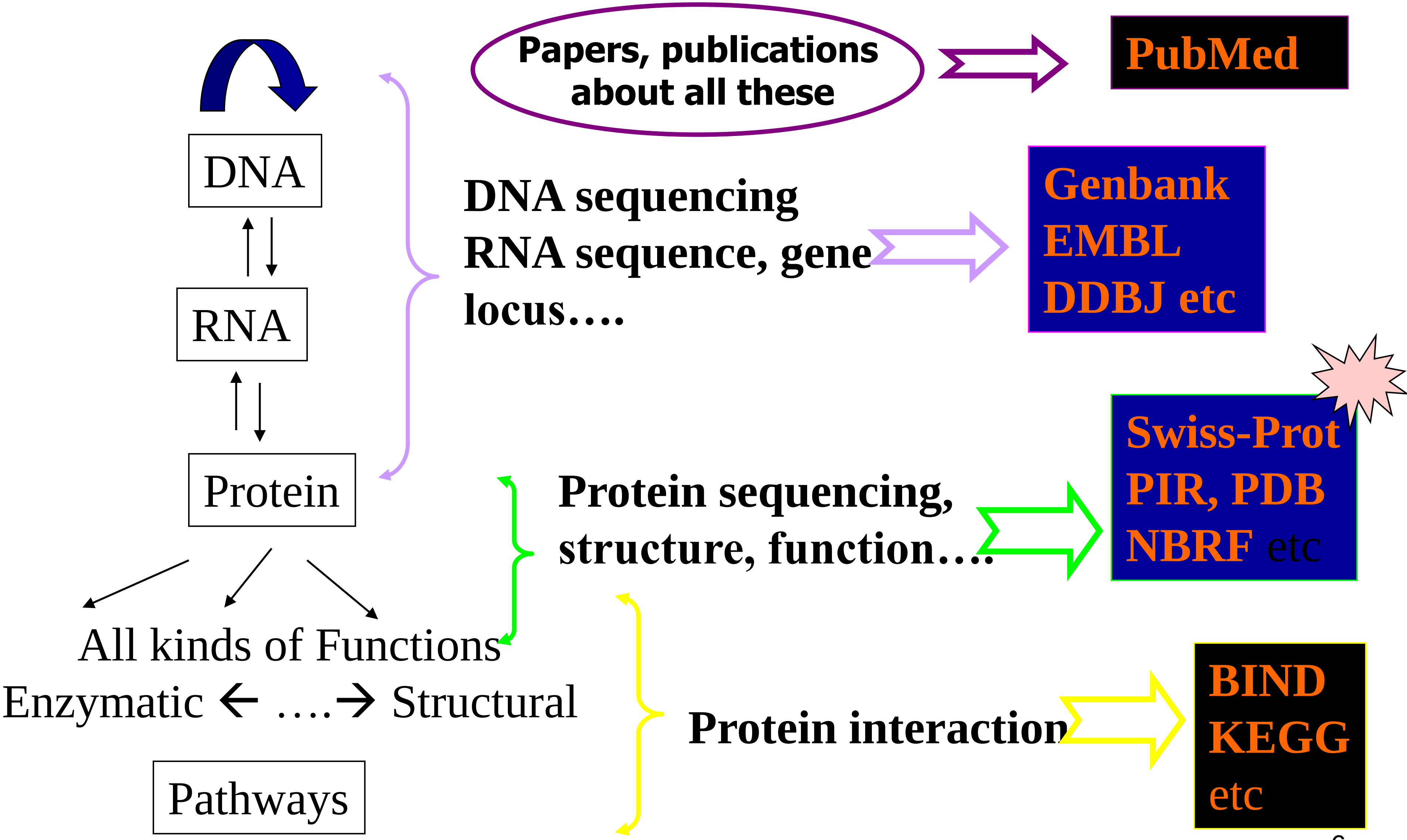
- **A database can be thought of as a large table, where the rows represent records and the columns represent fields.**

Field Record	Name	Length	Sequence	Enzyme
QA001	MTGA	243	MYQWI…	yes
QA002	Ribosomal protein L9	267	MAAPV…	no
QA003	Flagellin	374	GSSIL…	no
QA004	GDPMH	157	MFLRQ…	yes

The “perfect” database

1. Comprehensive, but easy to search.
2. Annotated, but not “too annotated”.
3. A simple, easy to understand structure.
4. Cross-referenced.
5. Minimum redundancy.
6. Easy retrieval of data.

Bioinformatics Databases



Papers, publications about all these

PubMed

DNA

RNA

Protein

DNA sequencing
RNA sequence, gene
locus….

Genbank
EMBL
DDBJ etc

Protein sequencing,
structure, function….

Swiss-Prot
PIR, PDB
NBRF etc

All kinds of Functions
Enzymatic ← ….→ Structural

Pathways

Protein interaction

BIND
KEGG
etc

生物信息学数据库简介

Part 1 outline:

1. Biological information and databases

- Overview and definition, types of biological databases

2. Popular databases, records, data format

- PubMed, Genbank, SwissProt, OMIM, PDB, KEGG, BIND, Pfam, PROSITE,

3. Accessing biological databases, retrieval systems

- Entrez, SRS

4. Searching biological databases

- Data quality, coverage, redundancy, errors

Biological Information

Nucleic acids:

- **DNA sequence, genes, gene products (proteins), mutation, gene coding, distribution patterns, motifs**
- **Genomics: genome, gene structure and expression, genetic map, genetic disorder**
- **RNA sequence, secondary structure, 3D structure, interactions**

Proteins:

- **Protein sequence, corresponding gene, secondary structure, 3D structure, function, motifs, homology, interactions**
- **Proteomics: expression profile, proteins in disease processes etc.**
- **Ligands and drugs (inhibitors, activators, substrates, metabolites)**

Biological Information

Pathways:

- Molecular networks, biological chain events, regulation, feedback, kinetic data

Function:

- Binding sites, interactions, molecular action (binding, chemical reaction, etc.)
- Biological effect (signaling, transport, feedback, regulation, modification, etc.)
- Functional relationship, protein families, motifs, and homologs

Biological databases

Lists of biological databases

- INFOBIOGEN Catalog of Databases
  http://www.infobiogen.fr/services/dbcat/

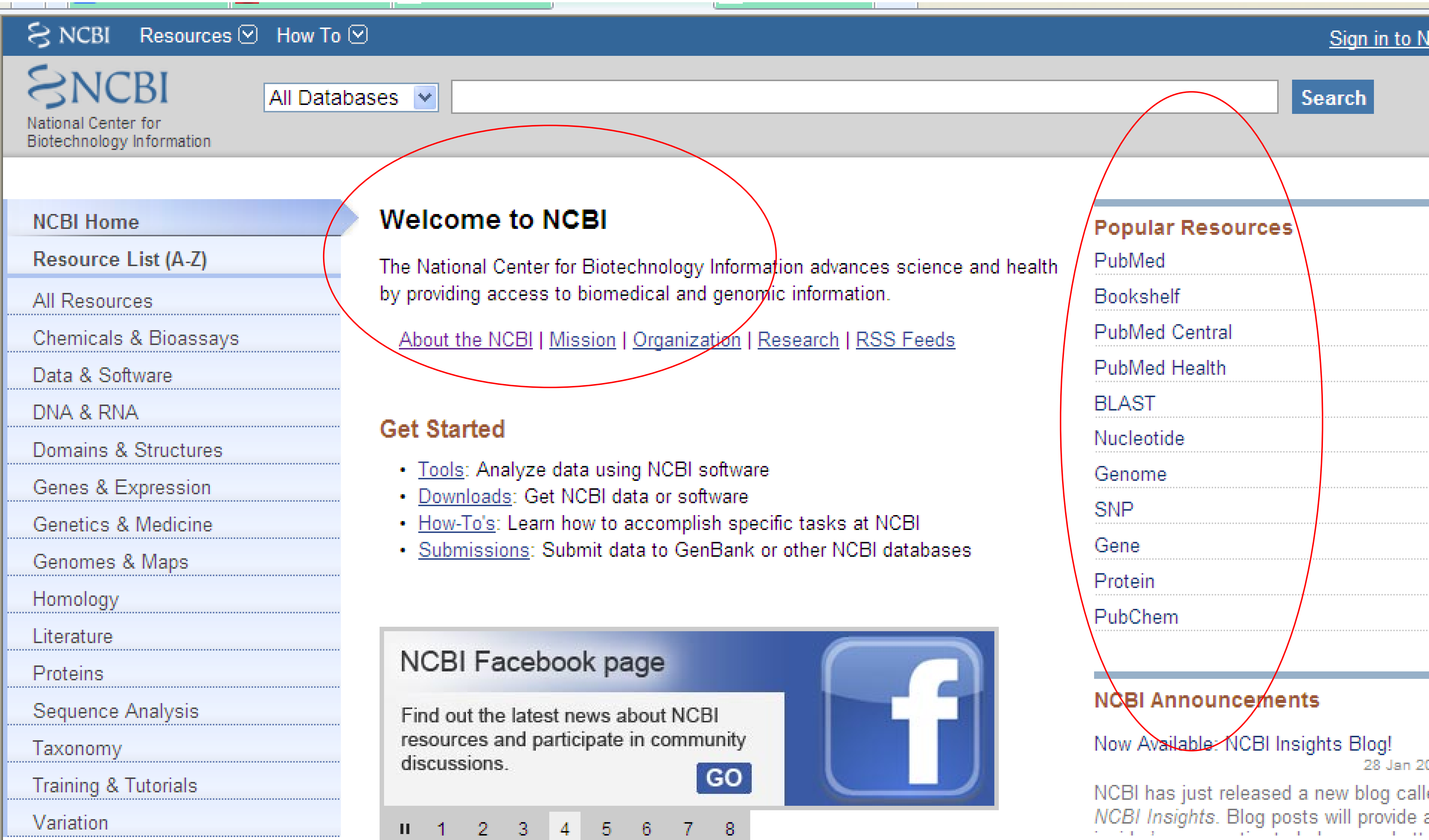
- Nucleic Acids Research Database Listing
  http://www3.oup.co.uk/nar/database/c/

  - These serve as starting point of biological databases.
  - More than 500 databases have been catalogued to date and those from the two listings satisfy minimal criteria for the content, access, and quality.
  - Other sites as a starting point.

Biological Database List

- **Database Categories List (**)
- Nucleotide Sequence Databases
- RNA sequence databases
- Protein sequence databases
- Structure Databases
- Genomics Databases (non-vertebrate)
- Metabolic and Signaling Pathways
- Human and other Vertebrate Genomes
- Human Genes and Diseases
- Micro－array Data and other Gene Expression Databases
- Proteomics Resources
- Other Molecular Biology Databases
- Organelle databases
- Plant databases
- Immunological databases

生物信息资源简介: NCBI



Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

NCBI Facebook page

Find out the latest news about NCBI resources and participate in community discussions.

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI Announcements

Now Available: NCBI Insights Blog!

28 Jan 20

NCBI has just released a new blog call *NCBI Insights*. Blog posts will provide a

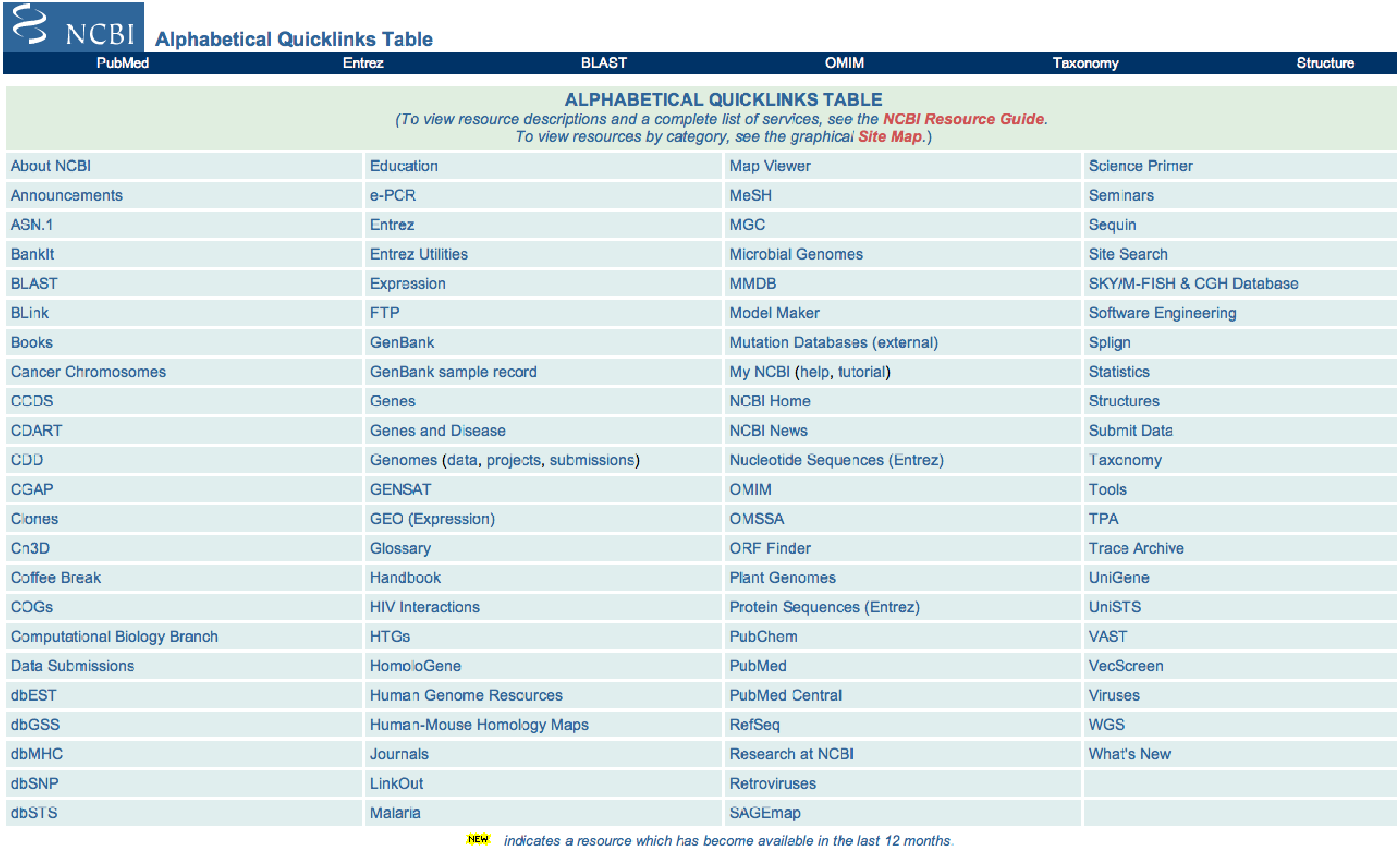
National Center for Biotechnology Information

- Established in 1988
- Creates public databases,
- Conducts research in computational biology,
- Develops software tools for analyzing genome data,
- Disseminates biomedical information.

➢ All for the better understanding of molecular processes affecting human health and disease.

