# Entrez Sequences Help

Last Updated: 70, 2.02

National Center for Biotechnology Information (US)
Bethesda (MD)

National Center for Biotechnology Information (US), Bethesda (MD)

NLM Citation: Entrez Sequences Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-.

This book contains information on the Entrez Protein, Nucleotide, Expressed Sequence Tag (EST), and Genome Survey Sequence (GSS) databases. The instructions here should allow you to quickly begin searching and using the features of the Entrez sequence databases.

# Table of Contents

# Entrez Sequences Quick Start

Peter Cooper,[1] Melissa Landrum,[1] Ilene Mizrachi,[1] and Jane Weisemann[1]

Created: June 29, 2010; Updated: July 5, 2024.

This is a quick start guide for the Entrez Protein, and Nucleotide databases. The instructions here should allow you to quickly begin searching and using the features of the Entrez sequence databases.

- How do I use a simple query, such as a word or a phrase?
- How can I make my search more specific with Boolean operators (AND, OR, NOT)?
- How do I restrict my search to specific subsets of records such as those from a specific organism, molecule type or source database?
- How can I change the format, number, or sorting order of records displayed?
- How do I download sequence records to a file on my computer?
- How can I change the information that is shown such as optional biological features or sequence?
- How do I analyze the sequence data directly or find additional related data?
- How can I search for a sub-sequence, or pattern in a protein or nucleotide sequence?
- How can I locate and highlight a biological feature in a protein or nucleotide sequence?

## How do I use a simple query, such as a word or a phrase?

You can use a protein name, gene name, or gene symbol directly. Searching with a submitter or author name in the following format will produce the best results.

Smith JR (last name followed by initials, no punctuation)

Database identifiers such as accession numbers or gi numbers will directly retrieve the full sequence record.

```
CAA79696
NP_778203
263191547
BC043443
NM_002020
```

To find a match to an exact phrase, enclose it in quotation marks.

```
"contactin associated protein"
"duchenne muscular dystrophy"
```

## How can I make my search more specific with Boolean operators (AND, OR, NOT)?

Use the Boolean operator AND to find records that contain every one of your search terms, the intersection of search results.

```
contactin AND neurofascin          Protein          Nucleotide
```

**Author Affiliation:** 1 NCBI; Email: cooper@ncbi.nlm.nih.gov.

Use the Boolean operator OR to find records that include one of several search terms, the union of search results.

`contactin OR neurofascin` Protein Nucleotide

Use the Boolean operator NOT to exclude records matching a search term

`contactin NOT neurofascin` Protein Nucleotide

# How do I restrict my search to specific subsets of records such as those from a specific organism, molecule type or source database?

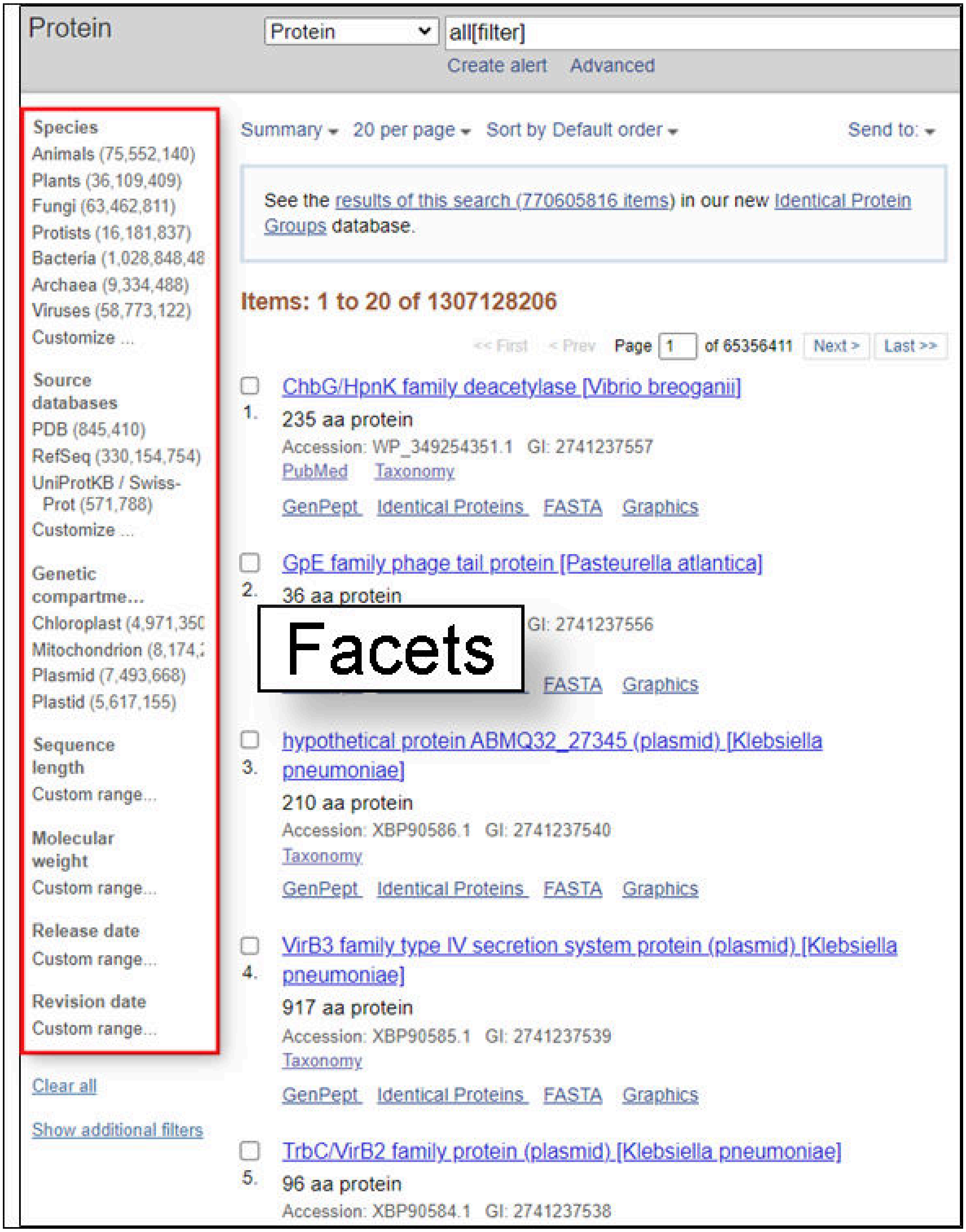
You can use the *Facets* on the left-hand side of the page to show only certain kinds of records. Follow these links to jump to the Facet of interest: organism, molecule type, source database.

## Facets

Use the facets on the left-hand side of any of the Protein, Nucleotide results pages to restrict the types of records shown.

### Organism

To get records from a specific organism or group of organisms click the appropriate Species filter. You can use the Customize option to add a filter for a particular organism or group or organisms. You use the common or scientific name of a species, strain, or higher taxon as a Filter term. Examples: human, *Mus musculus*, *Drosophila similis*, green plants, bacteria.



Protein

Protein ▾ all[filter]

Create alert Advanced

Species
Animals (75,552,140)
Plants (36,109,409)
Fungi (63,462,811)
Protists (16,181,837)
Bacteria (1,028,848,48
Archaea (9,334,488)
Viruses (58,773,122)
Customize ...

Source databases
PDB (845,410)
RefSeq (330,154,754)
UniProtKB / Swiss-Prot (571,788)
Customize ...

Genetic compartme...
Chloroplast (4,971,350
Mitochondrion (8,174,
Plasmid (7,493,668)
Plastid (5,617,155)

Sequence length
Custom range...

Molecular weight
Custom range...

Release date
Custom range...

Revision date
Custom range...

Clear all

Show additional filters

Summary ▾ 20 per page ▾ Sort by Default order ▾ Send to: ▾

See the results of this search (770605816 items) in our new Identical Protein Groups database.

