The National Library of Medicine

August 3, 1956 Senators Lister Hill (D-Alabama) and John F. Kennedy (D-Massachusetts) sponsored the National Library of Medicine Act (Public Law 84-941) which placed the Armed Forces Medical Library under the Public Health Service and renamed it the National Library of Medicine.

October 22, 1965 The Medical Library Assistance Act of 1965 (Public Law 89-291) was signed into law, authorizing NLM’s extramural programs of grant assistance.

August 3, 1968 Public Law 90-456 authorized the establishment of the Lister Hill National Center for Biomedical Communications as the research arm of the NLM. (Building 38 was dedicated December 1961.)

Representative Claude Pepper (D-Florida) and Senator Lawton Chiles (D-Florida) introduced legislation that established the National Center for Biotechnology Information within the Department of Health and Human Services as a component of the National Library of Medicine. Congress finally passed the bill which was signed into law by President Reagan on November 4, 1988.

The legislation mandated that the National Center for Biotechnology Information:

- Establish public databases
- Develop software tools for sequence analysis
- Perform research in computational biology
- Disseminate biomedical information

Selected Moments in NCBI History:

1990 BLAST
1992 GeneBank moves to NCBI
1993 GenBank/Entrez on CD-ROM
1994 GenBank/Entrez version 1.0
1995 OMIM version & wwwBLAST
1997 PubMed
1999 dbSNP & new NCBI webpage design
2000 RefSeq
2001 NLM Logo
2002 Human Genome Sequence declared “complete”
2004 Gene
2005 MyNCBI
2006 PubChem
NCBI Web Traffic

Christmas and New Year's Days

US

Internet Users

World Internet Users

Taxonomy
Genomes
Nucleotides
Genes
Expression
Proteins
Structures
Compounds

A Few Key Databases
**Databases & Search Tools**

- **Text Entrez**
- **Sequence BLAST**
- **Protein Structure VAST**
- **Small Molecule Structure PubChem**
- **Structure Search**

---

**Essential Components of Entrez Databases**

- Each record is stored in ASN.1 format, which can be easily displayed in a number of formats as requested.

- All Molecular Database entries are organized by organism (Taxonomy Database).

- Each record is assigned a UID, a “unique integer identifier” for internal tracking.

- Each record is indexed by data fields: [author], [title], [organism], and many others.

- Each record is given a Document Summary, a summary of the record's content (DocSum) shown in list form.

- Each record is manually or computationally assigned links to biologically related UIDs in and across databases.
Searching the NCBI Databases

How to Query a Database

Example query:

- Organism
- Journal
- User compounds
- Author

Examples of tag delimiters:

1. term1 term2
2. Brauninger a c-src kinase

Boolean operators MUST be in ALL CAPS!

op = AND, OR, NOT
# Finding Fields & Indexing Terms

<table>
<thead>
<tr>
<th>Accession</th>
<th>All Fields</th>
<th>Author</th>
<th>EC/RN Number</th>
<th>Feature Key</th>
<th>Filter</th>
<th>Gene Name</th>
<th>Journal</th>
<th>Keyword</th>
<th>Modification Date</th>
<th>Molecular Weight</th>
<th>Organism</th>
<th>Page Number</th>
<th>Primary Accession</th>
<th>Protein Name</th>
<th>Publication Date</th>
<th>SeqID String</th>
<th>Sequence Length</th>
<th>Substance Name</th>
<th>Text Word</th>
<th>Title</th>
<th>Volume</th>
</tr>
</thead>
</table>

Most useful search field: [Organism] human OR bacteria

Useful search terms in [Properties]:
- srcdb: "source database" srcdb_refseq
- gbdiv: "genbank division" gbdiv_est
- biomol: "biomolecular type" biomol_mrna

**EXAMPLE: Searching with Entrez**

**Finding information on Sickle Cell Anemia...**

- "How do I retrieve the mRNA record NM_000311?"
- "How do I find a information on genes expressed in mouse pancreas?"
- "How do I get a structure of the HIV-1 Reverse transcriptase?"

**EXAMPLE: Searching with Entrez**

- "How do I download the human proteome?"
Searching the NCBI Databases

Why do we need sequence similarity searching?

- To identify and annotate sequences with...
  - incomplete (or no) annotations (GenBank)
  - incorrect annotations
- To assemble genomes
- To explore evolutionary relationships by...
  - finding homologous molecules
  - developing phylogenetic trees

NOTE: Similar sequences may NOT have similar function!
Basic Local Alignment Search Tool

- Calculates similarity for biological sequences
- Finds best local alignments:
  
  Detects exact matches for “words” and then extends them as far as they can reasonably match.
  
  Provides statistics about the likelihood of finding a sequence match by random chance (“e-value”).

EXAMPLE: Searching with BLAST

“How do I find out what my human mRNA sequence is?”

“How do I find a human homolog for my mouse protein sequence?”
EXAMPLE: Searching with BLAST

“How do I find out which bacteria is causing the infection with an amplified 16S rRNA sequence?”

