A Field Guide

part 2

- Entrez searching
- Genome resources
- Sequence similarity

GenBank Records

The Flatfile Format

- Header
- Feature Table
- Sequence

A Typical GenBank Record

LOCUS NM_019570  4279 bp mRNA  linear  ROD 28-OCT-2004
DEFINITION Mus musculus REV1-like (S. cerevisiae) (Rev1l) mRNA
ACCESSION NM_019570
VERSION NM_019570.3 GI:50811869
KEYWORDS

GenBank Record: Feature Table

FEATURES
LOCATIONS/QUALIFIERS
    Source  1..4279 /organism="Mus musculus" /mol_type="mRNA" /db_xref="taxon:10050" /GChromosome="1"
    Date	28-OCT-2004
    Protein	Rev1l /db_xref="protein:NM_019570" /db_xref="RefSeq:NM_019570.3" /gnote="Rev1l protein"

VERSION gi only change when sequence changes

NCBI FieldGuide
...AND srcdb refseq[Properties]

...AND srcdb ddbj/embl/genbank[Properties]
‘Properties’ Search Field

<table>
<thead>
<tr>
<th>Query</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1 hfe</td>
<td>137</td>
</tr>
<tr>
<td>#2 hfe[title] AND human[orgn]</td>
<td>42</td>
</tr>
<tr>
<td>#3 AND srcdb refseq[prop]</td>
<td>11</td>
</tr>
<tr>
<td>#4 AND srcdb ddbj/embl/genbank[prop]</td>
<td>31</td>
</tr>
<tr>
<td>#5 AND gbdiv pri[prop]</td>
<td>29</td>
</tr>
<tr>
<td>#4 AND gbdiv est[prop]</td>
<td>2</td>
</tr>
</tbody>
</table>

Primate division gbdiv pri[prop]
EST division gbdiv est[prop]

More Queries…

Fields are database-specific

Entrez Nucleotide

Reviewed RefSeqs with transcript variants.
srcdb refseq reviewed[prop] AND transcript[title] AND variant[title]

Entrez Gene

Topoisomerase genes from Archaea:
topoisorase[genename] AND archaea[organism]

Genes on human chromosome 2 with OMIM links

Membrane proteins linked to cancer:
“integral to plasma membrane”[gene ontology] AND cancer[dis]
Other Entrez Databases

**UniGene**: rat clusters that have at least one mRNA
  - `rat[organism] NOT @mRNA count`

**SNP**: uniquely mapped microsatellites on human chr2

**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 0.00:0.002.00[resolution]`
Homologene

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

 orthologs

gene duplication
early globin

A-chain gene

Rice Homolog

Genome Resources

Genomic Biology

Homologene

Map Viewer

Entrez Gene
Mouse MapViewer: Gene Filter

Search results for query 'adar AND (gene[obj_type])': 4 hits

<table>
<thead>
<tr>
<th>Chr</th>
<th>Assembly Match</th>
<th>Map element</th>
<th>Type</th>
<th>Map</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>reference</td>
<td>all matches</td>
<td>Adar</td>
<td>Arly</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Adar</td>
<td>Arly</td>
</tr>
<tr>
<td>3</td>
<td>Celera</td>
<td>Adar</td>
<td>Adar</td>
<td>Arly</td>
</tr>
<tr>
<td>3</td>
<td>MGC268</td>
<td>Adar</td>
<td>Adar</td>
<td>Arly</td>
</tr>
</tbody>
</table>

Maps & Options:

- Sequence maps
- Cytogenetic maps
- Phylogenetic maps
- Ideogram
- Map
- Gene
- Genes
- Variations
- RefSeq Transcripts
- Variation
- RNA
- MHC
- STS

Gene annotations:
- Exon
- 3' UTR
- RefSeq RNA
- Tiling path
- Variations
Web BLAST

- pre-computed results
- how BLAST works
  - words; scoring matrices; statistics
- specialized BLAST algorithms
- what’s new, or important
- example oligo search
Precomputed BLAST Services

- Nucleotide or protein: Related Sequences
- BLAST link: Blink
- Transcript clusters: UniGene
- Protein homologs: Homologene