Items: 1 to 20 of 1307128206

<< First < Prev Page 1 of 65356411 Next > Last >>

□ 1. ChbG/HpnK family deacetylase [Vibrio breoganii]
235 aa protein
Accession: WP_349254351.1 GI: 2741237557
PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

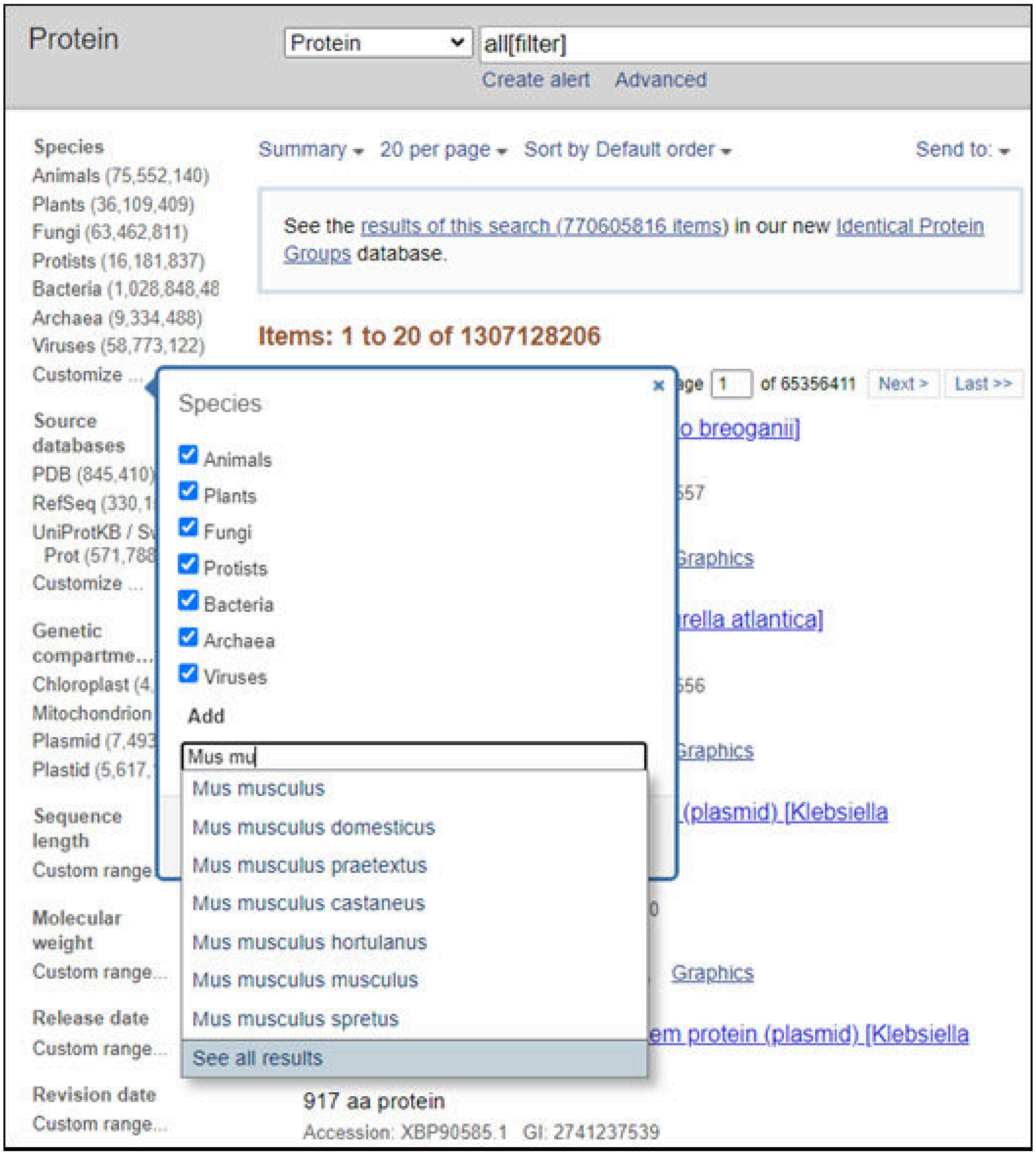
□ 2. GpE family phage tail protein [Pasteurella atlantica]
36 aa protein
GI: 2741237556
FASTA Graphics

Facets

□ 3. hypothetical protein ABMQ32_27345 (plasmid) [Klebsiella pneumoniae]
210 aa protein
Accession: XBP90586.1 GI: 2741237540
Taxonomy
GenPept Identical Proteins FASTA Graphics

□ 4. VirB3 family type IV secretion system protein (plasmid) [Klebsiella pneumoniae]
917 aa protein
Accession: XBP90585.1 GI: 2741237539
Taxonomy
GenPept Identical Proteins FASTA Graphics

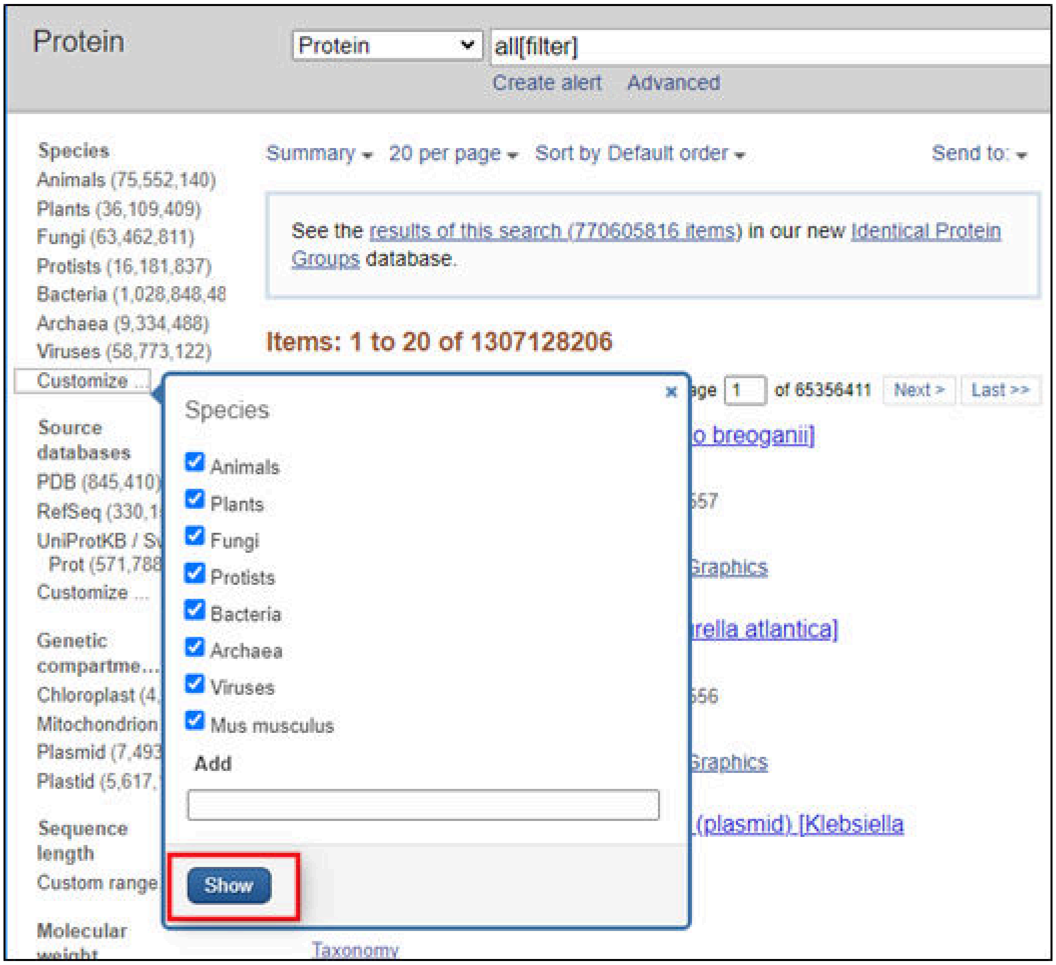
□ 5. TrbC/VirB2 family protein (plasmid) [Klebsiella pneumoniae]
96 aa protein
Accession: XBP90584.1 GI: 2741237538



See the results of this search (770605816 items) in our new Identical Protein Groups database.

Items: 1 to 20 of 1307128206

917 aa protein
Accession: XBP90585.1 GI: 2741237539



Protein

Protein ▾ all[filter]

Create alert Advanced

Species
Animals (75,552,140)
Plants (36,109,409)
Fungi (63,462,811)
Protists (16,181,837)
Bacteria (1,028,848,48
Archaea (9,334,488)
Viruses (58,773,122)
Customize ...

Source databases
PDB (845,410)
RefSeq (330,1
UniProtKB / Sw
Prot (571,788
Customize ...

Genetic compartme...
Chloroplast (4,
Mitochondrion
Plasmid (7,493
Plastid (5,617,

Sequence length
Custom range

Molecular weight

Summary ▾ 20 per page ▾ Sort by Default order ▾ Send to: ▾

See the results of this search (770605816 items) in our new Identical Protein Groups database.