➢Advances science and health by providing access to biomedical and genomic information.

http://www.ncbi.nlm.nih.gov/Sitemap/AlphaList.html



NCBI Alphabetical Quicklinks Table

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

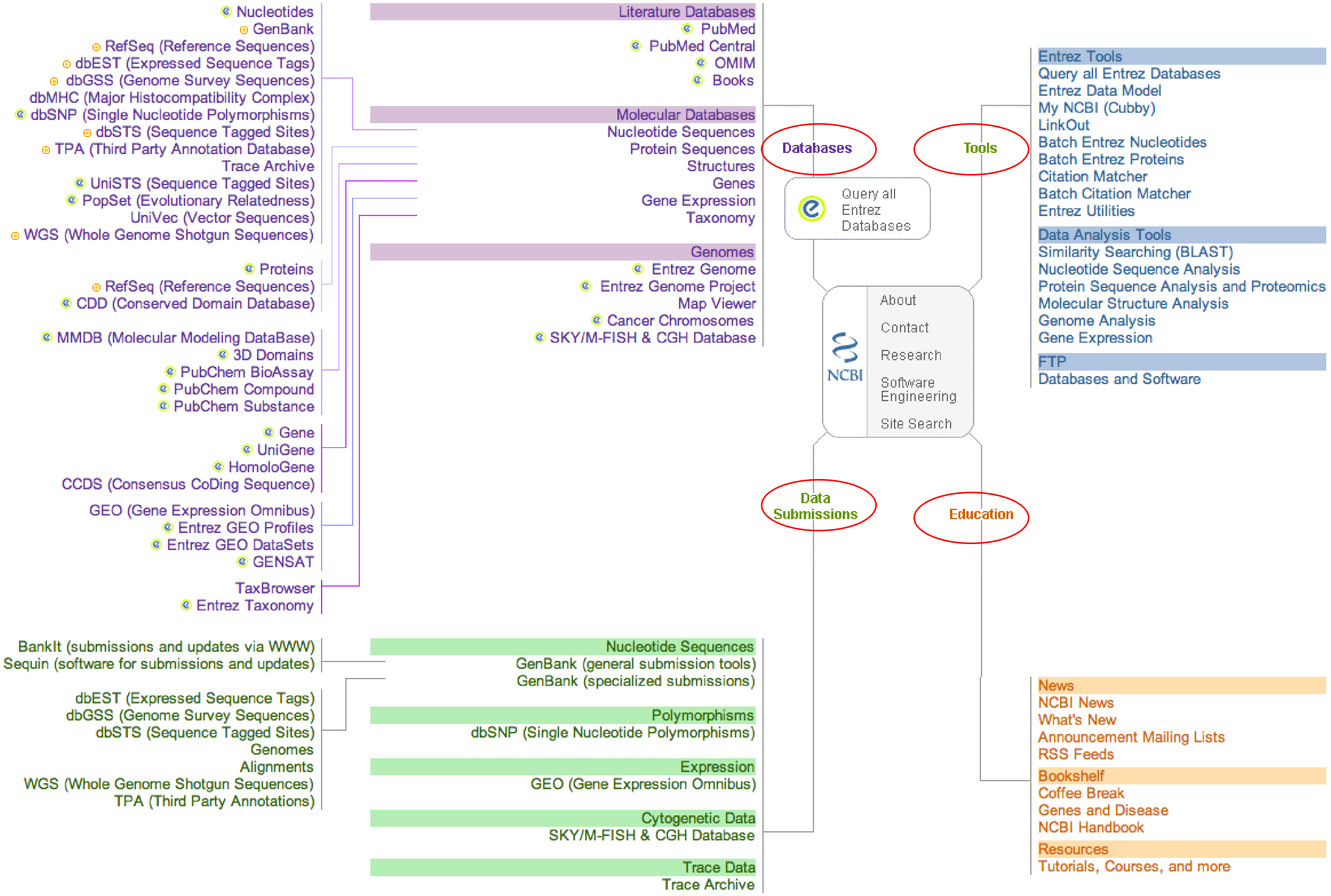
ALPHABETICAL QUICKLINKS TABLE

*(To view resource descriptions and a complete list of services, see the **NCBI Resource Guide**.
To view resources by category, see the graphical **Site Map**.)*

- About NCBI
- Announcements
- ASN.1
- BankIt
- BLAST
- BLink
- Books
- Cancer Chromosomes
- CCDS
- CDART
- CDD
- CGAP
- Clones
- Cn3D
- Coffee Break
- COGs
- Computational Biology Branch
- Data Submissions
- dbEST
- dbGSS
- dbMHC
- dbSNP
- dbSTS
- Education
- e-PCR
- Entrez
- Entrez Utilities
- Expression
- FTP
- GenBank
- GenBank sample record
- Genes
- Genes and Disease
- Genomes (data, projects, submissions)
- GENSAT
- GEO (Expression)
- Glossary
- Handbook
- HIV Interactions
- HTGs
- HomoloGene
- Human Genome Resources
- Human-Mouse Homology Maps
- Journals
- LinkOut
- Malaria
- Map Viewer
- MeSH
- MGC
- Microbial Genomes
- MMDB
- Model Maker
- Mutation Databases (external)
- My NCBI (help, tutorial)
- NCBI Home
- NCBI News
- Nucleotide Sequences (Entrez)
- OMIM
- OMSSA
- ORF Finder
- Plant Genomes
- Protein Sequences (Entrez)
- PubChem
- PubMed
- PubMed Central
- RefSeq
- Research at NCBI
- Retroviruses
- SAGEmap
- Science Primer
- Seminars
- Sequin
- Site Search
- SKY/M-FISH & CGH Database
- Software Engineering
- Splign
- Statistics
- Structures
- Submit Data
- Taxonomy
- Tools
- TPA
- Trace Archive
- UniGene
- UniSTS
- VAST
- VecScreen
- Viruses
- WGS
- What's New

NEW *indicates a resource which has become available in the last 12 months.*

http://www.ncbi.nlm.nih.gov/Sitemap/index.html



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)

Proteins
RefSeq (Reference Sequences)
CDD (Conserved Domain Database)

MMDB (Molecular Modeling DataBase)
3D Domains
PubChem BioAssay
PubChem Compound
PubChem Substance

Gene
UniGene
HomoloGene
CCDS (Consensus CoDing Sequence)

GEO (Gene Expression Omnibus)
Entrez GEO Profiles
Entrez GEO DataSets
GENSAT

TaxBrowser
Entrez Taxonomy

Literature Databases
PubMed
PubMed Central
OMIM
Books

Molecular Databases
Nucleotide Sequences
Protein Sequences
Structures
Genes
Gene Expression
Taxonomy

Genomes
Entrez Genome
Entrez Genome Project
Map Viewer
Cancer Chromosomes
SKY/M-FISH & CGH Database

Databases

Tools

Query all Entrez Databases

NCBI

About
Contact
Research
Software Engineering
Site Search

Data Submissions

Education

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

BankIt (submissions and updates via WWW)
Sequin (software for submissions and updates)

dbEST (Expressed Sequence Tags)
dbGSS (Genome Survey Sequences)
dbSTS (Sequence Tagged Sites)
Genomes
Alignments
WGS (Whole Genome Shotgun Sequences)
TPA (Third Party Annotations)

Nucleotide Sequences
GenBank (general submission tools)
GenBank (specialized submissions)

Polymorphisms
dbSNP (Single Nucleotide Polymorphisms)

Expression
GEO (Gene Expression Omnibus)

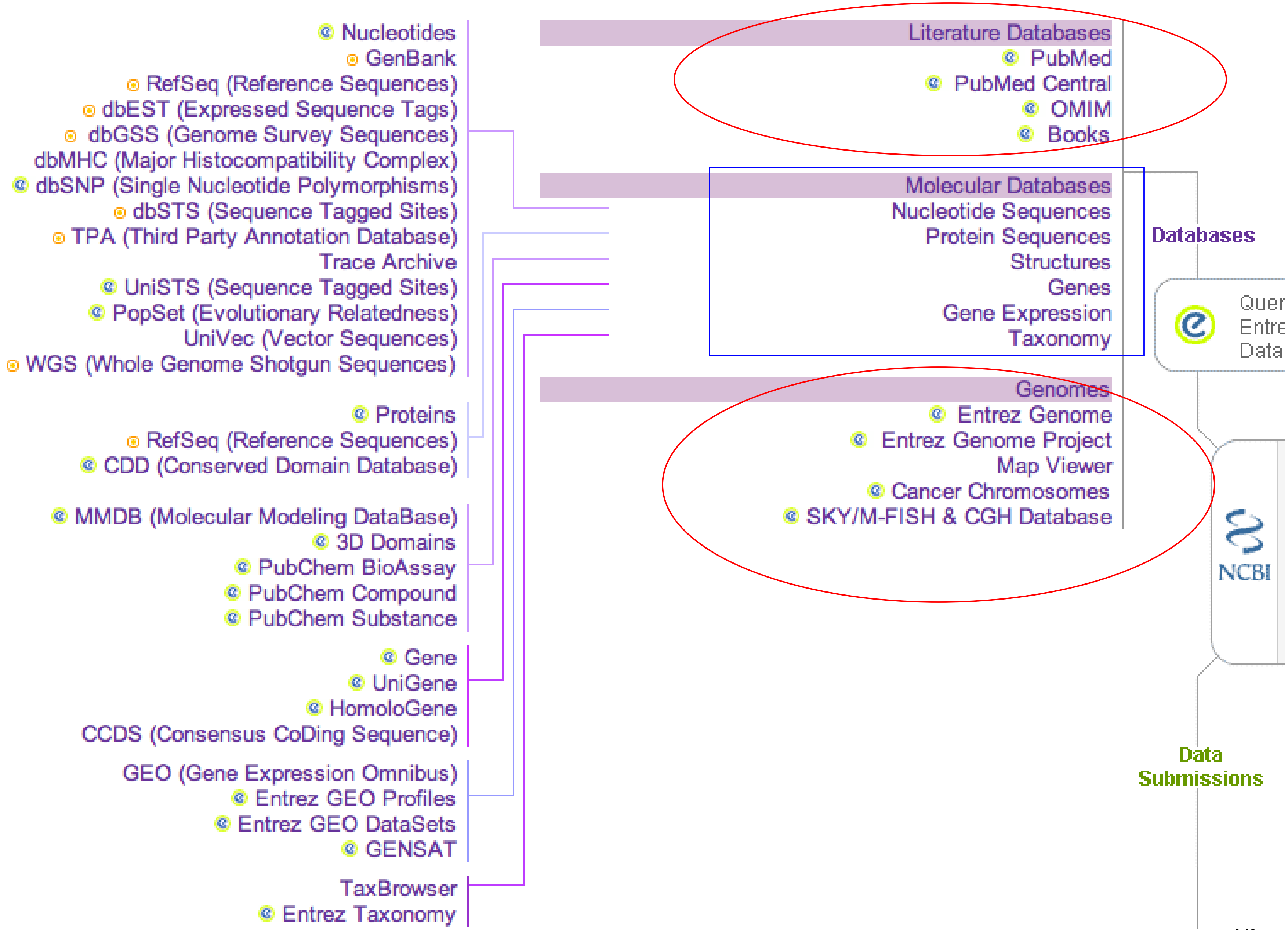
Cytogenetic Data
SKY/M-FISH & CGH Database

Trace Data
Trace Archive

News
NCBI News
What's New
Announcement Mailing Lists
RSS Feeds

Bookshelf
Coffee Break
Genes and Disease
NCBI Handbook

Resources
Tutorials, Courses, and more



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)

Proteins
RefSeq (Reference Sequences)
CDD (Conserved Domain Database)

MMDB (Molecular Modeling DataBase)
3D Domains
PubChem BioAssay
PubChem Compound
PubChem Substance

Gene
UniGene
HomoloGene
CCDS (Consensus CoDing Sequence)

GEO (Gene Expression Omnibus)
Entrez GEO Profiles
Entrez GEO DataSets
GENSAT

TaxBrowser
Entrez Taxonomy

Literature Databases
PubMed
PubMed Central
OMIM
Books

Molecular Databases
Nucleotide Sequences
Protein Sequences
Structures
Genes
Gene Expression
Taxonomy

Genomes
Entrez Genome
Entrez Genome Project
Map Viewer
Cancer Chromosomes
SKY/M-FISH & CGH Database

Databases

Quer
Entre
Data

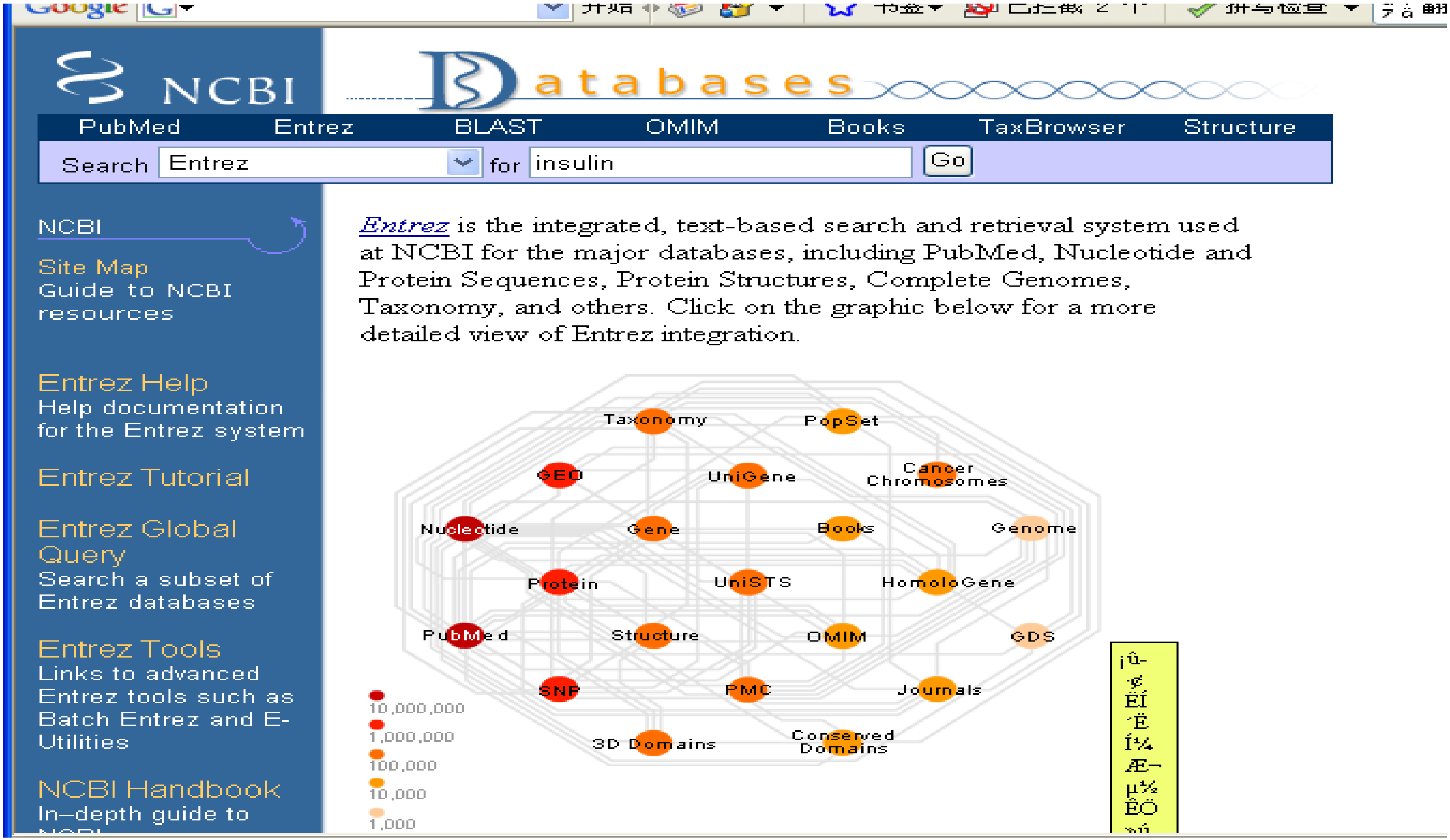
NCBI

Data Submissions

这么多数据库如何组织在一起?

- **Entrez** - provides integrated access to nucleotide and protein sequence data from differnt organisms, along with 3D protein structures, genomic mapping information, PubMed MEDLINE, and more.