Nucleotide > Somewhat similar sequences (BLASTn)
Query: the 16S rRNA sequence
DB: “Others”
 blastp: Bacteria[organism] AND rRNA[properties]
Program: Somewhat similar sequences (BLASTn)
[You can also go to the organism-specific genomic BLAST page]

“How do I find out what protein sequence my mRNA might encode?”

Nucleotide > Somewhat similar sequences (BLASTn)
Query: primer1-(N10-15)-primer2
DB: “human genomic + transcript”
Program: Somewhat similar sequences (BLASTn)
[You can also go to the organism-specific genomic BLAST page]

Searching the NCBI Databases

Text
Entrez
Sequence
BLAST
Structure
VAST
Small Molecule
Structure
PubChem
Structure
Search

in-lieu-of BLAST

Precomputed BLASTn&p Services

If you don’t want to run a BLAST search...
We’ve already done some for you!

• Precomputed Entrez Links: “Related Sequences”
• The Transcript Cluster Database: “UniGene”
• The Protein Homolog Database: “HomoloGene”
• The “Conserved Domain” Database & CDART a.k.a. “Domain Relatives”
• The Multifunctional Blast Link: “BLink”
EXAMPLE: Finding Pre-Computed BLAST Data

"How do I find in which rainbow trout tissues G6PD is expressed?"

"How do I find mammalian homologs for MYC?"

"How do I find functional regions in Shiga toxin?"

Searching the NCBI Databases

Text \( \text{Entrez} \) Sequence \( \text{BLAST} \) Structure \( \text{VAST} \) Small Molecule Structure \( \text{PubChem Structure Search} \)

Structure-based Neighbors:
Vector Alignment Search Tool

For each protein chain:
- locate secondary structure elements,
- represent them as individual vectors,
- and compare these with precomputed vectors of database structures.
EXAMPLE: Using VAST to find binding sites

“T have a methyltransferase with a weird fold bound to the substrate s-adenosylmethionine. Are there any other proteins that have this fold?”

The Crystal Structure Of E. coli m1G37 tRNA Methyltransferase (1P9P)
& A protein of unknown function from Thermotoga maritima (1O6D)
- 2.4Å rmsd
- 11% sequence identity

EXAMPLE: Using PubChem Structure Search

“We isolated a compound from grape skins. What is it and has it been found to have any biological activity?”
**Searching the NCBI Databases**

**Linking!**

**Following Links**

Follow links to related data in the same database or in others!

*“Hard” Links: Curated links based on biology for example:*
- nucleotide → taxonomy (based on organism identifier)
- protein → domain relatives (based on domain assignment)
- domains → pubmed (based on supporting literature)

*“Soft” Links: Pre-computed analyses for example:*
- nucleotide → related sequences (BLASTn neighbors)
- protein → conserved domains (RPS-BLAST search)
- gene → map viewer (map position of annotated gene)

**Examples of Soft Linking in Entrez**

- Term Frequency Statistics
- Phylogeny
- Taxonomy
- Genomes
- mmdbs (3D structure)
- Structure Similarity (VAST)
- Sequence Similarity (BLASTn)
Summary pages of curated information (LINKS!) for genetic loci of RefSeq organisms.

¾ Summary
¾ Genomic regions, transcripts and products
¾ Genomic context
¾ Bibliography
¾ HIV-1 protein interactions
¾ Interactions
¾ General gene information
¾ Markers
¾ Genotypes
¾ Phenotypes
¾ Pathways
¾ Homology
¾ GeneOntology (function, process, component)
¾ General protein information
¾ Reference Sequences
¾ mRNA and Protein(s)
¾ Reference and Alternate Genomic assemblies
¾ Related Sequences
¾ Additional Links

Free Courses & Course Materials!
- Field Guide: a general review course of all NCBI
- Field Guide Plus: a 2-day in-depth course
- Medical Library Association Courses: 2 courses for librarians
- MiniCourses: short, targeted example-driven courses
  1. Making Sense of DNA and Protein Sequences
  2. Unmasking Genes in Human DNA
  3. Identification of Disease Genes
  4. Correlating Disease Genes and Phenotypes
  5. BLAST Quick Start
  6. EntrezGene Quick Start
  7. Structure Analysis Quick Start
  8. ALIGNER Quick Start
  9. GenBank Quick Start
  10. Entrez Quick Start
  11. Microbial Genomes Quick Start

- Structure Course: structural analysis with Cn3D & VAST
- PubChem Course: small molecule information & analysis
- PowerTools: PowerScripting: a 4-day course to automate search and retrieval via scripting
The Goal @ NCBI

Not only to provide access to the data with a way to find it, but also to help users answer a question/solve a puzzle!

-In many cases,...
the data is already there.

For More Information...
E-mail addresses:

General Help     info@ncbi.nlm.nih.gov
BLAST Help       blast-help@ncbi.nlm.nih.gov

Help Line:       (301) 496-2475     (8:30am - 5:30pm EST)

The (free!) NCBI Newsletter
http://www.ncbi.nih.gov/About/newsletter.html

The NCBI Handbook
Follow the link from the NCBI Home Page under "Hot Spots"

The NCBI Education Page