Items: 1 to 20 of 1307128206

age 1 of 65356411 Next > Last >>

o breoganii]

557

Graphics

rella atlantica]

556

Graphics

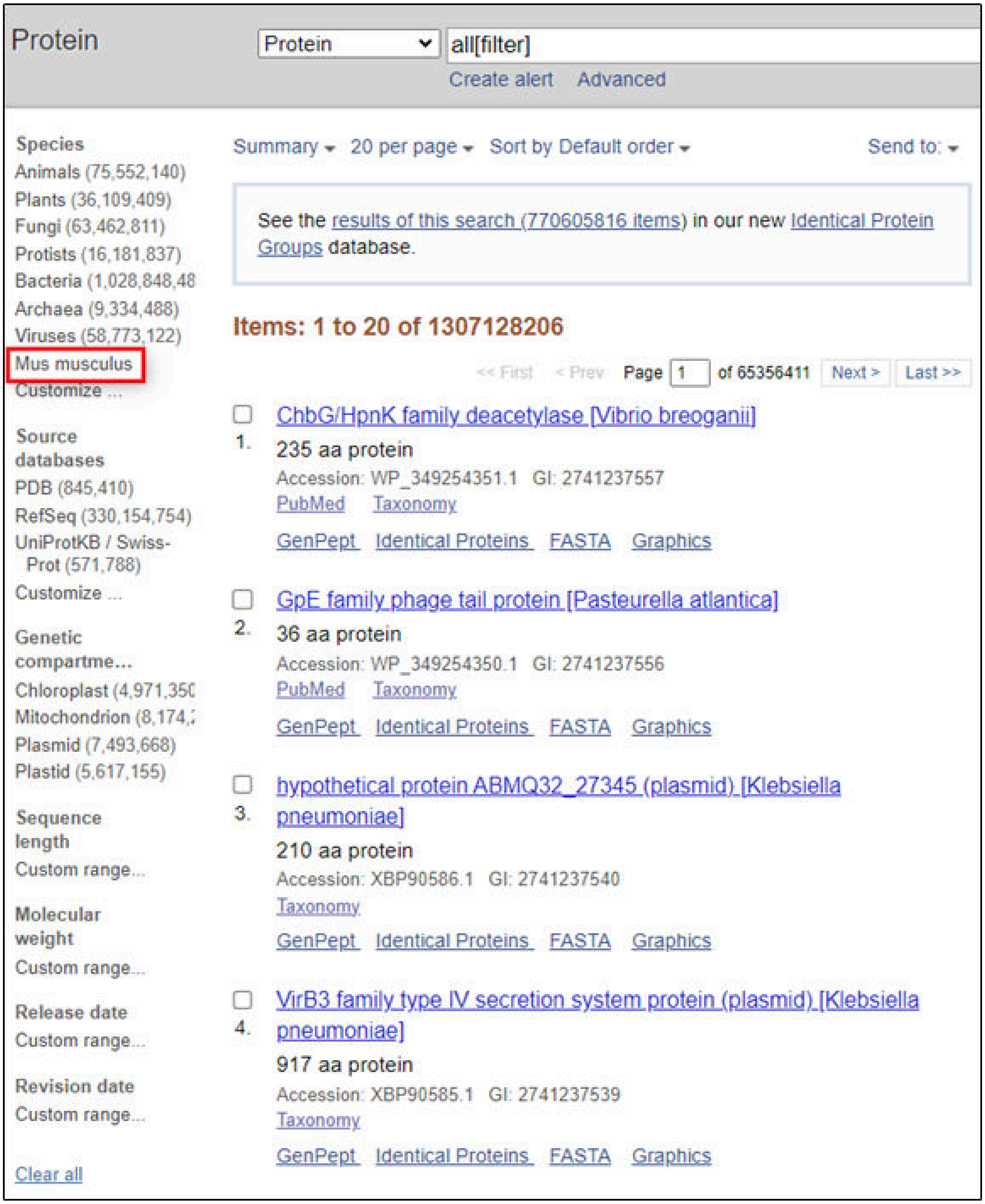
(plasmid) [Klebsiella

Taxonomy

Species ×

☑ Animals
☑ Plants
☑ Fungi
☑ Protists
☑ Bacteria
☑ Archaea
☑ Viruses
☑ Mus musculus

Add



See the results of this search (770605816 items) in our new Identical Protein Groups database.

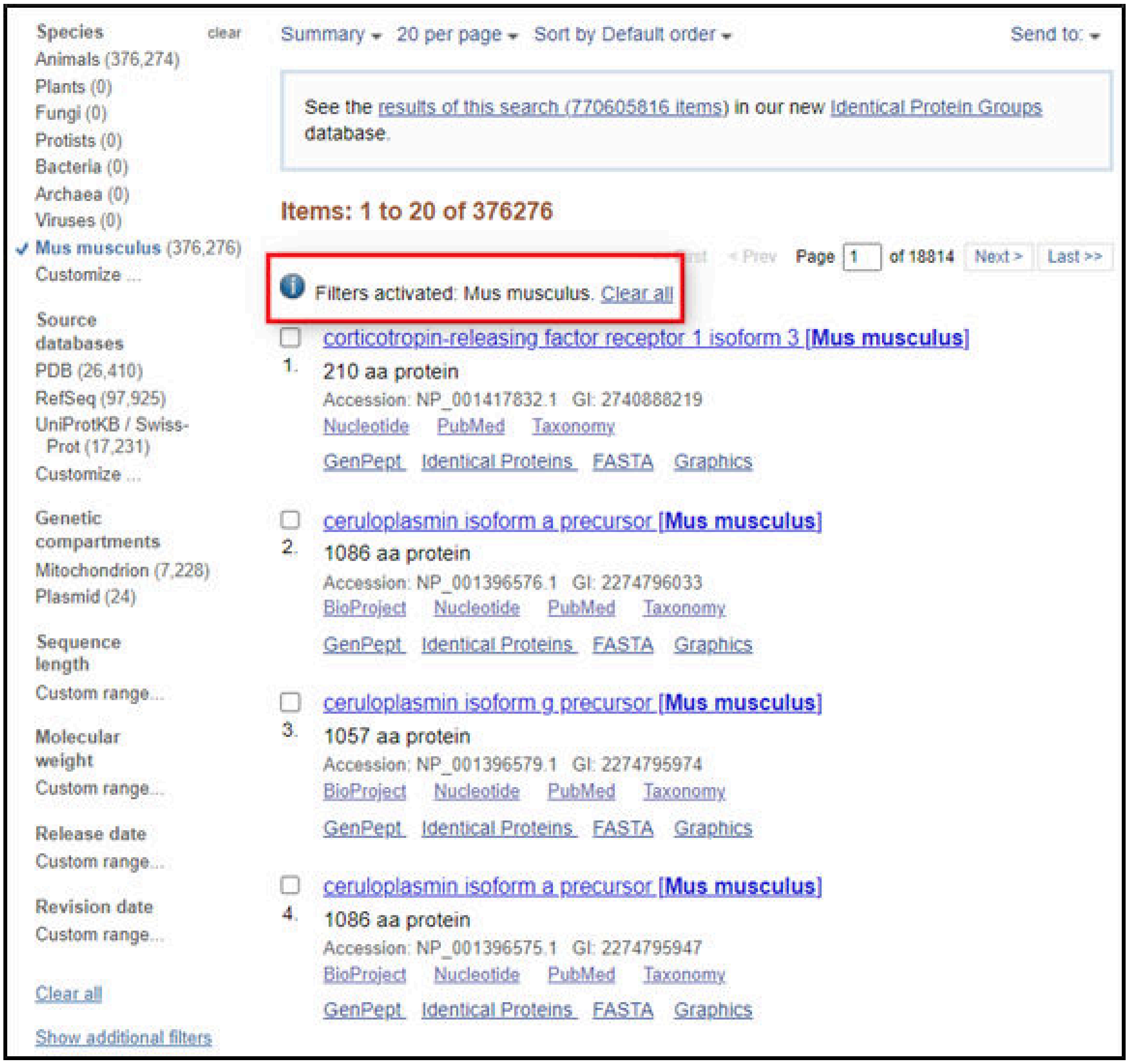
Items: 1 to 20 of 1307128206

□ 1. ChbG/HpnK family deacetylase [Vibrio breoganii]
235 aa protein
Accession: WP_349254351.1 GI: 2741237557
PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

□ 2. GpE family phage tail protein [Pasteurella atlantica]
36 aa protein
Accession: WP_349254350.1 GI: 2741237556
PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

□ 3. hypothetical protein ABMQ32_27345 (plasmid) [Klebsiella pneumoniae]
210 aa protein
Accession: XBP90586.1 GI: 2741237540
Taxonomy
GenPept Identical Proteins FASTA Graphics

□ 4. VirB3 family type IV secretion system protein (plasmid) [Klebsiella pneumoniae]
917 aa protein
Accession: XBP90585.1 GI: 2741237539
Taxonomy
GenPept Identical Proteins FASTA Graphics



Species clear
Animals (376,274)
Plants (0)
Fungi (0)
Protists (0)
Bacteria (0)
Archaea (0)
Viruses (0)
✓ Mus musculus (376,276)
Customize ...

Source databases
PDB (26,410)
RefSeq (97,925)
UniProtKB / Swiss-Prot (17,231)
Customize ...

Genetic compartments
Mitochondrion (7,228)
Plasmid (24)

Sequence length
Custom range...

Molecular weight
Custom range...

Release date
Custom range...

Revision date
Custom range...

