NCBI Molecular Biology Resources
A Field Guide

May 26, 2004

NCBI Resources

- About NCBI
- NCBI Sequence Databases
- Other NCBI Databases
- Entrez Databases and Text Searching
- Genomic Resources
- BLAST Services

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
Types of Databases

- Primary Databases
  - Original submissions by experimentalists
  - Content controlled by the submitter
  - Examples: GenBank, SNP, GEO
- Derivative Databases
  - Built from primary data
  - Content controlled by third party (NCBI)
  - Examples: Refseq, TPA, RefSNP, UniGene, NCBI Protein, Structure, Conserved Domain

The Entrez System

Entrez Nucleotides

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary GenBank / EMBL / DDBJ</td>
<td>38,859,434</td>
</tr>
<tr>
<td>Derivative RefSeq</td>
<td>284,238</td>
</tr>
<tr>
<td>Third Party Annotation</td>
<td>4,255</td>
</tr>
<tr>
<td>PDB</td>
<td>5,000</td>
</tr>
<tr>
<td>Total</td>
<td>39,153,111</td>
</tr>
</tbody>
</table>
**Entrez Protein**

- GenPept (GB, EMBL, DDBJ) 3,240,758
- RefSeq 933,905
- Third Party Annotation 4,676
- Swiss Prot 151,524
- PIR 282,819
- PRF 12,079
- Total 4,728,703
- BLAST nr 1,829,482

**What is GenBank?**

NCBI’s Primary Sequence Database

- Nucleotide only sequence database
- Archival in nature
- GenBank Data
  - Direct submissions (traditional records)
  - Batch submissions (EST, GSS, STS)
  - ftp accounts (genome data)
- Three collaborating databases
  - GenBank
  - DNA Database of Japan (DDBJ)
  - European Molecular Biology Laboratory (EMBL) Database

**International Sequence Database Collaboration**
GenBank: NCBI's Primary Sequence Database

- Release 141 April 2004
- 33,676,218 Records
- 38,989,342,565 Nucleotides
- >140,000 Species
- 146 Gigabytes 606 files

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


The Growth of GenBank

Organization of GenBank: GenBank Divisions

Records are divided into 17 Divisions.

11 Traditional
6 Bulk

Traditional Divisions:
- Direct Submissions (Sequin and BankIt)
- Accurate
- Well characterized

PRI (27) Primate
PLN (11) Plant and Fungal
BCT (8) Bacterial and Archaeal
INV (6) Invertebrate
ROD (12) Rodent
VRL (4) Viral
VRT (6) Other Vertebrate
MAM (1) Mammalian (ex. ROD and PRI)
PHG (1) Phage
SYN (1) Synthetic (cloning vectors)
UNA (1) Unannotated

Entrez query: gbdiv_xxx[Properties]
Organization of GenBank: GenBank Divisions

Records are divided into 17 Divisions.

- 11 Traditional
- 6 Bulk

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

Entrez query: gbdiv_xxx[Properties]

A Traditional GenBank Record

**LOCUS**

AF062069  3808 bp mRNA    linear   INV 23-OCT-2002

**DEFINITION**

Limulus polyphemus myosin III mRNA, complete cds.

**REFERENCE**


**TITLE**

A myosin III from Limulus eyes is a clock-regulated phosphoprotein

**JOURNAL**


**MEDLINE**

98279067

**PUBMED**

9614231

**REFERENCE**

2  (bases 1 to 3808)

**AUTHORS**


**TITLE**

Direct Submission

**JOURNAL**

Submitted (29-APR-1998) Whitney Laboratory, University of Florida, 9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA

**REFERENCE**

3  (bases 1 to 3808)

**AUTHORS**


**TITLE**

Direct Submission

**JOURNAL**

Submitted (02-MAR-2000) Whitney Laboratory, University of Florida, 9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA

**REMARK**

Sequence update by submitter

**COMMENT**

On Mar 2, 2000 this sequence version replaced gi:3132700.

GenBank Record: Locus

**LOCUS**

AP042069  3808 bp mRNA    linear   INV 23-OCT-2002

**DEFINITION**

Limulus polyphemus myosin III mRNA, complete cds.

