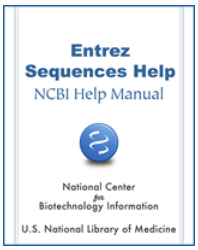
U.S. National Library of Medicine
National Center for Biotechnology Information

**NLM Citation:** Romiti M. Entrez Nucleotide and Entrez Protein FAQs. 2006 Oct 1 [Updated 2010 May 12]. In: Entrez Sequences Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-.
**Bookshelf URL:** https://www.ncbi.nlm.nih.gov/books/




# Entrez Nucleotide and Entrez Protein FAQs

Monica Romiti[1]

Created: October 1, 2006; Updated: May 12, 2010.

## Section A. GenBank nucleotide records, GenPept protein records, and fields within records

**1. Why are there records that duplicate mine with NM_*, XM_*, and XP_* accession numbers?**

The records that have NM_* or XM_* or other two-letter underscore 6 or 12+ digit formats, are reference sequences or RefSeqs. RefSeqs are curated from single or multiple sequence records that have been already directly submitted to GenBank. For a complete explanation that will include all of the accession number prefixes, click here for context on RefSeqs and a key to the RefSeq accessions.

**2. My record needs to be updated. How do I correct it? What should I do if I find an error in a GenBank or RefSeq sequence record?**

Follow the instructions at Updating Information on GenBank Records to update your own NCBI direct submission(s). To update your EST, STS, or GSS record, e-mail the update request to batch-sub@ncbi.nlm.nih.gov. If you have comments on or updates to a record that does not belong to you, please e-mail the general NCBI Service Desk at info@ncbi.nlm.nih.gov. In all cases be sure to provide the accession number of the record(s) on which you are commenting.

**3. What does the date in the upper right-hand corner of a GenBank record mean?**

The date in the upper right-hand corner of a GenBank record, to the far right on the LOCUS line, is the date of last modification. In some cases, it might correspond to the first release date into GenBank or when the record was last updated, but there is no way to tell simply from the data in the record. See corresponding FAQ 4. Refer to the Sample GenBank Record for field descriptions.

**4. How do I find out when a sequence record was released to the GenBank public database?**

To find out the approximate date on which a GenBank record was first released, e-mail a message, including the accession number(s) of interest, to the NCBI general Service Desk address which is info@ncbi.nlm.nih.gov.

**5. What is LinkOut?**

LinkOut allows publishers, aggregators, libraries, biological databases, sequence centers, and other Web resources to display links to their sites on items from the Entrez databases. These links can take you to the provider's site to obtain the full-text of articles or related resources, e.g., consumer health information or

**Author Affiliation:** 1 Email: romiti@ncbi.nlm.nih.gov

genome centers. There may be a charge to access the text or information. Click to the current complete list of all LinkOut Providers.

**6. Where can I find a description of the various fields in a GenBank record?**

To see a description of the various fields in a GenBank record, click to the Sample GenBank Record.

**7. If a sequence has been updated, is it possible to retrieve earlier versions of it?**

Earlier versions of a GenBank record are available. If there was a change in the sequence, there will be a link within the GenBank record COMMENT field stating that the current sequence replaces or is replaced by GI number xxxxx. If the change was not to the actual bases of the sequence, the older version(s) of the GenBank records are accessed from the Sequence Revision History under the More Formats menu. Example: U12345

Select More Formats→ Revision History.

**8. What are the sources of the Protein database sequences?**

The protein sequences in the NCBI Protein database come from several different sources. There are GenPept translations for each of the coding sequences within the GenBank Nucleotide database. That means that there can be more than one protein sequence associated with a corresponding Nucleotide sequence record.

Example: DQ489526

Scroll to the Features section and note the coding regions.

There are records from other databases that are loaded periodically when builds become available, such as UniProt. A simple search to limit records to a specific component database within the Entrez Protein database is:

srcdb_swiss prot [prop]

**9. What is the “calculated Molecular Weight” that is displayed in protein records?**

The calculated molecular weight '/calculated_mol_wt= ' as seen in protein records is calculated as part of the indexing process for protein records in Entrez. Entrez's molecular weight is an average molecular weight, not monoisotopic. Masses are rounded to the nearest integer. The weights are present only in the Molecular Weight index and are not shown explicitly on the protein sequence records. If completely unknown amino acids (e.g., X) are found, a molecular weight is not calculated. Ambiguous amino acids are calculated as one of their possible forms:

B means D or N -- molecular weight is calculated as D

Z means E or Q -- molecular weight is calculated as E

**10. What is the 'DBSOURCE' field within a Protein record?**

The 'DBSOURCE' field within a Protein record shows the source of protein records imported from other databases.

**11. What do these symbols '<' and '>' mean when used in the features section of a nucleotide or protein record?**

The '<' and '>' symbols used in the features section of a nucleotide record, as in DQ882243 for example, mean partial on the 5' and 3' ends, in the case below the start and stop codon are missing:

gene <1..>270

/gene="HLA-DRB1"

/allele="HLA-DRB1*1449 variant"

mRNA <1..>270

In a protein record, ABI31835 which is the GenPept translation for the DQ882243 nucleotide record, the '<' and '>' symbols mean the protein translation is 5' partial and 3' partial.

