NCBI Molecular Biology Resources
A Field Guide
part 1

April 5, 2004

The National Center for Biotechnology Information

Created in 1988 as a part of the National Library of Medicine at NIH

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

Number of Users and Hits Per Day

Currently averaging 10,000,000 to 35,000,000 hits per day!

Countries of Origin

U.S. 40%
Japan 6%
Italy 4%
Canada 3%
Germany 3%
United Kingdom 3%
Netherlands 2%
Spain 2%
Brazil 2%
Sweden 1%
Switzerland 1%
Belgium 1%
Other 14%

Literature Databases

PubMed
National Library of Medicine
OMIM
Online Mendelian Inheritance in Man
Types of Databases

- **Primary Databases**
  - Archival in nature
  - Original submissions by experimentalists
  - Database staff review and may organize the data, but NCBI does not add/modify information
  - Submitters retain editorial control of records
  - Examples: GenBank, SNP, GEO

- **Derivative Databases**
  - Curated by NCBI staff by hand, computer or both from primary data
  - NCBI retains editorial control of records
  - Record content is updated continually

**GenBank**

- Nucleotide only sequence database
- Archival in nature
- Submission of GenBank Data to NCBI
  - Direct submissions of individual records via Web (BankIt, Sequin)
  - Batch submissions of bulk sequences via Email (EST, GSS, STS)
  - FTP accounts for Sequencing Centers
The International Sequence Database Collaboration

NIH

Entrez

NCBI

GenBank

EMBL

DDBJ

EMBL

Entrez

SRS

NIG

getentry

CIB

NIH

• Submissions
• Updates

Sequence Records (millions)

Total Base Pairs (billions)

Release 140: 32.5 million records
37.9 billion nucleotides
Average doubling time ~ 12 months

'83 '84 '85 '86 '87 '88 '89 '90 '91 '92 '93 '94 '95 '96 '97 '98 '99 '00 '01 '02 '03 '04

GenBank

Release 140 February 2004
32,549,400 Records
37,893,844,733 Nucleotides
>140,000 Species
143 Gigabytes 591 files

• full release every two months
• incremental and cumulative updates daily
• available only through internet

Organization of GenBank: GenBank Divisions (gbdiv)

Records are divided into 17 Divisions.
- 1 Patent (11 files)
- 5 High Throughput
- 11 Traditional

**BULK Divisions:**
- Batch Submission (Email and FTP)
- Inaccurate
- Poorly characterized

**PRI** (27) Primate
**PLN** (10) Plant and Fungal
**BCT** (8) Bacterial and Archaeal
**INV** (6) Invertebrate
**ROD** (11) Rodent
**VRL** (3) Viral
**VRT** (4) Other Vertebrate
**MAM** (1) Mammalian (ex. ROD and PRI)
**PHQ** (1) Phage
**SYN** (1) Synthetic (cloning vectors)
**UNA** (1) Unannotated

Finding Nucleotide Records: Accession Numbers

- GenBank
- DDBJ
- EMBL
- RefSeq NCBI
- TPA NCBI
- PDB

Summary Brief
- ASN.1
- FASTA
- GenPept
- GI list

*Other links:*
- Data Links
- Gene Link
- Primer Link
- Nucleotide Links
- Protein Links
- Structured Link
- Supporting Links
Each sequence is represented by a text record called a flat file.