Related Sequences

<table>
<thead>
<tr>
<th>BLAST Link</th>
<th>Description</th>
<th>Link</th>
<th>Link Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniRef100</td>
<td>Mouse cholinergic receptor (CHR), dRNA</td>
<td><img src="#" alt="Link" /></td>
<td>Related Sequences</td>
</tr>
<tr>
<td>dbJBrowse</td>
<td>Mouse choroid plexus choroid plexus (CP), dRNA</td>
<td><img src="#" alt="Link" /></td>
<td>Related Sequences</td>
</tr>
<tr>
<td>LocusLink</td>
<td>Mouse choroid plexus choroid plexus (CP), dRNA</td>
<td><img src="#" alt="Link" /></td>
<td>Related Sequences</td>
</tr>
<tr>
<td>RefSeq</td>
<td>Mouse choroid plexus choroid plexus (CP), dRNA</td>
<td><img src="#" alt="Link" /></td>
<td>Related Sequences</td>
</tr>
</tbody>
</table>

BLink (BLAST Link)

- BLAST Link: Mouse choroid plexus choroid plexus (CP), dRNA | ![Link](#) | Related Sequences |
- Links: Gene, Genome, Homologene, Full text in PMC, Related Sequences, Map Viewer, GEO Profiles, Domain Relat, Genicview in dBSNP, UniGene, Related Sequences, LocusLink, Map Viewer, Nucleotide, Domain, Published, SNP, Taxonomy, UniGene, Links
**BLink Output**

**Query:** globin

**3D structures**

**CDD-Search**

---

**Basic Local Alignment Search Tool**

- local, isolated, “surprising” regions of similarity
- breaks the query sequence into “words”
- word hits to database sequences become “seeds” for alignment extension

---

**How BLAST Works**

1. Make lookup table of “words” for query
2. Scan database for hits
3. Extend alignment both directions
   - Ungapped extensions of hits (initial HSPs)
   - Gapped extensions (no traceback)
   - Gapped extensions (traceback = alignment details)
Nucleotide Words

Make a lookup table based on the word size.

11-mer
ATGCTGCTAGTCGATGACGTAGCTA
ATGCTGCTAGT
TGCTGCTAGTC
GCTGCTAGTCG
...

Protein Words

AIEKCYTGCTLAEADDTA
AIE
IEK
LEK, IDK, IQK, IER, IDR, etc
EKC
KCY
CYT
...

Lookup table, including neighborhood words, is based on word size, score matrix, and threshold.

Scoring Systems - Proteins (BLOSUM62)

| A | 4 |
| B | -1 5 |
| C | 0 -3 -3 -3 9 |
| D | -1 -1 -3 -3 -3 6 |
| E | 0 0 -2 -2 -2 6 |
| F | -2 0 1 -1 -3 0 0 -2 8 |
| G | -1 -3 -3 -3 -3 -3 -4 3 |
| H | -2 -3 -3 -3 -3 -3 -3 3 |
| I | -1 2 0 -1 -3 1 1 -2 -1 |
| J | 0 -2 -2 -2 -2 -2 -2 -2 -2 |
| K | -1 -1 -1 -1 -1 -1 -1 -1 -1 |
| L | 2 -2 -2 -2 -2 -2 -2 -2 -2 |
| M | 1 1 1 1 1 1 1 1 1 |
| N | -2 -2 -2 -2 -2 -2 -2 -2 -2 |
| O | -1 0 0 0 0 0 0 0 0 |
| P | -3 -3 -3 -3 -3 -3 -3 -3 -3 |
| Q | -2 -2 -2 -2 -2 -2 -2 -2 -2 |
| R | -1 0 0 0 0 0 0 0 0 |
| S | 3 3 3 3 3 3 3 3 3 |
| T | -2 -2 -2 -2 -2 -2 -2 -2 -2 |
| U | -1 0 0 0 0 0 0 0 0 |
| V | -1 -1 -1 -1 -1 -1 -1 -1 -1 |
| W | 2 2 2 2 2 2 2 2 2 |
| X | 1 1 1 1 1 1 1 1 1 |