Clear all

Show additional filters

Summary ▾ 20 per page ▾ Sort by Default order ▾ Send to: ▾

See the results of this search (770605816 items) in our new Identical Protein Groups database.

Items: 1 to 20 of 376276

First < Prev Page 1 of 18814 Next > Last >>

Filters activated: Mus musculus. Clear all

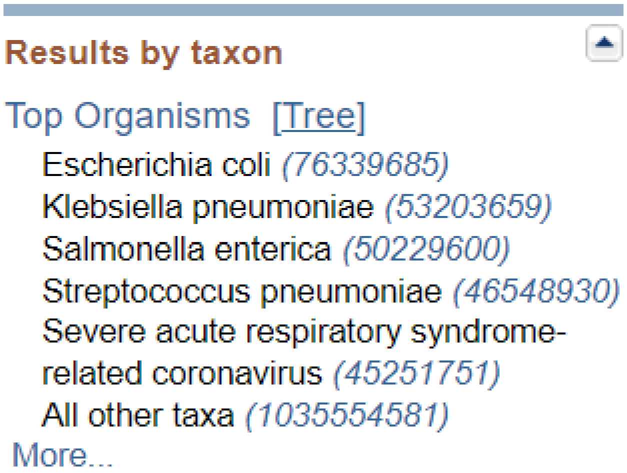
□ corticotropin-releasing factor receptor 1 isoform 3 [Mus musculus]
1. 210 aa protein
Accession: NP_001417832.1 GI: 2740888219
Nucleotide PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

□ ceruloplasmin isoform a precursor [Mus musculus]
2. 1086 aa protein
Accession: NP_001396576.1 GI: 2274796033
BioProject Nucleotide PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

□ ceruloplasmin isoform g precursor [Mus musculus]
3. 1057 aa protein
Accession: NP_001396579.1 GI: 2274795974
BioProject Nucleotide PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

□ ceruloplasmin isoform a precursor [Mus musculus]
4. 1086 aa protein
Accession: NP_001396575.1 GI: 2274795947
BioProject Nucleotide PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

You can also use the linked numbers in the Top Organisms list in the right-hand column of search results to filter select records from specific organisms from your results.



Results by taxon

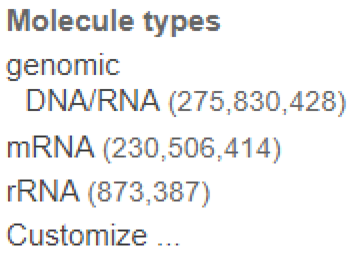
Top Organisms [Tree]

| | |
|---|---|
| Escherichia coli | *(76339685)* |
| Klebsiella pneumoniae | *(53203659)* |
| Salmonella enterica | *(50229600)* |
| Streptococcus pneumoniae | *(46548930)* |
| Severe acute respiratory syndrome-related coronavirus | *(45251751)* |
| All other taxa | *(1035554581)* |

More...

## Molecule type

In the Nucleotide database you can use the Molecule types facet to limit results to a particular molecule type.



**Molecule types**

genomic
DNA/RNA (275,830,428)

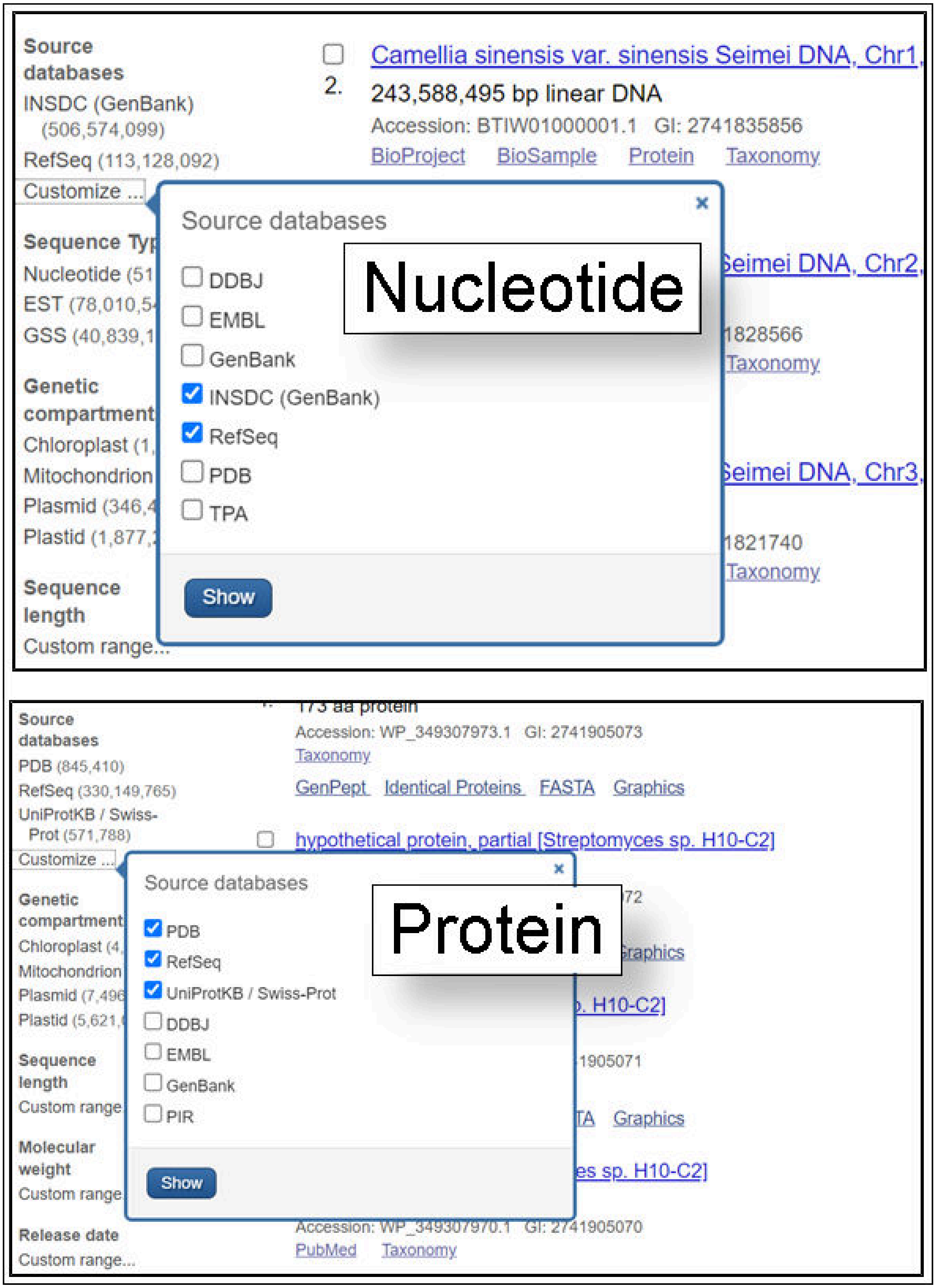
mRNA (230,506,414)

rRNA (873,387)

Customize ...

## Source database

The Source databases facet allows you to limit to results from a particular database.



Nucleotide

Source databases
INSDC (GenBank) (506,574,099)
RefSeq (113,128,092)
Customize ...

Source databases

☐ DDBJ
☐ EMBL
☐ GenBank
☑ INSDC (GenBank)
☑ RefSeq
☐ PDB
☐ TPA

Show

Protein

Source databases
PDB (845,410)
RefSeq (330,149,765)
UniProtKB / Swiss-Prot (571,788)
Customize ...

Source databases

☑ PDB
☑ RefSeq
☑ UniProtKB / Swiss-Prot
☐ DDBJ
☐ EMBL
☐ GenBank
☐ PIR

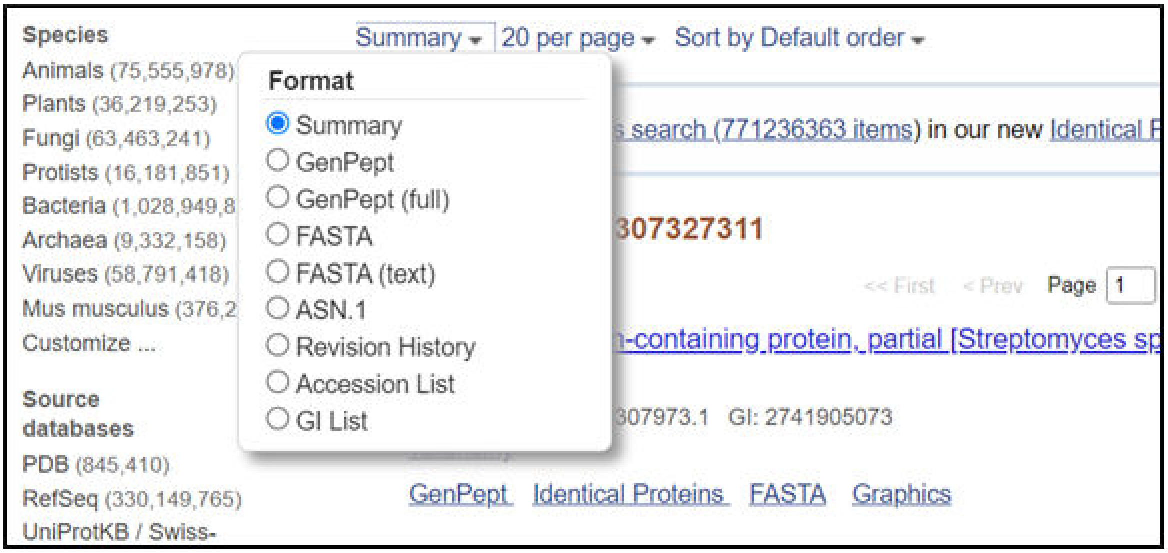
Show

The source databases for NCBI nucleotide and protein sequences are listed below.