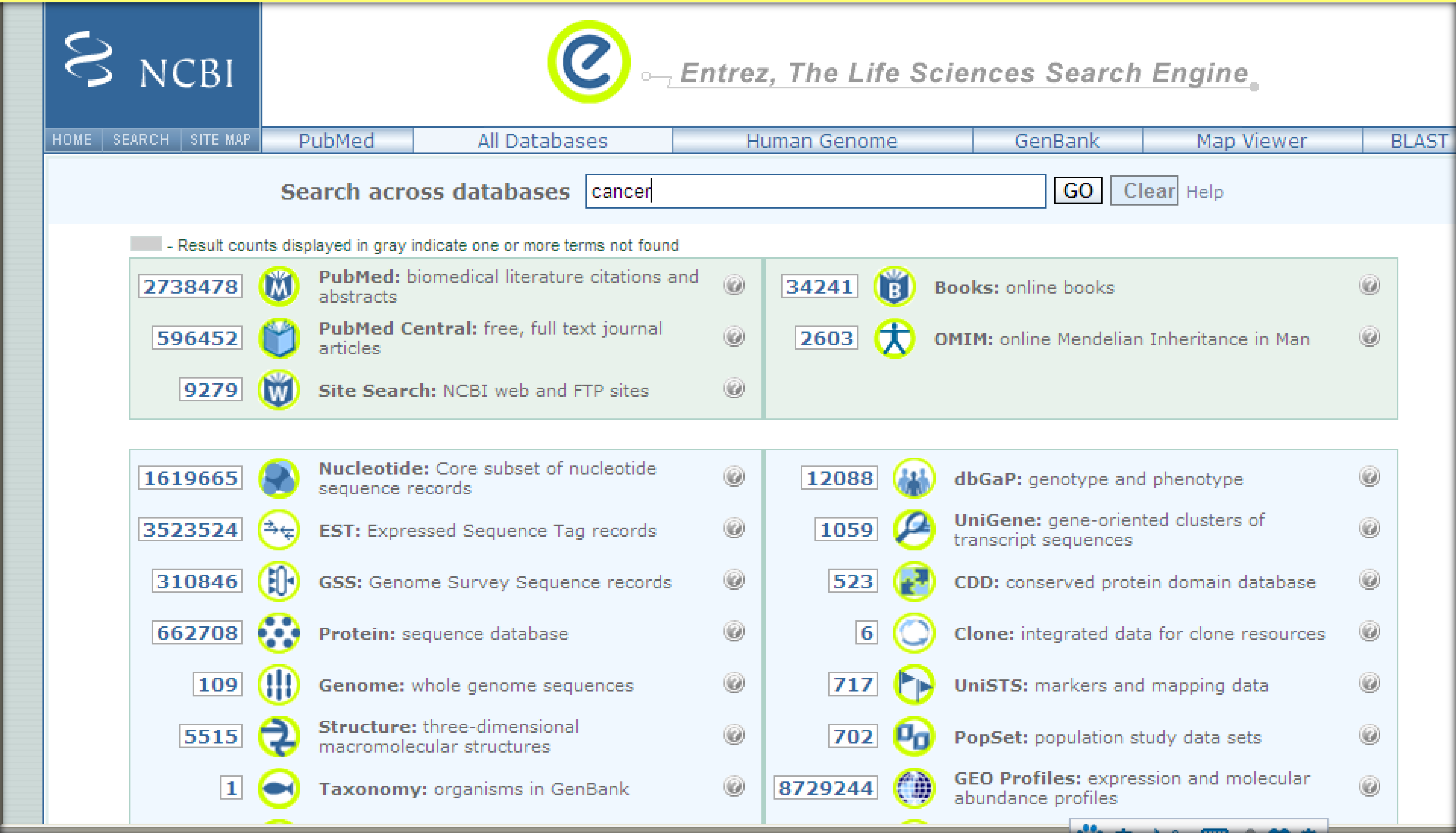
ENTREZ



Search Entrez for insulin

Entrez is the integrated, text-based search and retrieval system used at NCBI for the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, and others. Click on the graphic below for a more detailed view of Entrez integration.

ENTREZ results



Search across databases: cancer

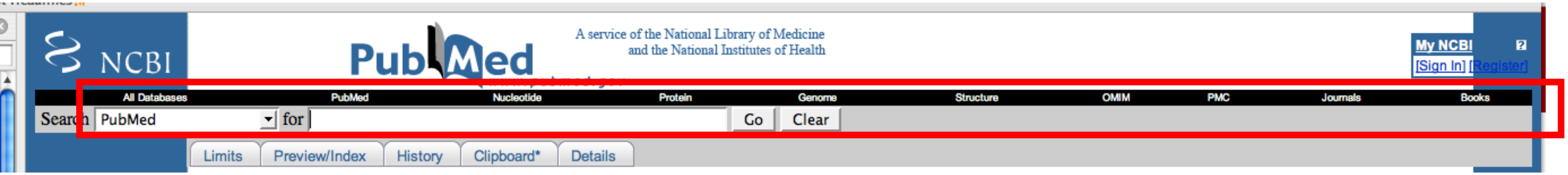
▨ - Result counts displayed in gray indicate one or more terms not found

2738478	**PubMed:** biomedical literature citations and abstracts
596452	**PubMed Central:** free, full text journal articles
9279	**Site Search:** NCBI web and FTP sites
34241	**Books:** online books
2603	**OMIM:** online Mendelian Inheritance in Man

1619665	**Nucleotide:** Core subset of nucleotide sequence records
3523524	**EST:** Expressed Sequence Tag records
310846	**GSS:** Genome Survey Sequence records
662708	**Protein:** sequence database
109	**Genome:** whole genome sequences
5515	**Structure:** three-dimensional macromolecular structures
1	**Taxonomy:** organisms in GenBank
12088	**dbGaP:** genotype and phenotype
1059	**UniGene:** gene-oriented clusters of transcript sequences
523	**CDD:** conserved protein domain database
6	**Clone:** integrated data for clone resources
717	**UniSTS:** markers and mapping data
702	**PopSet:** population study data sets
8729244	**GEO Profiles:** expression and molecular abundance profiles

1. PubMed

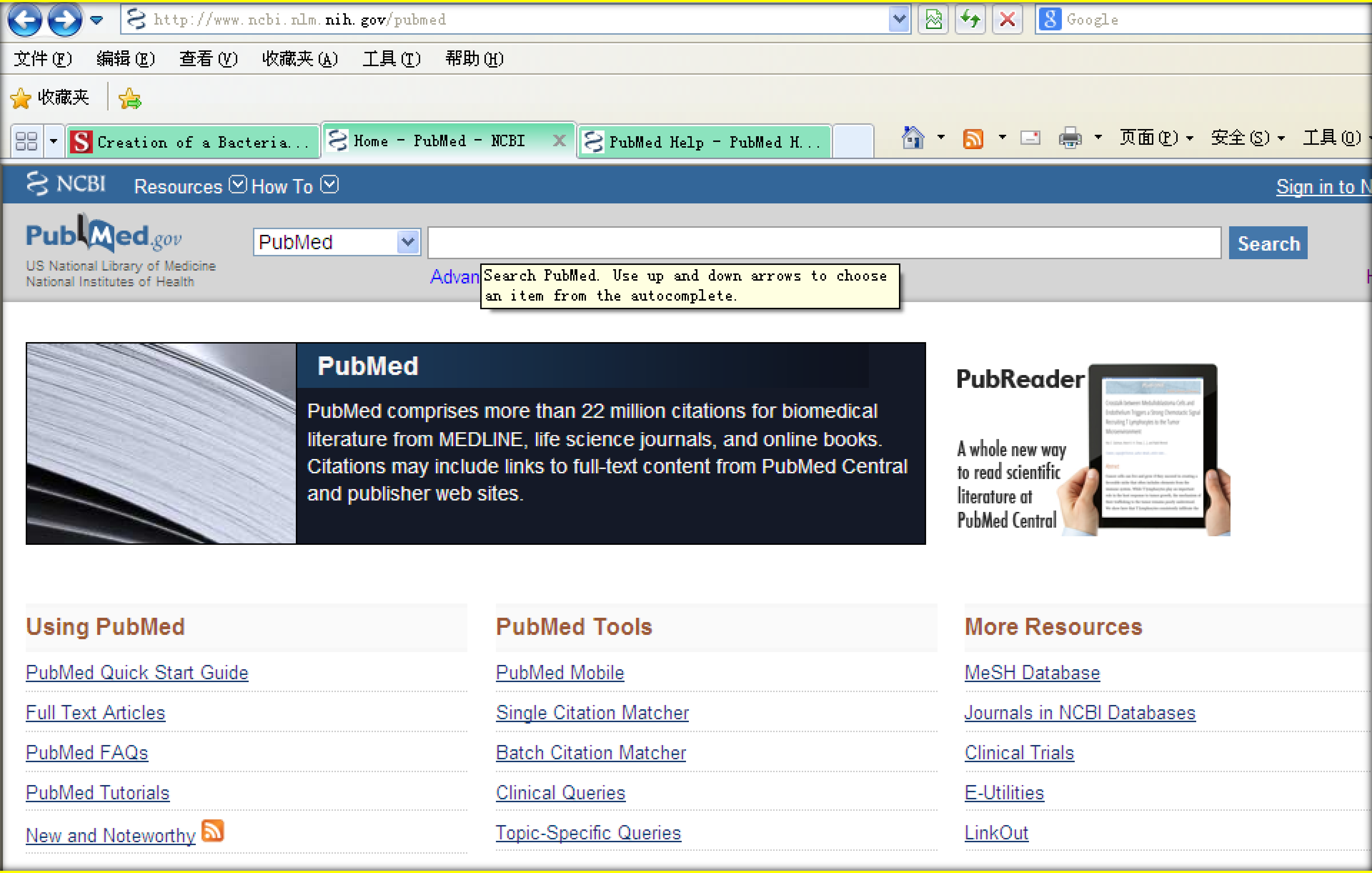
PubMed is one of the literature databases in the NCBI family.



In the NCBI family with Nucleotide, Protein, Genome, Structure, OMIM, PMC, Journals, Books and more

1. includes millions of citations from MEDLINE and other life science journals for biomedical articles back to the 1950s.
2. **PubMed** includes links to full text articles and other related resources.

Search terms may be topics, authors or journals.



http://www.ncbi.nlm.nih.gov/pubmed
Google
文件(F) 编辑(E) 查看(V) 收藏夹(A) 工具(T) 帮助(H)
收藏夹
Creation of a Bacteria...
Home - PubMed - NCBI
PubMed Help - PubMed H...
页面(P) 安全(S) 工具(O)
NCBI Resources How To
Sign in to N
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
Search
Advan

Search PubMed. Use up and down arrows to choose an item from the autocomplete.

PubMed

PubMed comprises more than 22 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

PubReader
A whole new way to read scientific literature at PubMed Central

Using PubMed

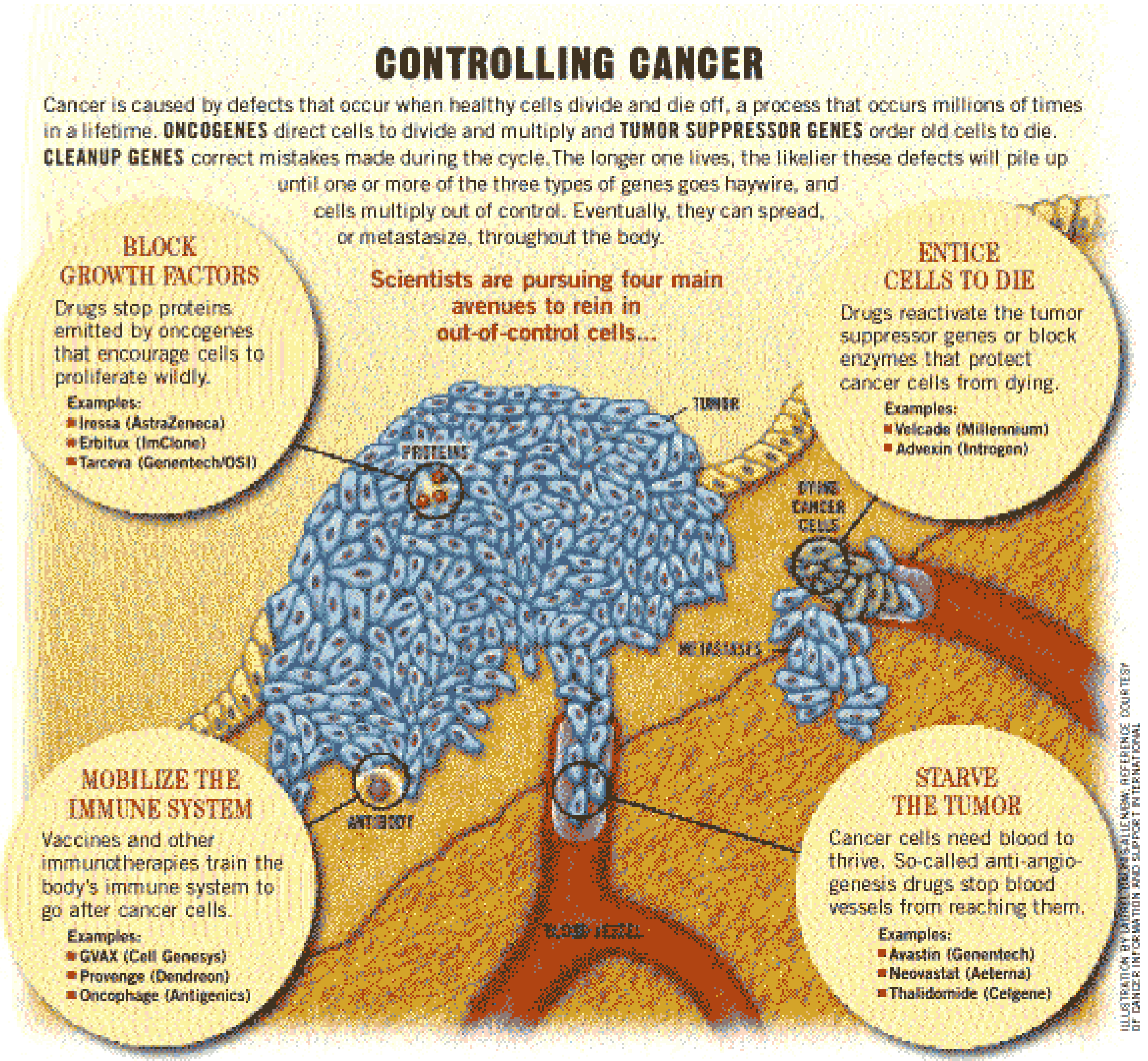
- PubMed Quick Start Guide
- Full Text Articles
- PubMed FAQs
- PubMed Tutorials
- New and Noteworthy

PubMed Tools

- PubMed Mobile
- Single Citation Matcher
- Batch Citation Matcher
- Clinical Queries
- Topic-Specific Queries

More Resources

- MeSH Database
- Journals in NCBI Databases
- Clinical Trials
- E-Utilities
- LinkOut



CONTROLLING CANCER

Cancer is caused by defects that occur when healthy cells divide and die off, a process that occurs millions of times in a lifetime. **ONCOGENES** direct cells to divide and multiply and **TUMOR SUPPRESSOR GENES** order old cells to die. **CLEANUP GENES** correct mistakes made during the cycle. The longer one lives, the likelier these defects will pile up until one or more of the three types of genes goes haywire, and cells multiply out of control. Eventually, they can spread, or metastasize, throughout the body.

Scientists are pursuing four main avenues to rein in out-of-control cells...

BLOCK GROWTH FACTORS

Drugs stop proteins emitted by oncogenes that encourage cells to proliferate wildly.

Examples:

- **Iressa (AstraZeneca)**
- **Erbitux (ImClone)**
- **Tarceva (Genentech/OSI)**

ENTICE CELLS TO DIE

Drugs reactivate the tumor suppressor genes or block enzymes that protect cancer cells from dying.

Examples:

- **Velcade (Millennium)**
- **Advexin (Introgen)**

MOBILIZE THE IMMUNE SYSTEM

Vaccines and other immunotherapies train the body's immune system to go after cancer cells.

Examples:

- **GVAX (Cell Genesys)**
- **Provenge (Dendreon)**
- **Oncophage (Antigenics)**

STARVE THE TUMOR

Cancer cells need blood to thrive. So-called anti-angiogenesis drugs stop blood vessels from reaching them.

