NCBI Molecular Biology Resources

A Field Guide

Part 1

March 2, 2006
University of Medicine and Dentistry of New Jersey

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

NCBI Web Traffic

Users per day

World Internet Users

US Internet Users

Christmas and New Year's Day
NCBI FieldGuide

Literature Databases

PubMed

NCBI Books

PubMed Central

OMIM

Catalog

Journals

Part 1. The Databases

Part 2. Data Flow and Processing

Part 3. Querying and Linking the Data

Part 4. User Support

A part of the NCBI Bookshelf
The (ever expanding) Entrez System

What is Entrez?

- A system of 29 linked databases
- A tool for finding biologically linked data
- A text search and retrieval engine
- A virtual workspace for manipulating large datasets

Entrez Databases

- Each record is assigned a UID.
  - A "unique integer identifier" for internal tracking
- All Molecular Database entries are organized by organism (Taxonomy Database).
- 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)
- Each record is assigned links to biologically related UIDs.
Examples of Database Integration at NCBI

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

Following Links

NCBI FieldGuide
Examples of Database Integration at NCBI

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

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

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

NCBI FieldGuide
Examples of Database Integration at NCBI
Types of Molecular Databases

- **Primary Databases**
  - Raw and redundant data submitted, "owned" and updated by experimentalists
  - Examples: GenBank, SNP, GEO, PubChem Substance & BioAssay

- **Derivative Databases**
  - *Human-curated* (compilation and curation of data)
    - Examples: GEO Datasets, Structure & Literature databases
  - *Computationally-Derived*
    - Example: UniGene, HomoloGene, PubChem Compound
  - *Combination*
    - Examples: RefSeq, Gene, Genome Assembly, Conserved Domain and Structure databases

---

### Primary vs. Derivative Sequence Databases

- **Primary Databases** (updated continually by NCBI)
  - GenBank
    - Updated only by submitters
  - RefSeq
  - UniGene

- **Derivative Databases**
  - Curators
  - Labs
  - Sequencing Centers
  - Algorithms
  - Genome Assembly
  - Updated continually by NCBI
**GenBank**

- Nucleotide only sequence database
- Archival in nature
- Each record is assigned a stable accession number
- 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
- Three collaborating databases and other sources of data

---

**The International Sequence Database Collaboration**

- NIH
- EMBL
- NCBI
- EBI
- DDBJ
- CIB
- SRS
- Entrez

---

**GenBank**

<table>
<thead>
<tr>
<th>Release 152</th>
<th>February 2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>54,584,635</td>
<td>Records</td>
</tr>
<tr>
<td>59,750,386,305</td>
<td>Nucleotides</td>
</tr>
<tr>
<td>&gt;140,000</td>
<td>Species</td>
</tr>
<tr>
<td>216 Gigabytes</td>
<td>908 files</td>
</tr>
</tbody>
</table>

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

### GenBank Divisions

<table>
<thead>
<tr>
<th>Code</th>
<th>Division</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRI</td>
<td>Primate</td>
</tr>
<tr>
<td>ROD</td>
<td>Rodent</td>
</tr>
<tr>
<td>PLN</td>
<td>Plant and Fungal</td>
</tr>
<tr>
<td>BCT</td>
<td>Bacterial/Archeal</td>
</tr>
<tr>
<td>VRT</td>
<td>Other Vertebrate</td>
</tr>
<tr>
<td>INV</td>
<td>Invertebrate</td>
</tr>
<tr>
<td>VRL</td>
<td>Viral</td>
</tr>
<tr>
<td>MAM</td>
<td>Mammalian</td>
</tr>
<tr>
<td>PHG</td>
<td>Phage</td>
</tr>
<tr>
<td>SYN</td>
<td>Synthetic</td>
</tr>
<tr>
<td>UNA</td>
<td>Unannotated</td>
</tr>
</tbody>
</table>

### Traditional
- Direct Submissions (Sequin/Bankit)
- Accurate (~1 error per 10,000 bp)
- Well characterized
- Organized by taxonomy

### Bulk
- From sequencing projects
- Batch submissions (ftp/email)
- Inaccurate
- Poorly characterized
- Organized by sequence type

### Derivative Sequence Database
- The curated "best representative" sequences
- Standardized nomenclature and record structure
- Added annotation (references, sequence features)

### RefSeq Curation Processes

- Curated Model mRNA (XM) → Model protein (XP)
- Curated mRNA (NM) → Protein (NP)
- Curated genomic DNA (NC, NT, NW)
Curated RefSeq Records

**LOCUS** ADSS
**DEFINITION** Homo sapiens adenylosuccinate synthase (ADSS), mRNA
**ACCESSION** NM_001126
**VERSION** NM_001126.1
**GI** 4557270
**RefSeq Nucleotide**

