Sample GenBank Record

PubMed Entrez BLAST OMIM Taxonomy Structure

GenBank Flat File Format

Click on any link in this sample record to see a detailed description of that data element or field. All of the descriptions are included on this page, so it can be printed as a single document. You can also return to the Alphabetical Quicklinks Table or Resource Guide

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
     CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
                     /function="required for axial budding pattern of S.
                     cerevisiae"
                     /product="Axl2p"
                     /protein_id="AAA98666.1"
                     /db_xref="GI:1293615"
                     /translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
                     TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN
                     VILEGTDSADSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE
                     VFNVTFDRSMFTNEESIVSYYGRSQLYNAPLPNWLFFDSGELKFTGTAPVINSAIAPE
                     TSYSFVIIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLIINVTDTGNVSYDLPLNYV
                     YLDDDPISSDKLGSINLLDAPDWVALDNATISGSVPDELLGKNSNPANFSVSIYDTYG
                     DVIYFNFEVVSTTDLFAISSLPNINATRGEWFSYYFLPSQFTDYVNTNVSLEFTNSSQ
                     DHDWVKFQSSNLTLAGEVPKNFDKLSLGLKANQGSQSQELYFNIIGMDSKITHSNHSA
                     NATSTRSSHHSTSTSSYTSSTYTAKISSTSAAATSSAPAALPAANKTSSHNKKAVAIA
                     CGVAIPLGVILVALICFLIFWRRRRENPDDENLPHAISGPDLNNPANKPNQENATPLN
                     NPFDDDASSYDDTSIARRLAALNTLKLDNHSATESDISSVDEKRDSLSGMNTYNDQFQ
                     SQSKEELLAKPPVQPPESPFFDPQNRSSSVYMDSEPAVNKSWRYTGNLSPVSDIVRDS
                     YGSQKTVDTEKLFDLEAPEKEKRTSRDVTMSSLDPWNSNISPSPVRKSVTPSPYNVTK
                     HRNRHLQNIQDSQSGKNGITPTTMSTSSSDDFVPVKDGENFCWVHSMEPDRRPSKKRL
                     VDFSNKSNVNVGQVKDIHGRIPEML"
     gene            complement(3300..4037)
                     /gene="REV7"
     CDS             complement(3300..4037)
                     /gene="REV7"
                     /codon_start=1
                     /product="Rev7p"
                     /protein_id="AAA98667.1"
                     /db_xref="GI:1293616"
                     /translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ
                     FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD
                     KDDQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR
                     RVDSLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
                     LISGDDKILNGVYSQYEEGESIFGSLF"
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
      241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
      301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
      361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
      421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
      481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc
      541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta gaagaacaga
      601 acaattactt aatagaaaaa ttatatcttc ctcgaaacga tttcctgctt ccaacatcta
      661 cgtatatcaa gaagcattca cttaccatga cacagcttca gatttcatta ttgctgacag
      721 ctactatatc actactccat ctagtagtgg ccacgcccta tgaggcatat cctatcggaa
      781 aacaataccc cccagtggca agagtcaatg aatcgtttac atttcaaatt tccaatgata
      841 cctataaatc gtctgtagac aagacagctc aaataacata caattgcttc gacttaccga
      901 gctggctttc gtttgactct agttctagaa cgttctcagg tgaaccttct tctgacttac
      961 tatctgatgc gaacaccacg ttgtatttca atgtaatact cgagggtacg gactctgccg
     1021 acagcacgtc tttgaacaat acataccaat ttgttgttac aaaccgtcca tccatctcgc
     1081 tatcgtcaga tttcaatcta ttggcgttgt taaaaaacta tggttatact aacggcaaaa
     1141 acgctctgaa actagatcct aatgaagtct tcaacgtgac ttttgaccgt tcaatgttca
     1201 ctaacgaaga atccattgtg tcgtattacg gacgttctca gttgtataat gcgccgttac
