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

Molecular Databases

- Sequence
- Structure
- Expression

NCBI FieldGuide

NCBI Molecular Biology Resources

September 2006

NCBI Resources

- About NCBI
- NCBI Databases
- Tools
- Entrez
- Examples
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

GenBank: NCBI’s Primary Sequence Database

- Release 155
- August 2006
- 79,093,266 Records
- 145,739,069,776 Total Bases
- 240 Gigabytes (non-WGS)
- 1010 files (non-WGS)

- Full release every two months
- Incremental updates daily
- Available only via ftp


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        |

Bulk Divisions

- Batch Submission and htg (email and ftp)
- Inaccurate
- Poorly Characterized
- Largest component of GenBank (non-WGS)

- Expressed Sequence Tag
  - 1st pass single read cDNA
- Genome Survey Sequence
  - 1st pass single read gDNA
- High Throughput Genomic
  - Incomplete sequences of genomic clones
- Sequence Tagged Site
  - PCR-based mapping reagents
NCBI FieldGuide

GenBank Records

- LOCUS AY198415
- ACCESSION AY198415
- DEFINITION Homo sapiens PR domain-containing protein 1 beta (PRDM1) mRNA, complete cds.
- ACCESSION AI583931
- VERSION AI583931.1
- DEFINITION Zebrafish WashU MPIMG EST Danio rerio cDNA clone IMAGE:3720022 3', mRNA sequence.
- ACCESSION AC027700
- VERSION AC027700.7
- DEFINITION Mus musculus chromosome 10 clone RP23-145G7 map 10, *** SEQUENCING IN PROGRESS ***, 3 unordered pieces.
- ACCESSION AADN01000358
- VERSION AADN01000358.1
- DEFINITION Gallus gallus chromosome 3 cont0.358, whole genome shotgun sequence.
- ACCESSION NM_007548
- VERSION NM_007548.1
- DEFINITION Mus musculus chromosome 10 genomic contig, strain C57BL/6J.
- ACCESSION XP_228320
- VERSION XP_228320.2
- DEFINITION Homo sapiens chromosome 6, complete sequence.
- ACCESSION NC_000006
- VERSION NC_000006.9

RefSeq: NCBI’s Derivative Sequence Database

- Curated transcripts and proteins [NM_, NP_] – reviewed
  - human, mouse, rat, cow, fruit fly, zebrafish, arabidopsis
- Model transcripts and proteins [XM_, XP_] – draft human genome
- Assembled Genomic Regions [NT_, NW_] – mouse genome
- Chromosome records [NC_] – microbial – organelle

NCBI’s SNP Database

- Primary Database and Derivative (RefSNP)
- Single Nucleotide Polymorphism
- Repeat polymorphisms
- Insertion-Deletion Polymorphisms
- 35 species
- Over 87 million submissions
- Over 50 million RefSNP clusters
RefSNP HFE

Substitution eliminates a disulfide bond

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 known genes and uncharacterized ESTs
- Useful for gene discovery and selection of mapping reagents

EST hits: Human mRNA

5' EST hits

3' EST hits

Albumin mRNA

UniGene August 2006

Chordates

Plants

Invertebrates

Fungi et al.
**UniGene: Expressed Sequences**

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

**Expression Data**

- Structure neighbors determined by Vector Alignment Search Tool (VAST)

**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, and other information
  - Conversion to ASN.1 data description language

**Cn3D 4.1: C-SRC**
Protein Domains

- Structural Domain
  - Discrete independently folding unit of a protein
- Conserved Domain (sequence-based)
  - Protein region with recognizable position specific pattern of sequence conservation
- Sequence-based domains often roughly correspond to structural domains
- Domains often have distinct, identifiable functions

NCBI’s Conserved Domain Database

- PSI-BLAST –based score matrices
- Searchable with RPS-BLAST
- Sources
  - SMART
  - PFAM
  - COGs
  - NCBI curated domains
    - structure informed alignments
Src Domains

<table>
<thead>
<tr>
<th>Description</th>
<th>Source</th>
<th>Taxonomy</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ovarial Structure of Human Tyrosine Protein Kinase C-Src</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Structure vs Conserved Domain