Word Hits & Extensions

Nucleotide: one exact match
ATGCTGCTAGTCGATGACGTAGCTA

Protein: two matches within 40 residues
PHAIEKCYTGCTLAEADDTA
BLASTP Summary

Example query words:

Query: IETVYAAYLPKNTHPFL

Sbjct 287 LEETYAKYLHKGASYFV

Query: LEISPQNVDVNVHPTKHEV

Sbjct 287 LEETYAKYLHKGASYFV

Drop-off score = 

Highest score - current score

-X X dropoff value for gapped alignment (in bits)

blast x, megablast x, tblastx, all others 15

BLASTP Summary (gapped extension)

Query 1: IETVYAAYLPKNTHPFL

Sbjct 287 LEETYAKYLHKGASYFV

Query: LEISPQNVDVNVHPTKHEV

Sbjct 287 LEETYAKYLHKGASYFV

Drop-off score = 

Highest score - current score

-X X dropoff value for gapped alignment (in bits)

blast x, megablast x, tblastx, all others 15

Scoring Systems

Distribution of all Blast Hits on the Query Sequence

Identity matrix

A G C T

T -3 -3 -3 -3

C -3 0 -3 -3

G -3 -3 -3 -3

A -3 -3 -3 -3

raw score = 19-(6+7)* = 6*

Scoring Systems - Nucleotides

CAGTGGCGATCTGTCTG

raw score = 19-(6+7)* = 6*

CAGTGGCGATCTGTCTG

* gap costs -G 5 -E 2
### Scoring Systems - Proteins (BLOSUM62)

| A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
| -4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| -1 | 5 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | -2 | 1 | 6 | -2 | 1 | 6 | -2 | 1 | 6 | -2 | 1 | 6 | -2 | 1 | 6 | -2 | 1 | 6 | -2 | 1 |
| -1 | 0 | 0 | -3 | 9 | -1 | 0 | 0 | -3 | 9 | -1 | 0 | 0 | -3 | 9 | -1 | 0 | 0 | -3 | 9 | -1 |
| -2 | 0 | 0 | -2 | 4 | 2 | 5 | 0 | 0 | -2 | 4 | 2 | 5 | 0 | 0 | -2 | 4 | 2 | 5 | 0 | 0 |
| 0 | -2 | 0 | -1 | 3 | -3 | -6 | 0 | -2 | 0 | -1 | 3 | -3 | -6 | 0 | -2 | 0 | -1 | 3 | -3 |
| -2 | 0 | 1 | -3 | 0 | 0 | -2 | 8 | -2 | 0 | 1 | -3 | 0 | 0 | -2 | 8 | -2 | 0 | 1 | -3 | 0 |
| 1 | -1 | -3 | -3 | -3 | -4 | -3 | 4 | 1 | -1 | -3 | -3 | -4 | -3 | 4 | 1 | -1 | -3 | -3 | -4 | -3 |
| -1 | 0 | -3 | -4 | -3 | -4 | -3 | 4 | -1 | 0 | -3 | -4 | -3 | -4 | -3 | 4 | -1 | 0 | -3 | -4 | -3 |
| -1 | 2 | 0 | -3 | 1 | 1 | 2 | -3 | 5 | -1 | 2 | 0 | -3 | 1 | 1 | 2 | -3 | 5 | -1 | 2 |
| 0 | -1 | -2 | -1 | -2 | -2 | -1 | -2 | 7 | 0 | -1 | -2 | -1 | -2 | -2 | -1 | -2 | 7 | 0 | -1 |
| -1 | 0 | -1 | -1 | -1 | -2 | -1 | -1 | 4 | -1 | 0 | -1 | -1 | -1 | -2 | -1 | -1 | 4 | -1 | 0 |
| -3 | -4 | -4 | -3 | -2 | -3 | -2 | -1 | -2 | -3 | -4 | -3 | -2 | -3 | -2 | -1 | -2 | -3 | -4 | -3 |
| -2 | -2 | -3 | -3 | -3 | -1 | -3 | -3 | -1 | -2 | -3 | -3 | -3 | -1 | -3 | -3 | -1 | -2 | -3 | -3 |