     1261 ccaattggct gttcttcgat tctggcgagt tgaagtttac tgggacggca ccggtgataa
     1321 actcggcgat tgctccagaa acaagctaca gttttgtcat catcgctaca gacattgaag
     1381 gattttctgc cgttgaggta gaattcgaat tagtcatcgg ggctcaccag ttaactacct
     1441 ctattcaaaa tagtttgata atcaacgtta ctgacacagg taacgtttca tatgacttac
     1501 ctctaaacta tgtttatctc gatgacgatc ctatttcttc tgataaattg ggttctataa
     1561 acttattgga tgctccagac tgggtggcat tagataatgc taccatttcc gggtctgtcc
     1621 cagatgaatt actcggtaag aactccaatc ctgccaattt ttctgtgtcc atttatgata
     1681 cttatggtga tgtgatttat ttcaacttcg aagttgtctc cacaacggat ttgtttgcca
     1741 ttagttctct tcccaatatt aacgctacaa ggggtgaatg gttctcctac tatttttgc
     1801 cttctcagtt tacagactac gtgaatacaa acgtttcatt agagtttact aattcaagcc
     1861 aagaccatga ctgggtgaaa ttccaatcat ctaatttaac attagctgga gaagtgccca
     1921 agaatttcga caagctttca ttaggtttga aagcgaacca aggttcacaa tctcaagagc
     1981 tatattttaa catcattggc atggattcaa agataactca ctcaaaccac agtgcgaatg
     2041 caacgtccac aagaagttct caccactcca cctcaacaag ttcttacaca tcttctactt
     2101 acactgcaaa aatttcttct acctccgctg ctgctacttc ttctgctcca gcagcgctgc
     2161 cagcagccaa taaaacttca tctcacaata aaaaagcagt agcaattgcg tgcggtgttg
     2221 ctatcccatt aggcgttatc ctagtagctc tcatttgctt cctaatattc tggagacgca
     2281 gaagggaaaa tccagacgat gaaaacttac cgcatgctat tagtggacct gatttgaata
     2341 atcctgcaaa taaaccaaat caagaaaacg ctacaccttt gaacaacccc tttgatgatg
     2401 atgcttcctc gtacgatgat acttcaatag caagaagatt ggctgctttg aacactttga
     2461 aattggataa ccactctgcc actgaatctg atatttccag cgtggatgaa aagagagatt
     2521 ctctatcagg tatgaataca tacaatgatc agttccaatc ccaaagtaaa gaagaattat
     2581 tagcaaaacc cccagtacag cctccagaga gcccgttctt tgacccacag aataggtctt
     2641 cttctgtgta tatggatagt gaaccagcag taaataaatc ctggcgatat actggcaacc
     2701 tgtcaccagt ctctgatatt gtcagagaca gttacggatc acaaaaaact gttgatacag
     2761 aaaaactttt cgatttagaa gcaccagaga aggaaaaacg tacgtcaagg gatgtcacta
     2821 tgtcttcact ggacccttgg aacagcaata ttagcccttc tcccgtaaga aaatcagtaa
     2881 caccatcacc atataacgta acgaagcatc gtaaccgcca cttacaaaat attcaagact
     2941 ctcaaagcgg taaaaacgga atcactccca caacaatgtc aacttcatct tctgacgatt
     3001 ttgttccggt taaagatggt gaaaattttt gctgggtcca tagcatggaa ccagacagaa
     3061 gaccaagtaa gaaaaggtta gtagattttt caaataagag taatgtcaat gttggtcaag
     3121 ttaaggacat tcacggacgc atcccagaaa tgctgtgatt atacgcaacg atattttgct
     3181 taattttatt ttcctgtttt atttttatt agtggtttac agatacccta tattttattt
     3241 agtttttata cttagagaca tttaatttta attccattct tcaaatttca tttttgcact
     3301 taaaacaaag atccaaaaat gctctcgccc tcttcatatt gagaatacac tccattcaaa
     3361 attttgtcgt caccgctgat taatttttca ctaaactgat gaataatcaa aggccccacg
     3421 tcagaaccga ctaaagaagt gagttttatt ttaggaggtt gaaaaccatt attgtctggt
     3481 aaattttcat cttcttgaca tttaacccag tttgaatccc tttcaatttc tgctttttcc
     3541 tccaaactat cgaccctcct gtttctgtcc aacttatgtc ctagttccaa ttcgatcgca
     3601 ttaataactg cttcaaatgt tattgtgtca tcgttgactt taggtaattt ctccaaatgc
     3661 ataatcaaac tatttaagga agatcggaat tcgtcgaaca cttcagtttc cgtaatgatc
     3721 tgatcgtctt tatccacatg ttgtaattca ctaaaatcta aaacgtattt ttcaatgcat
     3781 aaatcgttct ttttattaat aatgcagatg gaaaatctgt aaacgtgcgt taatttagaa
     3841 agaacatcca gtataagttc ttctatatag tcaattaaag caggatgcct attaatggga
     3901 acgaactgcg gcaagttgaa tgactggtaa gtagtgtagt cgaatgactg aggtgggtat
     3961 acatttctat aaaataaaat caaattaatg tagcatttta agtataccct cagccacttc
     4021 tctacccatc tattcataaa gctgacgcaa cgattactat tttttttttc ttcttggatc
     4081 tcagtcgtcg caaaaacgta taccttcttt ttccgacctt ttttttagct ttctggaaaa
     4141 gtttatatta gttaaacagg gtctagtctt agtgtgaaag ctagtggttt cgattgactg
     4201 atattaagaa agtggaaatt aaattagtag tgtagacgta tatgcatatg tatttctcgc
     4261 ctgtttatgt ttctacgtac ttttgattta tagcaagggg aaaagaaata catactattt
     4321 tttggtaaag gtgaaagcat aatgtaaaag ctagaataaa atggacgaaa taaagagagg