- GenBank/GenPept (useful for scientists)
- FASTA (the simplest format)
- ASN.1 & XML (useful for programmers)
GenBank & GenPept Files

FASTA Format
Abstract Syntax Notation: ASN.1

```
Seq-entry ::= set {
  level 1,
  class desc {
    title "Medicago sativa glucose-6-phosphate
    and translated products",
    source {
      org {
        taxname "Medicago sativa",
        source {
          db {
            tag id 56147
          }
          orgname {
            name binomial {
              genus "Medicago",
              species "sativa",
              subspecies "subsp. sativa"
            }
          }
        }
      }
    }
  }
}
```

Redundancy in the Primary Databases

- **Primary Databases**
  - Raw and redundant Data...submitted and “owned” by experimentalists
    - Examples: GenBank, SNP, GEO

- **Derivative Databases**
  - Human-curated (compilation and correction of data)
    - Examples: LocusLink, OMIM & Literature databases
  - Computationally-Derived
    - Example: UniGene
  - Combination
    - Examples: RefSeq, Genome Assembly
Primary vs. Derivative Sequence Databases

NCBI's Derivative Sequence Database

GenBank

RefSeq

genomes  transcripts  proteins

RELEASE 4 IS NOW AVAILABLE ON THE FTP SITE!

- Forming the “best representative” sequence
- Standardizing nomenclature and record structure
- Adding annotation (references, sequence features)
RefSeq Curation Processes

Curated genomic DNA (NC, NT, NW)

Curated Model mRNA (XM) → Model protein (XP)

Curated mRNA (NM) → Protein (NP)

Sequence & Genome

Curated RefSeq Records

<table>
<thead>
<tr>
<th>LOCUS</th>
<th>ADSS</th>
<th>1348 bp mRNA</th>
<th>linear</th>
<th>PRI 27-MAY-2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locus</td>
<td>ADSS</td>
<td>1348 bp mRNA</td>
<td>linear</td>
<td>PRI 27-MAY-2002</td>
</tr>
<tr>
<td>RefSeq Nucleotide</td>
<td>ADSS</td>
<td>1348 bp mRNA</td>
<td>linear</td>
<td>PRI 27-MAY-2002</td>
</tr>
<tr>
<td>RefSeq Protein</td>
<td>ADSS</td>
<td>455 aa</td>
<td>linear</td>
<td>PRI 27-MAY-2002</td>
</tr>
</tbody>
</table>

COMMENT

This record has been curated by NCBI staff. The reference sequence was derived from X66503.1.

Summary: Adenylosuccinate synthetase catalyzes the first committed step in the conversion of IMP to AMP.

LOCUS | ADSS | 1348 bp mRNA | linear | PRI 27-MAY-2002 |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Definiton</td>
<td>Homo sapiens adenylosuccinate synthase (ADSS), mRNA.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accession</td>
<td>NM_001126</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Version</td>
<td>NM_001126.1 GI:4557270</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

RefSeq Nucleotide

<table>
<thead>
<tr>
<th>LOCUS</th>
<th>ADSS</th>
<th>455 aa</th>
<th>linear</th>
<th>PRI 27-MAY-2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definiton</td>
<td>Adenylosuccinate synthetase (Aden(-)H-complementing) Homo sapiens</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accession</td>
<td>NP_001117</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Version</td>
<td>NP_001117.1 GI:4557271</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DBSOURCE | REFSEQ: accession NM_001126.1 |

Third Party Annotation (TPA) Database

NCBI now accepts the submission of new annotations of existing GenBank sequences.

• Submissions must be published in a peer-reviewed journal.
• Facilitates the annotation of sequences by experts.

Examples of seq

- Annotation of features on gene and/or mRNA sequences
- Assembled "full length" genes and/or mRNAs

What should not be submitted to TPA:

- Synthetic constructs (such as cloning vectors) that use well-characterized, publicly available genes, promoters, or terminators
- Updates or changes to existing sequence data
- Sequence annotations without experimental evidence

Example 1:

<table>
<thead>
<tr>
<th>LOCUS</th>
<th>F80261</th>
<th>1954 bp mRNA</th>
<th>linear</th>
<th>PRI 13-OCT-2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definiton</td>
<td>Homo sapiens adenylsuccinate synthetase (F80261), sequence annotation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accession</td>
<td>F80261.1 GI:5240978</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example 2:

<table>
<thead>
<tr>
<th>LOCUS</th>
<th>F80261</th>
<th>1954 bp mRNA</th>
<th>linear</th>
<th>PRI 13-OCT-2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definiton</td>
<td>Homo sapiens adenylsuccinate synthetase (F80261), sequence annotation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accession</td>
<td>F80261.2 GI:5240978</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example 3:

<table>
<thead>
<tr>
<th>LOCUS</th>
<th>F80261</th>
<th>1954 bp mRNA</th>
<th>linear</th>
<th>PRI 13-OCT-2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definiton</td>
<td>Homo sapiens adenylsuccinate synthetase (F80261), sequence annotation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accession</td>
<td>F80261.3 GI:5240978</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example 4:

<table>
<thead>
<tr>
<th>LOCUS</th>
<th>F80261</th>
<th>1954 bp mRNA</th>
<th>linear</th>
<th>PRI 13-OCT-2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definiton</td>
<td>Homo sapiens adenylsuccinate synthetase (F80261), sequence annotation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accession</td>
<td>F80261.4 GI:5240978</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**UniGene:**
Sets of expressed sequences clustered by BLAST similarity

Summary pages of curated information about expressed gene transcripts.

---

**A Cluster of ESTs:**
Arabidopsis serine protease

Distribution of EST Blast Hits on the Query Sequence

- 5' EST hits
- 3' EST hits

---
### UniGene Collections

**as of February 2004**

#### Chordata
- **Mammalia**
  - *Bos taurus* (cow)
  - *Homo sapiens* (human)
  - *Mus musculus* (mouse)
  - *Rattus norvegicus* (rat)
  - *Sus scrofa* (pig)
- **Aves**
  - *Gallus gallus* (chicken)
- **Amphibia**
  - *Silurana tropicalis* (western clawed frog)
  - *Xenopus laevis* (african clawed frog)
- **Actinopterygii**
  - *Danio rerio* (zebra fish)
  - *Oncorhynchus mykiss* (rainbow trout)
  - *Oryzias Latipes* (japanese rice fish)
  - *Salmo salar* (salmon)
- **Ascidiacea**
  - *Ciona intestinalis* (sea squirt)
- **Arthropoda Insecta**
  - *Anopheles gambiae* (malaria mosquito)
  - *Drosophila melanogaster* (fruit fly)
  - *Bombyx mori* (silkworm)
- **Mycetozoa**
  - *Dictyosteliid* *Dictyostelium discoideum* (slime mold)
- **Chlorophyta Chlorophyceae**
  - *Chlamydomonas reinhardii*
- **Apicomplexa Coccidia**
  - *Toxoplasma gondii*
- **Embryophyta Cycadopsida**
  - *Pinus taeda* (loblolly pine)
- **Bryopsida**
  - *Physcomitrella patens*
- **Eudicotyledons**
  - *Arabidopsis thaliana* (thale cress)
  - *Glycine max* (soybean)
  - *Lycopersicon esculentum* (tomato)
  - *Medicago truncatula* (barrel medic)
  - *Solanum tuberosum* (potato)
  - *Vitis vinifera* (wine grape)
  - *Helianthus annuus* (sunflower)
- **Liliopsida**
  - *Hordeum vulgare* (barley)
  - *Oryza sativa* (rice)
  - *Saccharum officinarum* (sugar cane)
  - *Sorghum bicolor* (sorghum)
  - *Triticum aestivum* (bread wheat)
  - *Zea mays* (corn)