Examples:

- **Avastin (Genentech)**
- **Neovastat (Aeterna)**
- **Thalidomide (Celgene)**

ILLUSTRATION BY [illegible] SALIENOW; REFERENCE COURTESY OF CANCER INFORMATION AND SUPPORT INTERNATIONAL

PubMed Search

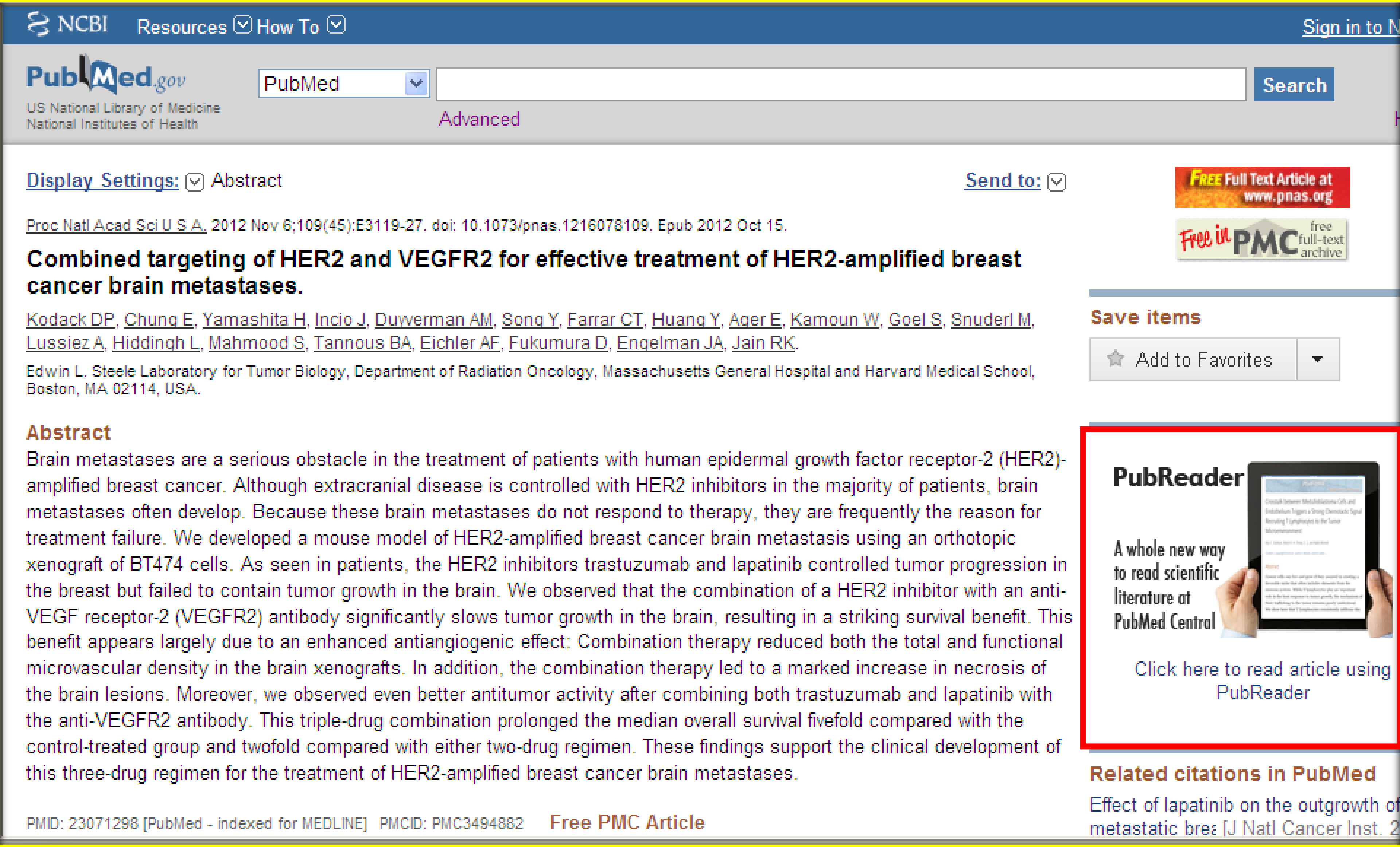
Cancer treatment by targeting blood supply:

Cancer growth depends on blood supply (why?) and thus requires the growth of new blood vessels – angiogenesis

Proteins involved in angiogenesis may be potential anticancer targets

You can find some of these targets by searching Pubmed

Key Word	**No. of Entries**
Cancer	
Cancer Blood supply	
Cancer Blood supply Protein	
Cancer Blood supply Enzyme	
Cancer Blood supply Enzyme Drug	


NCBI
Resources
How To
Sign in to N
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
Search
Advanced
Display Settings:
Send to:
FREE Full Text Article at www.pnas.org
Free in PMC free full-text archive
Save items
Add to Favorites

Display Settings: Abstract

Proc Natl Acad Sci U S A. 2012 Nov 6;109(45):E3119-27. doi: 10.1073/pnas.1216078109. Epub 2012 Oct 15.

Combined targeting of HER2 and VEGFR2 for effective treatment of HER2-amplified breast cancer brain metastases.

Kodack DP, Chung E, Yamashita H, Incio J, Duyverman AM, Song Y, Farrar CT, Huang Y, Ager E, Kamoun W, Goel S, Snuderl M, Lussiez A, Hiddingh L, Mahmood S, Tannous BA, Eichler AF, Fukumura D, Engelman JA, Jain RK.

Edwin L. Steele Laboratory for Tumor Biology, Department of Radiation Oncology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA.

Abstract

Brain metastases are a serious obstacle in the treatment of patients with human epidermal growth factor receptor-2 (HER2)-amplified breast cancer. Although extracranial disease is controlled with HER2 inhibitors in the majority of patients, brain metastases often develop. Because these brain metastases do not respond to therapy, they are frequently the reason for treatment failure. We developed a mouse model of HER2-amplified breast cancer brain metastasis using an orthotopic xenograft of BT474 cells. As seen in patients, the HER2 inhibitors trastuzumab and lapatinib controlled tumor progression in the breast but failed to contain tumor growth in the brain. We observed that the combination of a HER2 inhibitor with an anti-VEGF receptor-2 (VEGFR2) antibody significantly slows tumor growth in the brain, resulting in a striking survival benefit. This benefit appears largely due to an enhanced antiangiogenic effect: Combination therapy reduced both the total and functional microvascular density in the brain xenografts. In addition, the combination therapy led to a marked increase in necrosis of the brain lesions. Moreover, we observed even better antitumor activity after combining both trastuzumab and lapatinib with the anti-VEGFR2 antibody. This triple-drug combination prolonged the median overall survival fivefold compared with the control-treated group and twofold compared with either two-drug regimen. These findings support the clinical development of this three-drug regimen for the treatment of HER2-amplified breast cancer brain metastases.

PMID: 23071298 [PubMed - indexed for MEDLINE] PMCID: PMC3494882 Free PMC Article

PubReader

A whole new way to read scientific literature at PubMed Central

Click here to read article using PubReader

Related citations in PubMed

Effect of lapatinib on the outgrowth of metastatic brea [J Natl Cancer Inst. 2

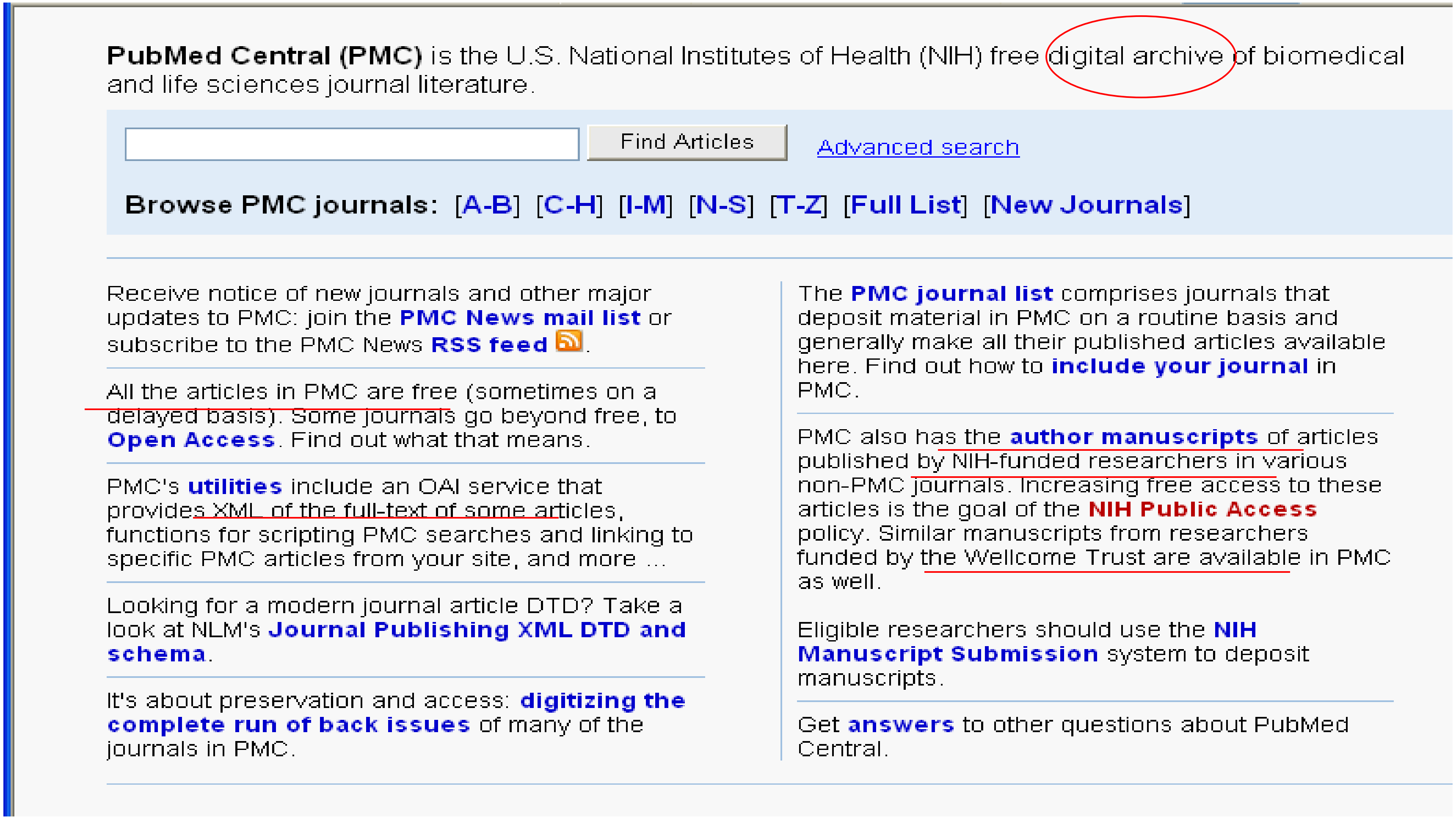
PubMed searching skills

- keyword
- 连词的应用 “and”“or” “not”
- limits
- 同义词...

My NCBI

- **My NCBI** is a central place to customize NCBI Web services. To use it, you must first register, and your browser must accept cookies.
- You can use **My NCBI** to:
  - Save searches
  - Set up e-mail alerts for new content
  - Display links to Web resources (LinkOut)
  - Choose filters that group search results

PubMed Central (PMC)



PubMed Central (PMC) is the U.S. National Institutes of Health (NIH) free digital archive of biomedical and life sciences journal literature.

Browse PMC journals: [A-B] [C-H] [I-M] [N-S] [T-Z] [Full List] [New Journals]

Receive notice of new journals and other major updates to PMC: join the **PMC News mail list** or subscribe to the PMC News **RSS feed** .

All the articles in PMC are free (sometimes on a delayed basis). Some journals go beyond free, to **Open Access**. Find out what that means.

PMC's **utilities** include an OAI service that provides XML of the full-text of some articles, functions for scripting PMC searches and linking to specific PMC articles from your site, and more ...

Looking for a modern journal article DTD? Take a look at NLM's **Journal Publishing XML DTD and schema**.

It's about preservation and access: **digitizing the complete run of back issues** of many of the journals in PMC.

The **PMC journal list** comprises journals that deposit material in PMC on a routine basis and generally make all their published articles available here. Find out how to **include your journal** in PMC.

PMC also has the **author manuscripts** of articles published by NIH-funded researchers in various non-PMC journals. Increasing free access to these articles is the goal of the **NIH Public Access** policy. Similar manuscripts from researchers funded by the Wellcome Trust are available in PMC as well.

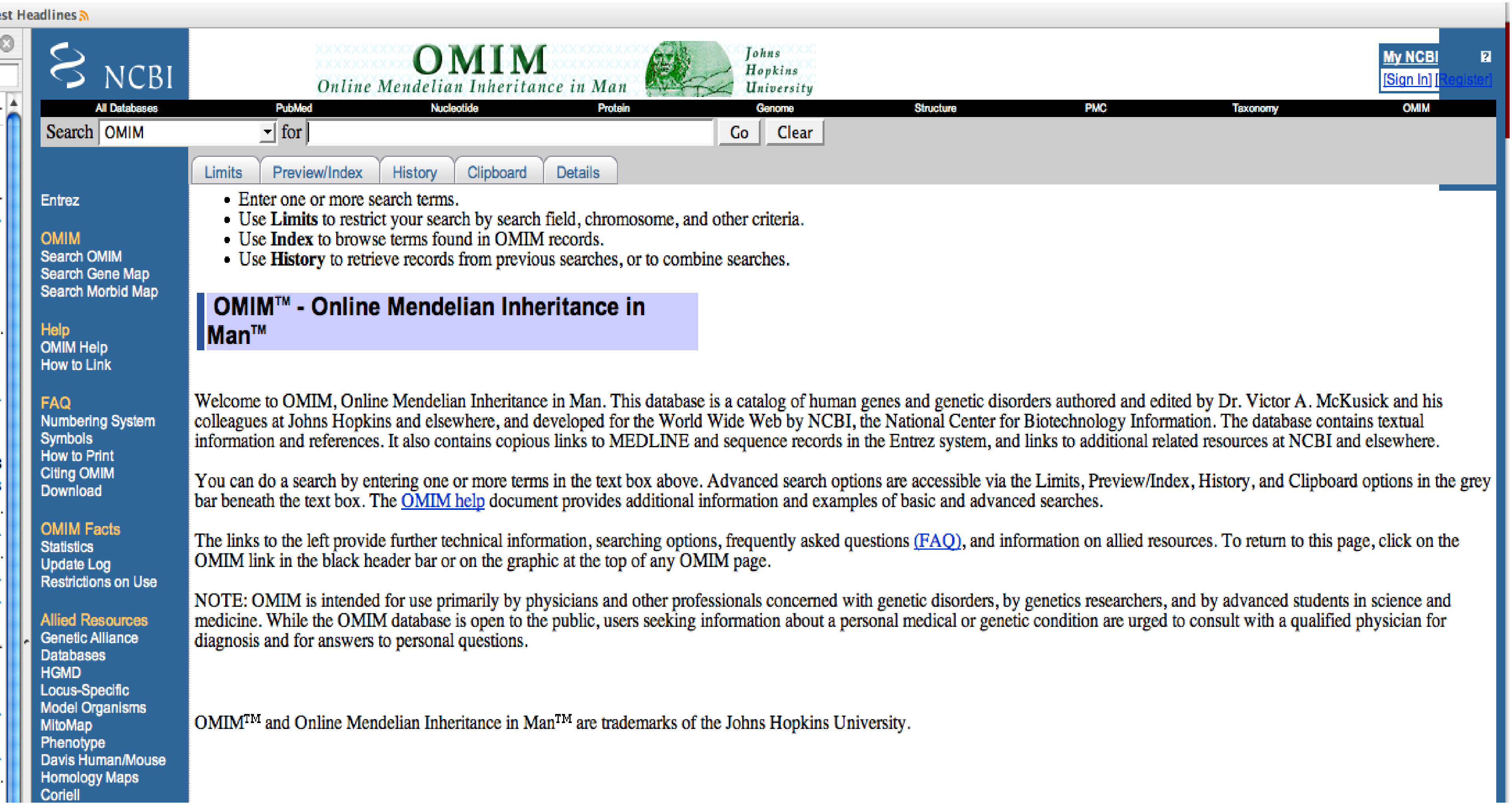
Eligible researchers should use the **NIH Manuscript Submission** system to deposit manuscripts.