     4381 cttagttcat ctttttttcca aaaagcaccc aatgataata actaaaatga aaaggatttg
     4441 ccatctgtca gcaacatcag ttgtgtgagc aataataaaa tcatcacctc cgttgccttt
     4501 agcgcgtttg tcgtttgtat cttccgtaat tttagtctta tcaatgggaa tcataaattt
     4561 tccaatgaat tagcaatttc gtccaattct tttgagctt cttcatattt gctttggaat
     4621 tcttcgcact tcttttccca ttcatctctt tcttcttcca aagcaacgat ccttctaccc
     4681 atttgctcag agttcaaatc ggcctctttc agtttatcca ttgcttcctt cagtttggct
     4741 tcactgtctt ctagctgttg ttctagatcc tggtttttct tggtgtagtt ctcattatta
     4801 gatctcaagt tattggagtc ttcagccaat tgctttgtat cagacaattg actctctaac
     4861 ttctccactt cactgtcgag ttgctcgttt ttagcggaca aagatttaat ctcgttttct
     4921 ttttcagtgt tagattgctc taattctttg agctgttctc tcagctcctc atatttttct
     4981 tgccatgact cagattctaa ttttaagcta ttcaatttct ctttgatc
//
```

The corresponding live record for U49845 can be viewed in Entrez.
Examples of other records that show a range of biological features are listed below.

FIELD **COMMENTS**

LOCUS

The LOCUS field contains a number of different data elements, including locus name, sequence length, molecule type, GenBank division, and modification date. Each element is described below.

- **Locus Name**

The locus name in this example is SCU49845.

The locus name was originally designed to help group entries with similar sequences: the first three characters usually designated the organism; the fourth and fifth characters were used to show other group designations, such as gene product; for segmented entries, the last character was one of a series of sequential integers. (See GenBank release notes section 3.4.4 for more info.)

However, the 10 characters in the locus name are no longer sufficient to represent the amount of information originally intended to be contained in the locus name. The only rule now applied in assigning a locus name is that it must be unique. For example, for GenBank records that have 6-character accessions (e.g., U12345), the locus name is usually the first letter of the genus and species names, followed by the accession number. For 8-character character accessions (e.g., AF123456), the locus name is just the accession number.

The RefSeq database of reference sequences assigns formal locus names to each record, based on gene symbol. RefSeq is separate from the GenBank database, but contains cross-references to corresponding GenBank records.

Entrez Search Field: Accession Number [ACCN]
Search Tip: It is better to search for the actual accession number rather than the locus name, because the accessions are stable and locus names can change.

- **Sequence Length**

Number of nucleotide base pairs (or amino acid residues) in the sequence record. In this example, the sequence length is 5028 bp.

There is no maximum limit on the size of a sequence that can be submitted to GenBank. You can submit a whole genome if you have a contiguous piece of sequence from a single molecule type. However, there is a limit of 350 kb on an individual GenBank record (with some exceptions, as noted in section 1.3.2 of the release notes for GenBank 112.0 target="one"). That limit was agreed upon by the international collaboratoring sequence databases to facilitate handling of sequence data by various software programs. (For more information, see NCBI News articles on Complete Genomes and GenBank Enters Megabase Era.) The minimum length required for submission is 50 bp, although there might be some shorter records from past years.

Entrez Search Field: Sequence Length [SLEN]
Search Tips: (1) To retrieve records within a range of lengths, use the colon as the range operator, e.g., 2500:2600[SLEN]. (2) To retrieve all sequences shorter than a certain number, use 2 as the lower bound, e.g., 2:100[SLEN]. (3) To retrieve all sequences longer than a certain number, use a series of 9's as the upper bound, e.g., 325000:99999999[SLEN].