**REFERENCE**


**TITLE**

A myosin III from Limulus eyes is a clock-regulated phosphoprotein

**JOURNAL**


**MEDLINE**

98279067

**PUBMED**

9614231

**REFERENCE**

2  (bases 1 to 3808)

**AUTHORS**


**TITLE**

Direct Submission

**JOURNAL**

Submitted (29-APR-1998) Whitney Laboratory, University of Florida, 9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA

**REFERENCE**

3  (bases 1 to 3808)

**AUTHORS**


**TITLE**

Direct Submission

**JOURNAL**

Submitted (02-MAR-2000) Whitney Laboratory, University of Florida, 9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA

**COMMENT**

On Mar 2, 2000 this sequence version replaced gi:3132700.
GenPept: FASTA format

>gi|7144485|gb|AAC16332.2| myosin III [Limulus polyphemus]
MEYKCISEHLPFETLPDPGDRFEVQELVGTGTYATVYSAIDKQANKKVALKIIGHIAENLLDIETEYRIY
KAVNGIQFFPEFRGAFFKRGERESDNEVWLGIEFLEEGTAADLLATHRRFGIHLKEDLIALIIKEVVRAV
QYLHENSIIHRDIRAANIMFSKEGYVKLIDFGLSASVKNTNGKAQSSVGSPYWMAPEVISCDCLQEPYNY
TCDVWSIGITAIELADTVPSLSDIHALRAMFRINRNPPPSVKRETRWSETLKDFISECLVKNPEYRPCIQ
EIPQHPFLAQVEGKEDQLRSELVDILKKNPGEKLRNKPYNVTFKNGHLKTISGQPHEKIYVDDLAFLDSP
TEEVVLENLEQRYRKGEIYTFAGDVLLTLNPGKVLPLYGDQTAVKYCERGRSDNPPHVFAVADRAYQQML
HHKSPQAVILSGVSGSGKSFCTHQVIRHLAFLGAQNKEGMREKLEYLCPLLDTLGNAYTSTNPNSSHFVK
ILEVTFTKTGKITGAILFTFLLEARRLTDIPKGERNFHVFYYFYEGLRSEGRLKEFGLEEKNYRYLPELK
SSNSPEYVKGYQQFLRALTSLAFTEEEIFAIQKVLAAILLLGETEIQNSAAFKLLGAESSELENTLTQDV
HAKSYVANIKLRELEKLNVAFQNLGSLRELQGVTCLPQYVTVCPGQLPGQLHIQSGYQGKAVNIDLEQGK
IQQTFFKELEETVGDHVPSLAVNKQGOTLKURNKMSQIQGQELTAISATQHGEEHHMSQGCAS
DELETVGNHGRPEPGPASCTKTVALTQHCNGDKQGQIAPINHKSQYSPSFL
SPRLSNGMELFFVTQYDASLSASVHAIKSTTEDFSPLSLCMMPQREXIMOVYHEKUQVQLVAKLAV
YITLILQSFARKSPSFPFL1W1VQLAEFHMPPLCPYELYTVQVCSQVQIQLSYCEY
LSBTVYN1KEVYCVQJAIEMRTYTVQVCSQVQIQLSYCEY

Abstract Syntax Notation: ASN.1

```ascend
Seq-entry ::= set {
  class nuc-prot ,
  descr {
    title "Limulus polyphemus mRNA, complete cds." ,
    source {
      org {
        taxname "Limulus polyphemus" ,
        common "Atlantic horseshoe crab" ,
        db {
          { db "taxon" ,
            tag id 6850 } } ,
        orgname {
          name {
            genus "Limulus" ,
            species "polyphemus" } ,
          lineage "Eukaryota; Metazoa; Arthropoda; Chelicerata; Merostomata; Xiphosura; Limulidae; Limulus" ,
          gcode 1 ,
          mgcode 5 ,
          div "INV" } } ,
    } 
}
```

NCBI Toolbox

```
Toolbox Sources

ftp> open ftp.ncbi.nih.gov
ftp> cd toolbox
cd ncbi_tools
ftp> ls
```

```bash
```

```
# License: () 1993-2013 NCBI
# Source: /usr/bin/accid1
# Source: /usr/bin/nwutil
# Source: /usr/bin/objutil
# Source: /usr/bin/lsqfetch
# Source: /usr/bin/explore
# Source: /usr/bin/subutil
# Source: /usr/bin/objcode
# Source: /usr/bin/objall
# Source: /usr/bin/lsqutil
# Source: /usr/bin/accutil
# Source: /usr/bin/tcpdump
```

```
7
```
### Bulk Divisions

