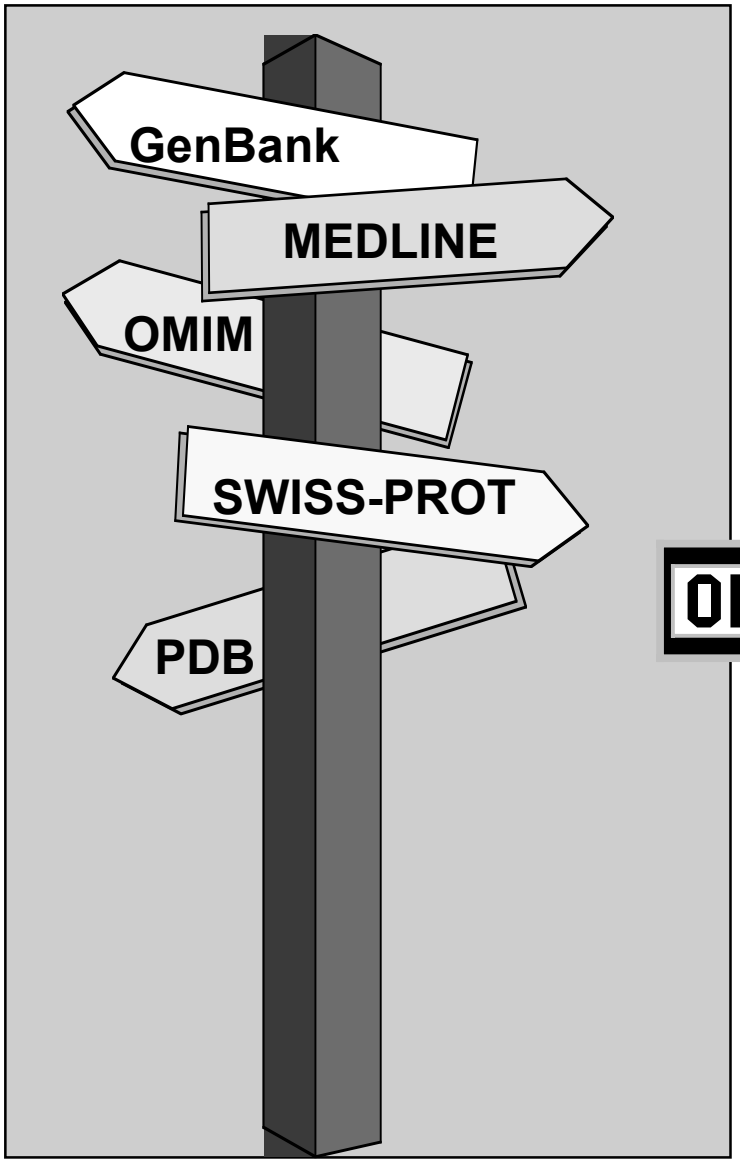
# Biological Databases: Information Retrieval