Get **answers** to other questions about PubMed Central.

2. OMIM

OMIM is another literature database in the NCBI family.

It is the online version of a catalog of human genes and genetic disorders.



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

OMIM™ - Online Mendelian Inheritance in Man™

Welcome to OMIM, Online Mendelian Inheritance in Man. This database is a catalog of human genes and genetic disorders authored and edited by Dr. Victor A. McKusick and his colleagues at Johns Hopkins and elsewhere, and developed for the World Wide Web by NCBI, the National Center for Biotechnology Information. The database contains textual information and references. It also contains copious links to MEDLINE and sequence records in the Entrez system, and links to additional related resources at NCBI and elsewhere.

You can do a search by entering one or more terms in the text box above. Advanced search options are accessible via the Limits, Preview/Index, History, and Clipboard options in the grey bar beneath the text box. The OMIM help document provides additional information and examples of basic and advanced searches.

The links to the left provide further technical information, searching options, frequently asked questions (FAQ), and information on allied resources. To return to this page, click on the OMIM link in the black header bar or on the graphic at the top of any OMIM page.

NOTE: OMIM is intended for use primarily by physicians and other professionals concerned with genetic disorders, by genetics researchers, and by advanced students in science and medicine. While the OMIM database is open to the public, users seeking information about a personal medical or genetic condition are urged to consult with a qualified physician for diagnosis and for answers to personal questions.

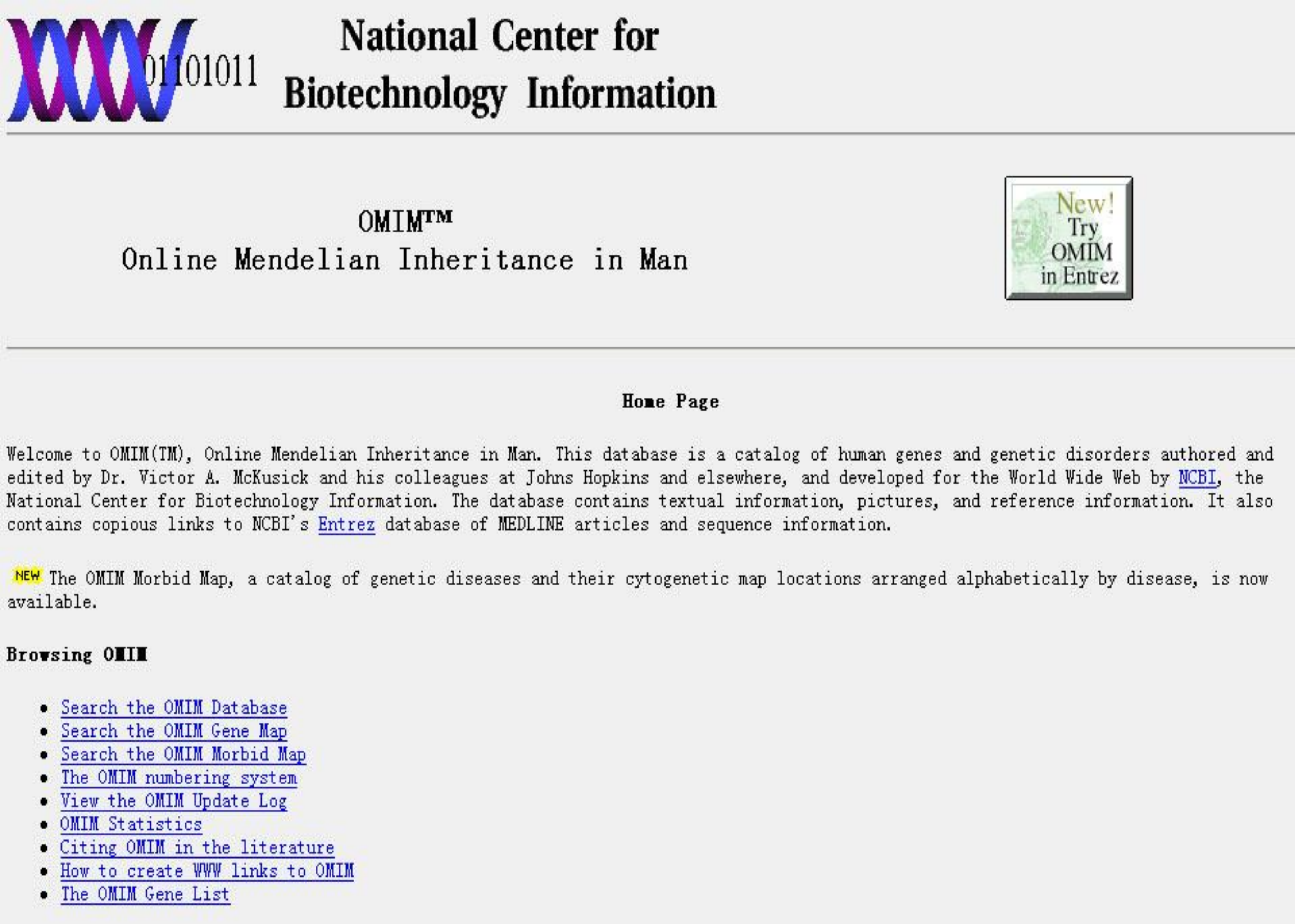
OMIM™ and Online Mendelian Inheritance in Man™ are trademarks of the Johns Hopkins University.

Biological databases: OMIM Online Mendelian Inheritance in Man (http://www.ncbi.nlm.nih.gov/Omim/)

- The OMIM database contains abstracts and texts describing genetic disorders to support genomics efforts and clinical genetics. It provides gene maps, and known disorder maps in tabular listing formats. Contains keyword search.

Hamosh A. *et* al. Online Mendelian Inheritance in Man (OMIM), a knowledge base of human genes and genetic disorders *Nucleic Acids Res*. 2002 30: 52-55.

Biological databases: OMIM web-page



National Center for Biotechnology Information

OMIM™

Online Mendelian Inheritance in Man

New! Try OMIM in Entrez

Home Page

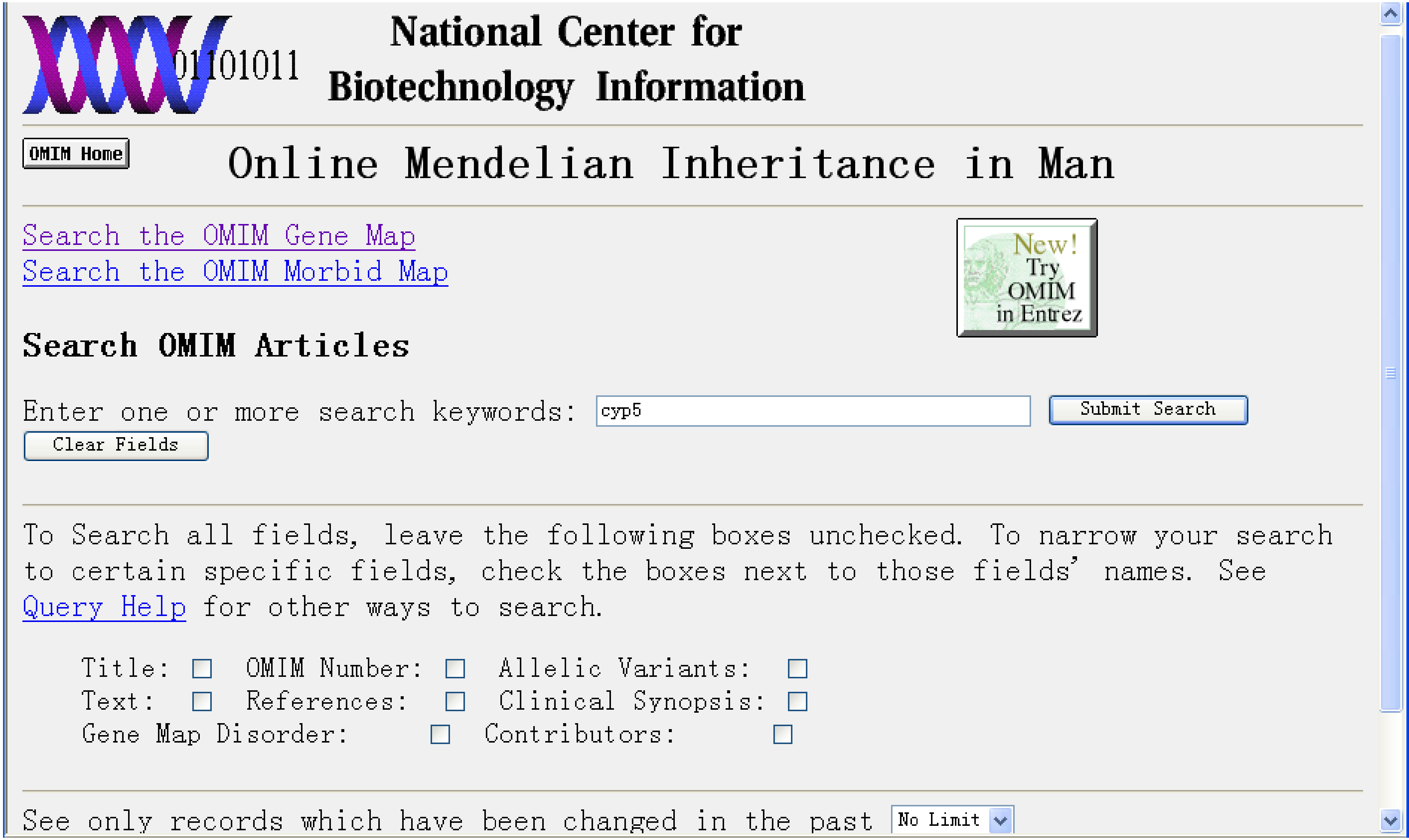
Welcome to OMIM(TM), Online Mendelian Inheritance in Man. This database is a catalog of human genes and genetic disorders authored and edited by Dr. Victor A. McKusick and his colleagues at Johns Hopkins and elsewhere, and developed for the World Wide Web by NCBI, the National Center for Biotechnology Information. The database contains textual information, pictures, and reference information. It also contains copious links to NCBI's Entrez database of MEDLINE articles and sequence information.

NEW The OMIM Morbid Map, a catalog of genetic diseases and their cytogenetic map locations arranged alphabetically by disease, is now available.

Browsing OMIM

- Search the OMIM Database
- Search the OMIM Gene Map
- Search the OMIM Morbid Map
- The OMIM numbering system
- View the OMIM Update Log
- OMIM Statistics
- Citing OMIM in the literature
- How to create WWW links to OMIM
- The OMIM Gene List

Biological databases: OMIM search engine



National Center for Biotechnology Information

Online Mendelian Inheritance in Man

Search the OMIM Gene Map
Search the OMIM Morbid Map

Search OMIM Articles

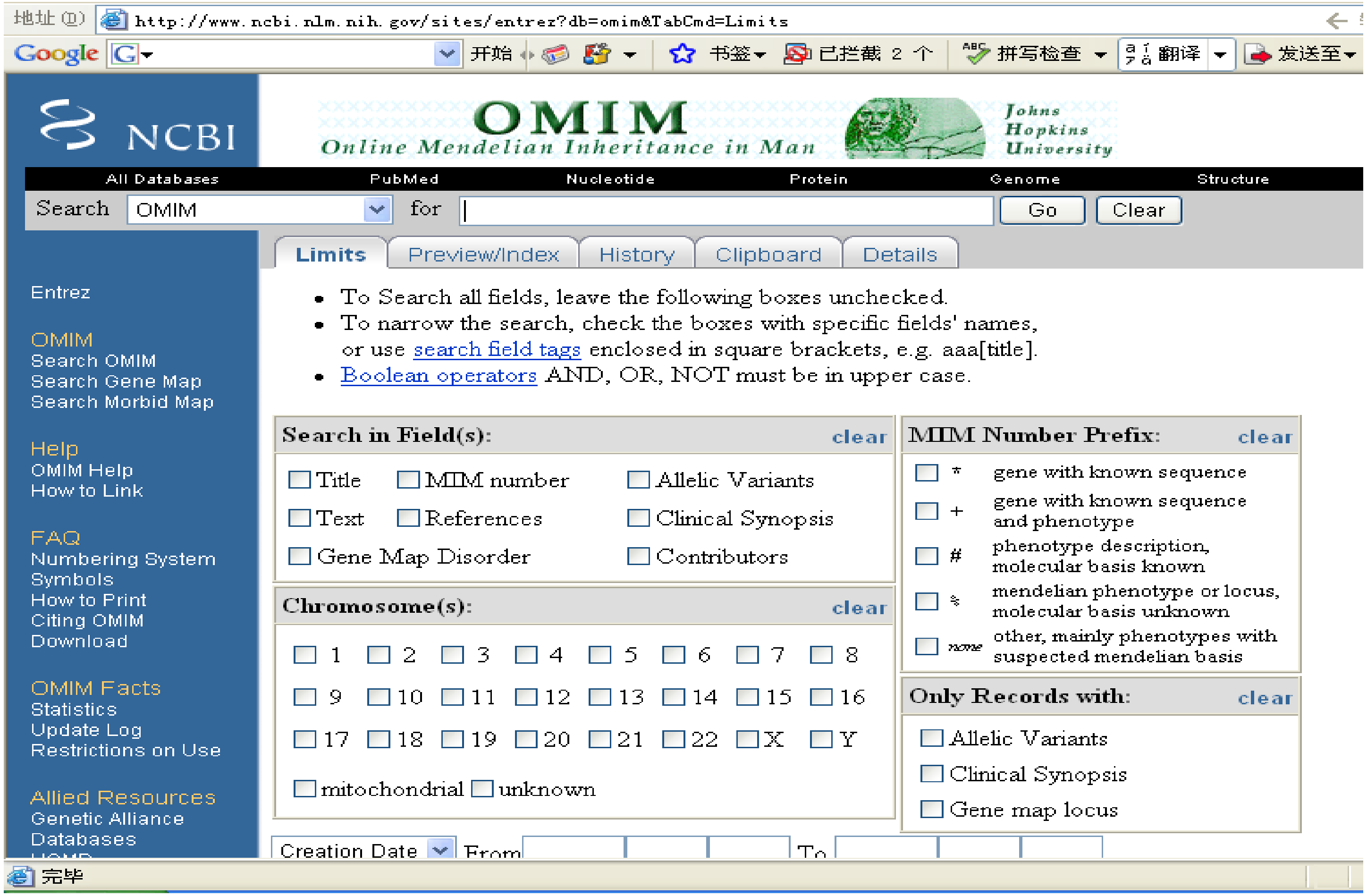
Enter one or more search keywords: cyp5

To Search all fields, leave the following boxes unchecked. To narrow your search to certain specific fields, check the boxes next to those fields' names. See Query Help for other ways to search.

Title: □ OMIM Number: □ Allelic Variants: □
Text: □ References: □ Clinical Synopsis: □
Gene Map Disorder: □ Contributors: □

See only records which have been changed in the past No Limit

Search OMIM



Search OMIM for

- To Search all fields, leave the following boxes unchecked.
- To narrow the search, check the boxes with specific fields' names, or use search field tags enclosed in square brackets, e.g. aaa[title].
- Boolean operators AND, OR, NOT must be in upper case.