- **Molecule Type**

The type of molecule that was sequenced. In this example, the molecule type is DNA.

Each GenBank record must contain contiguous sequence data from a single molecule type. The various molecule types are described in the Sequin documentation and can include genomic DNA, genomic RNA, precursor RNA, mRNA (cDNA), ribosomal RNA, transfer RNA, small nuclear RNA, and small cytoplasmic RNA.

Entrez Search Field: Properties [PROP]
Search Tip: Search term should be in the format: **biomol_genomic**, **biomol_mRNA**, etc. For more examples, view the Properties field in the Index mode.

- **GenBank Division**

The GenBank division to which a record belongs is indicated with a three letter abbreviation. In this example, GenBank division is PLN.

The GenBank database is divided into 18 divisions:

1. PRI - primate sequences
2. ROD - rodent sequences
3. MAM - other mammalian sequences
4. VRT - other vertebrate sequences
5. INV - invertebrate sequences
6. PLN - plant, fungal, and algal sequences
7. BCT - bacterial sequences
8. VRL - viral sequences
9. PHG - bacteriophage sequences
10. SYN - synthetic sequences
11. UNA - unannotated sequences
12. EST - EST sequences (expressed sequence tags)
13. PAT - patent sequences
14. STS - STS sequences (sequence tagged sites)
15. GSS - GSS sequences (genome survey sequences)
16. HTG - HTG sequences (high-throughput genomic sequences)
17. HTC - unfinished high-throughput cDNA sequencing
18. ENV - environmental sampling sequences

Some of the divisions contain sequences from specific groups of organisms, whereas others (EST, GSS, HTG, etc.) contain data generated by specific sequencing technologies from many different organisms. The **organismal divisions are historical and do not reflect the current NCBI Taxonomy**. Instead, they merely serve as a convenient way to divide GenBank into smaller pieces for those who want to FTP the database. Because of this, and because sequences from a particular organism can exist in technology-based divisions such as EST, HTG, etc., **the NCBI Taxonomy Browser should be used for retrieving all sequences from a particular organism**.

The divisions are also listed in section 3.3 of the GenBank release notes.

The RNA division of GenBank was removed in release 113.0 (August 1999). Sequences that were previously in the RNA division have been moved to the appropriate organismal division. (See section 1.3.2 of the GenBank 113.0 release notes for additional information.)

The HTC division was added to GenBank in release 123.0 (April 2001) and is described in Section 1.3.3 of the GenBank 123.0 release notes.

Another division, called CON, was added in release 115.0 (December 1999) but is not listed above because it records in that division contain no sequence data. Instead, they contain sequence assembly instructions on how to construct contigs from multiple GenBank records. See the Fall 1999 NCBI News and section 1.3.3 of GenBank 115.0 release notes for details.

Entrez Search Field: Properties [PROP]
Search Tip: Search term should be in the format: **gbdiv_pri**, **gbdiv_est**, etc. For more examples, view the Properties field in the Index mode. For example, to eliminate all sequences from a particular division, such as all ESTs, you can use a Boolean query formatted such as:

```
human[ORGN] NOT gbdiv_est[PROP]
```

For the reasons noted above, **do not use GenBank divisions to retrieve all sequences from a specific organism. Instead, use the NCBI Taxonomy Browser**.

- **Modification Date**

The date in the LOCUS field is the **date of last modification**. The sample record shown here was last modified on 21-JUN-1999.

In some cases, the modification date might correspond to the release date, but there is no way to tell just by looking at the record. If you need to know the first date of public availability for a specific sequence record, send a message to info@ncbi.nlm.nih.gov. We will check the history of the record for you, and let you know the date of first public release. If the sequence was originally submitted to our collaborators at DDBJ or EMBL, rather than to GenBank, we will ask them to send the release date information to you. (See also notes re: date in the Direct Submission reference.)

Entrez Search Field: Modification Date [MDAT]
Search Tips: (1) Enter search term in the format: yyyy/mm/dd, e.g., 1999/07/25. (2) To retrieve records modified between two dates, use the colon as a range operator, e.g., 1999/07/25:1999/07/31[MDAT]. (3) You can use the Publication Date [PDAT] field of Entrez to limit search results by the date on which records were added to the Entrez system. Publication date can be in the form of a range, just like the Modification Date.

