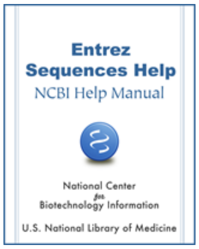
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: Romiti M, Cooper P. Entrez Nucleotide and Entrez Protein FAQs. 2006 Oct 1 [Updated 2024 Jul 2]. In: Entrez Sequences Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



Entrez Sequences Help
NCBI Help Manual
National Center for Biotechnology Information
U.S. National Library of Medicine

Entrez Nucleotide and Entrez Protein FAQs

Monica Romiti[1] and Peter Cooper[2]

Created: October 1, 2006; Updated: July 2, 2024.

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 prefix followed by an underscore and 6 or 12+ digit formats, are reference sequences or RefSeqs. These include curated records that are generated from single or multiple sequence records that have been already directly submitted to GenBank or other members of the International Sequence Database Collaboration (INSDC). RefSeqs also include transcript and protein sequences from genome annotation pipelines. See the Reference Sequences page for more information on RefSeqs and accession number prefixes.

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). If you have comments on or updates to a record that you did not submit, 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 first (LOCUS) line, is the date of last modification. It may correspond to the first release date into GenBank or the 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?

You can also see the sequence revision history including the approximate date of first appearance by selecting the Revision History format from the format list in the upper left of the record in Entrez. See the Revision History for BA000005 as an example. To get any additional information about the date that a GenBank record was first released, e-mail a message, including the accession number(s) of interest, to the NCBI general Service Desk address, info@ncbi.nlm.nih.gov.

5. What is LinkOut?

Author Affiliations: 1 Email: romiti@ncbi.nlm.nih.gov 2 Email: cooper@ncbi.nlm.nih.gov

LinkOut is a system that allows publishers, aggregators, libraries, biological databases, sequence centers, and other web resources (e.g., consumer health information or genome centers) 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. 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?

The Sample GenBank Record has a description of the various fields in a GenBank record.

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

Earlier versions of a sequence record are available. If there was a change in the sequence, there will be a link within the record COMMENT field stating that the current sequence replaces or is replaced by GI number xxxxx. You can also access older version(s) of sequence records from the sequence Revision History under the available formats in the upper-left corner.

Example: Revision History for BA000005

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

The protein sequences in the NCBI Protein database come from several different sources including coding region translations of INSDC (DDBJ, ENA/EMBL, GenBank) and NCBI RefSeq records. There are also protein-only records from outside databases such as the Swiss-Prot portion of UniProt, PIR, and sequences of protein chains from Protein Data Bank (PDB) by way of NCBI's Structure resource. Here are some simple searches that will retrieve records for some of these sources:

srcdb_ddbj/embl/genbank[PROP]

srcdb_refseq[PROP]

srcdb_swiss prot [PROP]

srcdb_pdb[PROP]

9. What is the "calculated Molecular Weight" that is displayed in protein records?

The calculated molecular weight "/calculated_mol_wt= " that is indexed for protein records is an average molecular weight rounded to the nearest integer. The molecular weight is calculated for the amino acid portion of the protein only and does not include posttranslational modifications that may be present on the protein in living systems. Ambiguous amino acids are calculated as one of their possible forms:

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

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

No molecular weight is calculated if the sequence contains unknown amino acids (symbol X).

The weights are available only in the Molecular Weight index and are not shown explicitly on the protein sequence records.

You can search by molecular weight in the Protein database by limiting to the [Molecular Weight] or [MOLWT] field.