Search in Field(s): clear

□ Title □ MIM number □ Allelic Variants
□ Text □ References □ Clinical Synopsis
□ Gene Map Disorder □ Contributors

Chromosome(s): clear

□ 1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □ 8
□ 9 □ 10 □ 11 □ 12 □ 13 □ 14 □ 15 □ 16
□ 17 □ 18 □ 19 □ 20 □ 21 □ 22 □ X □ Y
□ mitochondrial □ unknown

MIM Number Prefix: clear

□	*	gene with known sequence
□	+	gene with known sequence and phenotype
□	#	phenotype description, molecular basis known
□	%	mendelian phenotype or locus, molecular basis unknown
□	*none*	other, mainly phenotypes with suspected mendelian basis

Only Records with: clear

□ Allelic Variants
□ Clinical Synopsis
□ Gene map locus

Creation Date From To

Nucleic Acids databases

What info are in these databases:

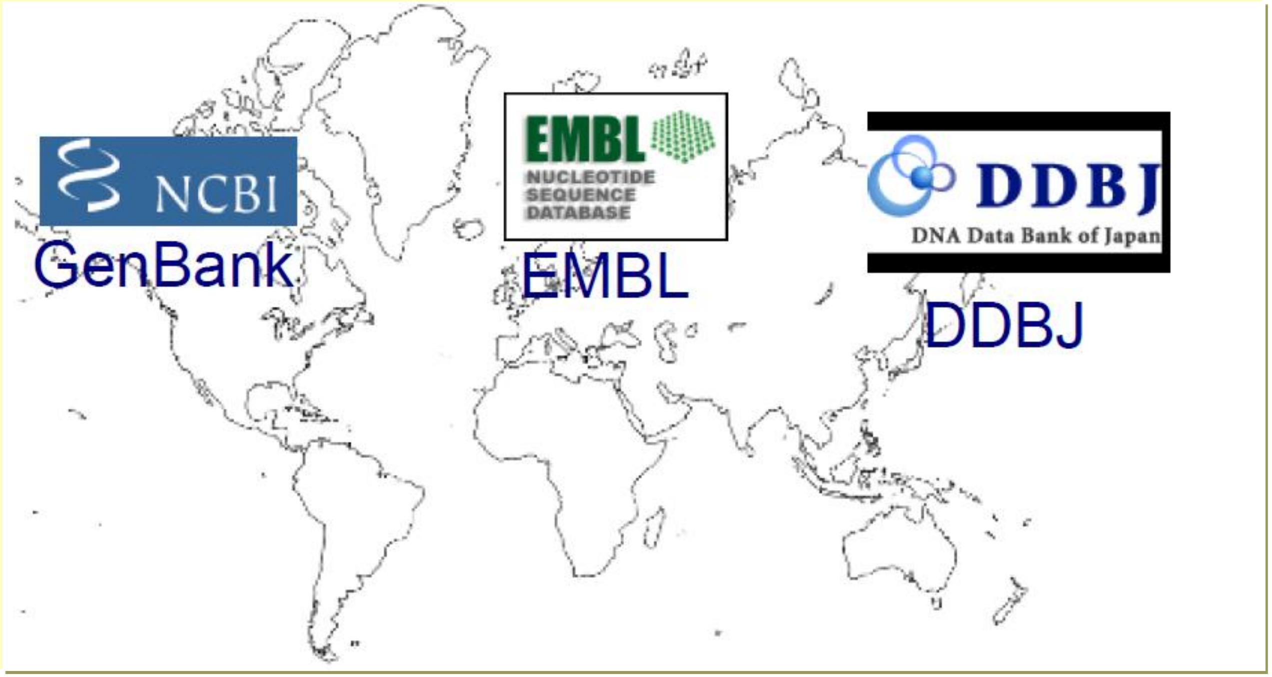
- DNA sequence, genes, gene products (proteins), mutation, gene coding, distribution patterns, motifs
- Genomics: genome, gene structure and expression, genetic map, genetic disorder
- RNA sequence, secondary structure, 3D structure, interactions

GenBank

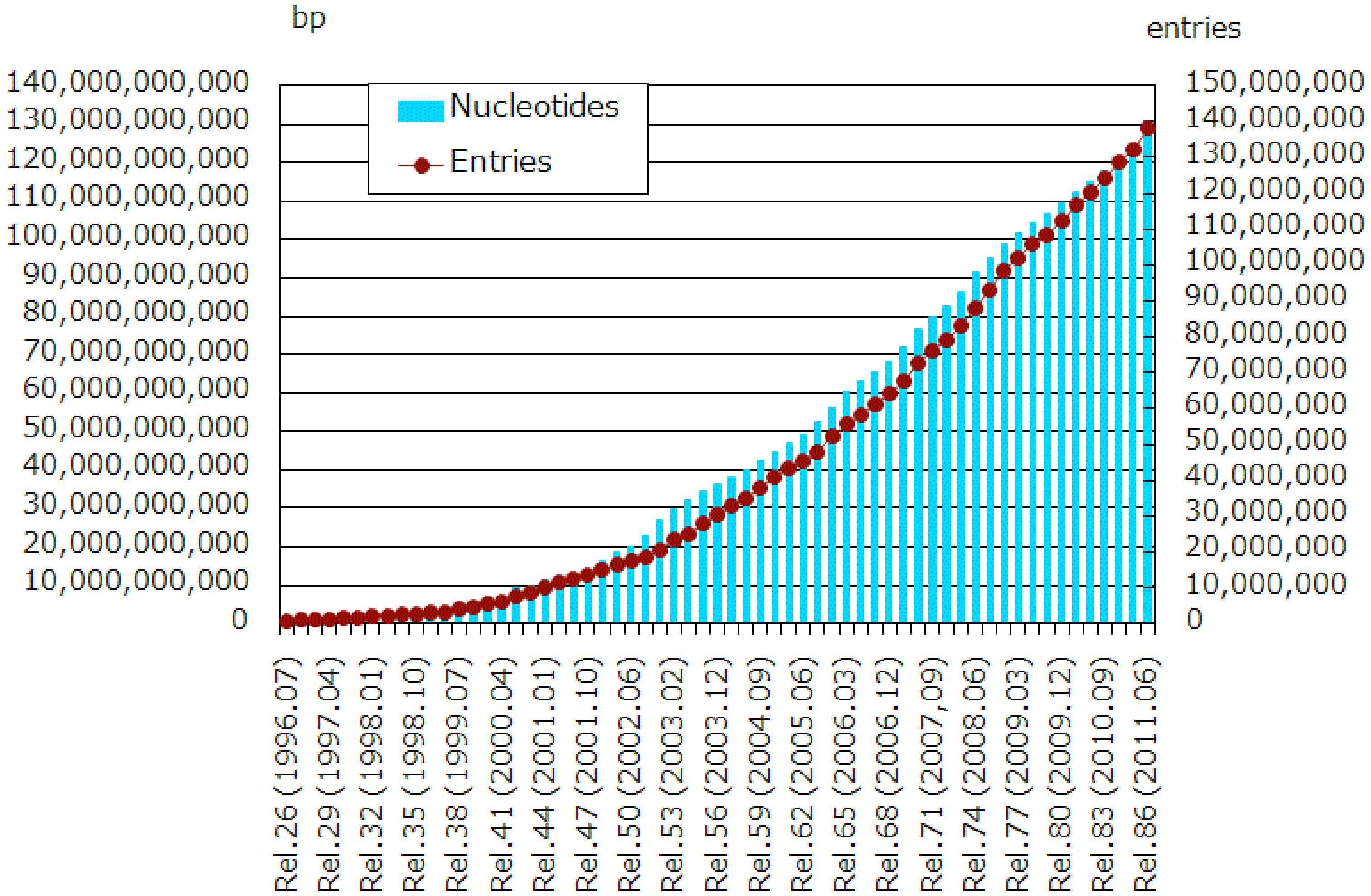
- The complete release notes for the current version of GenBank are available on the NCBI ftp site.
- A new release is made every two months.
- GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

GenBank

- GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences.
-

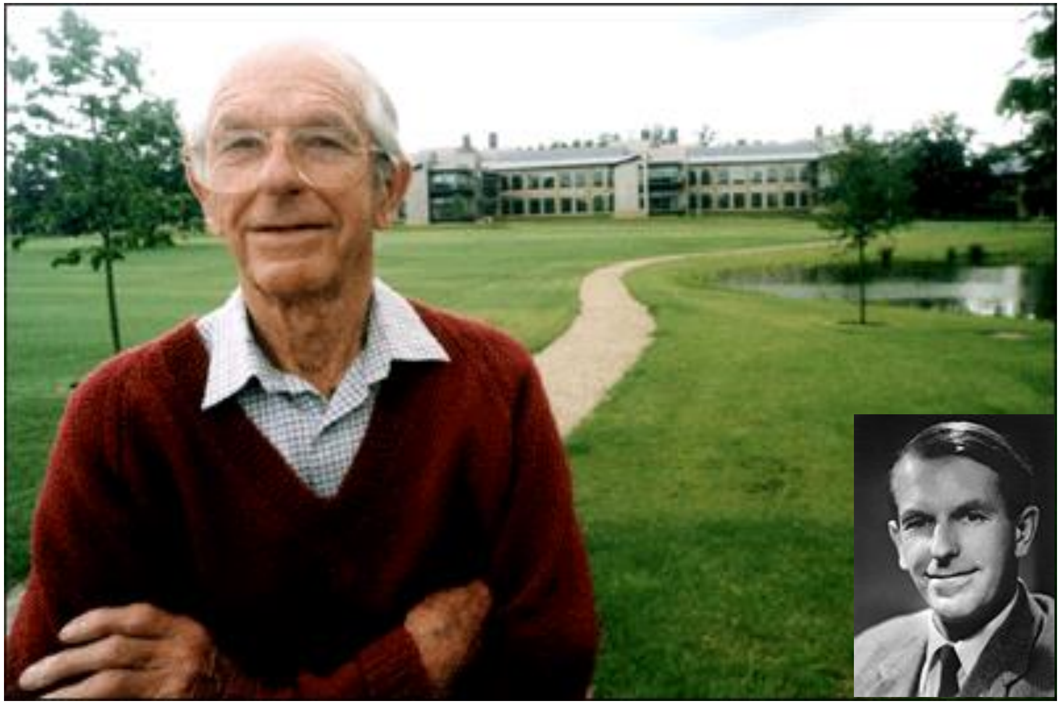


DDBJ/EMBL/GenBank database growth



Note: CON division is not counted in statistics of DDBJ periodical releases.

Where does large data come from? High-throughput techniques



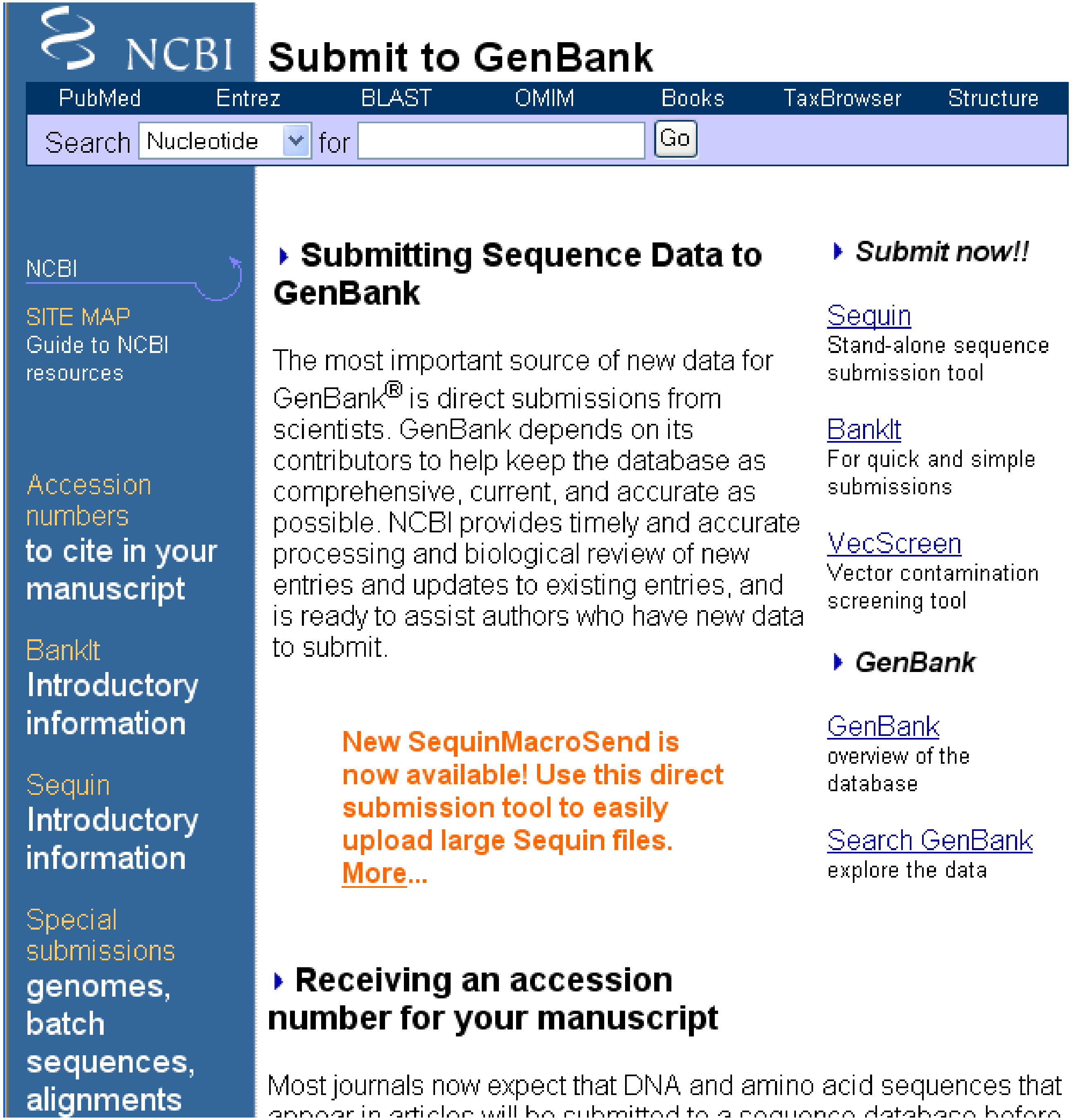
Fred Sanger

- Nobel prize in chemistry in 1958
"for his work on the **structure** of proteins, especially that of insulin"

- Nobel prize in chemistry in 1980
"for their contributions concerning the determination of **base sequences** in nucleic acids"

DNA databases:

GenBank
Web page



Submit to GenBank

NCBI

SITE MAP
Guide to NCBI resources

Accession numbers
to cite in your manuscript

BankIt
Introductory information

Sequin
Introductory information

Special submissions
genomes, batch sequences, alignments

Submitting Sequence Data to GenBank

The most important source of new data for GenBank® is direct submissions from scientists. GenBank depends on its contributors to help keep the database as comprehensive, current, and accurate as possible. NCBI provides timely and accurate processing and biological review of new entries and updates to existing entries, and is ready to assist authors who have new data to submit.

New SequinMacroSend is now available! Use this direct submission tool to easily upload large Sequin files. More...

Receiving an accession number for your manuscript

Most journals now expect that DNA and amino acid sequences that appear in articles will be submitted to a sequence database before

Submit now!!

Sequin
Stand-alone sequence submission tool

BankIt
For quick and simple submissions

VecScreen
Vector contamination screening tool

GenBank

GenBank
overview of the database

Search GenBank
explore the data

What might we want to know about a sequence?

- Is this sequence similar to any known genes? How close is the best match? Significance?
- What do we know about that gene?
  - Genomic (chromosomal location, allelic information, regulatory regions, etc.)
  - Structural (known structure? structural domains? etc.)
  - Functional (molecular, cellular & disease)
- Evolutionary information:
  - Is this gene found in other organisms?
  - What is its taxonomic tree?