---

**Sequence & Expression**

[Diagram of Gene Expression](#)
### Expression

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM827: FHCMV-T-1</td>
<td>Source: CMV infected fibroblasts</td>
</tr>
<tr>
<td>GSM825: FHCMV-T-2</td>
<td>Source: CMV infected fibroblasts</td>
</tr>
<tr>
<td>GSM828: FHCMV-T-3</td>
<td>Source: CMV infected fibroblasts</td>
</tr>
<tr>
<td>GSM829: FHCMV-H-1</td>
<td>Source: CMV infected fibroblasts</td>
</tr>
<tr>
<td>GSM830: FHCMV-H-2</td>
<td>Source: CMV infected fibroblasts</td>
</tr>
<tr>
<td>GSM831: FHCMV-H-3</td>
<td>Source: CMV infected fibroblasts</td>
</tr>
<tr>
<td>GSM832: CMV_AD169-2</td>
<td>Source: CMV infected fibroblasts</td>
</tr>
<tr>
<td>GSM833: CMV_AD169-3</td>
<td>Source: CMV infected fibroblasts</td>
</tr>
</tbody>
</table>

### Comparison of gene expression profiles of HFF cells infected with CMV strains

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE177: CMV infection of HFF cells</td>
<td>Comparison of gene expression profiles of HFF cells infected with CMV strains</td>
</tr>
</tbody>
</table>
Linking Protein Sequence, Structure and Function

- **Protein**
  - Protein sequences

- **Domain**
  - **CDD**: Conserved functional domains in proteins represented by a PSSM
    - PSI-BLAST, RPS-BLAST, CDART

- **Structure**
  - **MMDB**: Experimentally-derived 3D structure records from PDB

- **3D Domain**
  - Compact structural domains of protein folds

Sequence-based Neighbors: Domain Neighbors

- **Domains**
  - Conserved sequence elements that perform common functions

  Curation of protein multiple sequence alignments with known similar function by conversion to Position-Specific Scoring Matrices

  - “Conserved Domain Database” (CDD) Search
  - “Reverse Position Specific” Sequence Comparisons (RPS-BLAST)
**Sequence-based Neighbors:**

**Domain Relatives**

**Modular Architecture of Domains (CDART)**

- "Conserved Domain Architecture Retrieval Tool"
- Cartoon descriptions of protein sequences based on domain organization.
- Comparison with other proteins with the same Domain.

**Structure**

**Entrez Structure:**

**Molecular Modeling Database**

- Derived from experimentally determined PDB records
- Data is added to PDB records including:
  - Addition of explicit chemical bonding information
  - Validation and indexing of sequence
  - Inclusion of Taxonomy, Citation, and other information
  - Conversion to ASN.1 data description language
- Searching the Structure Databases:
  - Keyword search by Entrez
  - Sequence search by BLAST or BLink
  - Domain search by CDD/RPS-BLAST
  - Structure search by VAST

**Structure**
Structure Summary Page

CDD: Sequence, Structure, Function

Searching the NCBI Databases

VAST Structure Neighbors
For each protein chain:
- locate secondary structure elements,
- represent them as individual vectors,
- and compare these with precomputed vectors of database structures.

Human IL-4
Submitting a PDB File to VAST

- Choose the file format
- Remove all lines except ATOM

This is also the best way to convert PDB files to MMDB format for viewing with Cn3D!

Searching the NCBI Databases

- Entrez
- BLAST
- Structure VAST

Entrez, The Life Sciences Search Engine

Sequence Databases

- DDBJ/EMBL/GenBank
- PDB
- Protein Data Bank
- Protein Information Resource
How to Query a Particular Database

(term1 [tag delimiter] op term2 [tag delimiter] op ...)

- op = AND, OR, NOT
- Boolean operators MUST be in ALL CAPS!
- tag delimiter = Entrez indexing field

Examples of tag delimiters:
Organism, Journal, User compounds, Author

Sample Protein Query

Brauninger a c-src kinase

(brauninger a [Author] AND c-src kinase [All Fields])

Improving Queries with Tag Delimiters

Human adenylsuccinate synthetase

("Homo sapiens" [Organism] OR human [All Fields]) AND adenylsuccinate synthetase [All Fields]

Human [Organism] AND adenylsuccinate synthetase

("Homo sapiens" [Organism] AND adenylsuccinate synthetase [All Fields])
The Protein Limits Page