- Batch Submission and htg (email and ftp)
- Inaccurate
- Poorly Characterized

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expressed Sequence Tag</strong></td>
<td>1st pass single read cDNA</td>
</tr>
<tr>
<td><strong>Genome Survey Sequence</strong></td>
<td>1st pass single read gDNA</td>
</tr>
<tr>
<td><strong>High Throughput Genomic</strong></td>
<td>Incomplete sequences of genomic clones</td>
</tr>
<tr>
<td><strong>Sequence Tagged Site</strong></td>
<td>PCR-based mapping reagents</td>
</tr>
</tbody>
</table>

### EST Division: Expressed Sequence Tags

ESTs in Entrez

- Total: 21 million records
- Human: 5.6 million
- Mouse: 4.1 million
- Rat: 0.6 million

ESTs in Entrez

- gbdv_est [Properties]

<table>
<thead>
<tr>
<th>IMAGE:275615 5' mRNA sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>GACAGCATTCGGGCCGAGATGTCTCGCTCCGTGGCCTTAGCTGTGCTCGCGCTACTCTCTCTTTCTGGTGGAGGTATCCAGCGTACTCCAAAGATTCAGGTTTACTCACGTCATCCAGCAGAGAATGGAAAGTCAATTCCTGAATTGCTATGTGTCTGGGTTTCATCCATCCGACATTGAAGTTGACTTACTGAAGAATTGGAGA GAATTGAAAAAGTGGAGCATTCAGACTTGTCTTTCAGCAAGGACTGGTCTTTCTATCTCTTGTACTAC TGAATTCACCCCCACTGAAAAAGATGAGTATGCCTGCCGTGTTGAACCATGTNGACTTTGTCACAGNC AAGTTNAGTTTAAGTGGGNATCGAGACATGTAAGGCAGGCATCATGGGAGGTTTTGAAGNATGCCGCN TTGGATTGGGATGAATTCCAAATTTCTGGTTTGCTTGNTTTTTTAATATTGGATATGCTTTTG</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IMAGE:275615 3' mRNA sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>GACAGCATTCGGGCCGAGATGTCTCGCTCCGTGGCCTTAGCTGTGCTCGCGCTACTCTCTCTTTCTGGTGGAGGTATCCAGCGTACTCCAAAGATTCAGGTTTACTCACGTCATCCAGCAGAGAATGGAAAGTCAATTCCTGAATTGCTATGTGTCTGGGTTTCATCCATCCGACATTGAAGTTGACTTACTGAAGAATTGGAGA GAATTGAAAAAGTGGAGCATTCAGACTTGTCTTTCAGCAAGGACTGGTCTTTCTATCTCTTGTACTAC TGAATTCACCCCCACTGAAAAAGATGAGTATGCCTGCCGTGTTGAACCATGTNGACTTTGTCACAGNC AAGTTNAGTTTAAGTGGGNATCGAGACATGTAAGGCAGGCATCATGGGAGGTTTTGAAGNATGCCGCN TTGGATTGGGATGAATTCCAAATTTCTGGTTTGCTTGNTTTTTTAATATTGGATATGCTTTTG</td>
</tr>
</tbody>
</table>

80-100,000 unique cDNA clones in library
What is UniGene?

A gene-oriented view of sequence entries

- MegaBlast based automated sequence clustering
- Now informed by genome hits New!
- Nonredundant set of gene oriented clusters
- Each cluster a unique gene
- Information on tissue types and map locations
- Includes well-characterized genes and novel ESTs
- Useful for gene discovery and selection of mapping reagents

---

EST hits A.t. serine protease mRNA

Distribution of EST Hits on the Open Reading Frame

A. t. mRNA

5' EST hits

3' EST hits

---

UniGene

March 2004

<table>
<thead>
<tr>
<th>Species</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>23,977</td>
</tr>
<tr>
<td>Homo troglodytes</td>
<td>2,835</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>210,900</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>18,945</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>19,429</td>
</tr>
<tr>
<td>Canis familiaris</td>
<td>12,297</td>
</tr>
<tr>
<td>Sus scrofa</td>
<td>11,950</td>
</tr>
<tr>
<td>Equus caballus</td>
<td>1,130</td>
</tr>
<tr>
<td>Sus scrofa</td>
<td>1,227</td>
</tr>
<tr>
<td>Bovidae</td>
<td>12,800</td>
</tr>
<tr>
<td>Eutheria</td>
<td>2,660</td>
</tr>
<tr>
<td>Mammalia</td>
<td>1,560</td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>1,596</td>
</tr>
<tr>
<td>Macaca fascicularis</td>
<td>1,710</td>
</tr>
<tr>
<td>Canis familiaris</td>
<td>1,230</td>
</tr>
<tr>
<td>Hominidae</td>
<td>1,000</td>
</tr>
<tr>
<td>Cetacea</td>
<td>1,000</td>
</tr>
<tr>
<td>Chordata</td>
<td>1,000</td>
</tr>
</tbody>
</table>