DEFINITION

Brief description of sequence; includes information such as source organism, gene name/protein name, or some description of the sequence's function (if the sequence is non-coding). If the sequence has a coding region (CDS), description may be followed by a completeness qualifier, such as "complete cds". (See GenBank release notes section 3.4.5 for more info.)

Entrez Search Field: Title Word [TITL]
Search Tip: Although nucleotide definition lines follow a structured format, GenBank does not use a controlled vocabulary, and authors determine the content of their records. Therefore, if a search for a specific term does not retrieve the desired records, try other terms that authors might have used, such as synonyms, full spellings, or abbreviations. The "related records" (or "neighbors") function of Entrez also allows you to broaden your search by retrieving records with similar sequences, regardless of the descriptive terms used by the submitters.

ACCESSION

The unique identifier for a sequence record. An accession number applies to the complete record and is usually a combination of a letter(s) and numbers, such as a single letter followed by five digits (e.g., U12345) or two letters followed by six digits (e.g., AF123456). Some accessions might be longer, depending on the type of sequence record.

Accession numbers do not change, even if information in the record is changed at the author's request. Sometimes, however, an original accession number might become secondary to a newer accession number, if the authors make a new submission that combines previous sequences, or if for some reason a new submission supercedes an earlier record.

Records from the RefSeq database of reference sequences have a different accession number format that begins with two letters followed by an underscore bar and six or more digits, for example:

```
NT_123456   constructed genomic contigs
NM_123456   mRNAs
NP_123456   proteins
NC_123456   chromosomes
```

Note: compare accession number with Sequence Identifiers such as Version and GI for nucleotide sequences and protein_id and GI for amino acid sequences.

Entrez Search Field: Accession [ACCN]
Search Tip: The letters in the accession number can be written in upper- or lowercase. RefSeq accessions must contain an underscore bar between the letters and the numbers, e.g., NM_002111.

VERSION

A nucleotide sequence identification number that represents a single, specific sequence in the GenBank database. This identification number uses the accession.version format implemented by GenBank/EMBL/DDBJ in February 1999.

If there is any change to the sequence data (even a single base), the version number will be increased, e.g., U12345.1 → U12345.2, but the accession portion will remain stable.

The accession.version system of sequence identifiers runs parallel to the GI number system, i.e., when any change is made to a sequence, it receives a new GI number AND an increase to its version number.

For more information, see section 1.3.2 of the GenBank 111.0 release notes, and section 3.4.7 of the current GenBank release notes.

A Sequence Revision History tool is available to track the various GI numbers, version numbers, and update dates for sequences that appeared in a specific GenBank record (more information and example).

More details about sequence identification numbers and the difference between GI number and version are provided in Sequence Identifiers: A Historical Note.

Entrez Search Field: use the default setting of "All Fields"

- ### GI

"GenInfo Identifier" sequence identification number, in this case, for the nucleotide sequence. If a sequence changes in any way, a new GI number will be assigned.

A separate GI number is also assigned to each protein translation within a nucleotide sequence record, and a new GI is assigned if the protein translation changes in any way (see below).

GI sequence identifiers run parallel to the new **accession.version** system of sequence identifiers. For more information, see the description of Version, above, and section 3.4.7 of the current GenBank release notes.

A Sequence Revision History tool is available to track the various GI numbers, version numbers, and update dates for sequences that appeared in a specific GenBank record (more information and example).

More details about sequence identification numbers and the difference between GI number and version are provided in Sequence Identifiers: A Historical Note.

Entrez Search Field: use the default setting of "All Fields"

KEYWORDS

Word or phrase describing the sequence. If no keywords are included in the entry, the field contains only a period.

The Keywords field is present in sequence records primarily for historical reasons, and is not based on a controlled vocabulary. Keywords are generally present in older records. They are **not** included in newer records unless: (1) they are not redundant with any feature, qualifier, or other information present in the record; or (2) the submitter specifically asks for them to be added and #1 is true; or (3) the record contains a special type of sequence such as EST, STS, GSS, HTG, etc.

Entrez Search Field: Keyword [KYWD]
Search Tip: Because keywords are not present in many records, it is best not to search that field. Instead, search All Fields [ALL], the Text Word [WORD] field, or the Title Word [TITL] field, for progressively narrower retrieval.

SOURCE

Free-format information including an abbreviated form of the organism name, sometimes followed by a molecule type. (See section 3.4.10 of the GenBank release notes for more info.)