- **Protein:** SwissProt and PIR components of UniProt; Protein Research Foundation (PRF); protein chain sequences from the Protein Data Bank (PDB); and translations of coding regions on sequences in Entrez Nucleotide (RefSeq, International Sequence Database Collaboration – DDBJ / ENA-EMBL / GenBank, Third Party Annotation (TPA).
- **Nucleotide:** International Sequence Database Collaboration (DDBJ / ENA-EMBL / GenBank); NCBI Reference Sequences (RefSeq); Nucleotide sequences from PDB; Third Party Annotation (TPA).

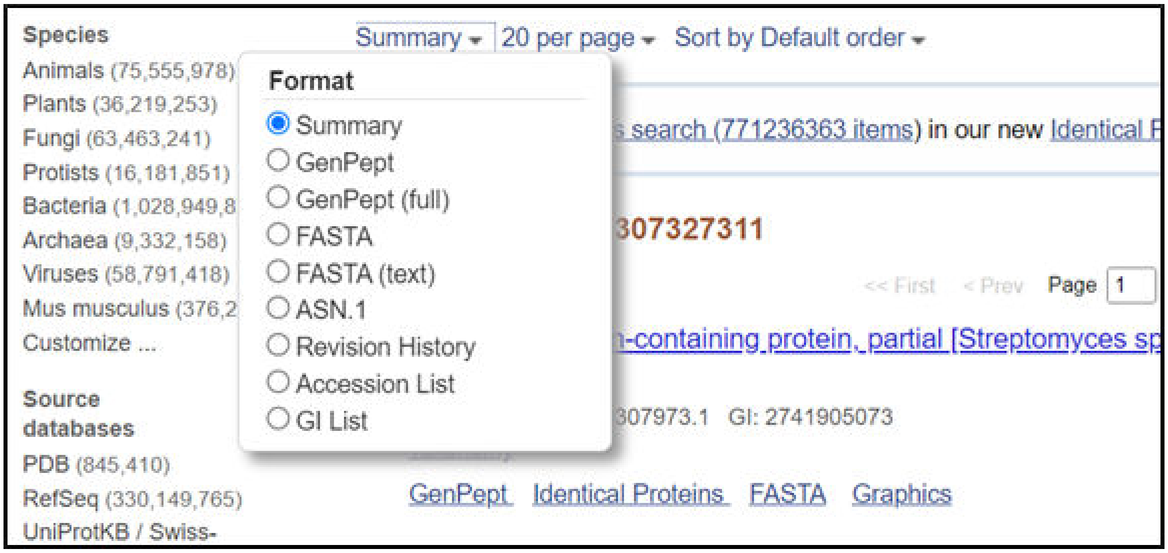
## How do I change the format, number, or sorting order of records displayed?

The menus at the upper left of the results page headed by *Summary*, *20 per page* and *Sort by Default Order* allow you to change the format displayed, the number of records and the sorting order respectively. Click any of these and select the desired format, items per page, or sorting order from the listed radio buttons. The new settings will apply automatically.



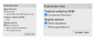
## How can I download sequence records to a file on my computer?

Click the *Send to* menu that appears at the upper right of document summaries or record views and select the file radio button. Then choose the desired format from the pull-down list. Click the *Create File* button to save the records.



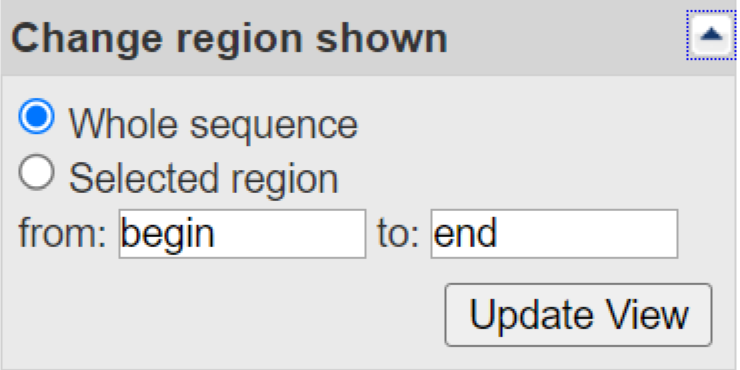
## How do I change the information that is shown such as optional biological features or sequence?

Open the *Customize View* dialog that appears in the right-hand column of a record display. You can change the kinds of biological features shown and toggle the sequence on or off using the radio buttons and check boxes. Click the *Update View* button to activate the changes.



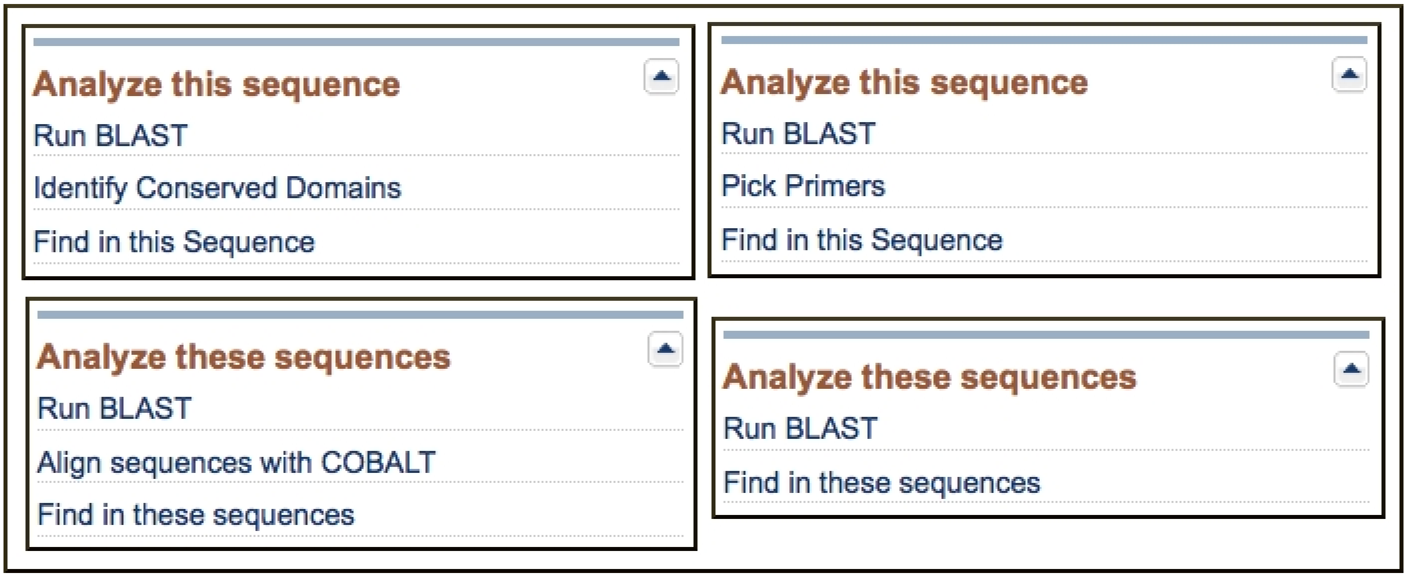
## How can I display a portion of the sequence?

Open the *Change region shown* dialog that appears in the right-hand column of a record display. You can change the kinds of biological features shown and toggle the sequence on or off using the radio buttons and check boxes. Click the *Update View* button to activate the changes.



## How do I analyze the sequence data directly or find additional related data?

There are direct links to analysis tools including BLAST, Primer-BLAST (Nucleotide), and Conserved Domain Database Search (Protein) in the right-hand column of displayed records.