Position-Specific Score Matrix

Serine/Threonine protein kinases
catalytic loop

### Position-Specific Score Matrix

<table>
<thead>
<tr>
<th>Serine/Threonine protein kinases</th>
<th>Catalytic loop</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PSSM scores</strong></td>
<td><strong>Catalytic loop</strong></td>
</tr>
<tr>
<td>1</td>
<td>R G I H R D E P N I L L E A</td>
</tr>
<tr>
<td>5</td>
<td>Q i g g A I A H R D P N I L L E A</td>
</tr>
<tr>
<td><strong>##</strong></td>
<td>A G I H R D P N I L L E A</td>
</tr>
<tr>
<td>4</td>
<td>P A I A H R D P N I L L E A</td>
</tr>
<tr>
<td><strong>##</strong></td>
<td>P A I A H R D P N I L L E A</td>
</tr>
<tr>
<td>4</td>
<td>T F L H R D P N I L L E A</td>
</tr>
<tr>
<td><strong>##</strong></td>
<td>P F Y V H R D P N I L L E A</td>
</tr>
<tr>
<td>4</td>
<td>V P Y H R D P N I L L E A</td>
</tr>
</tbody>
</table>

NCBI FlasterScope

### Scoring Systems - Proteins

**Position Independent Matrices**

PAM Matrices (Percent Accepted Mutation)
- Derived from observation; small dataset of alignments
- Implicit model of evolution
- All calculated from PAM1
- PAM250 widely used

BLOSUM Matrices (BLocK SUbstitution Matrices)
- Derived from observation; large dataset of highly conserved blocks
- Each matrix derived separately from blocks with a defined percent identity cutoff
- **BLOSUM62**: default matrix for BLAST
Expect Value

Local Alignment Statistics

Expect Value

$E = \text{number of database hits you expect to find by chance, } \leq S$

$E = K m^{\frac{1}{2}} S$ or $E = m n^{2 - S}$

K = scale for search space

$\lambda = \text{scale for scoring system}$

$S' = \text{bitscore} = (S - \ln k) / \ln 2$

m = query length

n = database length

E is dependent on m x n (search space)

More info: The Statistics of Sequence Similarity Scores

BLAST is a shortcut...

An alignment BLAST cannot make:

Reason:

no contiguous exact match of 7 bp.
An Alignment BLAST Can Make

Score = 290 hits (741), Expect = 7e-77
Identities = 147/331 (44%), Positives = 206/331 (61%), Gaps = 8/331 (2%)
Frame = +3

MegaBLAST: NCBI’s Genome Annotator

- Long alignments of similar DNA sequences
- Greedy algorithm
- Concatenation of query sequences
- Faster than blastn; less sensitive

<table>
<thead>
<tr>
<th>WORD SIZE</th>
<th>default</th>
<th>minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>blast</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>megablast</td>
<td>28</td>
<td>8</td>
</tr>
</tbody>
</table>

Other BLAST Algorithms

- MegaBLAST
- Discontiguous MegaBLAST
- PSI-BLAST
- PHI-BLAST

Discontiguous MegaBLAST

- Uses discontiguous word matches
- Better for cross-species comparisons
Templates for Discontiguous Words

<table>
<thead>
<tr>
<th>W</th>
<th>t</th>
<th>Coding</th>
<th>Non-Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>16</td>
<td>1010101010110111101010101010111</td>
<td>11111111111111110101010101010101010101</td>
</tr>
<tr>
<td>12</td>
<td>16</td>
<td>11111111111111110101010101010101010101</td>
<td>11111111111111110101010101010101010101</td>
</tr>
<tr>
<td>11</td>
<td>18</td>
<td>1010101010110111</td>
<td>11111111111111110101010101010101010101</td>
</tr>
<tr>
<td>12</td>
<td>18</td>
<td>11111111111111110101010101010101010101</td>
<td>11111111111111110101010101010101010101</td>
</tr>
<tr>
<td>11</td>
<td>21</td>
<td>1010101010110111</td>
<td>11111111111111110101010101010101010101</td>
</tr>
<tr>
<td>12</td>
<td>21</td>
<td>11111111111111110101010101010101010101</td>
<td>11111111111111110101010101010101010101</td>
</tr>
<tr>
<td>11</td>
<td>21</td>
<td>1010101010110111</td>
<td>11111111111111110101010101010101010101</td>
</tr>
<tr>
<td>12</td>
<td>21</td>
<td>11111111111111110101010101010101010101</td>
<td>11111111111111110101010101010101010101</td>
</tr>
</tbody>
</table>