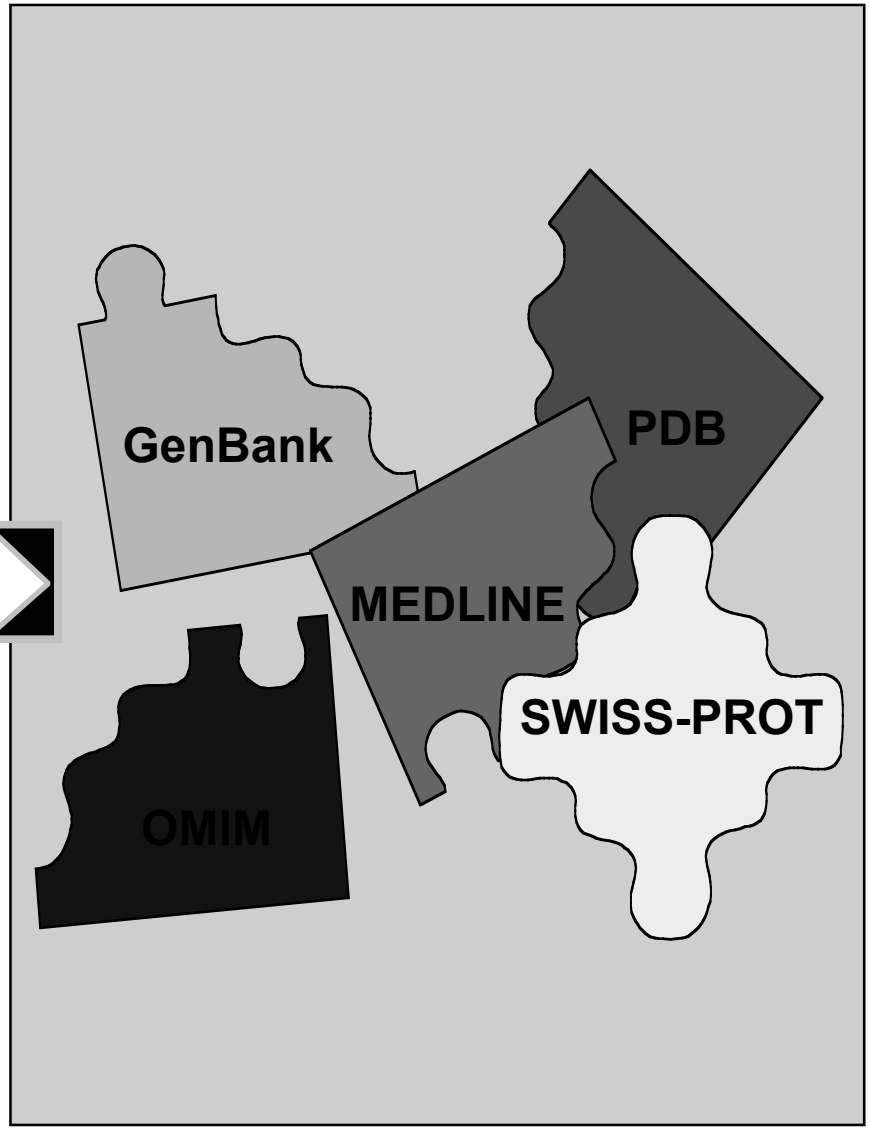
David Landsman

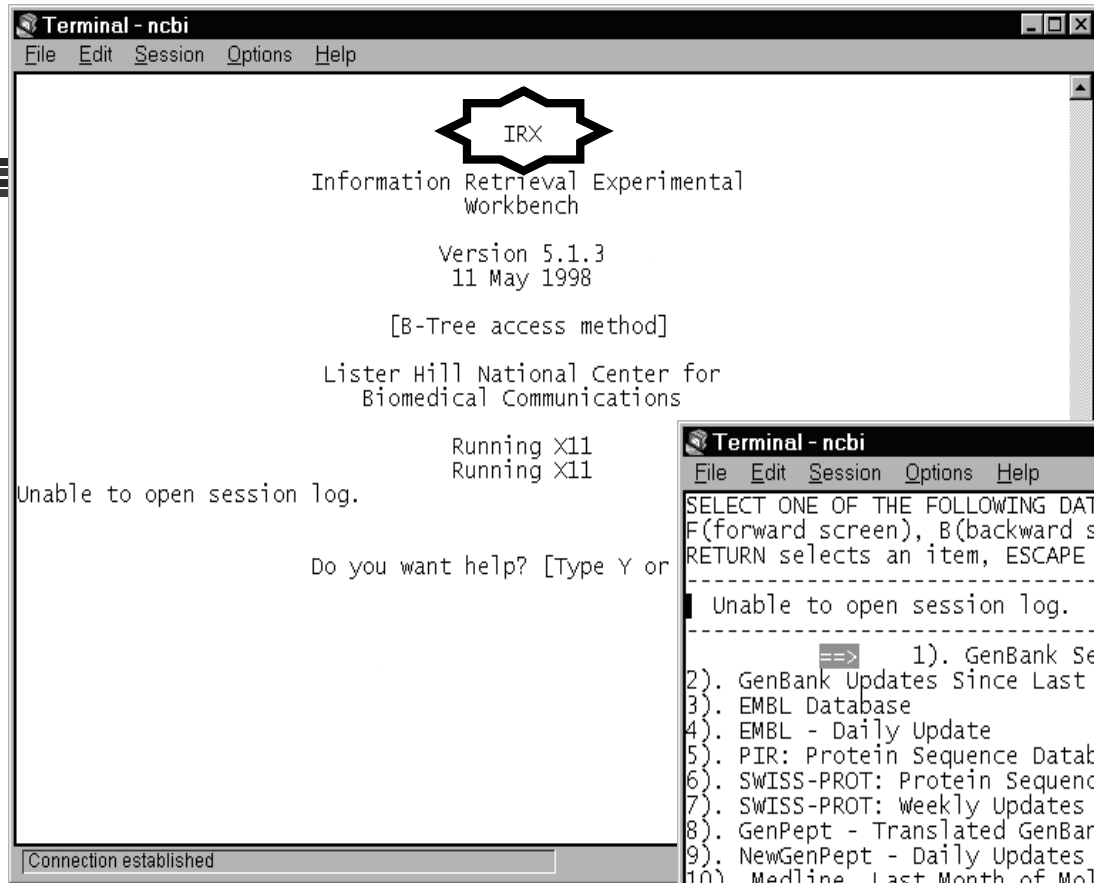
Computational Biology Branch

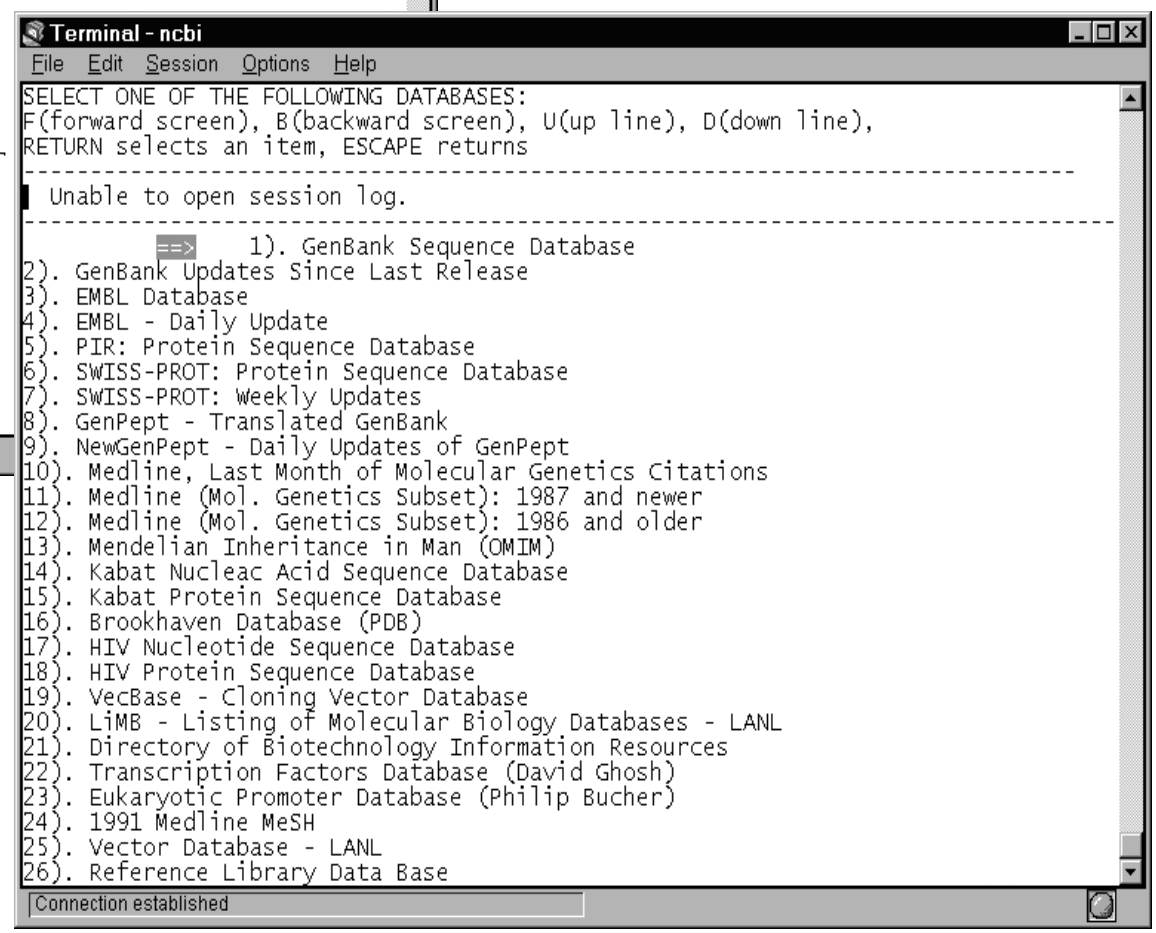
NCBI

# Information puzzles end abruptly?



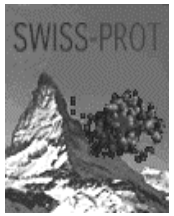



Terminal - ncbi
File Edit Session Options Help
IRX
Information Retrieval Experimental
Workbench
Version 5.1.3
11 May 1998
[B-Tree access method]
Lister Hill National Center for
Biomedical Communications
Running X11
Running X11
Unable to open session log.
Do you want help? [Type Y or
Connection established

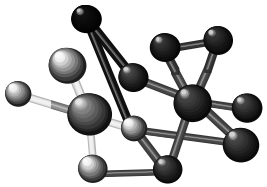

Terminal - ncbi
File Edit Session Options Help
SELECT ONE OF THE FOLLOWING DATABASES:
F(forward screen), B(backward screen), U(up line), D(down line),
RETURN selects an item, ESCAPE returns
Unable to open session log.
==> 1). GenBank Sequence Database
2). GenBank Updates Since Last Release
3). EMBL Database
4). EMBL - Daily Update
5). PIR: Protein Sequence Database
6). SWISS-PROT: Protein Sequence Database
7). SWISS-PROT: Weekly Updates
8). GenPept - Translated GenBank
9). NewGenPept - Daily Updates of GenPept
10). Medline, Last Month of Molecular Genetics Citations
11). Medline (Mol. Genetics Subset): 1987 and newer
12). Medline (Mol. Genetics Subset): 1986 and older
13). Mendelian Inheritance in Man (OMIM)
14). Kabat Nucleac Acid Sequence Database
15). Kabat Protein Sequence Database
16). Brookhaven Database (PDB)
17). HIV Nucleotide Sequence Database
18). HIV Protein Sequence Database
19). VecBase - Cloning Vector Database
20). LiMB - Listing of Molecular Biology Databases - LANL
21). Directory of Biotechnology Information Resources
22). Transcription Factors Database (David Ghosh)
23). Eukaryotic Promoter Database (Philip Bucher)
24). 1991 Medline MeSH
25). Vector Database - LANL
26). Reference Library Data Base
Connection established

# Vertical querying

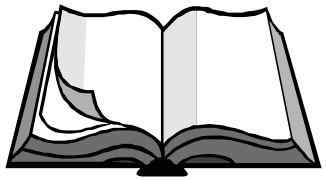
**GenBank**




**SWISS-PROT**

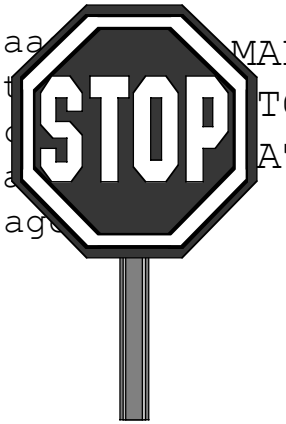


**PDB**

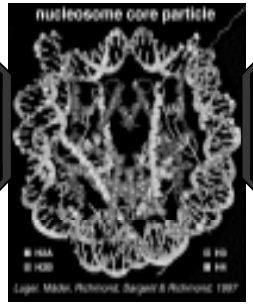


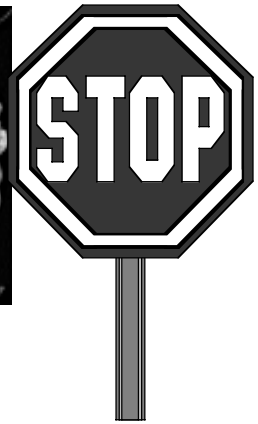
**MEDLINE**

```
atg gcc cga acc aa
act gct cgt aag t
ggt ggg aaa gcc c
aaa cag ctg gcc a
gcc gcc agg aaa agc
```




```
MARTKQTARI
TGGKAPRI
ATKAARKS
```



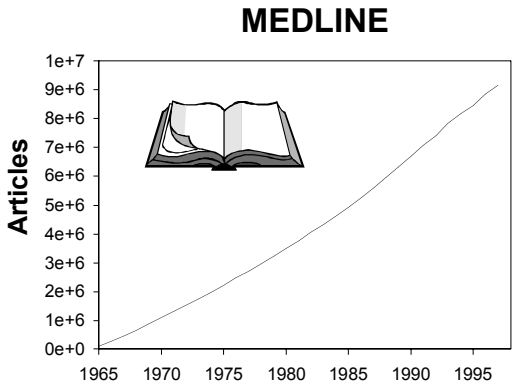



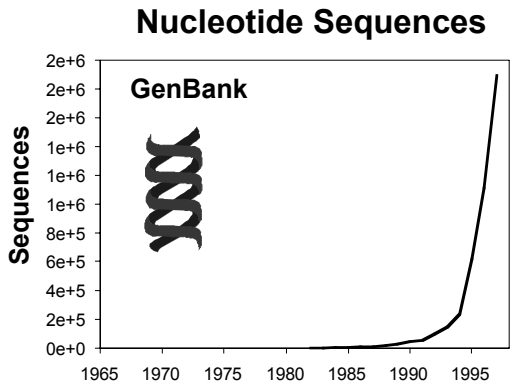

Nature
1997 Sep
18;389(6648):251-260

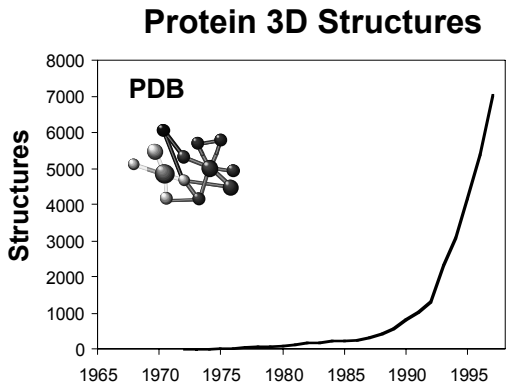
Crystal structure of the nucleosome core particle at 2.8 A resolution.