**COMMENT** REVIEWED: 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
**DEFINITION** Homo sapiens adenylosuccinate synthetase (Adenylsuccinate synthetase [Adenylsuccinate-synthesizing]) Homo sapiens
**ACCESSION** NP_001117
**VERSION** NP_001117.1
**GI** 4557271
**RefSeq Protein**

X records: Genome Annotation & Inferred or Predicted vs N records: Provisional, Reviewed or Validated

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 ORFs
- Assembled 'full' transcripts

What should not:
- 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

“Best representative” (reference) sequences
- Standardized nomenclature and record structure
- Added annotation (references, sequence features)

Mapping Genome Data on an Assembly:
(RefSeq: NC, NT, NW)
Transcript regions & ORFs (RefSeq: NM/NP, XM/XP)
Markers (STS)
Polymorphisms (SNP)
ESTs/Exons (UniGene)

• "Best representative" (reference) sequences
• Standardized nomenclature and record structure
• Added annotation (references, sequence features)
Organelles:
- Mitochondria (816)
- Plastids (50)
- Plasmids (871)
- Nucleomorphs (3)

- Viruses (2286)
- Archaeabacteria (26)
- Eubacteria (285)
- Eukaryotes
  (21 complete/86 assemblies)
Simple Genomes

- Full chromosomal sequences are provided
- Genes are annotated
- The annotation can be shown graphically and linked to sequence records

RefSeq Chromosomes: NC_

Base: 1
Start: 0
End: 4639221

### Annotation of Gene, CDS, and other features

<table>
<thead>
<tr>
<th>Base</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cgcttcatt gtcagacagc agaatttgta cgcgctgttc ggcttgttgt aatttggcct</td>
</tr>
<tr>
<td>61</td>
<td>gcccctgacg tgccagctgc acgccgcgtt cgaactcgtt cagcgcctct tccagcggca</td>
</tr>
<tr>
<td>121</td>
<td>ggtcgccact ttccagacgg gttacaatct gttccagctc gctcagcgcc ttttcaaagc</td>
</tr>
<tr>
<td>181</td>
<td>tggcgggcgc ctcatttttc ttcggcataa tgaatgtctg actctcaata tttttcgccc</td>
</tr>
<tr>
<td>241</td>
<td>cgtcatggta acggactcag ggcaaatagc aaataacgcg caatggtaag gtgatgtgca</td>
</tr>
<tr>
<td>301</td>
<td>cagcaaagcg atgttagtgg tatacttccg cgcctggatg cagccgcagg tgtgggctgc</td>
</tr>
<tr>
<td>361</td>
<td>tgtatttttc cctatacaag tcgcttaagg cttgccaacg aaccattgcc gccatgaagt</td>
</tr>
<tr>
<td>421</td>
<td>ttatcattaa attgttcccg gaaatcacca tcaaaagcca atctgtgcgc ttgcgcttta</td>
</tr>
<tr>
<td>481</td>
<td>taaaaatcct taccgggaac attcgtaacg ttttaaagca ctatgatgag acgctcgctg</td>
</tr>
<tr>
<td>541</td>
<td>tcgtccgcca ctgggataac atcgaagttc gcgcaaaaga tgaaaaccag cgtctggcta</td>
</tr>
<tr>
<td>601</td>
<td>ttcgcgacgc tctgacccgt attccgggta tccaccatat tctcgaagtc gaagacgtgc</td>
</tr>
<tr>
<td>661</td>
<td>cgtttaccga catgcacgat attttcgaga aagcgttggt tcagtatcgc gatcagctgg</td>
</tr>
<tr>
<td>721</td>
<td>aaggcaaaac cttctgcgta cgcgtgaagc gccgtggcaa acatgatttt agctcgattg</td>
</tr>
<tr>
<td>781</td>
<td>atgtggaacg ttacgtcggc ggcggtttaa atcagcatat tgaatccgcg cgcgtgaagc</td>
</tr>
<tr>
<td>841</td>
<td>tgaccaatcc ggatgtgact gtccatctgg aagtggaaga cgatcgtctc ctgctgatta</td>
</tr>
<tr>
<td>901</td>
<td>aaggccgcta cgaaggtatt ggcggtttcc cgatcggcac ccaggaagat gtgctgtcgc</td>
</tr>
<tr>
<td>961</td>
<td>tcatttccgg tggtttcgac tccggtgttt ccagttatat gttgatgcgt cgcggctgcc</td>
</tr>
</tbody>
</table>