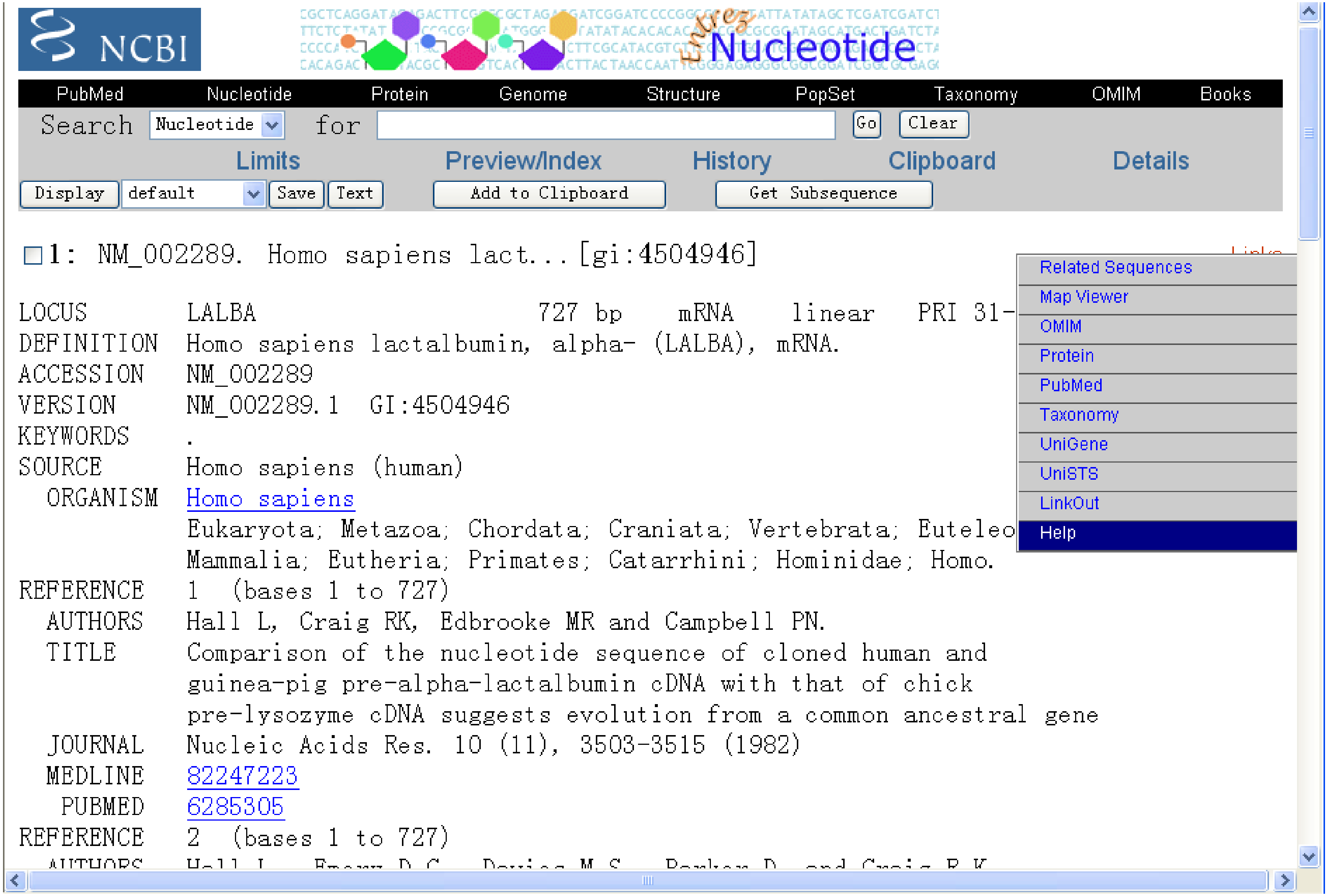
DNA databases

- An Example from GenBank– flat file

  - Human Alpha-Lactalbumin gene

This protein is a complex of 2 proteins A and B. In the absence of the B protein, the enzyme catalyzes the transfer of galactose from UDP-galactose to Nacetylglucosamine (cf. EC 2.4.1.90).

A GenBank entry – HEADER



□1: NM_002289. Homo sapiens lact...[gi:4504946]

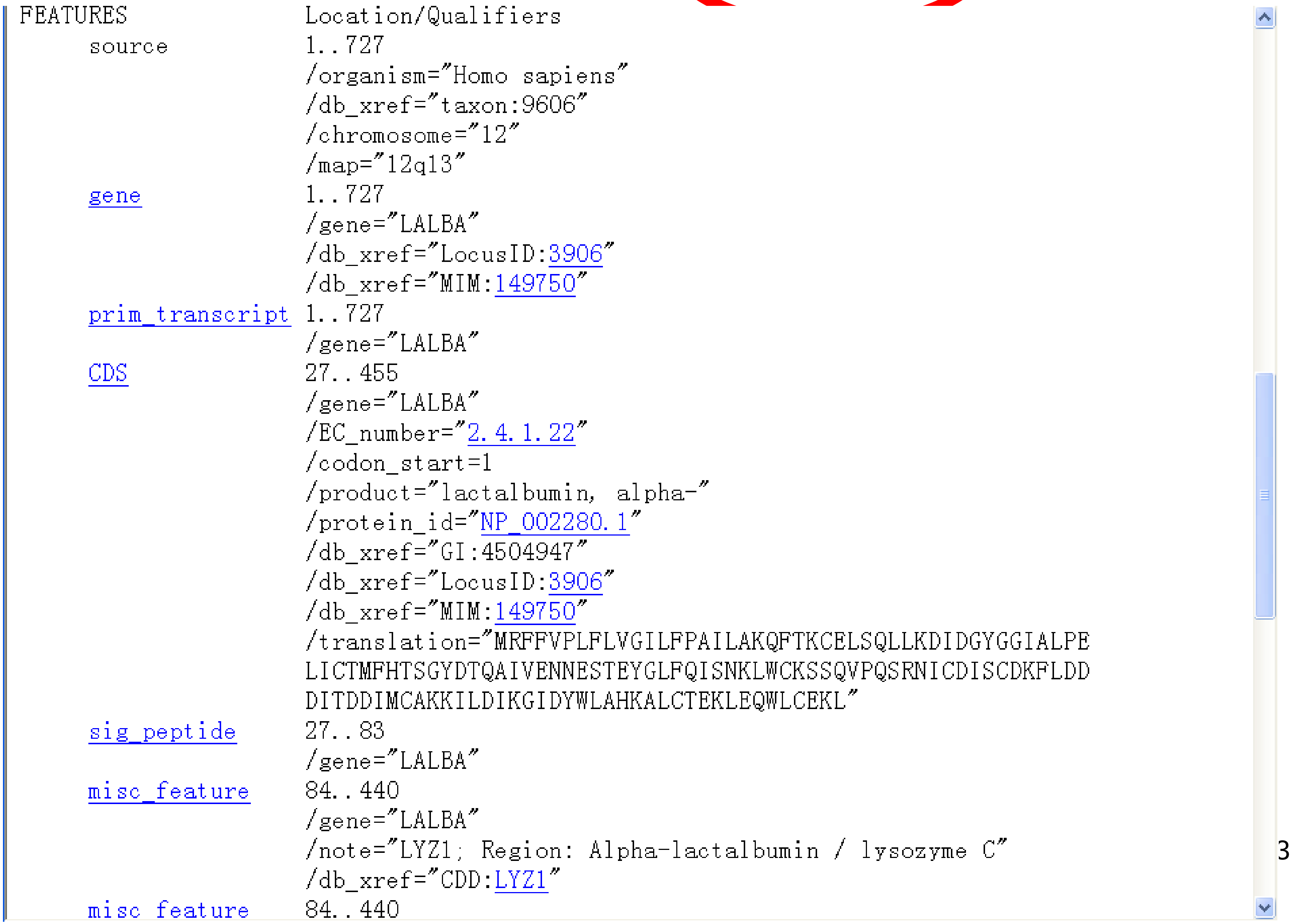
```
LOCUS       LALBA                  727 bp    mRNA    linear   PRI 31-
DEFINITION  Homo sapiens lactalbumin, alpha- (LALBA), mRNA.
ACCESSION   NM_002289
VERSION     NM_002289.1  GI:4504946
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleo
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 727)
  AUTHORS   Hall L, Craig RK, Edbrooke MR and Campbell PN.
  TITLE     Comparison of the nucleotide sequence of cloned human and
            guinea-pig pre-alpha-lactalbumin cDNA with that of chick
            pre-lysozyme cDNA suggests evolution from a common ancestral gene
  JOURNAL   Nucleic Acids Res. 10 (11), 3503-3515 (1982)
  MEDLINE   82247223
   PUBMED   6285305
REFERENCE   2  (bases 1 to 727)
  AUTHORS   Hall L., Emery D.C., Davies M.S., Parker D. and Craig R.K.
```

GenBank Entry – Links provided in the Header

- MapViewer – find the gene position in chromosome
- Related Sequences – other entries related to this gene (or sequence)
- OMIM– link to catalog of human genes and genetic disorders
- Protein – retrieve protein record from GenPept
- Medline and PubMed –literature abstracts related to this gene
- Taxonomy – Classification of organisms
- UniGene – Unified gene data
- UniSTS – Unified sequence tagged sites, marker and mapping data
- LinkOut – links to publishers, aggregators libraries, biological databases, sequence centers, and other Web resources
- REFSEQ – reference sequence standards

Note: These links are representative. Other links may also be found in GenBank entries.

GenBank entry - FEATURES



```
FEATURES             Location/Qualifiers
     source          1..727
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="12"
                     /map="12q13"
     gene            1..727
                     /gene="LALBA"
                     /db_xref="LocusID:3906"
                     /db_xref="MIM:149750"
     prim_transcript 1..727
                     /gene="LALBA"
     CDS             27..455
                     /gene="LALBA"
                     /EC_number="2.4.1.22"
                     /codon_start=1
                     /product="lactalbumin, alpha-"
                     /protein_id="NP_002280.1"
                     /db_xref="GI:4504947"
                     /db_xref="LocusID:3906"
                     /db_xref="MIM:149750"
                     /translation="MRFFVPLFLVGILFPAILAKQFTKCELSQLLKDIDGYGGIALPE
                     LICTMFHTSGYDTQAIVENNESTEYGLFQISNKLWCKSSQVPQSRNICDISCDKFLDD
                     DITDDIMCAKKILDIKGIDYWLAHKALCTEKLEQWLCEKL"
     sig_peptide     27..83
                     /gene="LALBA"
     misc_feature    84..440
                     /gene="LALBA"
                     /note="LYZ1; Region: Alpha-lactalbumin / lysozyme C"
                     /db_xref="CDD:LYZ1"
     misc_feature    84..440
```

GenBank Entry– Links provided in the Feature section

LocusID – locus and display of genomic and mRNA sequences

MIM – Link to OMIM description, other entries for this sequence

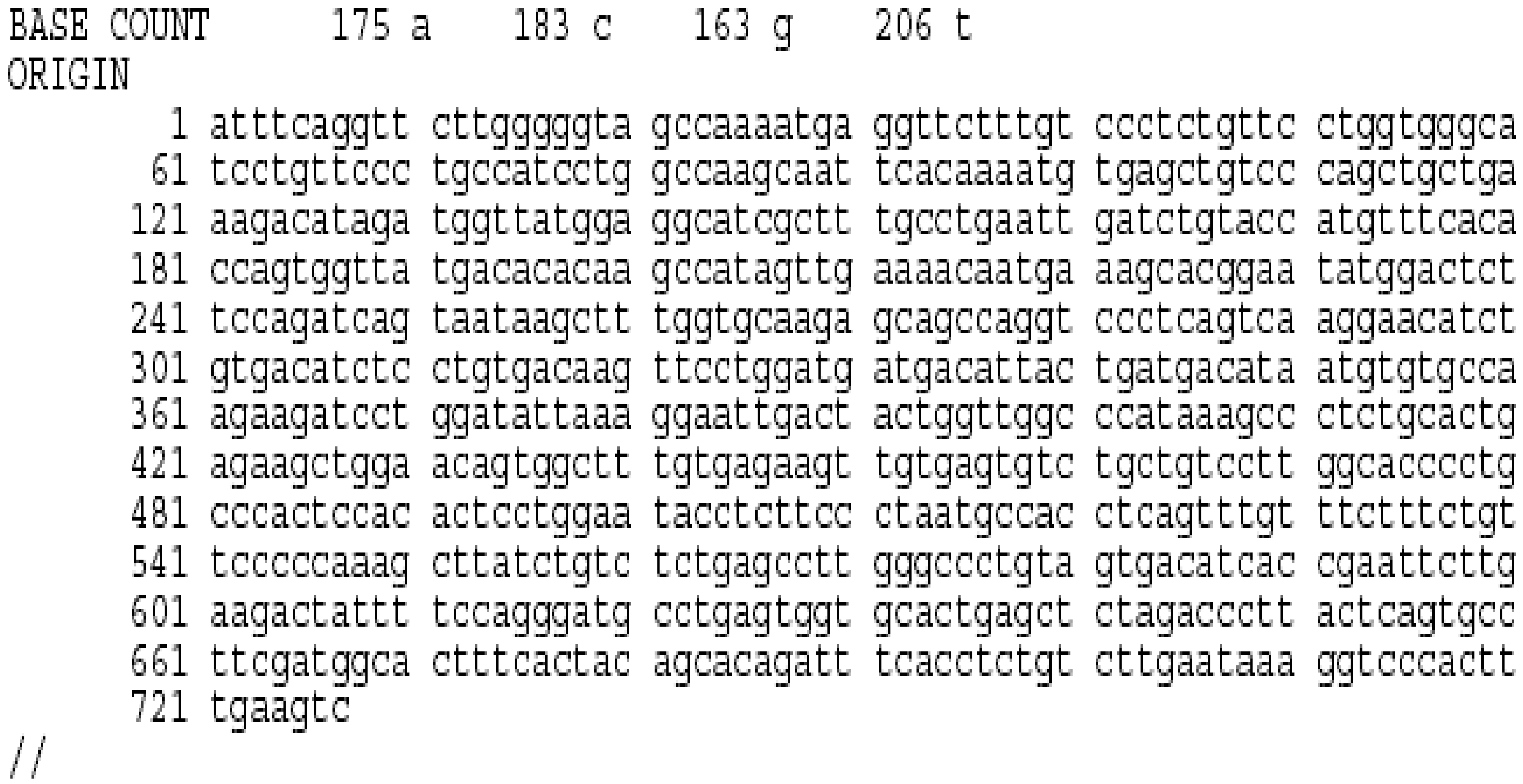
EC_number – link to the corresponding cataloged enzymes

Protein_id – retrieve protein record from GenPept

CD– conserved protein domain (SMART),

CDD – conserved protein domain (Pfam).