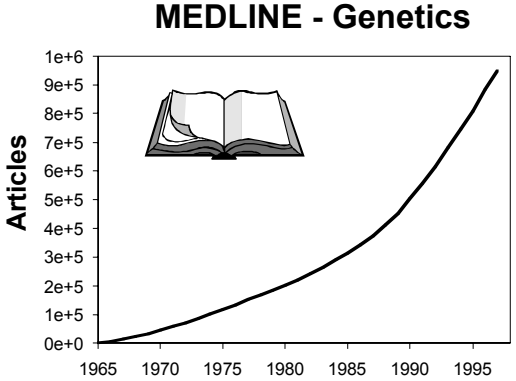
Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ

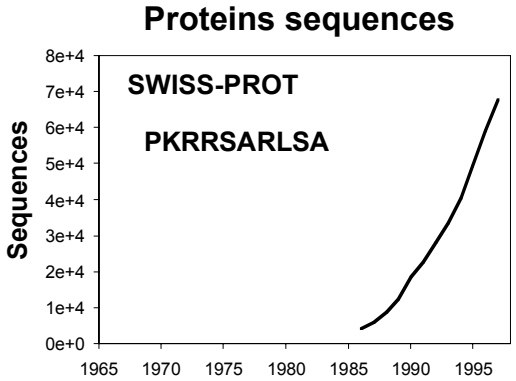
# The Biotechnology Information Explosion

