NCBI Molecular Biology Resources

PRIMO12: Short Course in Aquatic Toxicology

Peter Cooper
National Center for Biotechnology Information
May 9, 2003

Talk Outline

- About NCBI
- NCBI Databases
  - zebrafish sequences
- Integrated Access
  - Entrez
  - BLAST

The National Center for Biotechnology Information

- Created as part of NLM in 1988
  - Establish public databases
  - Perform research in computational biology
  - Develop software tools for sequence analysis
  - Disseminate biomedical information
Molecular Databases

- **Primary**
  - Data provided by experimentalist
  - Record maintained by submitter
  - GenBank

- **Derivative**
  - Value added to primary data
    - compilation
    - curation
    - assembly
  - Record maintained by database staff
  - NCBI Reference Sequences (RefSeq)
  - Protein Data

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

- **Release 135** April 2003
- 24,027,936 Records
- 31,099,264,455 Nucleotides
- 120,000 + Species

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


114 Gigabytes
The Growth of GenBank

Release 135: 24.0 million records
31.1 billion nucleotides
Doubling time = 12 months

"It's sink or swim as a tidal wave of data approaches"
Nature 399:517 10 June 1999

Traditional GenBank Divisions

- Direct Submissions (Sequin and BankIt)
- Accurate
- Well characterized

**BCT** Bacterial and Archaeal
**INV** Invertebrate
**MAM** Mammalian (ex. ROD and PRI)
**PHG** Phage
**PLN** Plant and Fungal
**PRI** Primate
**ROD** Rodent
**SYN** Synthetic (vectors, synth. genes)
**VRL** Viral
**VRT** Other Vertebrate

Traditional GenBank Records
Traditional GenBank Records

- gbdiv_vrt[Properties]
- zebrafish: 3,816

Bulk GenBank Divisions

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

EST: Expressed Sequence Tag
STS: Sequence Tagged Site
GSS: Genome Survey Sequence
HTG: High Throughput Genomic
**Bulk GenBank Records: ESTs**

<table>
<thead>
<tr>
<th>Name</th>
<th>Date</th>
<th>Description</th>
<th>Accession</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>gbdiv_est</td>
<td>2023-01-01</td>
<td>ESTs from zebrafish adult tail</td>
<td>gbdiv_est[Properties]</td>
<td>1000</td>
</tr>
<tr>
<td>zebrafish ESTs</td>
<td>2023-01-02</td>
<td>ESTs from zebrafish tissues</td>
<td>gbdiv_est[Properties]</td>
<td>2000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference</th>
<th>Authors</th>
<th>Title</th>
<th>Journal</th>
<th>PubMed ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen</td>
<td>Chen et al.</td>
<td>A gene-oriented view of sequence entries</td>
<td>Nucleic Acids Research</td>
<td>30</td>
</tr>
</tbody>
</table>

**What is UniGene?**

A gene-oriented view of sequence entries

- MegaBlasT based automated sequence clustering
- Non-redundant 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

### UniGene Collections: Animals

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
<th>Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>4,135,222</td>
<td>111,064</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>3,442,382</td>
<td>90,444</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>346,154</td>
<td>63,253</td>
</tr>
<tr>
<td>Bos taurus</td>
<td>130,603</td>
<td>10,675</td>
</tr>
<tr>
<td>Sus scrofa</td>
<td>62,004</td>
<td>14,390</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>125,827</td>
<td>5,042</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>217,019</td>
<td>15,707</td>
</tr>
<tr>
<td>Oryzias latipes</td>
<td>81,296</td>
<td>6,115</td>
</tr>
<tr>
<td>Xenopus laevis</td>
<td>177,814</td>
<td>19,045</td>
</tr>
<tr>
<td>Siluriana tropicalis</td>
<td>110,976</td>
<td>7,674</td>
</tr>
<tr>
<td>Ciona intestinalis</td>
<td>464,060</td>
<td>13,328</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>276,922</td>
<td>14,701</td>
</tr>
<tr>
<td>Anopheles gambizs</td>
<td>91,352</td>
<td>3,243</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>200,436</td>
<td>19,045</td>
</tr>
</tbody>
</table>