Protein <..>89

/product="MHC class II antigen"

CDS 1..89

## Section B. Searching tips

**1. Are there standard keywords in Entrez GenBank that should be used for searching? How do I limit my retrieval to a specific field name, organism like Xenopus laevis, to a biomolecule like genomic DNA, or to a specific GenBank division like expressed sequence tag (EST)?**

Use the Entrez Preview/Index option to view the different terms that are indexed in the GenBank records. This is necessary when searching Entrez as standard keywords are not required when submitting sequences. The Preview/Index option is available from the search toolbar on the Entrez database pages:

See the Nucleotide database toolbar. Select 'Preview/Index' and in the 'Add Terms to Query or Preview Index' section, enter the phrase "heat shock protein" without quotation marks, and select 'Index'. The resulting list contains the terms that are indexed in nucleotide records. Also try HSP in the 'Index' section to see that records can be indexed with synonymous terms. Note that PubMed (MEDLINE abstracts database) can be searched using Medical Subject Headings (MeSH).

**2. How do I search for a gene sequence?**

Search in Nucleotides using [gene] and organism qualifiers such as gene symbol[gene] AND genus species[organism]

e.g., brca1[gene] AND mouse[orgn]

or search in the Entrez Gene database with the following query to find links to nucleotide records, RefSeqs, and protein records:

gene_symbol[sym] AND genus_species[orgn]

**3. Can I retrieve a large dataset for a particular organism?**

For large datasets, you can formulate a search limited to organism, e.g., pig[orgn] in Entrez, display all the records in your desired format, and then save using the Send to file option from the toolbar. Confirm the message that asks if you want to download xxx number of records. You can also use Batch Entrez to download a database-specific file of accessions or GIs. You can download some organism-specific files from the NCBI FTP site, for example, genomes. You can also use the Entrez Utilities (E-Utils).

**4. How can I download data from the Nucleotide and Protein databases?**

You can get the current GenBank nucleotide release and daily updates from the NCBI FTP site in the GenBank directory. You can obtain the Refseq build from the NCBI RefSeq FTP site. See the BLAST FTP site for access to datasets available for download.

**5. Can I store a search, update the stored search, run the stored search multiple times, and then save those search results?**

Use My NCBI.

You will need to register for an account. Log into My NCBI, perform a search in the desired database, and click the 'Save Search' link to the right of the query box on the search toolbar.

This saves the search strategy. See MY NCBI FAQ.

**6. How do I make search URLs for retrieving accession numbers or GIs or other record identifiers?**

Use the E-Utils.

To link to specific Entrez pages from your Web page or application, select Linking to Entrez.

**7. My search keeps returning messages that a term is not found. What can I do?**

Select the Details tab from the search toolbar to see how the query is being translated from the search terms you entered. You can edit the search in the Details page or use Preview/Index to explore alternate search fields.

**8. How do I search for sequences annotated with a specific Enzyme Commission number?**

Start in either Nucleotide or Protein database and enter: the enzyme Commission number and field limiter [ecno]

Example: 1.1.1.53[ecno]

A more general search can be done of Enzyme Commission numbers by entering a truncated EC number by using the asterisk after the partial EC number.

Example: 1.1.1*[ecno]

**9. How can I perform a search to see all records in a specific Entrez database?**

Enter the following search in the search field for the specific database: all[filter] or all[filt]. This will provide the number of records for that database.

## Section C. Display of Records, format

**1. In what order are the resulting records displayed in Entrez and can I sort my results?**

GenBank records are displayed generally in a 'last into the database first displayed' order. In Nucleotide and Protein databases one can sort results retrieved by accession numbers by selecting the 'Sort By' pull down menu and choosing Accession.

**2. How do I display the sequence (bases) for some records that have only the join information instead of the whole sequence in the record?**

To display the sequence for a Contig record, a record where accession number join information has been provided in place of the sequence, select the FASTA format. This will provide the entire sequence without line numbers in a single web page.

An example is a Whole Genome Shotgun record: NW_001149201. Note the N's in the sequence which represent gaps.

CONTIG join(AANU01169770.1:1..10827,gap(29605),AANU01169771.1:1..7919,

gap(86),complement(AANU01169772.1:1..6773))

**3. Why are there N's in sequences in GenBank, example: NW_001149201?**

The N's represent a gap in a contig sequence. An example of a contig record is a Whole Genome Shotgun (WGS) sequence. Click the expand N's link to 'uncompress the N's' in order to see the entire sequence including the gap N's.

**4. What is the BLink option under the Links menu on the Document Summary or results page for a Protein database search like for protein record CAA36839?**

BLink means “BLAST link” and shows pre-calculated BLAST hits for protein sequences for protein sequence in the Entrez Proteins data domain. Blink shows graphical output of pre-computed blastp results against the protein non-redundant (nr) database. See the BLink Help document for further details.

## Section D. Entrez data

**1. How often are the Entrez Nucleotide and Protein databases updated?**

The Nucleotide database is updated every day. Records from the International Collaboration databases DDBJ and EMBL are added on a nightly build. The protein translations are added every night. For UniProt records, updates are processed when UniProt provides a new "cumulative update" at their FTP site, which is about twice per month.