**Analyze this sequence**
- Run BLAST
- Identify Conserved Domains
- Find in this Sequence

**Analyze this sequence**
- Run BLAST
- Pick Primers
- Find in this Sequence

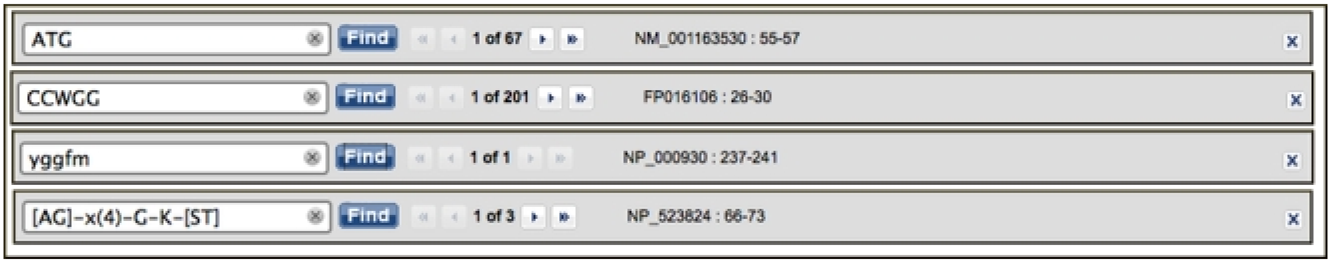
**Analyze these sequences**
- Run BLAST
- Align sequences with COBALT
- Find in these sequences

**Analyze these sequences**
- Run BLAST
- Find in these sequences

There are also links to related data in the right-hand column that may provide additional information and pre-computed analyses for the displayed records.

## How can I search for a sub-sequence, or pattern in a protein or nucleotide sequence?

You can access the *Find-in-sequence* feature in the Analysis tools in the right-hand column of single and multiple-record displays. This tool can find sub-sequences or patterns in displayed nucleotide or protein sequences. Clicking the **Find-in-this-Sequence** or **Find-in-these sequences** link opens a search box bar at the bottom of the page.



ATG Find 1 of 67 NM_001163530 : 55-57

CCWGG Find 1 of 201 FP016106 : 26-30

yggfm Find 1 of 1 NP_000930 : 237-241

[AG]-x(4)-G-K-[ST] Find 1 of 3 NP_523824 : 66-73

Find-in-sequence works with single and multiple sequence displays with any format that shows the sequence (GenBank, GenPept, FASTA). The tool can find sub-sequences and patterns typed in the box and works with standard (IUPAC) nucleotide and protein single letter and ambiguity codes as well as Prosite patterns that match motifs and domain signatures in protein sequences. Valid single letter codes are given below.

| Nucleotide Codes | | | |
|---|---|---|---|
| A | adenosine | Y | T or C |
| C | cytidine | M | A or C |
| G | guanine | W | A or T |
| T | thymidine | R | G or A |
| N | A, G, C, or T | B | G, T, or C |
| U | uridine (matches T) | D | G, A, or T |
| K | G or T | H | A, C, or T |

*Table continued from previous page.*

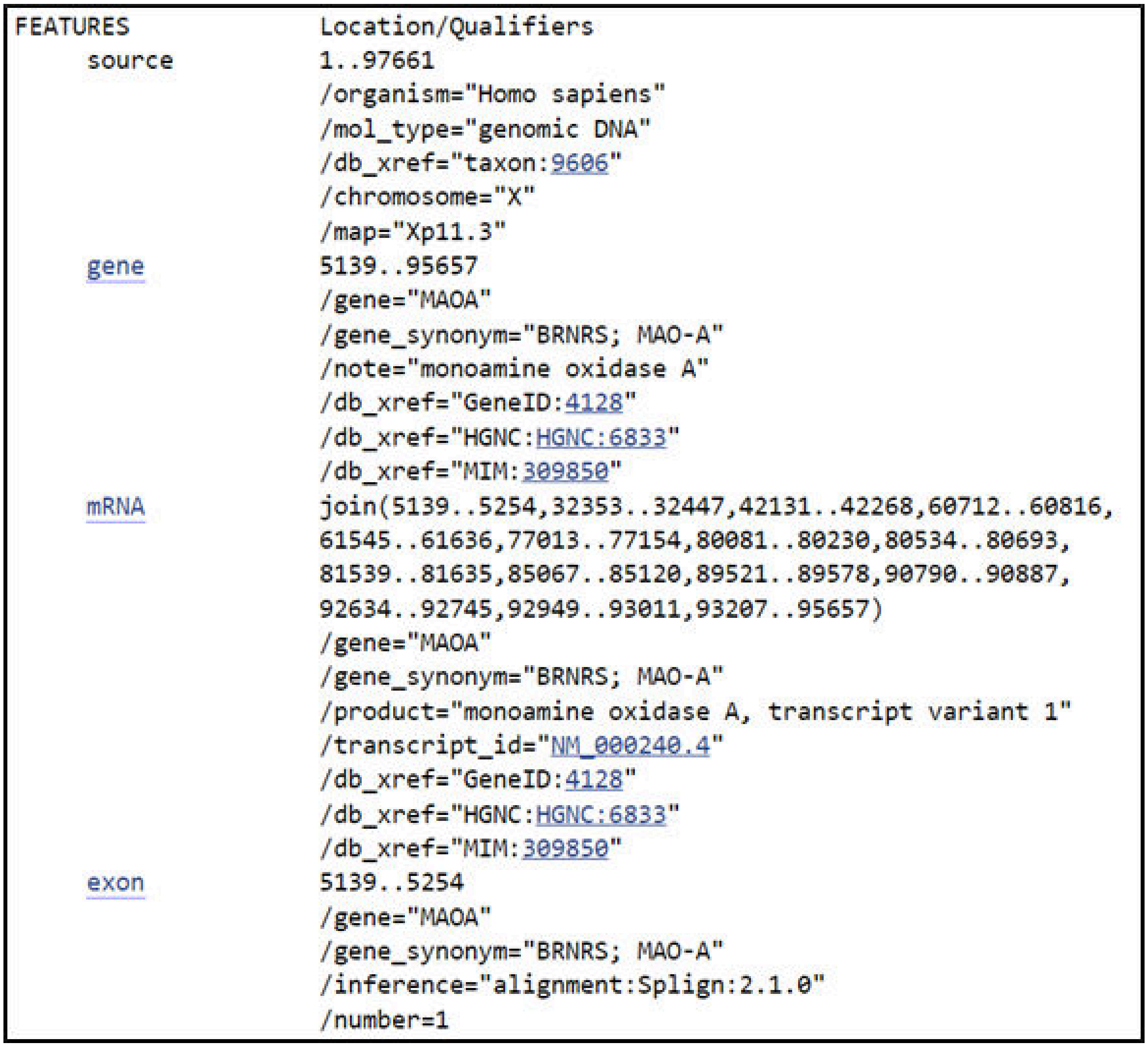
| | | | |
|---|---|---|---|
| S | G or C | V | G, C, or A |

| Amino Acid Codes | | | |
|---|---|---|---|
| A | alanine | N | asparagine |
| B | aspartate/asparagine | P | proline |
| C | cysteine | Q | glutamine |
| D | aspartate | R | arginine |
| E | glutamate | S | serine |
| F | phenylalanine | T | threonine |
| G | glycine | V | valine |
| H | histidine | W | tryptophan |
| I | isoleucine | Y | tyrosine |
| K | lysine | Z | glutamate/glutamine |
| L | leucine | X | any |
| M | methionine | | |

Find matches by clicking the find button. The first 500 matches are highlighted for each displayed sequence. The first or current match is highlighted in white text on a dark background in the sequence, and its position is shown in the search bar. The other matches are highlighted with a light blue background. The tool ignores spaces and line breaks in the formatted sequence. Clicking the arrow keys jumps to the next or previous match.

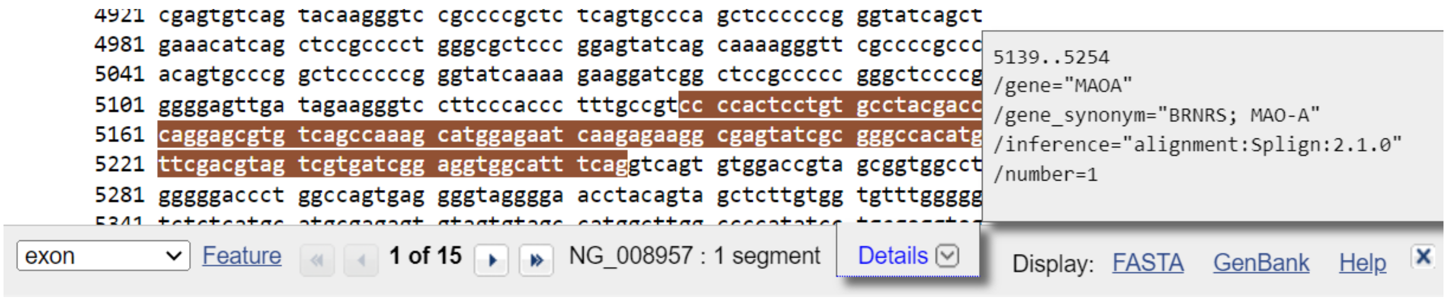
## How can I locate and highlight a biological feature in a protein or nucleotide sequence?