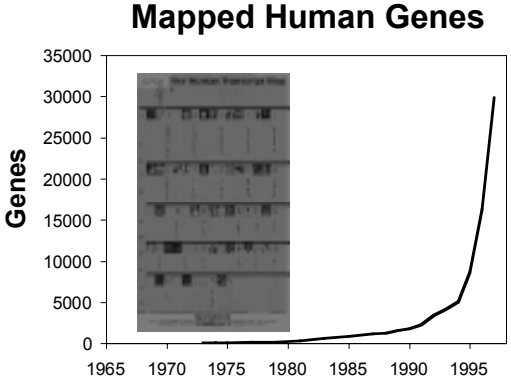


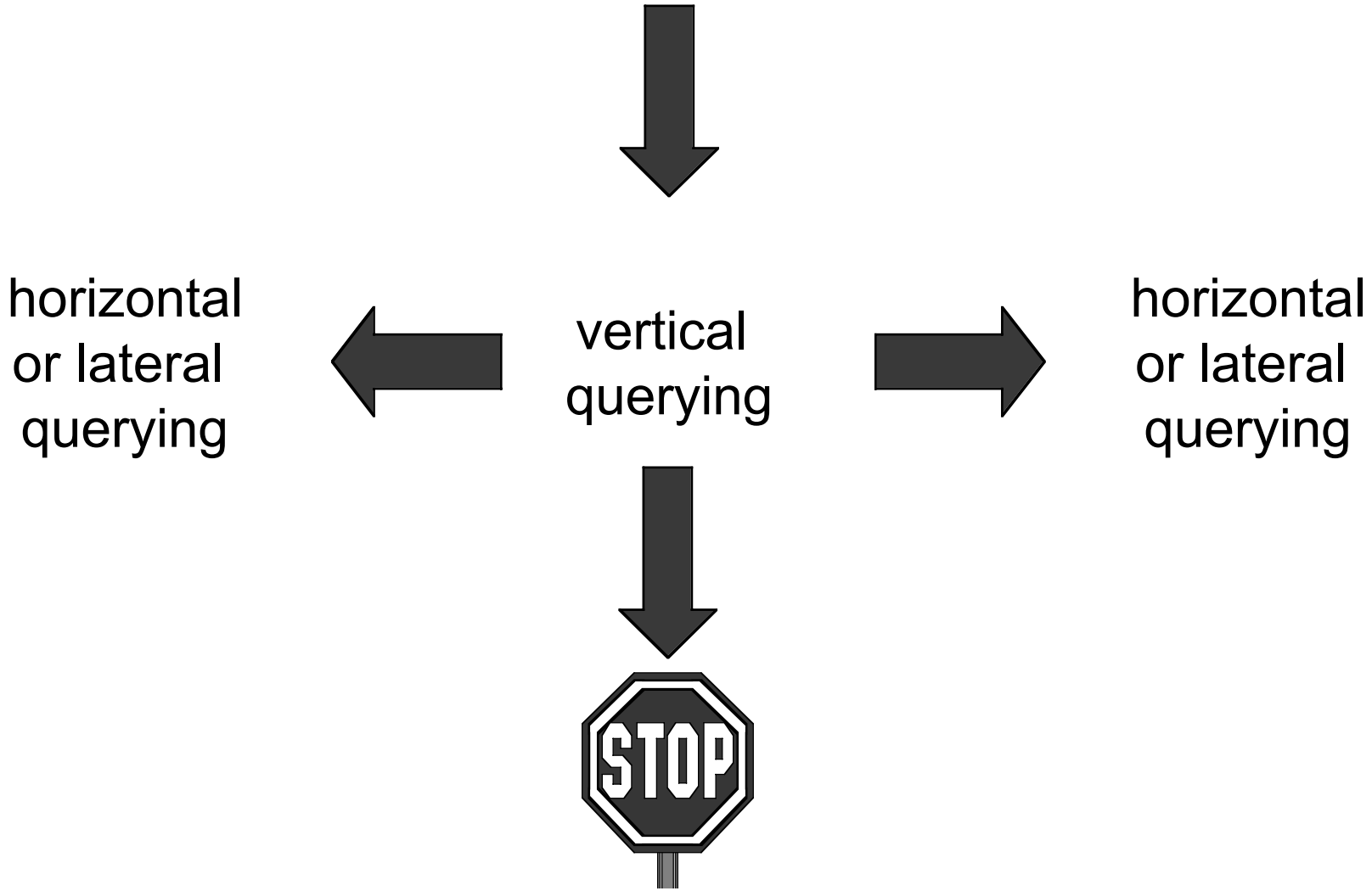



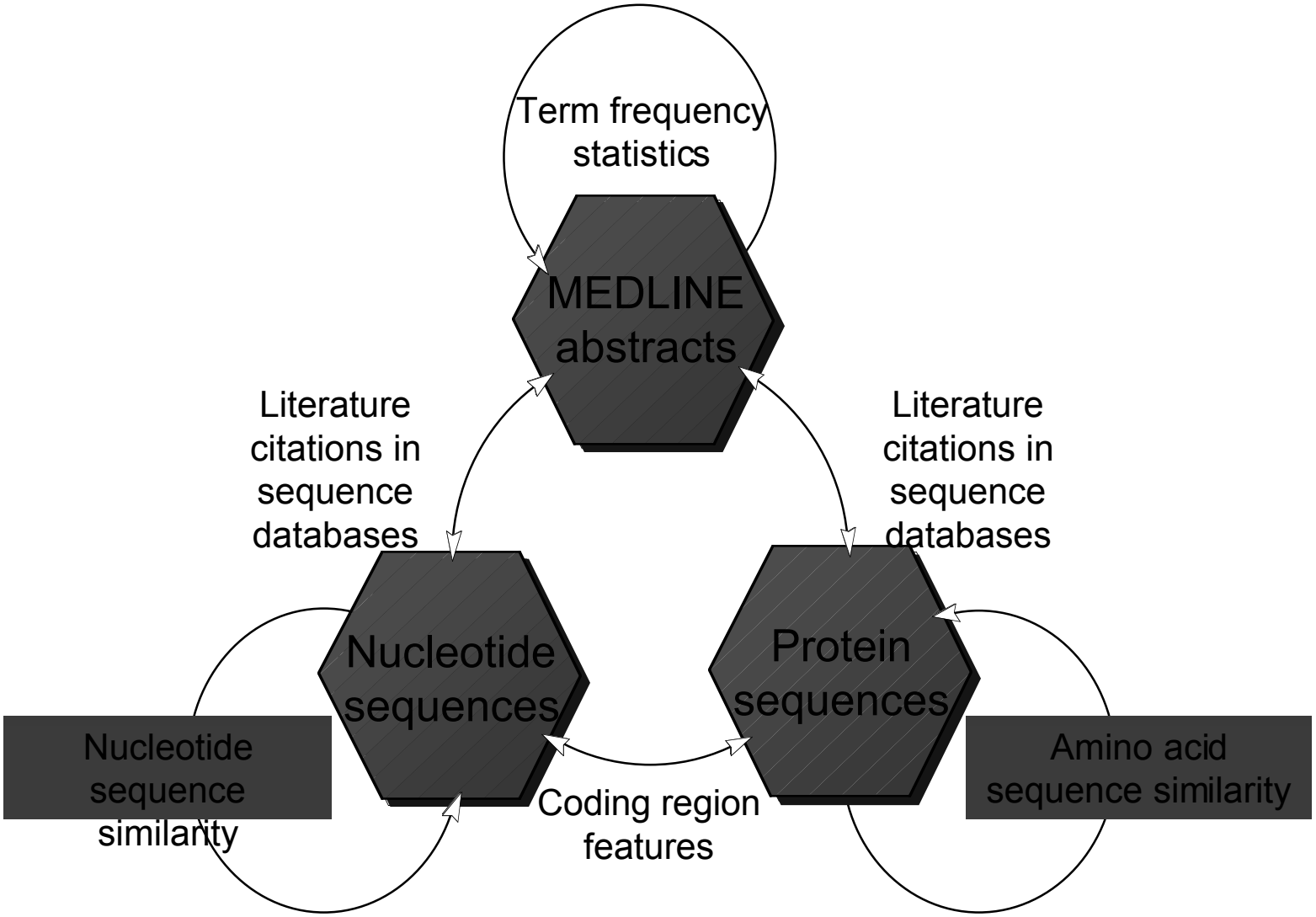



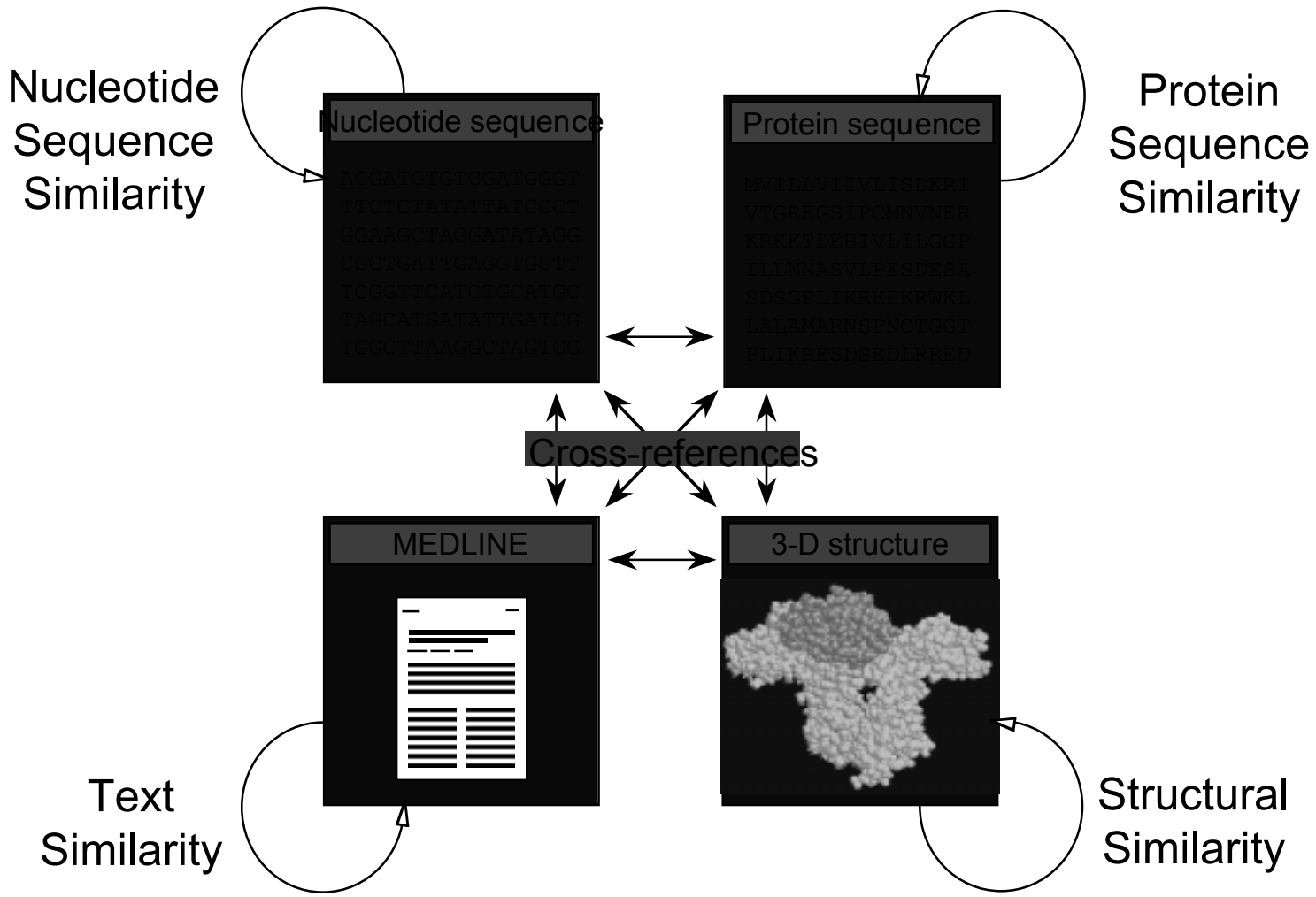
# Entrez and its databases



# Entrez (1992)


Term frequency statistics
MEDLINE abstracts
Literature citations in sequence databases
Literature citations in sequence databases
Nucleotide sequences
Protein sequences
Nucleotide sequence similarity
Amino acid sequence similarity
Coding region features

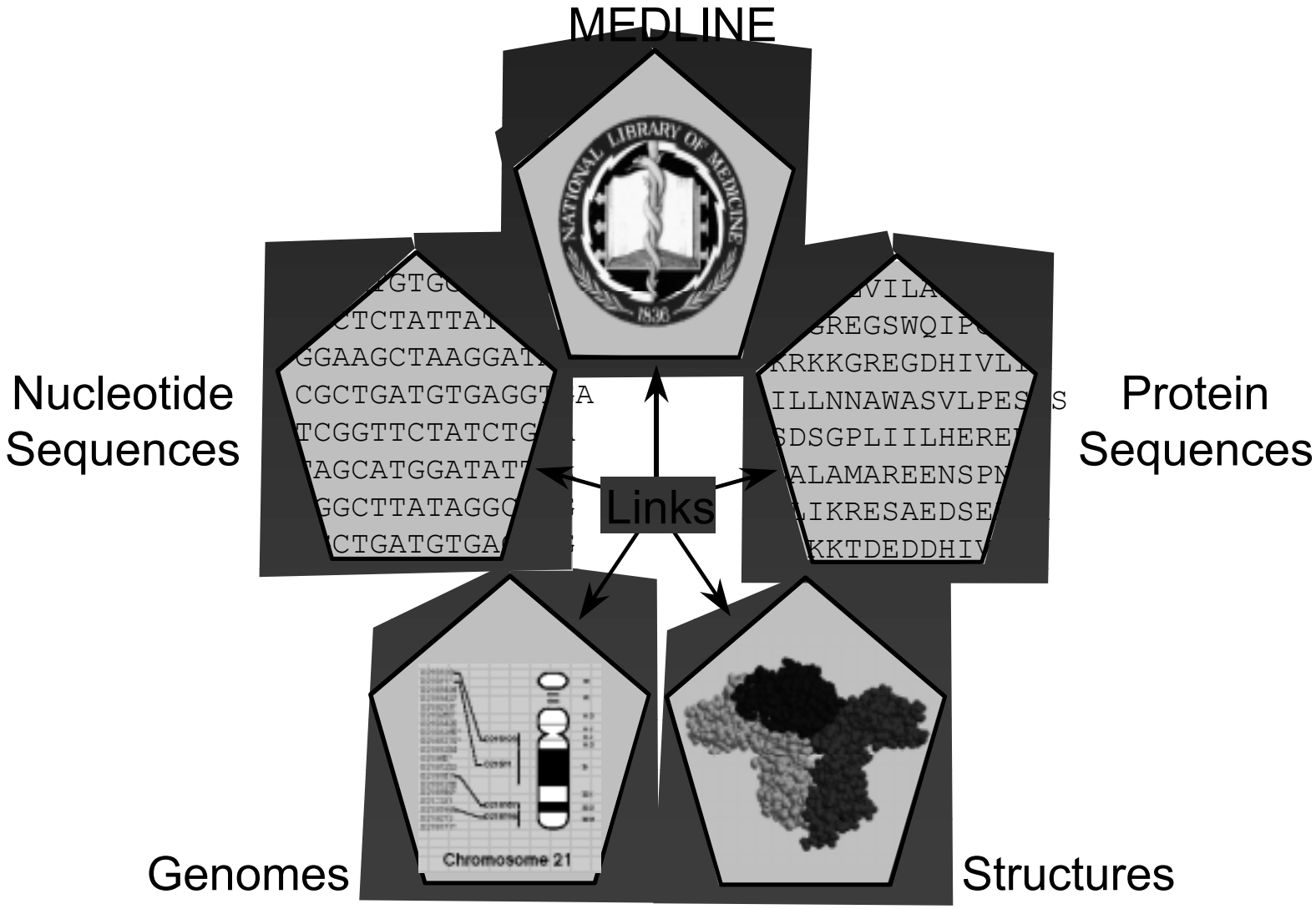
# Entrez (1994)