Examples:
3039[MOLWT]
25000:75000[MOLWT]

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 the less than ‘<’ and greater than ‘>’ symbols represent 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 codons 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 of the DQ882243 nucleotide record, the '<' and '>' symbols mean the protein translation is partial at the amino and carboxyl ends.

```
Protein         <..>89
                /product="MHC class II antigen"
CDS             1..89
```

Section B. Searching tips

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

Use the Advanced search page to view the different terms that are indexed for sequence records. The Advanced search interface is linked under the search box on all Entrez database pages:

In the Builder on the Advanced search page, you can see the indexed fields. To see the terms available for each, click the “Show Index List” link. For example, on the Nucleotide Advanced search page select the “Title” field and enter the phrase "heat shock protein" and click “Show Index.” The resulting list shows the terms that are indexed in nucleotide with this phrase and the number of records indexed. You can select any of the terms and click the “Search” button to run the search or you can use the “History” to combine with other searches.

2. How do I search for a gene sequence?

Search in Nucleotide using [Gene] and organism qualifiers:

gene symbol[Gene] AND organism name [Organism], or organism name [ORGN].

Example:

brca1[Gene] AND Mus musculus[ORGN]

You can also search in the Gene database with the following query to find a Gene record that will have links to nucleotide and protein records:

gene_symbol[SYM] AND organism name[ORGN]

Example:
brca1[SYM] AND rodents[ORGN]

3. Can I retrieve a set of sequences for a particular organism?

For small to medium sized downloads, you can formulate a search limited to organism — for example raccoon[ORGN] — in the Nucleotide or Protein database, display all the records in your desired format, and then save using the “Send to” file option from the upper-right of the results. You can also use Batch Entrez to upload a database-specific file of identifiers and download the corresponding sequence records.

For large sets of data you can use the Entrez Utilities (E-Utilities) or the Entrez Direct suite of command line scripts that access the E-Utilities. Use NCBI Datasets to download gene, genome sequences and metadata.

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

You can download small to medium-sized sets of results using the "Send to" menu in the Protein and Nucleotide databases. For access to genome sequences, associated annotations, and metadata use NCBI Datasets.

You can also download the current GenBank nucleotide release and daily updates from the NCBI FTP site in the GenBank directory. You can obtain the RefSeq release from the NCBI RefSeq FTP site.

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

You can set up a saved search when you're logged in to a My NCBI account. You will need to register for an account if you don't have one. Then, log in to My NCBI, perform a search in the desired database, and click the "Create Alert" link under the search toolbar.

The link will take you to a page with options for saving the search strategy and setting up a schedule for automatically running the search and sending e-email alerts when the search is run. See MY NCBI FAQ.

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

Use the Entrez API, the E-Utilities.

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

Look at the "Search details" box on the right-hand side of the search results to see how the query is being translated from the search terms you entered. You can edit the search in the "Search details" box or use Advanced search page 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 (EC) number and field limiter [ECNO].

Example:

1.1.1.53[ECNO]

You can make a broader search for related enzymes by entering a truncated EC number with an asterisk after the partial EC number.

Example:

1.1.1*[ECNO]

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

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

Section C. Display of Records, format

1. In what order are the records displayed in Nucleotide and Protein database results and can I sort my results?

Sequence records are displayed approximately in the order they were modified with the most recently modified shown first. In Nucleotide and Protein databases you can sort results by other criteria using the

"Sort by" pull down menu. Available sorts are Accession, Date Modified, Date Released, Organism Name, Taxonomy ID, and Sequence Length.

2. How do I display the sequence (bases) for some records such as NW_001799157.1 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 link at the top of the record. This will provide the entire sequence.

For example, NW_001799157.1 has a CONTIG join statement where the sequence would normally appear.

```
CONTIG      join(CAAL01000027.1:1..3194,gap(2767),CAAL01000028.1:1..3964)
```

You can retrieve the sequence in FASTA format by following the FASTA link at the top of the record

3. Why are there N's in nucleotide sequences?

The N's represent an unsequenced gap in a record. In cases with large gaps, you can click to expand N's link to show the entire sequence including all the N's.

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 other International Sequence Database Collaboration (INSDC) databases DDBJ and ENA/EMBL and their protein translations are added every night. For UniProt (Swiss-Prot) records, updates are processed when UniProt provides a new cumulative update on their FTP site.