- **Conserved phosphotyrosine binding residues**

Basic Local Alignment Search Tool

- Calculates similarity for biological sequences
- Finds best local alignments
- Heuristic approach based on Smith-Waterman algorithm
- Searches for matching "words" rather than individual residues
- Uses statistical theory to determine if a match might have occurred by chance

NCBI Tools

- BLAST
- VAST
**BLAST and BLAST-like programs**

- Traditional BLAST (blastall) nucleotide, protein, translations
  - blastn nucleotide query vs. nucleotide database
  - blastp protein query vs. protein database
  - blastx nucleotide query vs. protein database
  - tblastn protein query vs. translated nucleotide database
  - tblastx translated query vs. translated database
- Megablast nucleotide only
  - Contiguous megablast
  - Nearly identical sequences
  - Discontiguous megablast
- Cross-species comparison
- Position Specific BLAST Programs protein only
  - Position Specific Iterative BLAST (PSI-BLAST)
    - Automatically generates a position specific score matrix (PSSM)
  - Reverse PSI-BLAST (RPS-BLAST)
    - Searches a database of PSI-BLAST PSSMs

**The Flavors of BLAST**

- Position independent scoring
  - Standard BLAST
    - traditional contiguous word hit
    - nucleotide, protein and translations
    - Megablast
      - can use discontiguous words
      - nucleotide only
      - optimized for large batch searches
- Position dependent scoring
  - PSI-BLAST
    - constructs PSSMs automatically
    - searches protein database with PSSMs
  - RPS BLAST
    - searches a database of PSSMs
    - basis of conserved domain database

**The BLAST homepage**

- Standard databases
- Specialized Databases

**BLAST Search**

- Options for unrestricting
- Limiting results
- Search database
- E-value threshold
- Other options
For each protein chain, locate SSEs (secondary structure elements), and represent them as individual vectors.

align the vectors

Human IL-4

Finding a human PR domain
RefSeq

• Finding PRDM1
• Links and Neighbors with PRDM1
NCBI FieldGuide
Entrez: Database Integration
WWW Access
Entrez & BLAST
Nucleotide database now three parts
-EST expressed sequence tags
-GSS genome survey sequences
-CoreNucleotide everything else

Nucleotide sequences
Protein sequences
3-D Structure
Word weight
PubMed abstracts
BLAST
VAST
BLAST
Protein sequences
Nucleotide sequences
Structure
Genomes
Taxonomy
Phylogeny
Neighbors
Hard Link
Soft Link
PR Domain Global Search
PR Domain search results
Sometimes Quotes Help

“PR Domain” search results

PR Domain search results (continued)

What is *Pongo pygmaeus*?
Pongo pygmaeus: Taxonomy Browser

Refining Search

Limits

Title (Definition)
mRNA molecule type

RefSeq (no GenBank)
Limited Search Results

Add Terms to Query or View Index:
- Enter a term in the text box, use the pull-down menu to specify a search field.
- Click Preview to add terms to the query box and see the number of search results, or click Index to view terms within an index.
- Multiple terms selected from Index will be ORed, click AND to add to search.

Organism Field index is NCBI’s Taxonomy

Human PR domain RefSeqs

PRDM Reviewed RefSeqs

Organism Field index is NCBI’s Taxonomy
Protein Link –– Related Sequences

Protein Neighbors: Redundancy and homologs

Online Mendelian Inheritance in Man

• Zn Finger domains
• PR Domain
• Homologs

– BLIMP1 Mouse
– U-boot Zebrafish

Genomic context
The SET Domain

Zn finger: a DNA binding motif

SET Domain Structural Model

S-adenosyl homocysteine

catalytic tyrosine

Entrez Gene: One stop shopping
NCBI FieldGuide

HomoloGene: homologs in complete genomes

U-boot

NCBI FieldGuide

Map Viewer: PRDM1

GenBank BACs

RefSeq transcripts

RefSeq contigs

ESTs UniGene

NCBI Annotation

PRDM1: Potential Paralogs

Homo sapiens genome view

BLAST search the human genome

Search results for query "PRDM1": 17 hits