March 2003

### UniGene Collections: Plants and Others

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
<th>Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>221,113</td>
<td>26,813</td>
</tr>
<tr>
<td>Glycine max</td>
<td>201,710</td>
<td>9,084</td>
</tr>
<tr>
<td>Medicago truncatula</td>
<td>97,063</td>
<td>5,729</td>
</tr>
<tr>
<td>Lyccopensum esculentum</td>
<td>70,118</td>
<td>3,749</td>
</tr>
<tr>
<td>Triticum aestivum</td>
<td>297,961</td>
<td>20,376</td>
</tr>
<tr>
<td>Hordeum vulgare</td>
<td>263,407</td>
<td>10,272</td>
</tr>
<tr>
<td>Zea mays</td>
<td>163,541</td>
<td>13,512</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>94,780</td>
<td>19,191</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>17,742</td>
<td>802</td>
</tr>
<tr>
<td>Chlamydomonas reinhardtii</td>
<td>80,771</td>
<td>5,712</td>
</tr>
<tr>
<td>Dictyostelium discoideum</td>
<td>119,140</td>
<td>4,782</td>
</tr>
</tbody>
</table>

March 2003
Genome Sequencing - HTG, GSS, (WGS)

Whole BAC insert (or genome)

- Shredding
- Sequencing
- Cloning
- Isolating
- Assembly

GSS division

or trace archive

Draft Sequence (HTG division)

Bulk GenBank Records: GSS

- First pass, single read gDNA
- Surveys of BAC libraries
- GSS archive
- Zebrasfish GSS: 159,024

HTG Division: High Throughput Genome

Zebrafish PAC Clone

Phase 1

Acc = AC109580.1

Phase 2

Acc = AC109580.11

Phase 3

Acc = AC109580.14

40,000 to > 350,000 bp
HTG Division: High Throughput Genome

40,000 to > 350,000 bp

Whole Genome Shotgun

Whole Genome Shotgun in GenBank
Whole Genome Shotgun in GenBank

RefSeq Transcripts and Proteins

NCBI RefSeq: Gene Models XM_, XP_

Provisional → Reviewed
On April 11, 2003 this sequence was added to NCBI's Reference Sequence database. This contig is part of the RefSeq project, which provides a comprehensive set of non-redundant, selected, assembled gene sequences for a number of model organisms. The RefSeq project is coordinated by NCBI and is a collaboration with other Institutes.

Some Web Statistics

<table>
<thead>
<tr>
<th>July 2001</th>
<th>Users Per Weekday</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAST alone: currently 80,000 searches per day.</td>
<td></td>
</tr>
<tr>
<td>PubMed: up to 4,000,000 searches per day</td>
<td></td>
</tr>
</tbody>
</table>

Web Access

- BLAST
- Entrez
- Sequence
- VAST
- Structure
- Text
Database Searching with Entrez

- Using limits and field restriction to find fish RNA helicases
- Linking and neighboring with RNA helicases

Entrez Nucleotides

Document Summaries:

RNA Helicase[All Fields]
**Document Summaries: Limits**

- 17
- NCBI
- Ray Finned Fishes
- Adding Terms: Preview/Index
- Accession
- Author
- EC/RN Number
- Feature key
- Filter
- Gene Name
- Issue
- Journal
- Keyword
- Modification Date
- Organism
- Page Number
- Primary Accession
- Properties
- Protein Name
- Publication
- Date
- SeqID String
- Sequence
- Length
- Substance Name

**Fish RNA Helicases**
The human U2 snRNA-specific 10-kDa protein is an RS domain-containing, putative RNA binding with significant homology to the yeast spliceosome Pop2p.

Furthermore, a number of other proteins, including HSP90, actin, and the RS domain protein, have been identified as potential candidates for the yeast spliceosome Pop2p.

NCBI Handbook

NCBI Handbook

Part 1. The Database

1. PubMed: The Biochemical Sequence Database
2. Nucleotide: The Nucleotide Sequence Database
3. Genome Database: The Genome Database
4. Taxonomy Database
5. The Entrez Protein

Part 2. The Literature

1. NCBI Books
2. NCBI Online Books

Figure 3.4: The spliceosome mechanism. Spliceosome is composed of the small nuclear ribonucleoprotein (snRNP) complex, which is a multi-protein complex that recognizes and splices exons. The snRNP complex is composed of multiple subunits, including the U2 snRNP, which contains the U2 snRNA and the U2 snRNA-specific 10-kDa protein. The U2 snRNA-specific 10-kDa protein is a member of the RS domain family and has significant homology to the yeast spliceosome Pop2p.
## Genes Database: All Genomes

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Gene Symbol</th>
<th>Description</th>
<th>Chromosome</th>
<th>Genomic Location</th>
<th>Expression</th>
<th>Protein Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGP7743</td>
<td>glucose-6-phosphate dehydrogenase</td>
<td>[Escherichia coli A12]</td>
<td>2</td>
<td>0.0000 - 1.0000</td>
<td>transcribed and translated</td>
<td>protein ID: AGP7743, supported by mRNA: 22463, supported by mRNA: 29460190</td>
</tr>
</tbody>
</table>

### Service Addresses

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