Entrez Search Field: Organism [ORGN]
Search Tip: For some organisms that have well-established common names, such as baker's yeast, mouse, and human, a search for the common name will yield the same results as a search for the scientific name, e.g., a search for "baker's yeast" in the organism field retrieves the same number of documents as "Saccharomyces cerevisiae". This is true because the Organism field is connected to the NCBI Taxonomy Database, which contains cross-references between common names, scientific names, and synonyms for organisms represented in the Sequence databases.

- ### Organism

The formal scientific name for the source organism (genus and species, where appropriate) and its lineage, based on the phylogenetic classification scheme used in the NCBI

Taxonomy Database. If the complete lineage of an organism is very long, an abbreviated lineage will be shown in the GenBank record and the complete lineage will be available in the Taxonomy Database. (See also the /db_xref=taxon:nnnn Feature qualifer, below.)

Entrez Search Field: Organism [ORGN]
Search Tip: You can search the Organism field by any node in the taxonomic hierarchy, e.g., you can search for the term "Saccharomyces cerevisiae", "Saccharomycetales", "Ascomycota", etc. to retrieve all the sequences from organisms in a particular taxon.

REFERENCE

Publications by the authors of the sequence that discuss the data reported in the record. References are automatically sorted within the record based on date of publication, showing the oldest references first.

Some sequences have not been reported in papers and show a status of "unpublished" or "in press". When an accession number and/or sequence data has appeared in print, sequence authors should send the complete citation of the article to update@ncbi.nlm.nih.gov and the GenBank staff will revise the record.

Various classes of publication can be present in the References field, including journal article, book chapter, book, thesis/monograph, proceedings chapter, proceedings from a meeting, and patent.

The **last citation** in the REFERENCE field usually contains information about the submitter of the sequence, rather than a literature citation. It is therefore called the "**submitter block**" and shows the words "**Direct Submission**" instead of an article title. Additional information is provided below, under the header **Direct Submission**. Some older records do not contain a submitter block.

Entrez Search Field: The various subfields under References are searchable in the Entrez search fields noted below.

- **AUTHORS**

List of authors in the order in which they appear in the cited article.

Entrez Search Field: Author [AUTH]
Search Tip: Enter author names in the form: Lastname AB (without periods after the initials). Initials can be omitted. Truncation can also be used to retrieve all names that begin with a character string, e.g., Richards* or Boguski M*.

- **TITLE**

Title of the published work or tentative title of an unpublished work.

Sometimes the words "**Direct Submission**" instead of an article title. This is usually true for the last citation in the REFERENCE field because it tends to contain information about the submitter of the sequence, rather than a literature citation. The last citation is therefore called the "**submitter block**". Additional information is provided below, under the header **Direct Submission**. Some older records do not contain a submitter block.

Entrez Search Field: Text Word [WORD]
Note: For sequence records, the Title Word [TITL] field of Entrez searches the Definition Line, not the titles of references listed in the record. Therefore, use the Text Word field to search the titles of references (and other text-containing fields).
Search Tip: If a search for a specific term does not retrieve the desired records, try other terms that authors might have used, such synonyms, full spellings, or abbreviations. The 'related records' (or 'neighbors') function of Entrez also allows you to broaden your search by retrieving records with similar sequences, regardless of the descriptive terms used by the submitters.

- **JOURNAL**

MEDLINE abbreviation of the journal name. (Full spellings can be obtained from the Entrez Journals Database.)

Entrez Search Field: Journal Name [JOUR]
Search Tip: Journal names can be entered as either the full spelling or the MEDLINE abbreviation. You can search the Journal Name field in the Index mode to see the index for that field, and to select one or more journal names for inclusion in your search.

- **PUBMED**

PubMed Identifier (PMID).

References that include PubMed IDs contain links from the sequence record to the corresponding PubMed record. Conversely, PubMed records that contain accession number(s) in the SI (secondary source identifier) field contain links back to the sequence record(s).

Entrez Search Field: It is not possible to search the Nucleotide or Protein sequence databases by PubMed ID. However, you can search the PubMed (literature) database of Entrez for the PubMed ID, and then link to the associated sequence records.

- **Direct Submission**

Contact information of the submitter, such as institute/department and postal address. This is always the last citation in the References field. Some older records do not contain the "Direct Submission" reference. However, it is required in all new records.

The Authors subfield contains the submitter name(s), Title contains the words "Direct Submission", and Journal contains the address.