### Genome sequence

- Annotation details
- CDS sequences
- Gene locations
Complex Genomes

- Sequences are provided complete or we help assemble
- Heavy annotation:
  Genes, transcript regions & ORFs, sequence variations & markers, clones, ESTs, etc.
- The annotation can be shown graphically and linked to other databases using the MapViewer

A database for retrieval and analysis of karyotype data: Cancer Chromosomes

RefSeq Records Contig: NT_ & Chromosome: NC_
NCBI FieldGuide
Higher Genome MapViews

NCBI FieldGuide
Higher Genome MapViews

Gene Expression Databases

A new database for localization of proteins in Mouse Brains: GenSat
A new database for information on Expression Reagents: Probe
Submit and update data

Query the database:
- gene identifiers
- field information
- sequence

Browse datasets

Download data

Gene Expression Omnibus

Submitted by Manufacturer

GSM
Platform descriptions
Raw/processed spot intensities from a single slide/chip

GSE
Grouping of slide/chip data
"a single experiment"

GDS
Grouping of experiments

Curated by NCBI

Entrez GEO

GEO Datasets

GDS177: CMV infection of HFF cells
UniGene Collections

as of January 2006

SEVERAL Organisms
Expression oriented

A Cluster of ESTs:
Arabidopsis serine protease
**Probe: Expression probes**

**Pr196507.1**
Ribonucleic acid probe (riboprobe) Prr for Mus musculus gene prion protein (Prnp). Has been used in the GENSAT project for in situ hybridization.

**Pr186482.1**
Small hairpin RNA (shRNA) probe V2MM_66187 for Mus musculus gene prion protein (Prnp). Developed for RNA interference (RNAi). Reagent is available from Open Biosystems.

**Pr001034449.1**
Small interfering RNA (siRNA) probe for Mus musculus gene prion protein (Prnp). Has been used for RNA interference (RNAi).

**Probe: siRNAs & shRNAs**

**P001034499.1**
Small interfering RNA (siRNA) probe for Mus musculus gene prion protein (Prnp). Has been used for RNA interference (RNAi).
Linking Protein Sequence, Structure and Function

Protein sequences

**CDD**: Conserved functional domains in proteins represented by a PSSM

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

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

Sequence-based Neighbors: Domain Neighbors

-**Conserved 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

"Reverse-Position Specific" Sequence Comparisons (RPS-BLAST) a.k.a.
"Conserved Domain Database" (CDD) Search
Sequence-based Neighbors: Domain Relatives

“Conserved Domain Architecture Retrieval Tool” (CDART)

Modular Architecture ofDomains
• Cartoon descriptions of protein domain organization on the primary sequence
• Allows for comparison with other proteins with the same Domain

[Diagram showing cartoon descriptions of protein domain organization]
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 Summary Page

Structure-based Neighbors:
- Conserved Domains (CDD/RPS-BLAST)
- (3D Domains)

Entrez PubChem

PC Compound
- Derived database of known chemicals from PC Substance records

PC Substance
- Primary database of chemical samples

PC BioAssay
- Primary database of bioactivity screens of samples in PC Substance
PubChem: Compound, Substance, BioAssay

and more....
The Gene Summary Database

Summary pages of curated information about genetic loci for organisms in the RefSeq project.

- Graphics
- Gene information
- Bibliography (PubMed links)
- General gene information
- NCBI Reference Sequences
- Related sequences
- Additional Links

H.sapiens & B.taurus
G6PD
B.norvegicus
G6PDx
M.musculus
G6Pdx, G6pd1
D.melanogaster
Zw
A.thalia
At5g35970
S.pombe
SPAC3C7.13c, SPAC9.01, SPCC794.01c
B.anthracis
BA_3932
H.pylori
HP1101
E.coli, Salmonella, Yersinia, Neisseria

Default Display

Gene symbols: G6PD glucose-6-phosphate dehydrogenase

Gene information:
- Gene type
- Gene name
- Gene description
- RefSeq status
- Organism
- Lineage
- Gene Aliases
- Summary
- General protein information
- Graphics: Transcripts and products, genomic context

Bibliography (PubMed links)

General gene information:
- Gene Ontology
- Homology (Mouse, Rat, Human)
- Phenotypes
- Sequence Tagged Sites
- Pathways

NCBI Reference Sequences
- mRNA sequence
- Source sequence
- Product
- Conserved Domains
- Related sequences
- Additional Links

Default Display
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: BLAST, Cn3D, Sequin, Data format conversion scripts
  

E-Utilities: Guidelines for Entrez "URL calls" used to access data. Designed for use in scripts.

- Examples: ESearch, EPost, ESummary, EFetch and ELink
  

Caution: Overuse may result in blocked IPs!

Intermission