You can highlight a feature by clicking on linked feature in the FEATURES table of a displayed nucleotide or protein sequence. A portion of a FEATURES table is shown below for a nucleotide sequence (NG_008957).



```
FEATURES             Location/Qualifiers
     source          1..97661
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="X"
                     /map="Xp11.3"
     gene            5139..95657
                     /gene="MAOA"
                     /gene_synonym="BRNRS; MAO-A"
                     /note="monoamine oxidase A"
                     /db_xref="GeneID:4128"
                     /db_xref="HGNC:HGNC:6833"
                     /db_xref="MIM:309850"
     mRNA            join(5139..5254,32353..32447,42131..42268,60712..60816,
                     61545..61636,77013..77154,80081..80230,80534..80693,
                     81539..81635,85067..85120,89521..89578,90790..90887,
                     92634..92745,92949..93011,93207..95657)
                     /gene="MAOA"
                     /gene_synonym="BRNRS; MAO-A"
                     /product="monoamine oxidase A, transcript variant 1"
                     /transcript_id="NM_000240.4"
                     /db_xref="GeneID:4128"
                     /db_xref="HGNC:HGNC:6833"
                     /db_xref="MIM:309850"
     exon            5139..5254
                     /gene="MAOA"
                     /gene_synonym="BRNRS; MAO-A"
                     /inference="alignment:Splign:2.1.0"
                     /number=1
```

Clicking the feature activates the feature search bar that appears at the bottom of the display and highlights the corresponding residues in the display as shown below for an exon feature in the RefSeq gene record for the MAOA gene (NG_008957).

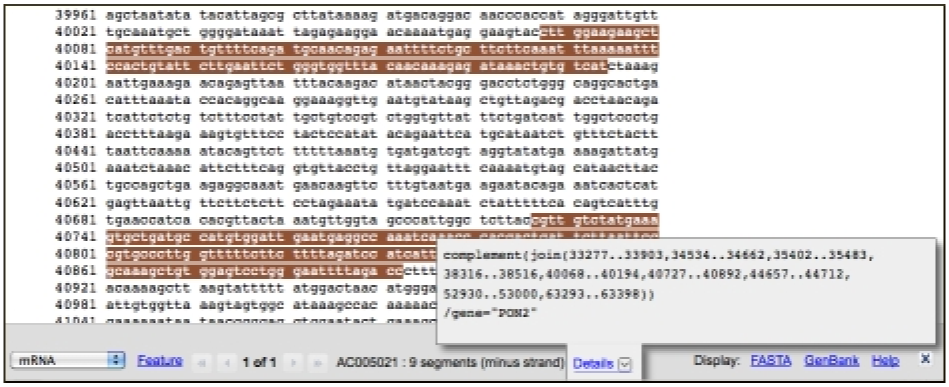


```
4921 cgagtgtcag tacaagggtc cgccccgctc tcagtgccca gctccccccg ggtatcagct
4981 gaaacatcag ctccgcccct gggcgctccc ggagtatcag caaaagggtt cgccccgccc
5041 acagtgcccg gctccccccg ggtatcaaaa gaaggatcgg ctccgccccc gggctccccg
5101 ggggagttga tagaagggtc cttcccaccc tttgccgtcc ccactcctgt gcctacgacc
5161 caggagcgtg tcagccaaag catggagaat caagagaagg cgagtatcgc gggccacatg
5221 ttcgacgtag tcgtgatcgg aggtggcatt tcaggtcagt gtggaccgta gcggtggcct
5281 gggggaccct ggccagtgag gggtagggga acctacagta gctcttgtgg tgtttggggg
```

```
5139..5254
/gene="MAOA"
/gene_synonym="BRNRS; MAO-A"
/inference="alignment:Splign:2.1.0"
/number=1
```

The "Details" box that shows the annotation from the FEATURES table for the highlighted location can be collapsed if desired by clicking the link. Clicking the "Details" link again re-opens the box.

Discontigous features that have multiple segments such as mRNA alignments on genomic DNA can also be highlighted. In all cases the number of segments is shown at the right of the sequence accession. Opposite strand features are indicated with the notation "minus strand" to the right of the number of segments of the bar. The image below shows mRNA minus strand feature for the PON2 gene from an annotated BAC clone sequence (AC005021).



```
39961 agctaatata tacattagcg cttataaaag atgacaggac aacccaccat agggattgtt
40021 tgcaaatgct ggggataaat tagagaagga acaaaatgag gaagtacctt ggaagaagct
40081 catgtttgac tgttttcaga tgcaacagag aattttctgc ttcttcaaat ttaaaaattt
40141 ccactgtatt cttgaattct gggtggttta caacaaagag ataaactgtg tcatctaaag
40201 aattgaaaga acagagttaa tttacaagac ataactacgg gacctctggg caggcactga
40261 catttaaata ccacaggcaa ggaaaggttg aatgtataag ctgttagacg acctaacaga
40321 tcattctctg tctttcctat tgctgtccgt ctggtgttat ttctgatcat tggctccctg
40381 acctttaaga aagtgtttcc tactccatat acagaattca tgcataatct gtttctactt
40441 taattcaaaa atacagttct tttttaaatg tgatgatcgt aggtatatga aaagattatg
40501 aaatctaaac attctttcag gtgttacctg ttaggaattt caaaatgtag cataacttac
40561 tgccagctga agaggcaaat gaacaagttc tttgtaatga agaatacaga aatcactcat
40621 gagttaattg ttcttctctt cctagaaata tgatccaaat ctatttttca cagtcatttg
40681 tgaaccatca cacgttacta aatgttggta gcccattggc tcttaccgtt gtctatgaaa
40741 gtgctgatgc catgtggatt gaatgaggcc aaatca
40801 cgtgcccttg gtttttcttc ttttagatcc atcatt
40861 gcaaagctgt ggagtcctgg gaatttagaa ccettt
40921 acaaaagctt aagtattttt atggactaac atggga
40981 attgtggtta aagtagtggc ataaagccac aaaaac
```

```
complement(join(33277..33903,34534..34662,35402..35483,
38316..38516,40068..40194,40727..40892,44657..44712,
52930..53000,63293..63398))
/gene="PON2"
```

mRNA | Feature | 1 of 1 | AC005021 : 9 segments (minus strand) | Details | Display: FASTA GenBank Help | ×

## Navigating Using the Feature Highlight Bar

If there is more than one feature of the same type, as in the first example shown above, the navigational arrows on the bar allow jumping to the next, previous, first, and last instances of that feature. The Feature pull-down list at the right-hand side of the bar allows selecting other available feature types. The highlight moves to the next available instance of the selected feature type. The "Feature" link returns the display to the corresponding position in the FEATURES table of the record.

## Displaying Highlighted Regions as Separate Sequences

The FASTA and GenBank links on the right-hand side of the bar present the highlighted sub-sequence in the these formats in the Nucleotide or Protein Entrez system and provide a simple means to display and download the corresponding sequence or to forward it to the analysis available analysis tools: BLAST, Primer-BLAST, Find in this Sequence, and Identify Conserved Domains (protein only).

# 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?

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.

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

### 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.

# Search Field Descriptions for Sequence Database

Monica Romiti, M.L.S.[1] and Peter Cooper, Ph.D.[2]

Created: December 3, 2010; Updated: July 5, 2024.

**Table 1.** Fields available for Nucleotide and Protein Sequence Databases.

| Search Field | Short Field Specifier† | Definition |
|---|---|---|
| **[Accession]** | **[ACCN]** | The accession number assigned by NCBI.<br>*Examples:*<br>AF123456[ACCN] Nucleotide<br>NP_000240[ACCN] Protein |
| **[All Fields]** | **[ALL]** | All terms from all search fields in the database.<br>*Example:*<br>human[All Fields] Nucleotide Protein<br>(Compare with human[Organism], see [Organism] entry in this table.) |
| **[Author]** | **[AU]**<br>**[AUTH]** | All authors from all references in the records. The format is last name [space] first initial(s), without punctuation.<br>*Example:*<br>venter jc[AUTH] Nucleotide Protein |
| **[EC/RN Number]** | **[ECNO]** | Enzyme Commission (EC) number for an enzyme activity.<br>*Example:*<br>5.3.1.9[ECNO]) Protein Nucleotide<br>(glucose-6-phosphate isomerase) |
| **[Feature Key]** (Nucleotide, Protein, GSS) | **[FKEY]** | Biological features listed in the Feature Table of the sequence records.<br>*Examples:*<br>3 utr[FKEY] Nucleotide<br>nonstdres[FKEY] Protein<br>The GenBank feature table definition has more information on available features. |