---
Human UniGene

- 136,416 mRNAs
- 5,187 models
- 7,415 HTC
- 1,416,936 EST, 3' reads
- 2,092,475 EST, 5' reads
- 775,747 EST, other/unknown
- 4,434,176 total sequences in clusters

Final Number of Clusters (sets)

<table>
<thead>
<tr>
<th>Total</th>
<th>mRNA</th>
<th>3,000,000,000 bp</th>
<th>HTC</th>
<th>EST</th>
<th>75% excess</th>
</tr>
</thead>
<tbody>
<tr>
<td>106,219</td>
<td>28,126</td>
<td>5,750</td>
<td>104,890</td>
<td>26,839</td>
<td></td>
</tr>
</tbody>
</table>

Unigene Build 170
Apr. 24, 2004

---

Genome Sequencing – HTG, GSS (WGS)

Whole BAC insert (or genome)

- Shredding
- Sequencing
- Cloning/Isolating
- GSS division or trace archive
- Assembly
- Whole genome shotgun assemblies (traditional division)

Draft Sequence (HTG division)

HTG Division: Honeybee Draft Sequences

- Unfinished sequences of BACs
- Gaps and unordered pieces
- Finished sequences move to traditional GenBank division
Other Genome Sequencing Products

Trace Archive
Whole Genome Shotgun

Whole Genome Shotgun Projects

• Traditional GenBank Divisions
• 164 + projects
  – Virus
  – Bacteria
  – Environmental sequences
  – Archaea
  – 44 Eukaryotes featuring:
    • Chicken, Rat, Mouse, Dog, Chimpanzee, Human
    • Pufferfish (2)
    • Honeybee, Anopheles, Fruit Flies (2), Silkworm
    • Nematode (C. briggsae)
    • Yeasts (8), Aspergillus (2)
    • Rice
Derivative Sequence Databases

RefSeq
TPA

NCBI Derivative Sequence Data

RefSeq: NCBI’s Derivative Sequence Database

- Curated transcripts and proteins
  - reviewed
  - human, mouse, rat, fruit fly, zebrafish, arabidopsis
- Model transcripts and proteins
- Assembled Genomic Regions (contigs)
  - human genome
  - mouse genome
- Chromosome records
  - Human genome
  - microbial
  - organelle

RefSeq Benefits

- non-redundancy
- explicitly linked nucleotide and protein sequences
- updates to reflect current sequence data and biology
- data validation
- format consistency
- distinct accession series
- stewardship by NCBI staff and collaborators

RefSeq Accession Numbers

<table>
<thead>
<tr>
<th>mRNAs and Proteins</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_123456</td>
<td>Curated mRNA</td>
</tr>
<tr>
<td>NP_123456</td>
<td>Curated Protein</td>
</tr>
<tr>
<td>NR_123456</td>
<td>Curated non-coding RNA</td>
</tr>
<tr>
<td>XM_123456</td>
<td>Predicted mRNA</td>
</tr>
<tr>
<td>XP_123456</td>
<td>Predicted Protein</td>
</tr>
<tr>
<td>XR_123456</td>
<td>Predicted non-coding RNA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene Records</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NG_123456</td>
<td>Reference Genomic Sequence</td>
</tr>
<tr>
<td>NC_123455</td>
<td>Microbial replicons, organelle</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assemblies</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NT_123456</td>
<td>Contig</td>
</tr>
<tr>
<td>NW_123456</td>
<td>WGS Supercontig</td>
</tr>
</tbody>
</table>

Third Party Annotation (TPA) Database

- Annotations of existing GenBank sequences
- Allows for community annotation of genomes
- Direct submissions
  - BankIt  tpa[Properties]
  - Sequin
TPA record: WGS Assembly

Human Nucleotide Sequences

Other NCBI Databases

- dbSNP: nucleotide polymorphism
- Geo: Gene Expression Omnibus microarray and other expression data
- Gene: gene records
  - Unifies LocusLink and Microbial Genomes
- Structure: imported structures (PDB)
  - Cn3D viewer, NCBI curation
- CDD: conserved domain database
  - Protein families (COGs and KOGs)
  - Single domains (PFAM, SMART, CD)
NCBI Structures and Domains

