Entrez Sequences Help

Last Updated: April 22, 2016

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: April 22, 2016.

This is a quick start guide for 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.

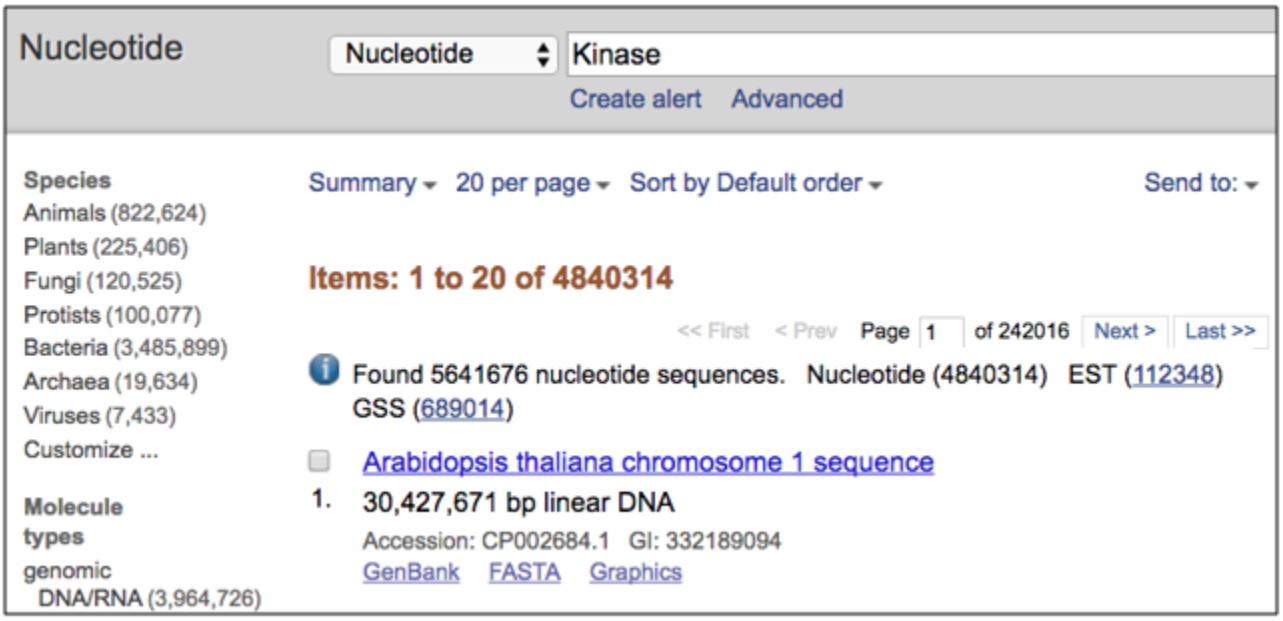
- Which of the three databases (nucleotide, EST, or GSS) should I search?
- How do I use a simple query, such as a word or a phrase?
- How can I make my search more specific with Boolean operators (AND, OR, NOT)?
- How do I restrict my search to specific subsets of records such as those from a specific 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?

Which of the three databases containing nucleic acid sequence (Nucleotide, EST, or GSS) should I search?

The Nucleotide, Genome Survey Sequence (GSS), and Expressed Sequence Tag (EST) database all contain nucleic acid sequences. The data in GSS and EST are from two large bulk sequence divisions of GenBank. GSS and EST data are typically uncharacterized, short genomic (GSS) or cDNA (EST) sequences.

Searching any of the three databases will provide links to results in the other. Unless you know that you are trying to find a specific set of EST or GSS sequences, searching the Nucleotide database with general text queries will produce the most relevant results. You can always follow links to results in EST and GSS from the Nucleotide database results.