Nucleotide Sequence Similarity
Nucleotide sequence
ACGATGTGTCGATGGGT
TTCTCTATATTATCCCT
GGAAGCTAGGATATAGG
CGCTGATTGAGGTGGTT
TCGGTTCATCTGCATGC
TAGCATGATATTGATCG
TGGCTTAAGGCTAGTCG
Protein sequence
MVILLVIIVLISDKRI
VTGREGSIPCMNVNER
KRKKTDEHIVLILGGP
ILLNNASVLPESDESA
SDSGPLIKRKEKRWKL
LALAMARNSPNCTGGT
PLIKRESDSEDLRRED
Protein Sequence Similarity
Cross-references
MEDLINE
3-D structure
Text Similarity
Structural Similarity

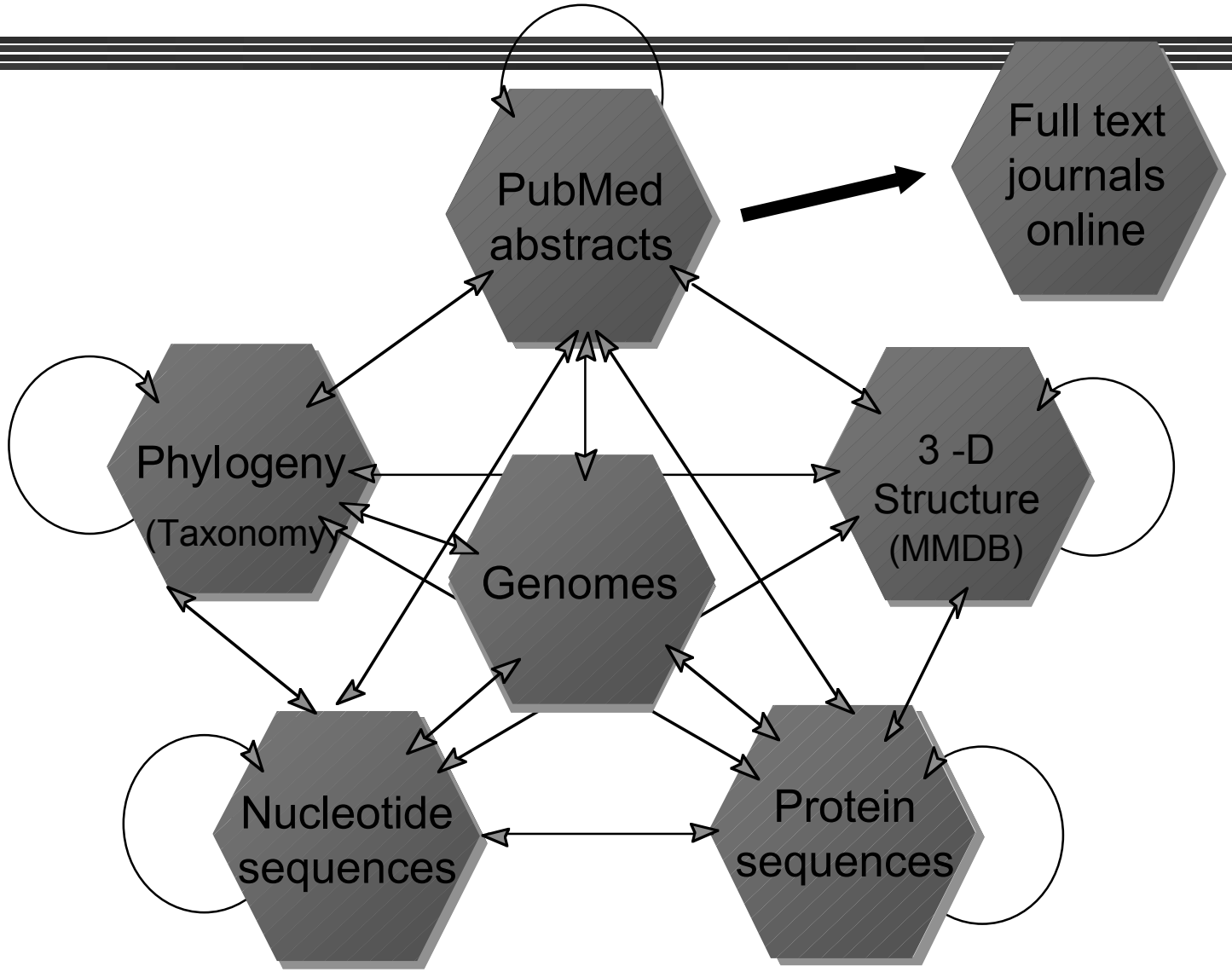
# Entrez BigMed Data Set (Replaced by PubMed)

- Version 27.58; April 22, 1997
- 1,392,146 MEDLINE entries
  - All citations indexed under the MeSH term “Genetics” and its descendants ( *i.e.* G5 tree)
  - Smaller subsets having to do with molecular sequences
  - Compare to total MEDLINE with ~ 8 million records
- 421,797 protein entries (“nr”)
- 1,273,697 nucleotide entries (“nr”)
- 4,873 structure entries (MMDB)
- 47,213 genome  entries
- linked journals: *J. Biol. Chem., PNAS, Science, J.M.B.,* ++++

# Entrez: Genotype to Phenotype (1996)



# Entrez Increases Discovery Space 1998



# Entrez Implementations

- CD-ROM subscriptions (stand-alone)
  - Impractical data currency
    - 6,000 new peer-reviewed articles per month
    - Sequence databases double ~20 months
  - Slow access speed
  - Capacity limited to 650 MB per CD-ROM

# Entrez Implementations

- **Network Entrez** (client-server)
  - Connect to dispatcher
  - Fastest Entrez implementation
  - Software installations and updates maintained by user
- **Web Entrez** (WWW client-server)
  - Universality of WWW
  - Can serve users without access to GUI environment (Lynx)
  - Linking and neighboring information can be expressed as hypertext
  - Ability to link to external data sources
  - WWW browsers are supported by third-party developers
  - Web implementation maintained by NCBI

# Entrez Implementations

## Network Entrez Version 6.6

Query

File Edit Options Misc

Database: MEDLINE | Field: Author Name | Mode: Selection | Accept

Term: Vincent A

Term Selection

Vincent A
Vincent A@
Vincent Ablazey
Vincent Ablazey M
Vincent Ablazey M
Vincent AC
Vincent AE

& serendipity
Vincent A

Retrieve 0 Doc

Document

File Edit Options Misc

Format: Summary Abstract Citation MEDLINE PubMed ID

Ruez C, et al. — Transcriptional control of Drosophila bicoid by Serendipity delta: cooperative binding sites, promoter context, and co-evolution. Mech Dev. 1998 Nov;78(1-2):125-34.

Ibnsouda S, et al. — Conservation of read-through transcription of the Drosophila serendipity genes during evolution is gratuitous. Mol Gen Genet. 1998 Sep;259(5):484-90.

Payre F, et al. — Two types of zinc fingers are required for dimerization of the serendipity delta transcriptional activator. Mol Cell Biol. 1997

Neighbor 101 | Target: MEDLINE | Prev

Refine 1 | Select: All None Parents | Next

## Web Entrez

Entrez Browser - Netscape

File Edit View Go Communicator Help

NCBI

Entrez Help

The Entrez Databases

Network Entrez

Retrieve large data sets

Making WWW Links to Entrez

NCBI Home

NLM Home

NIH Home

# Entrez

## Search WWW Entrez at NCBI

- Nucleotides
- Proteins
- 3D structures
- Genomes
- Taxonomy
- Literature - PubMed

The Entrez Browser is provided by the National Center for Biotechnology Information. NCBI also builds, maintains, and distributes the GenBank Sequence Database.

# NCBI's Home Page

The National Center for Biotechnology Information - Netscape

File Edit View Go Communicator Help

## National Center for Biotechnology Information

National Library of Medicine / National Institutes of Health

PubMed | Entrez | BLAST | BankIt | OMIM | Taxonomy | Structure

**Welcome to NCBI**

- What's New
  - Site Search
  - Organism-specific BLAST
- Programs and Activities
- NCBI Newsletter
- NCBI Exhibit Schedule
- Service Addresses
- User Comments

**GenBank Sequence Database**

- Overview
- Searching GenBank
- Submitting Sequences
  - BankIt
  - Sequin

**Database Services**

- PubMed MEDLINE
- Entrez Search System
- BLAST Sequence Similarity Searching
- E-mail Servers
- Anonymous FTP

**Other NCBI Resources**

- Cancer Genome Anatomy Project
- Gene Map of the Human Genome
- Genes and Disease
- UniGene: Unique Gene Sequence Collection for Human, Mouse, and Rat
- Human/Mouse Homology Maps
- HGSI: Human Genome Sequencing Index
- Clusters of Orthologous Groups
- OMIM: Online Mendelian Inheritance in Man
- dbEST: Database of Expressed Sequence Tags
- dbGSS: Database of Genome Survey Sequences
- dbSTS: Database of Sequence Tagged Sites
- dbSNP: Database of Single Nucleotide Polymorphisms
- Electronic PCR
- MMDB: Molecular Modelling Database
- NCBI Taxonomy
- ORF Finder
- Malaria Genetics and Genomics
- Retrovirus Resources