W = word size; if matches in template
= template length


Discontiguous Word Options

Options for advanced BLAST

- Query length
- Template type
- Template length
- Discontiguous word size
- Other advanced options
Megablast in New Interface

Disco. Megablast Example . . .

Query: NM_078651
Drosophila melanogaster CG18582-PA (mbt) mRNA, (3244 bp)
\(\text{note= mushroom bodies tiny; synonyms: Pak2, STE20, dPAK2} \)

Database: \(\text{nr (nt), Mammalia[orgn]}\)

- MegaBLAST = poor hits
- Discontiguous megaBLAST = numerous hits . . .
Example: Discontiguous MegaBLAST

**Distribution of 96 Blast Hits on the Query Sequence**

<table>
<thead>
<tr>
<th>Query Score</th>
<th>&lt;=40</th>
<th>40-60</th>
<th>60-80</th>
<th>80-100</th>
<th>&gt;100</th>
</tr>
</thead>
</table>

Color key for alignment scores:

- Blue: <=40
- Green: 40-60
- Yellow: 60-80
- Orange: 80-100
- Red: >100

Score = 365 bits (190), Expect = 2e-97
Identities = 624/926 (74%), Gaps = 2/926 (0%)
Strand=Plus/Plus

Example: N4/N6 Cytosine Methyltransferase

**PSI-BLAST**

Position-specific Iterated BLAST

Search Parameter Settings:
- Query: P11409
- Subsequence Database: swissprot
- PSI-BLAST selected
- Max Hits: 1000
- Expect: 1
- Inclusion: 0.005
Convergence

Results of PSI-Blast iteration 6
No new sequences were found above the 0.005 threshold!

Query Specific PSSM vs BLOSUM62

Last position-specific scoring matrix computed, weighted observed perc:

Nucleotide BLAST Databases
- **nr (nt)**
  - Traditional GenBank Divisions
  - NM, XM, RefSeqs
- **refseq_rna**
  - NM, XM, NR
- **refseq_genomic**
  - NC, NT, NC
- **est**
  - EST Division
- **wgs**
  - HTC division
- **dbst**
  - STS Division
- **combined, for human and mouse**
  - chromosome
    - NC genomic records
  - **gss**
    - GSS division
  - **wgs**
    - wgs entries from traditional divisions
  - **pdb**
    - Nucleotide sequences from structures
  - **env_nt**
    - environmental samples
Protein BLAST Databases

- nr: traditional GenBank records
- refseq = NP_, XP_
- swissprot
- pdb
- pat
- env_nr

New Nucleotide Databases

Two new Human and Mouse databases combine genomic plus transcript alignments in a single report. You can also choose from Others to use nr 1 or an existing database.

New Formatter

Select lower case
Select red

BLAST Output: Alignments & Filter

low complexity sequence filtered
Example: Mapping Oligos Onto a Genome

Example: Mapping Oligos Onto a Genome
Genome BLAST Results

RID: 107695772-31416-1775591455.61AP9Q8
Query: CCTAGGGACCCCTGGGAAAAACERERERRRRRRRRRRRCAACCCCCTGGGCT(TG)GGG
(52 letters)
Database: contig
g98 sequence(s): 5,005,300,271 total letters

If you have any problems or questions with the results of this search
please refer to the BLAST FAQ.

Primer Alignments

MapViewer

MapViewer
Sequence View (sv)

Service Addresses

• BLAST  blast-help@ncbi.nlm.nih.gov
• General Help  info@ncbi.nlm.nih.gov
• Wayne Matten  matten@ncbi.nlm.nih.gov