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



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

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

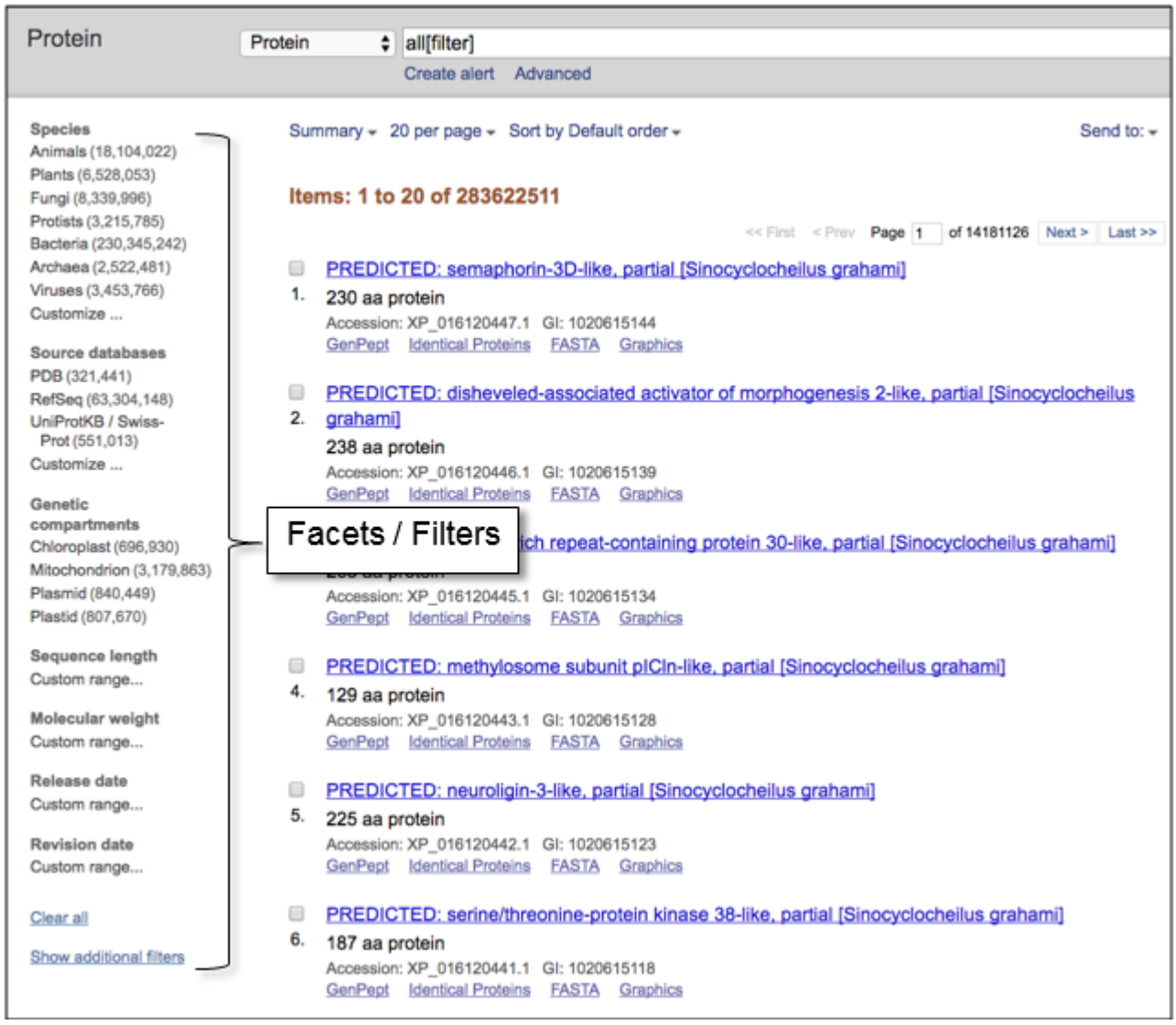
`contactin NOT neurofascin` Protein Nucleotide

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.