**Research**

- Research Projects
- Seminar Schedule
- Staff Bibliography
- Full Text of Selected Staff Publications

Revised March 11, 1999

www.ncbi.nlm.nih.gov

# Advanced PubMed querying

http://www.ncbi.nlm.nih.gov/PubMed/medline.html

Advanced PubMed Search - Netscape

File Edit View Go Communicator Help

Overview

Help / FAQs

New/Noteworthy

Clinical Alerts

Basic Search

Clinical Queries

Journal Browser NEW

MeSH Browser

Citation Matcher

Internet Grateful Med

NLM Home

NCBI Home

NIH Home

Credits

Restrictions on use

# PubMed

## Advanced MEDLINE Search

Search Field: All Fields Mode: Automatic

Affiliation
All Fields
Author Name
EC/RN Number
Entrez Date
Issue
Journal Name
Language
MeSH Major Topic
MeSH Terms
Page
Publication Date
Publication Type
Subheading
Substance Name
Text Word
Title Word
Volume
MEDLINE ID
PubMed ID

Search

- Enter o ... s.
- Author ... ed in the form Smith JB, but initials are optional.
- Journal ... full, as valid MEDLINE abbreviations, or as ISSN numbers (see Journal ... rmation).
- Use pu ... fy fields and search mode.
- Boolea ... also be entered directly in the search box, using AND, OR, and NOT 1 ... olean Search page for more information.

Questions or ... Help Desk.

NOTICE: The PubMed data are for personal use only. Users are responsible for complying with all copyright and licensing restrictions associated with data.

PubMed Clinical Queries - Netscape

File Edit View Go Communicator Help

NATIONAL LIBRARY OF MEDICINE

PubMed Home

Overview

Help

New/Noteworthy

NEW PubMed 2.0

Clinical Alerts

Basic Search

Advanced Search

Clinical Queries

Journal Browser

MeSH Browser - Netscape

File Edit View Go Communicator Help

NCBI PubMed MeSH Browser PubMed ?

Enter another MeSH ter

Histones [Detailed

Small chromoson
DNA in cell nuclei by io
based on the relative am

Note! The term 'Histone

Jump to: location 2

MeSH Tree Location 1:

Top of MeSH Tree
Amino Acids, Pepti
Proteins
Nuclear Pro
Histone

MeSH Tree Location 2:

Top of MeSH Tree
Amino Acids, Pepti
Proteins
Nucleoprot
Histone

Journal Database Browser - Netscape

File Edit View Go Communicator Help

NCBI Journal Database Browser Entrez ?

Found 123 journal(s)

Prev. matches

| Title |
| --- |
| PROGRESS IN FOOD AND NUTRITION SCIENCE |
| PROTEIN SCIENCE |
| QUARTERLY JOURNAL OF MICROSCOPICAL SCI |
| RESEARCH IN VETERINARY SCIENCE |
| ROGERIAN NURSING SCIENCE NEWS |
| SCANDINAVIAN JOURNAL OF MEDICINE AND SC |
| SCIENCE |
| SCIENCE AND JUSTICE |
| SCIENCE ET RECHERCHE ODONTOSTOMATOLOG |
| SCIENCE IN CHINA. SERIES B, CHEMISTRY, LIFE S SCIENCES |

Next matches

Enter the journal name, MEDLINE abbreviation or ISSN. Y
words in the title. Use an asterisk * at the end of a word to tr

All types science

Start

Comments and questions to the Help Desk
Credits: Andrei Smirnov

Document Done

Citation Matcher for Single Articles - Netscape

File Edit View Go Communicator Help

PubMed

Journal Browser

NLM Home Page

NCBI Home Page

NIH Home Page

PubMed

Citation Matcher for Single Articles

Enter information about the article you wish to find.

Journal: Science

Date:

Volume: Issue: First page:

Author's last name and initials (e.g., Smith BJ): Lowe T

Search Clear

Notes:

- You may omit any item if you wish
- Journal titles may be entered in full or as valid MEDLINE abbreviations
- For Date, you may enter yyyy, yyyy/mm, or yyyy/mm/dd. For example, 1998, 1998/03, or 1998/03/06
- Author names are automatically truncated to account for varying initials, e.g., smith j will also match on smith ja, smith jb, smith jc jr, etc. Enclose author names in double quotes to retrieve that exact match, e.g., "smith j"

Comments and questions to the Help Desk

Document Done

# Entrez Genomes Division

Entrez Browser - Netscape

File Edit View Go Communicator Help

NCBI

Entrez Help

The Entrez Databases

Network Entrez

Batch Entrez (Retrieve large data sets)

Making WWW Links to Entrez

NCBI Home

NLM Home

NIH Home

Credits

# Entrez

## Search WWW Entrez at NCBI

- Nucleotides
- Proteins
- 3D structures
- Genomes
- Taxonomy
- Literature - PubMed

The Entrez Browser is provided by the National Center for Biotechnology Information. NCBI also builds, maintains, and distributes the GenBank Sequence Database.

*Comments and questions to the NCBI help desk*

Entrez genomes - Netscape

File Edit View Go Communicator Help

NCBI

# Entrez Genomes

NCBI BLAST Nucleotides Proteins Structure Taxonomy PubMed Help

Search_for: All Fields

All Organisms

Prominent Organisms

Microbial genomes
- BLAST
- List of projects

Archaea
- Genome
- Plasmids

Bacteria
- Genome
- Plasmids

Eukaryota
- Genome
- Plasmids
- Organelles

Viruses
Viroids
Plasmids

## Prominent Organisms Taxonomy / List

**Complete Genomes**

- *Aquifex aeolicus*
- *Archaeoglobus fulgidus*
- *Bacillus subtilis*
- *Borrelia burgdorferi*
  - chromosome,
  - plasmids: cp26, cp9, lp17, lp25, lp28-1, lp28-2, lp28-3, lp28-4, lp36, lp38, lp54,
- *Chlamydia trachomatis*
- *Chlamydia pneumoniae* NEW
- *Escherichia coli*
- *Haemophilus influenzae*
- *Helicobacter pylori*
- *Helicobacter pylori J99*
- *Methanobacterium thermoautotrophicum*
- [3] *Methanococcus jannaschii*
  - chromosome
  - small extrachromosomal element
  - large extrachromosomal element
- *Mycobacterium tuberculosis*
- *Mycoplasma genitalium*
- *Mycoplasma pneumoniae*
- *Pyrococcus horikoshii*
- *Rickettsia prowazekii*
- [16] *Saccharomyces cerevisiae*
  chromosomes: I, II, III, IV, V, VI, VII, VIII, VIII, IX, X, XI, XII, XIII, XIV, XV, XVI,
- *Synechocystis PCC6803*
- *Treponema pallidum*

***Complete Sequences***

- *Leishmania major chromosome 1*
- *Plasmodium falciparum chromosome 2*

***Chromosome Maps***

- *[6] Caenorhabditis elegans*
  *chromosomes: I, II, III, IV, V, X,*
- *[5] Drosophila melanogaster*
  *chromosomes: 1, 2, 3, 4, X,*
- *[23] Homo sapiens*
  *chromosomes: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X,*
- *[21] Mus musculus*

Genomes - Netscape

File Edit View Go Communicator Help

NCBI

Entrez Genomes

BLAST Genomes Nucleotides Proteins Structure Taxonomy PubMed Help

Search by gene name:

CONTRIBUTING GENOME CENTERS:

U Wisc

OVERVIEW

DATA and ANALYSIS:

Graphical View

PROTABLE

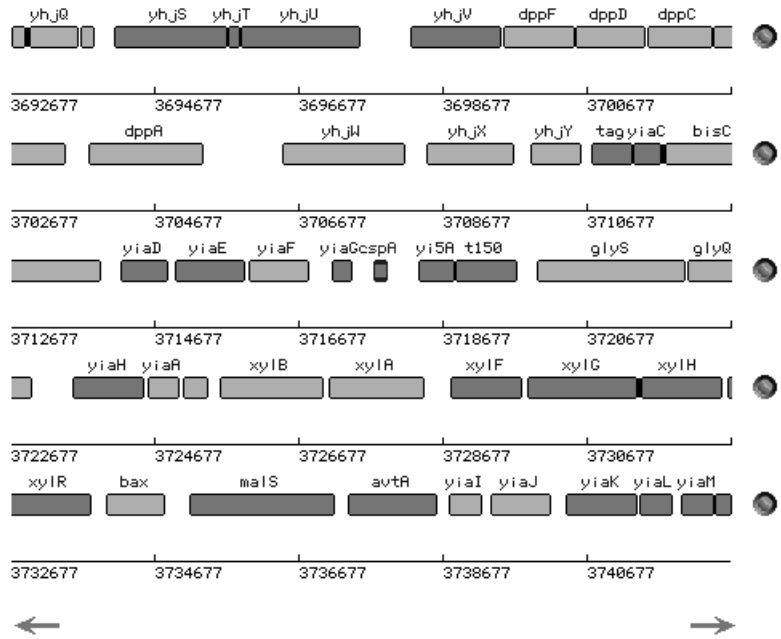
TAXTABLE

RNA Genes

GenBank FTP

BLAST WITH MICROBIAL GENOMES

## Escherichia coli K-12 MG1655 complete genome - 3692677..3742676




**Legend**

Find Open Reading Frames

Coding region on direct strand

Coding region on complementary strand

Site Name - Netscape

File Edit View Go Communicator Help

NCBI

# Entrez Genomes

NCBI Entrez Genomes BLAST Taxonomy Structure

Search by gene name: HHO1

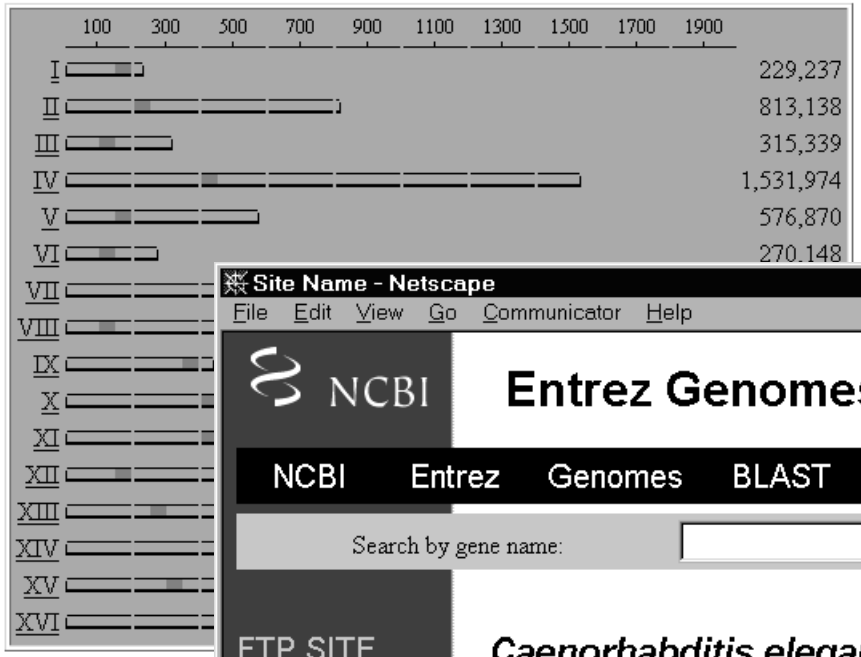
FTP site

MIPS

SGD

YPD

## *Saccharomyces cerevisiae* complete genome




Site Name - Netscape

File Edit View Go Communicator Help

NCBI

# Entrez Genomes

NCBI Entrez Genomes BLAST Taxonomy Structure

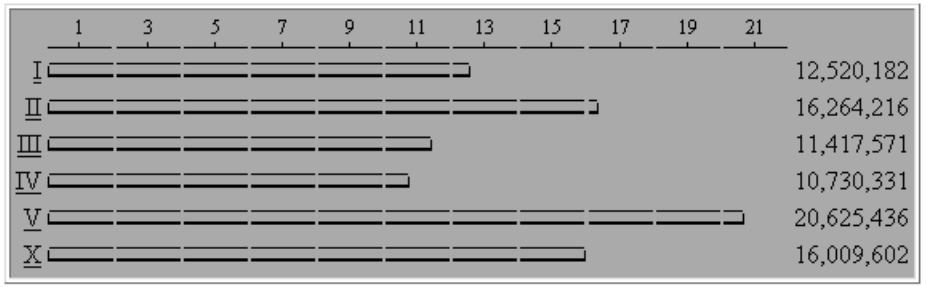
Search by gene name:

FTP SITE

Genome Sequencing:

Sanger Centre WashU

## *Caenorhabditis elegans* complete genome



# Structure searching ...

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

NCBI Structure Group - Netscape

File Edit View Go Communicator Help

NCBI Structure

PubMed Entrez BLAST OMIM Taxonomy Structure

Structure Links
NCBI Structure Group
What's New
MMDB
PDBeast
Cn3D
VAST
VAST Search
Research

Help Desk

## NCBI Structure Group

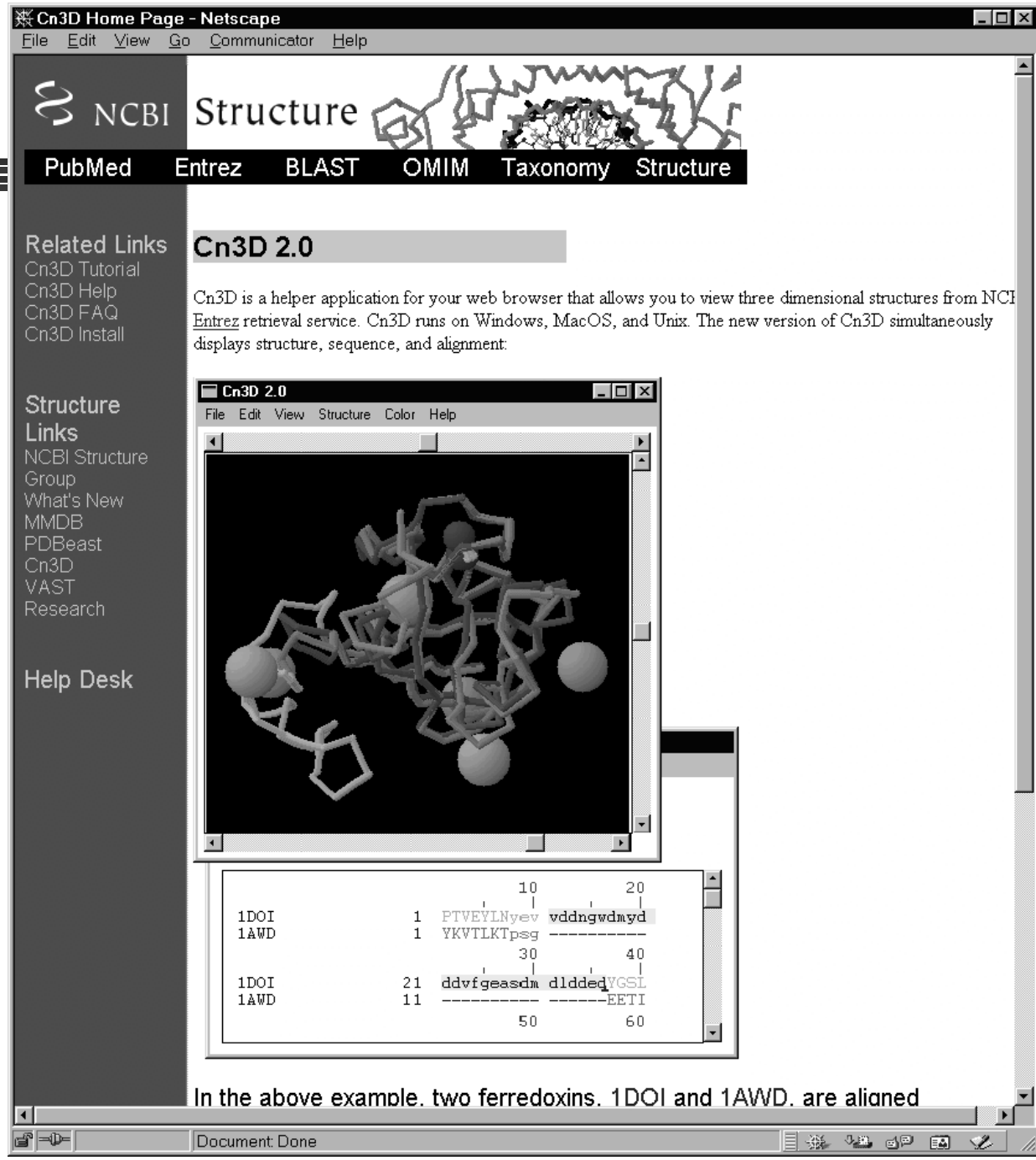
The Structure Group at NCBI has created an up-to-date database of 3-dimensional macromolecular structures along with a rich set of tools for comparing and visualizing structures and sequences, searching for similarities, and associating these data with other information sources such as taxonomy and literature databases.

The Molecular Modelling Database, MMDB, contains experimentally determined biopolymer 3D structures obtained from the Protein Data Bank (PDB). The structures contained in MMDB are stored in ASN.1 format, and make explicit representation of many structural features usually left only implicit by the PDB specifications. By making this information explicit, the extent of redundant and sometimes trivial computation required of structure comparison algorithms and molecular modeling packages is often greatly reduced. Cn3D is NCBI's 3D structure viewer specifically developed to read the structure files stored in MMDB. MMDB is fully integrated into Entrez, and in addition to the Cn3D structure viewer, provides users with a powerful search tool (VAST) for rapidly identifying similar 3-dimensional substructures.

In addition to the original structure files, MMDB maintains a pre-computed *n x n* record of "neighboring structures". All of the stored protein structures have been compared to each other with VAST (the Vector Alignment Search Tool), to identify similar 3-dimensional substructures. These neighbors often identify distant homologs. The associated structure alignments may be viewed in 3D by following the pre-computed "Structure Neighbor" links on MMDB's Structure Summary pages. There are currently two ways to retrieve a Structure Summary.