**Author Affiliations:** 1 Kevric Corp; Email: romiti@ncbi.nlm.nih.gov. 2 NCBI; Email: cooper@ncbi.nlm.nih.gov.

*Table 1. continued from previous page.*

| Search Field | Short Field Specifier† | Definition |
|---|---|---|
| **[Filter]** | **[FILT]**<br>**[SB]** | Filtered subsets of the database. An important kind of filter is based on the presence of links to other records. Other filters create useful subsets of data such as those set as Filters in the Discovery column of search results<br>*Examples:*<br>*Links*<br>nucleotide_protein[Filter] Nucleotide<br>protein_structure[Filter] Protein<br>*Organism or properties subsets*<br>all[filter] Nucleotide Protein<br>mrna[filter] Nucleotide<br>refseq[filter] Nucleotide Protein<br>mammals[filter] Nucleotide Protein |
| **[Gene Name]** | **[GENE]** | Gene names annotated on database records. For NCBI Reference Sequences, these names correspond to official nomenclature guidelines when possible. Submitters provide the gene names on GenBank/GenPept records. Gene names on submitted records may be historical names or vary from official guidelines for other reasons.<br>*Example:*<br>BRCA1[GENE] Nucleotide Protein |
| **[Bioproject]** | **[BPRJ]** | The numeric unique identifier for the BioProject that produced the sequence records.<br>*Examples:*<br>13139[Bioproject] Nucleotide Protein<br>(Oryza sativa Japonica)<br>21117[Bioproject] Nucleotide<br>(Pelagic Microbial Assemblages in the Oligotrophic Ocean) |
| **[Issue]** | **[ISS]** | The issue number of the journals cited on sequence records, not generally useful in sequence databases. |
| **[Journal]** | **[JOUR]** | The name of the journals cited on sequence records. Journal names are indexed in the database in abbreviated form although many full titles are mapped to their abbreviations. Journals are also indexed by their by International Standard Serial Number (ISSN).<br>*Examples:*<br>proceedings of the national academy of sciences of the united states of america[Journal] Nucleotide Protein<br>Proc Natl Acad Sci U S A[Journal] Nucleotide Protein<br>0027-8424[Journal] Nucleotide Protein |

*Table 1. continued from previous page.*

| Search Field | Short Field Specifier† | Definition |
|---|---|---|
| **[Keyword]** | **[KYWD]** | Keywords applied by submitter or from controlled vocabularies applied by NCBI or other databases. Except for specific kinds of records, such as the examples given below, the terms in this index are not well controlled. This field is unpopulated for many GenBank/GenPept records.<br>*Examples:*<br>BARCODE[KYWD] Nucleotide Protein<br>HTG[KYWD] Nucleotide<br>RefSeqGene[KYWD] Nucleotide<br>WGS_MASTER[KYWD] Nucleotide |
| **[Modification Date]** | **[MDAT]** | The date of most recent modification of a sequence record. The date format is YYYY/MM/DD. Only the year is required. The Modification Date is often used as a range of dates. The colon ( : ) separates the beginning and end of a date range.<br>*Examples:*<br>2023/01/08[MDAT] Nucleotide Protein<br>1995/09[MDAT] Nucleotide Protein<br>2022/01:2023/12/31[MDAT] Nucleotide Protein |
| **[Molecular Weight]** (Protein only) | **[MOLWT]** | The molecular weight in Daltons of the protein chain calculated from the amino acids only. This may not correspond to the molecular weight of the protein obtained from biological samples because of incomplete data or post-translational modifications of the protein in living systems. The colon ( : ) separates the beginning and end of a molecular weight range.<br>*Examples:*<br>3039[MOLWT] Protein<br>25000:75000[MOLWT] Protein |
| **[Organism]** | **[ORGN]** | The scientific and common names for the complete taxonomy of organisms that are the source of the sequence records.This vocabulary includes all available nodes in the NCBI taxonomy database.<br>*Examples:*<br>cellular organisms[ORGN] Nucleotide Protein<br>firmicutes[ORGN] Nucleotide Protein<br>human[ORGN] Nucleotide Protein<br>Escherichia coli O157:H7[ORGN] Nucleotide Protein |
| **[Page Number]** | **[PAGE]** | The page numbers of the articles that are cited on the sequence record, not generally useful in sequence databases. |
| **[Primary Accession]** | **[PACC]** | The primary accession number of the sequence record. This is the first one appearing on the ACCESSION line in the GenBank/GenPept format. Many records have additional secondary accessions representing records that have been merged. The Accession field indexes both primary and secondary accessions.<br>*Examples:*<br>U01317[PACC] Nucleotide<br>M18047[PACC] Nucleotide<br>(Compare: M18047[ACCN] Nucleotide, see [Accession] entry in this table.) |

*Table 1. continued from previous page.*

| Search Field | Short Field Specifier† | Definition |
|---|---|---|
| **[Primary Organism]** | **[PORGN]** | The primary organism when there is more than one source organism.<br>*Examples:*<br>human[PORGN] Nucleotide<br>(Compare with human[ORGN] Nucleotide, see [Organism] entry in this table.) |
| **[Properties]** | **[PROP]** | Molecular type, source database, and other properties of the sequence record. Terms indexed for this field are a useful classification system for sequence records.<br>*Examples:*<br>*Molecule type*<br>biomol_ncrna[PROP] Nucleotide<br>biomol_genomic[PROP] Nucleotide<br>biomol_mrna[PROP] Nucleotide<br>*Cellular location*<br>gene_in_genomic[PROP] Nucleotide Protein<br>gene_in_mitochondrion[PROP] Nucleotide Protein<br>gene_in_plastid[PROP] Nucleotide Protein<br>*GenBank division*<br>gbdiv_htg[PROP] Nucleotide<br>gbdiv_vrt[PROP] Nucleotide Protein<br>(These GenBank division queries must be combined with srcdb_genbank[PROP] to retrieve only GenBank records.)<br>*Database source*<br>srcdb_genbank[PROP] Nucleotide Protein<br>srcdb_ddbj/embl/genbank[PROP] Nucleotide Protein<br>srcdb_refseq[PROP] Nucleotide Protein<br>srcdb_pdb[PROP] Nucleotide Protein<br>srcdb_swiss-prot[PROP] Protein |
| **[Protein Name]** | **[PROT]** | The names of protein products as annotated on sequence records. The content of this field is not well controlled for GenBank/GenPept records and may contain inaccurate or incomplete information.<br>*Examples:*<br>aldolase[Protein Name] Nucleotide Protein |
| **[Publication Date]** | **[PDAT]** | The date that records were made public in Entrez. The date format is YYYY/MM/DD. The colon ( : ) separates the beginning and end of a date range.<br>*Examples:*<br>2023/01/08[PDAT] Nucleotide Protein<br>1995/09[PDAT] Nucleotide Protein<br>2022/01:2023/12/31[PDAT] Nucleotide Protein |

*Table 1. continued from previous page.*

| Search Field | Short Field Specifier† | Definition |
|---|---|---|
| **[SeqID String]** | **[SQID]** | The NCBI identifier string for the sequence record. This is a brief structured format used by NCBI software.<br>*Example:*<br>gnl asm gca 000000215 2 chr3 45328308[SeqID String] Nucleotide |
| **[Sequence Length]** | **[SLEN]** | The total length of the sequence – the number of nucleotides or amino acids in the sequence. The colon ( : ) separates the beginning and end of a length range.<br>*Examples:*<br>755[SLEN] Nucleotide Protein<br>100:1000[SLEN] Nucleotide Protein |
| **[Substance Name]** | **[SUBS]** | The names of chemical substances associated with a record. This field is only populated for sequences extracted from structure records – PDB derived sequences. The associated residue position is often included.<br>*Examples:*<br>mg, 1010[Substance Name] Nucleotide<br>atp[Substance Name] Protein |
| **[Text Word]** | **[WORD]** | Text on a sequence record that is not indexed in other fields. Terms indexed here are included in an All Fields search, not generally useful. |
| **[Title]** | **[TI] OR [TITL]** | Words and phrases found in the title of the sequence record. The title is the DEFINITION line of the GenBank/GenPept format of the record. This line summarizes the biology of the sequence and includes the organism, product name, gene symbol, molecule type, and sequence completeness.<br>complete cds[TI] Nucleotide<br>kinesin[TI] Nucleotide Protein<br>liver[TI] Nucleotide Protein<br>uncultured[TI] Nucleotide Protein |
| **[Volume]** | **[VOL]** | Contains the volume number of the journals in references on the sequence record, not generally useful in the sequence databases. |

† Queries using any term followed by the full name of the indexed field in square brackets will only retrieve records with the term indexed in that field. For example a search with apolipoprotein[Title] finds only records with "apolipoprotein" indexed for their Title field. Some fields have shorter names that can also be used instead of the full name. These are listed in the **Abbreviated Field Specifier** column of Table 1 when available.