MMDB: Molecular Modeling Data Base

- Derived from experimentally determined PDB records
- Value added to PDB records including:
  - Addition of explicit chemical graph information
  - Validation
  - Inclusion of Taxonomy, Citation,
  - Conversion to ASN.1 data description language
- Structure neighbors determined by Vector Alignment Search Tool (VAST)

Structure Summary

Conserved Domains  3D Domain Neighbors

Cn3D viewer Structure Neighbors
For each protein chain, locate SSEs (secondary structure elements), and represent them as individual vectors. Align the vectors.
NCBI FieldGuide

Cn3D: Simple Homology Modeling

NCBI’s Conserved Domain Database

- Multiple sequence alignments
- PSI-BLAST –based score matrices
- Sources SMART, PFAM, COGs,
  New NCBI curated domains
  – structure informed alignments
- Stats:
  - COGS 4,873
  - Pfam 5,193
  - Smart 653
  - NCBI CDD 316

NCBI CD: Tyrosine Kinase
Using Cn3D to model domains

Using Entrez
An integrated database search and retrieval system
Database Searching with Entrez

- Using limits and field restriction to find human MutL homolog
- Linking and neighboring with MutL
- Mapping SNPs onto structure and the genome
Human MutL Search Results

NM_000249: Links

GenBank Records

NM_000249: Links

Human MutL RefSeq

GenBank Records

NM_000249: Links

Human MutL Search Results
Literature Links

PubMed
OMIM

NM_000249: PubMed

Books Link
NCBI FieldGuide

Taxonomy Link

The Tax Browser
NCBI's Taxonomy

NCBI Protein Databases

- GenPept  GenBank, EMBL, DDBJ CDS translations
- RefSeq   mRNA based (NP_) and genome based (XP_)
- Swiss-Prot curated high quality protein reviews
- PIR      protein information resource Georgetown University
- PRF      protein resource foundation
- PDB      Protein Databank sequences from structures
BLink: non-redundant relatives

MLH1 Domain Structure: CDD

MLH1: ATPase Domain
Mapping Variation Onto Structure

Bacterial DNA mismatch repair proteins

Loads sequence alignment and structure in Cn3D

Mapping Variation Onto Structure

Conserved Asn

Ile - Val

NM_000249: Genome Links

[Genome Links interface]

[Text abstracts and links]

29
Synteny: Mammalian Genomes

The New Homologene
- No longer UniGene based
- Protein similarities first
- Guided by taxonomic tree
- Includes orthologs and paralogs

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of Genes</th>
<th>Homologene</th>
</tr>
</thead>
<tbody>
<tr>
<td>H. sapiens</td>
<td>21,450</td>
<td>16,793</td>
</tr>
<tr>
<td>M. musculus</td>
<td>23,650</td>
<td>15,743</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>23,650</td>
<td>15,743</td>
</tr>
<tr>
<td>C. elegans</td>
<td>18,710</td>
<td>14,366</td>
</tr>
<tr>
<td>M. musculus</td>
<td>23,650</td>
<td>15,743</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>23,650</td>
<td>15,743</td>
</tr>
<tr>
<td>C. elegans</td>
<td>18,710</td>
<td>14,366</td>
</tr>
<tr>
<td>S. pombe</td>
<td>6,500</td>
<td>3,269</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>6,500</td>
<td>3,269</td>
</tr>
<tr>
<td>A. thaliana</td>
<td>37,350</td>
<td>25,411</td>
</tr>
<tr>
<td>T. brucei</td>
<td>5,236</td>
<td>4,492</td>
</tr>
</tbody>
</table>

Orthologs and paralogs:
- chick A frog B

A-chain gene
### Microbial Genomes

### Related Proteins -> Genome Links

<table>
<thead>
<tr>
<th>Protein Name</th>
<th>Gene Name</th>
<th>Species</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>Gene1</td>
<td>Species1</td>
<td>Link1</td>
</tr>
<tr>
<td>Example 2</td>
<td>Gene2</td>
<td>Species2</td>
<td>Link2</td>
</tr>
<tr>
<td>Example 3</td>
<td>Gene3</td>
<td>Species3</td>
<td>Link3</td>
</tr>
</tbody>
</table>

### C.elegans homolog

<table>
<thead>
<tr>
<th>Protein Name</th>
<th>Gene Name</th>
<th>Species</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 4</td>
<td>Gene4</td>
<td>Species4</td>
<td>Link4</td>
</tr>
<tr>
<td>Example 5</td>
<td>Gene5</td>
<td>Species5</td>
<td>Link5</td>
</tr>
</tbody>
</table>
Entrez Genes: integrated gene-based access

Genes MLH1: Central Resource

Intermission