Biological databases: GenBank - SEQUENCE



```
BASE COUNT      175 a    183 c    163 g    206 t
ORIGIN
        1 atttcaggtt cttgggggta gccaaaatga ggttctttgt ccctctgttc ctggtgggca
       61 tcctgttccc tgccatcctg gccaagcaat tcacaaaatg tgagctgtcc cagctgctga
      121 aagacataga tggttatgga ggcatcgctt tgcctgaatt gatctgtacc atgtttcaca
      181 ccagtggtta tgacacacaa gccatagttg aaaacaatga aagcacggaa tatggactct
      241 tccagatcag taataagctt tggtgcaaga gcagccaggt ccctcagtca aggaacatct
      301 gtgacatctc ctgtgacaag ttcctggatg atgacattac tgatgacata atgtgtgcca
      361 agaagatcct ggatattaaa ggaattgact actggttggc ccataaagcc ctctgcactg
      421 agaagctgga acagtggctt tgtgagaagt tgtgagtgtc tgctgtcctt ggcacccctg
      481 cccactccac actcctggaa tacctcttcc ctaatgccac ctcagtttgt ttctttctgt
      541 tcccccaaag cttatctgtc tctgagcctt gggccctgta gtgacatcac cgaattcttg
      601 aagactattt tccagggatg cctgagtggt gcactgagct ctagaccctt actcagtgcc
      661 ttcgatggca ctttcactac agcacagatt tcacctctgt cttgaataaa ggtcccactt
      721 tgaagtc
//
```

GenBank - NOTES

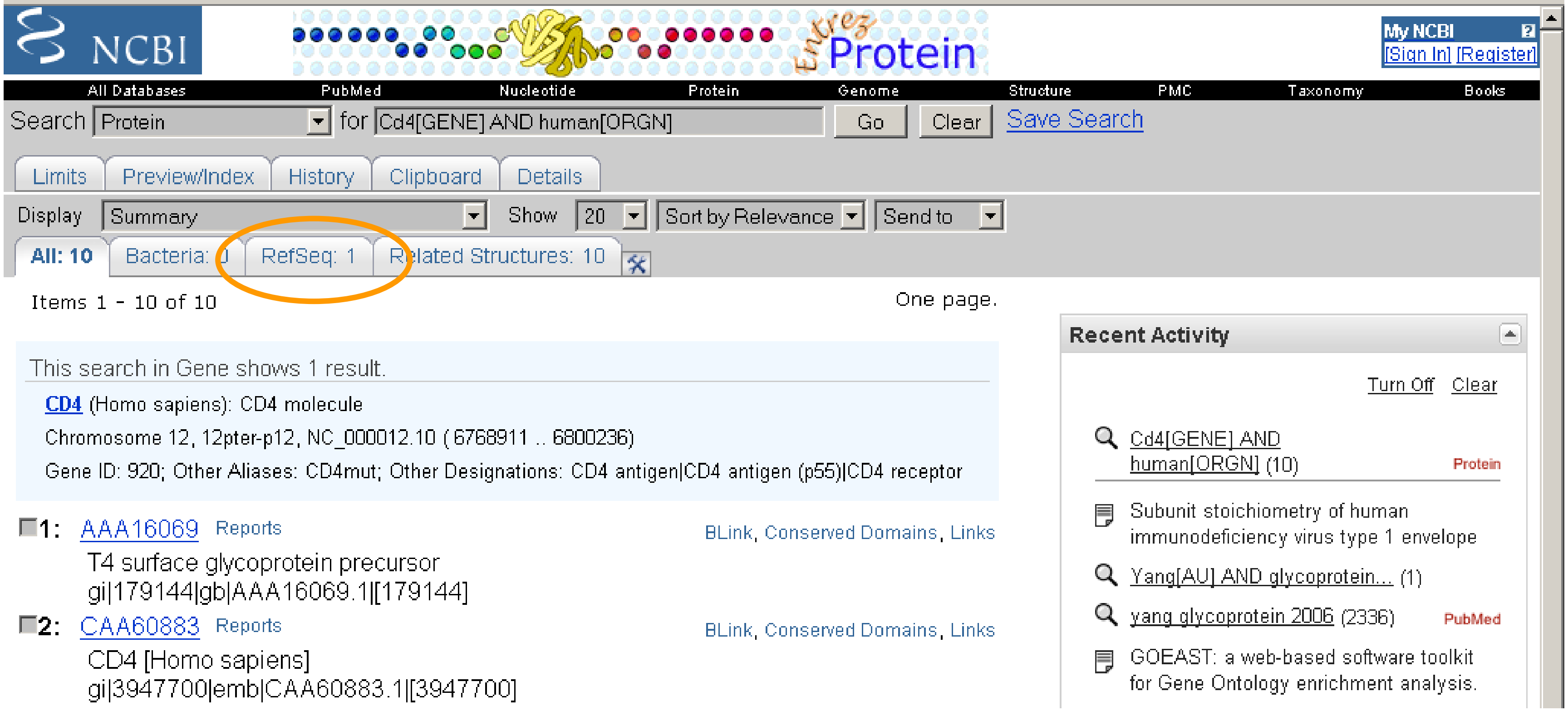
Majority of GenBank entries have similar form to our example.

When accessing the database, the following needs to be noticed:

- Some entries are huge, containing as much as 30,000 lines. (NT_021877 Homo sapiens chromosome 1 working draft sequence segment)
- Some entries have contig information instead of sequence information. (NT_021877 Homo sapiens chromosome 1 working draft sequence segment)
- Some entries are derived from cDNA sequences and thus represent putative genes/proteins. These should be used with caution.
- Some annotations are predicted using automated analysis. These should also be used with caution. (XM_131483 Mus musculus simi...[gi:20832685]).

RefSeq

- RefSeq: sub-collection of NCBI databases with only non-redundant, highly annotated entries (genomic DNA, transcript (RNA), and protein products)



Search Protein for Cd4[GENE] AND human[ORGN]

Display Summary Show 20 Sort by Relevance Send to

All: 10 Bacteria: 0 RefSeq: 1 Related Structures: 10

Items 1 - 10 of 10

One page.

This search in Gene shows 1 result.

CD4 (Homo sapiens): CD4 molecule
Chromosome 12, 12pter-p12, NC_000012.10 (6768911 .. 6800236)
Gene ID: 920; Other Aliases: CD4mut; Other Designations: CD4 antigen|CD4 antigen (p55)|CD4 receptor

□1: AAA16069 Reports BLink, Conserved Domains, Links
T4 surface glycoprotein precursor
gi|179144|gb|AAA16069.1|[179144]

□2: CAA60883 Reports BLink, Conserved Domains, Links
CD4 [Homo sapiens]
gi|3947700|emb|CAA60883.1|[3947700]

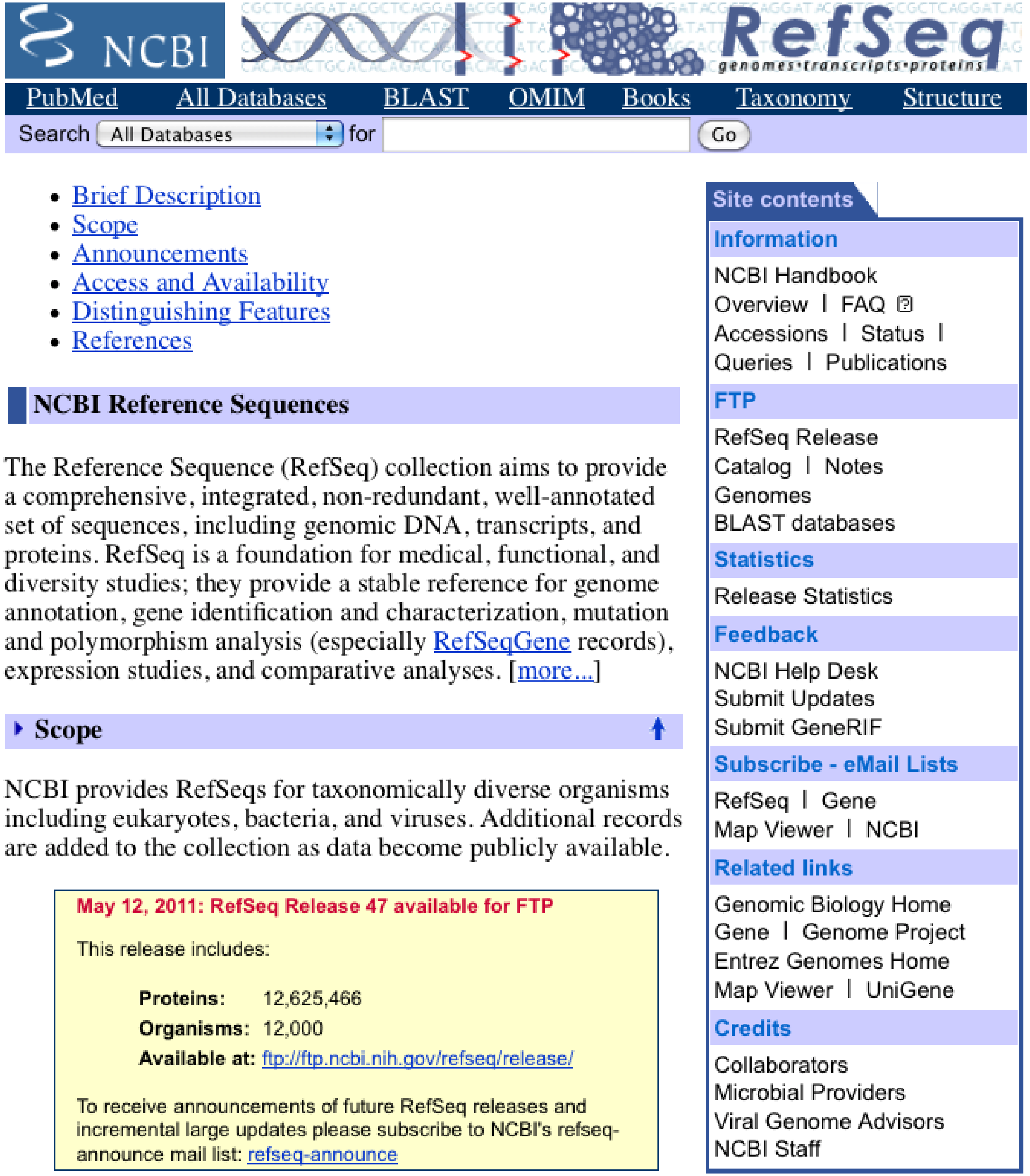
Cd4[GENE] AND human[ORGN] (10) Protein

Subunit stoichiometry of human immunodeficiency virus type 1 envelope

Yang[AU] AND glycoprotein... (1)

yang glycoprotein 2006 (2336) PubMed

GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis.



- Brief Description
- Scope
- Announcements
- Access and Availability
- Distinguishing Features
- References

NCBI Reference Sequences

The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins. RefSeq is a foundation for medical, functional, and diversity studies; they provide a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis (especially RefSeqGene records), expression studies, and comparative analyses. [more...]

Scope

NCBI provides RefSeqs for taxonomically diverse organisms including eukaryotes, bacteria, and viruses. Additional records are added to the collection as data become publicly available.

May 12, 2011: RefSeq Release 47 available for FTP

This release includes:

Proteins:	12,625,466
Organisms:	12,000
Available at:	ftp://ftp.ncbi.nih.gov/refseq/release/

To receive announcements of future RefSeq releases and incremental large updates please subscribe to NCBI's refseq-announce mail list: refseq-announce

The RefSeq Accession number format and molecule types

Accession	Molecule type
NC_xxxxxx	Complete genomic molecule
NG_xxxxxx	Genomic region
NM_xxxxxx	mRNA
NP_xxxxxx	Protein
NR_xxxxxx	RNA
NT_xxxxxx	computed Genomic contig
XM_xxxxxx	computed mRNA
XP_xxxxxx	computed Protein

Biological Databases

Database Searching

1. Most of the databases have a web-interface to search for data
2. Databases must have methods for accessing and extracting data stored.
3. The most basic search is keyword searching

   Keywords can be any word that occurs somewhere in the database records. It can be the name of the gene or protein (e.g. lactalbumin), species (e.g.*homo sapiens*, human), a taxonomy term (e.g.primates), or a word from the reference title (e.g. cancer)
4. Others include: Entry Id number, sequence
5. User can choose to view the data or save to your computer
6. Databases typically have hyperlinks that help to navigate from one database to another easily

Summary

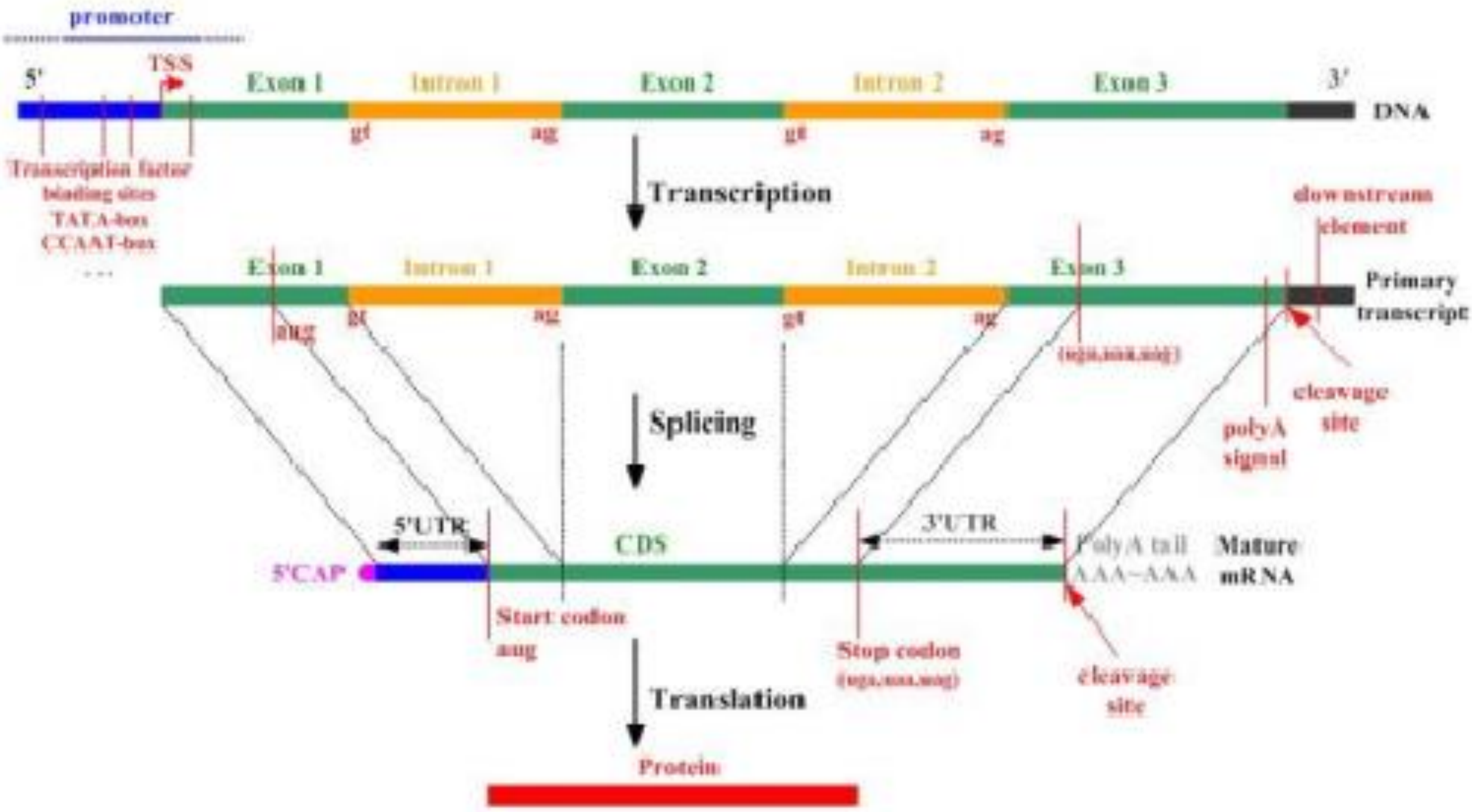
- what is a biological database?
- Why need bioinformatics database?
- Different types of bioinformatics database
- NCBI database
  - Pubmed
  - OMIM
  - GenBank

Homework

- What’s the latest amount of data for PubMed, OMIM and GenBank database?

- Explore NCBI database, choose 2 other database you are interested to explore details. Give a summary of them.

Gene


promoter
5'
TSS
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
3'
DNA
gt
ag
gt
ag
Transcription factor
binding sites
TATA-box
CCAAT-box
...
Transcription
downstream
element
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
Primary
transcript
aug
gt
ag
gt
ag
(uga,uaa,uag)
Splicing
cleavage
site
polyA
signal
5'UTR
3'UTR
CDS
PolyA tail
AAA~AAA
Mature
mRNA
5'CAP
Start codon
aug
Stop codon
(uga,uaa,uag)
cleavage
site
Translation
Protein