Accession
All Fields
Author
EC/RN Number
Filter
Gene Name
Issue
Keyword
Modification Date
Molecular Weight
Organism
Page Number
Primary Accession
Properties
Protein Name
Publication Date
SeqID String
Sequence Length
Substance Name
Text Word
Uid
Volume
Genomic
Mitochondrion
Chloroplast

RefSeq
GenBank
EMBL
DDBJ
PDB
SWISS-PROT
PIR
PRF

Improving Queries with the Limits Page

("Homo sapiens" [Organism]
AND cytochrome c oxidase [All Fields]
AND srcdb_refseq [Properties])

Complex searches you can do with Preview/Index

Terms used (and indexed) in Entrez fields
can be searched to gain useful information!

How many rat UniGene clusters contain at least one mRNA?

1) Select the UniGene database.
2) Find all the rat records.
3) Find those that have \( \geq 1 \) mRNAs. ("NOT")

UniGene is an experimental system for automatically partitioning GenBank sequences 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 tissues in which the gene has been expressed and its location.
Complex Queries with Preview/Index

The (ever expanding) Entrez System

Other Advanced Queries

Nucleotide: Non-genomic sequences from the PLN division of Genbank
gbdiv_pln [properties] NOT biomol_genomic [properties]

Protein: RefSeq sequences with molecular weights of 80 to 100 kDa
srcdb_refseq [properties] AND 080000:100000 [Molecular Weight]

SNP: True SNPs that are uniquely mapped on the mouse genome
Snp [SNP Class] AND 1 [Map Weight] AND mouse [organism]

UniSTS: Markers on the Genethon map of human chromosome 12
Genethon [Map Name] AND human [organism] AND 12 [chromosome]

Structure: Structures of bacterial kinases with resolutions below 2 Å
Bacteria [organism] AND kinase AND 000.00:002.00 [resolution]
Ming the NCBI Databases:

LINKING!

Following Links

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

**Soft Links:** Pre-computed analyses
- nucleotide \(\rightarrow\) related sequences (BLAST neighbors)
- protein \(\rightarrow\) domains (CDD/RPS-BLAST search)
- gene \(\rightarrow\) map viewer (map position of annotated gene)

**Hard Links:** Curated links based on biology
- nucleotide \(\rightarrow\) taxonomy (based on organism identifier)
- protein \(\rightarrow\) domain relatives (based on domain assignment)
- domains \(\rightarrow\) PubMed (based on supporting literature)
Examples of Database Integration in Entrez

- PubMed
- Phylogeny
- Taxonomy
- mmdb
- VAST
- Nucleotide sequences
- Protein sequences
- BLASTn
- BLASTp
- Word weight
- mmdb (3D structure)
- Phylogen
- Taxonomy
- Genomes
- mmdb
- VAST
- Nucleotide sequences
- Protein sequences
- BLASTn
- BLASTp
- Word weight

Exploring the Molecular Basis of Hemochromatosis via Links

- NCBI
- OMIM
- Examples of Database Integration in Entrez

OMIM Record → Link to Gene
Gene → Links to Everywhere (almost)

- Entrez Gene
- LocusLink → SNPs → Protein
- Protein Record → Domain (CDD) Search
CDD Record: IGc1 Domain Sequence → Structure!

FTP Downloads

Help for Programmers

E-Utilities: Guidelines for “URL calls” that provide access to data.
Designed for use in scripts.
• Examples: ESearch, EPost, ESummary, EFetch and ELink

NCBI Toolbox: In-house source code useful for incorporating NCBI-like functionality into their programs.
Three main parts: Data Model, Data Encoding and Programming Libraries.
• Examples: Cn3D, BLAST, Sequin, Data format conversion scripts

Caution: Overuse may result in blocked IPs!
Intermission

To come in Part 2:

- Genomes & Genomic Resources
- Searching Sequences with BLAST