The date in the Journal subfield is the date on which the author prepared the submission. In many cases, it is also the date on which the sequence was received by the GenBank staff, but it is not the date of first public release. If you need to know the latter, send a message to info@ncbi.nlm.nih.gov. We will check the history of the record for you.

Entrez Search Field: Use the Author Field [AUTH] if searching for the author name. Use All Fields [ALL] if searching for an element of the author's address (e.g., Yale University). Note, however, that retrieved records might contain the institution name in a field such as Comment, rather than in the Direct Submission reference, so you might get some false hits.
Search Tip: It is sometimes helpful to search for both the full spelling and an abbreviation, e.g., "Washington University" OR "WashU", because the spelling used by authors might vary.

FEATURES

Information about genes and gene products, as well as regions of biological significance reported in the sequence. These can include regions of the sequence that code for proteins and RNA molecules, as well as a number of other features. (See section 3.4.12 of the GenBank release notes for more info.)

A **complete list of features** is available in the following places:

- Appendix III: Feature keys reference of the DDBJ/EMBL/GenBank Feature Table provides definitions, optional qualifiers, and comments for each feature. An alphabetical list is also available. Appendix IV: Summary of qualifiers for feature keys provides definitions for the Feature qualifiers.
- Sequin Help documentation (scroll down to 'Features' in the table of contents to see an alphabetical list of features with links to descriptions)
- section 3.4.12.1 of the GenBank release notes

The **location of each feature** is provided as well, and can be a single base, a contiguous span of bases, a joining of sequence spans, and other representations. If a feature is located on the complementary strand, the word "complement" will appear before the base span. If the "**<**" symbol precedes a base span, the sequence is partial on the 5' end (e.g., CDS <1..206). If the "**>**" symbol follows a base span, the sequence is partial on the 3' end (e.g., CDS 435..915>).

For more information about feature locations, see the Sequin Help Documentation and section 3.4.12.2 of the

GenBank release notes.

The sample record shown here only includes a small number of features (source, CDS, and gene, all of which are described below). The Other Features section, below, provides links to some GenBank records that show a variety of additional features.

Entrez Search Field: Feature Key [FKEY]
Search Tip: To scroll through the list of available features, view the Feature Key field in Index mode. You can then select one or more features from the index to include in your query. For example, you can limit your search to records that contain both primer_bind and promoter features.

- **source**

Mandatory feature in each record that summarizes the length of the sequence, scientific name of the source organism, and Taxon ID number. Can also include other information such as map location, strain, clone, tissue type, etc., if provided by submitter.

Entrez Search Field: All Fields [ALL] can be used to search for some elements in the source field, such as strain, clone, tissue type.

Use the Sequence Length [SLEN] field to search by length and the Organism [ORGN] field to search by organism name.

Because map location is written as free text and can be represented in a number of ways (e.g., chromosome number, cytogenetic location, marker name, physical map location), it is not directly searchable in the Entrez Nucleotide or Protein databases. However, there are a number of resources that allow you to browse and/or search the maps of various genomes.

Taxon

A stable unique identification number for the taxon of the source oganism. A taxonomy ID number is assigned to each taxon (species, genus, family, etc.) in the NCBI Taxonomy Database. See also the Organism field, above.

Entrez Search Field: The Taxonomy ID number is not seachable in the Organism search field of Entrez but is searchable in the Taxonomy Browser.

Note: The /db_xref qualifier is one of many that can be applied to various features. A complete list is available in Appendix IV: Summary of qualifiers for feature keys of the DDBJ/EMBL/GenBank Feature Table, and in section 3.4.12.3 of the GenBank release notes. Appendix III: Feature keys reference shows which qualifiers can be used with specific features (see alphabetical list).

- **CDS**

Coding sequence; region of nucleotides that corresponds with the sequence of amino acids in a protein (location includes start and stop codons). The CDS feature includes an amino acid translation. Authors can specify the nature of the CDS by using the qualifier "/evidence=experimental" or "/evidence=not_experimental".

Submitters are also encouraged to annotate the mRNA feature, which includes the 5' untranslated region (5'UTR), coding sequences (CDS, exon), and 3' untranslated region (3'UTR).

Entrez Search Field: Feature Key [FKEY]
Search Tip: You can use this field to limit your search to records that contain a particular feature, such as CDS. To scroll through the list of available features, view the Feature Key field in Index mode. A complete list of features is also available from the resources noted above.