The Entrez interface to MMDB allows you to perform a text-based search to locate 3D structure files containing the keywords entered:

**Search MMDB text:** 1AAB Find Clear


Cn3D Home Page - Netscape
File Edit View Go Communicator Help
NCBI Structure
PubMed Entrez BLAST OMIM Taxonomy Structure
Related Links
Cn3D Tutorial
Cn3D Help
Cn3D FAQ
Cn3D Install
Structure Links
NCBI Structure Group
What's New
MMDB
PDBeast
Cn3D
VAST
Research
Help Desk
Cn3D 2.0
Cn3D is a helper application for your web browser that allows you to view three dimensional structures from NCI
Entrez retrieval service. Cn3D runs on Windows, MacOS, and Unix. The new version of Cn3D simultaneously
displays structure, sequence, and alignment:
Cn3D 2.0
File Edit View Structure Color Help
10 20
1DOI 1 PTVEYLNyev vddngwdmyd
1AWD 1 YKVTLKTpsg ----------
30 40
1DOI 21 ddvfgeasdm dlddedYGSL
1AWD 11 ---------- ------EETI
50 60
In the above example, two ferredoxins, 1DOI and 1AWD, are aligned
Document: Done

MMDB Structure Summary - Netscape

File Edit View Go Communicator Help

NCBI *MMDB* STRUCTURE SUMMARY Entrez ?

MMDB Id: 4475 PDB Id: 1AAB

**Protein Chains:** (single chain)
**MEDLINE:** PubMed
**Taxonomy:** Rattus norvegicus

**PDB Authors:** C.H.Hardman, R.W.Broadhurst, A.R.C.Raine, K.D.Grasser, J.O.Thomas & E.D.Laue
**PDB Deposition:** 28-Oct-95
**PDB Class:** Dna-Binding
**PDB Title:** Nmr Structure Of Rat Hmg1 Hmga Fragment

**Sequence Neighbors:** (single chain)
**Structure Neighbors:** (single chain)

View / Save Structure NEW *Get Cn3D 2.0 Now!*

| Options: | Viewer: | Complexity: | |
|---|---|---|---|
| Launch Viewer | Cn3D v2.0 (asn.1) | Cn3D Subset | Up to 5 Models |
| See File | Cn3D v1.0 (asn.1) | Virtual Bond Model | Up to 10 Models |
| Save File | Mage | All Atom Model | All Models |
| | RasMol (PDB) | | |

Help MMDB, Cn3D, Viewing Options, VAST, PDB, NCBI Structure

Vast Results - Netscape

File Edit View Go Communicator Help

NCBI *VAST* STRUCTURE NEIGHBORS Entrez ?

# Structures similar to MMDB 4475, 1AAB

Nmr Structure Of Rat Hmg1 Hmga Fragment

View / Save Alignments NEW *Get Cn3D 2.0 Now!*

**Options:**
- (•) Launch Viewer
- ( ) See File
- ( ) Save File

**Viewer:**
- (•) Cn3D v2.0 (asn.1)
- ( ) Mage (Kinemage)
- ( ) (PDB)

**Complexity:**
- ( ) Aligned Chains only
- (•) All Chains
- (•) Alpha Carbons only
- ( ) All Atoms

**Structure neighbors 1-3 out of 3 displayed. Page 1 of 1.**

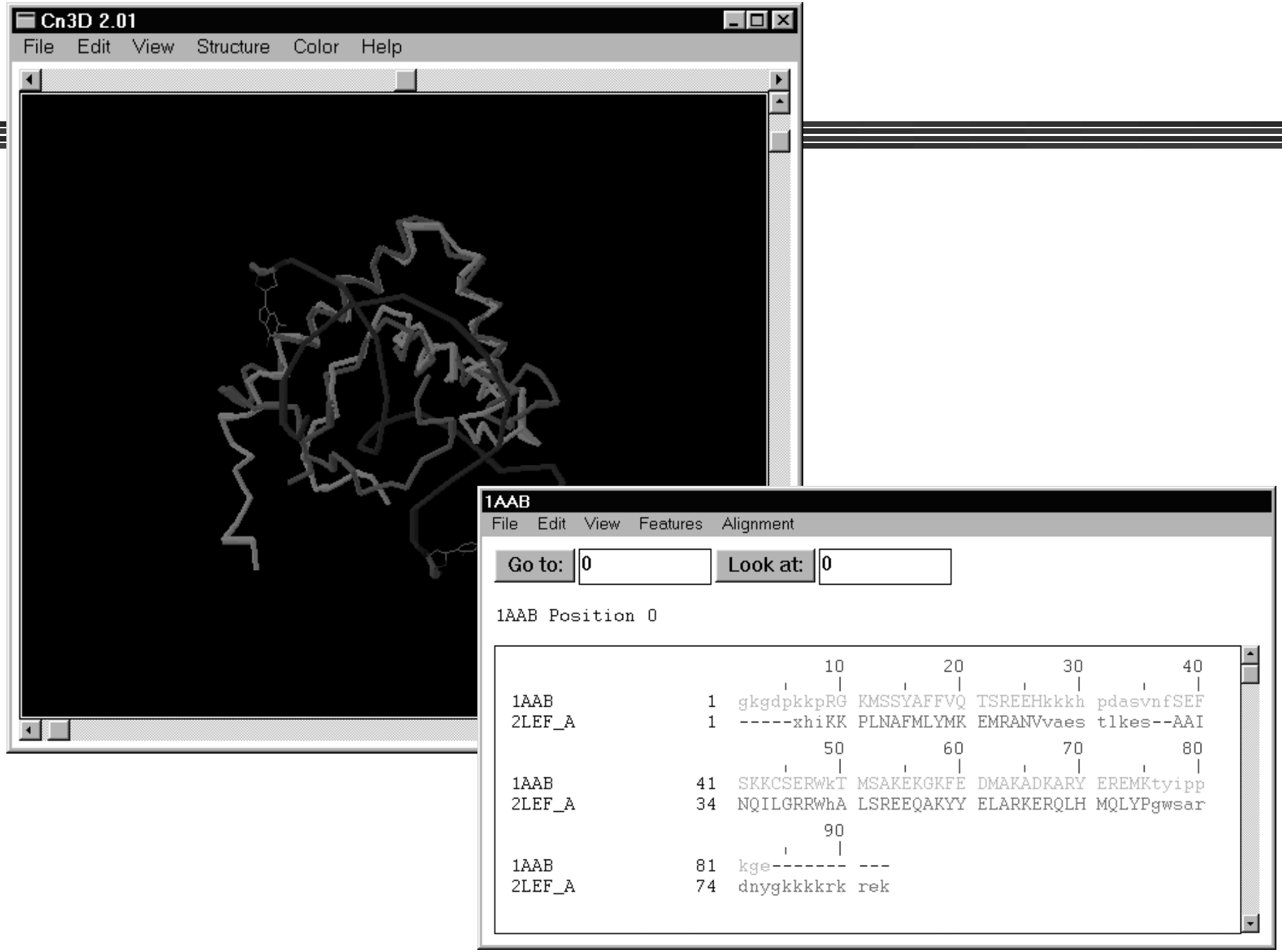
| | PDB C D | RMSD | NRES | %Id | Description |
|---|---|---|---|---|---|
| ☐ | 1HJO A 3 | 2.6 | 43 | 14.0 | Heat-Shock 70kd Protein 42kd Atpase N-Terminal Domain |
| ☐ | 1A81 C 5 | 1.8 | 26 | 3.8 | Crystal Structure Of The Tandem Sh2 Domain Of The Syk Kinase Bound To A Dually Tyrosine-Phosphorylated Itam |
| ☑ | 2LEF A | 1.7 | 55 | 12.7 | Lef1 Hmg Domain (From Mouse), Complexed With Dna (15bp), Nmr, 12 Structures |

Display / Sort Hits page number: 1 Hits to display per page: 20 choose between 20-100 neighbors per page.

**Display Subset:**
- (•) Non-redundant; BLAST p-value 10e-7
- ( ) Non-redundant; BLAST p-value 10e-40
- ( ) Non-redundant; BLAST p-value 10e-80
- ( ) Non-identical sequences
- ( ) All of MMDB

**Sorted by:**
- (•) VAST Score
- ( ) VAST P-value
- ( ) Rmsd
- ( ) Aligned residues
- ( ) Identities

**Column Format:**
- (•) RMSD, NRES, %Id
- ( ) All values


Cn3D 2.01
File Edit View Structure Color Help
1AAB
File Edit View Features Alignment
Go to: 0
Look at: 0
1AAB Position 0
10 20 30 40
1AAB 1 gkgdpkkpRG KMSSYAFFVQ TSREEHkkkh pdasvnfSEF
2LEF_A 1 -----xhiKK PLNAFMLYMK EMRANVvaes tlkes--AAI
50 60 70 80
1AAB 41 SKKCSERWkT MSAKEKGKFE DMAKADKARY EREMKtyipp
2LEF_A 34 NQILGRRWhA LSREEQAKYY ELARKERQLH MQLYPgwsar
90
1AAB 81 kge------- ---
2LEF_A 74 dnygkkkkrk rek