- **<1..206**

Base span of the biological feature indicated to the left, in this case, a CDS feature. (The CDS feature is described above, and its base span includes the start and stop codons.) **Features can be complete, partial on the 5' end, partial on the 3' end, and/or on the complementary strand.** Examples:

1. **complete** feature is simply written as ***n*..*m***

   Example: 687..3158
   The feature extends from base 687 through base 3158 in the sequence shown

2. **<** indicates **partial on the 5' end**

   Example: <1..206
   The feature extends from base 1 through base 206 in the sequence shown, and is partial on the 5' end

3. **>** indicates **partial on the 3' end**

   Example: 4821..5028>
   The feature extends from base 4821 through base 5028 and is partial on the 3' end

4. **(complement)** indicates that the feature is on the complementary strand

   Example: complement(3300..4037)
   The feature extends from base 3300 through base 4037 but is actually on the complementary strand. It is therefore read in the opposite direction on the reverse complement sequence. (For an example, see the third

CDS feature in the sample record shown on this page. In this case, the amino acid translation is generated by taking the reverse complement of bases 3300 to 4037 and reading that reverse complement sequence in its 5' to 3' direction.)

protein_id

↑

A protein sequence identification number, similar to the Version number of a nucleotide sequence. Protein IDs consist of three letters followed by five digits, a dot, and a version number. If there is any change to the sequence data (even a single amino acid), the version number will be increased, but the accession portion will remain stable (e.g., AAA98665.1 will change to AAA98665.2).

The accession.version format of protein sequence identification numbers was implemented by GenBank/EMBL/DDBJ in February 1999 and runs parallel to the GI number system. More details about sequence identification numbers and the difference between GI number and version are provided in Sequence Identifiers: A Historical Note.

Entrez Search Field: use the default setting of "All Fields"

GI

↑

"GenInfo Identifier" sequence identification number, in this case, for the protein translation.

The **GI** system of sequence identifiers runs parallel to the **accession.version** system, which was implemented by GenBank, EMBL, and DDBJ in February 1999. Therefore, if the protein sequence changes in any way, it will receive a new GI number, and the suffix of the protein_id will be incremented by one.

For more information, see the description of protein_id, above, section 1.3.2 of the GenBank 111.0 release notes, and section 3.4.7 of the current GenBank release notes.

More details about sequence identification numbers and the difference between GI number and version are provided in Sequence Identifiers: A Historical Note.

Entrez Search Field: use the default setting of "All Fields"

translation

↑

The amino acid translation corresponding to the nucleotide coding sequence (CDS). In many cases, the translations are conceptual. Note that authors can indicate whether the CDS is based on experimental or non-experimental evidence.

Entrez Search Field: It is not possible to search the translation subfield using Entrez. If you want use a string of amino acids as a query to retrieve similar protein sequences, use BLAST instead.

- **gene**

A region of biological interest identified as a gene and for which a name has been assigned. The base span for the gene feature is dependent on the furthest 5' and 3' features. Additional examples of records that show the relationship between gene features and other features such as mRNA and CDS are AF165912 and AF090832.

Entrez Search Field: Feature Key [FKEY]
Search Tip: You can use this field to limit your search to records that contain a particular feature, such as a gene. To scroll through the list of available features, view the Feature Key field in Index mode. A complete list of features is also available from the resources noted above.

complement

Indicates that the feature is located on the complementary strand.

- **Other Features**

Examples of other records that show a variety of biological features; a graphic format is also available for each sequence record and visually represents the annotated features:

- **AF165912** (gene, promoter, TATA signal, mRNA, 5'UTR, CDS, 3'UTR) GenBank flat file
- **AF090832** (protein bind, gene, 5'UTR, mRNA, CDS, 3'UTR) GenBank flat file
- **L00727** (alternatively spliced mRNAs) GenBank flat file

A complete list of features is available from the resources noted above.

ORIGIN

The ORIGIN may be left blank, may appear as "Unreported," or may give a local pointer to the sequence start, usually involving an experimentally determined restriction cleavage site or the genetic locus (if available). This information is present only in older records.

The sequence data begin on the line immediately below ORIGIN. To view/save the sequence data only, display the record in FASTA format. A description of FASTA format is accessible from the BLAST Web pages.

Help Desk NCBI NLM NIH Credits

Revised October 23, 2006
Questions about NCBI resources to info@ncbi.nlm.nih.gov
Comments about site map to Renata Geer renata@ncbi.nlm.nih.gov

Disclaimer Privacy statement