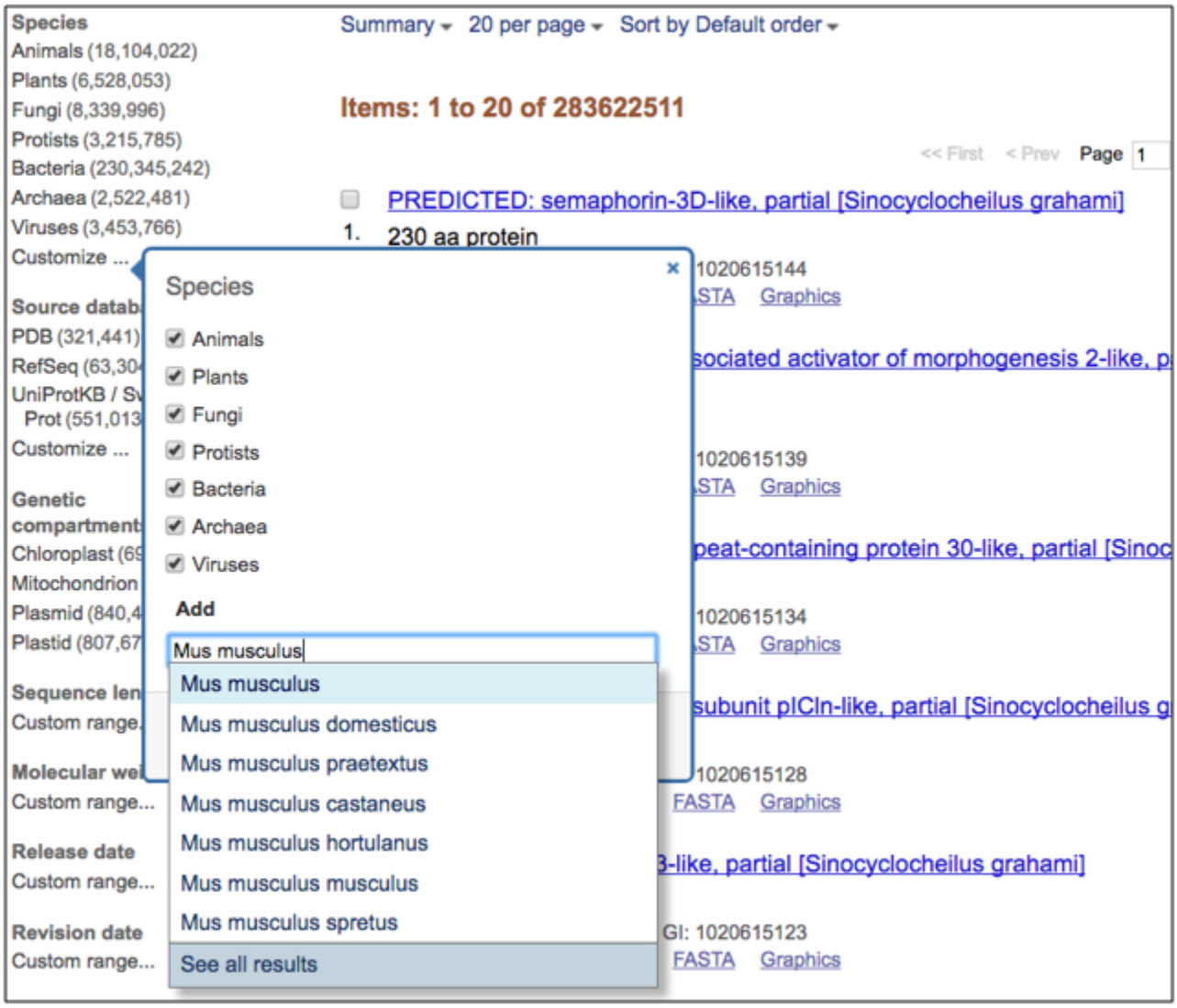
Facets

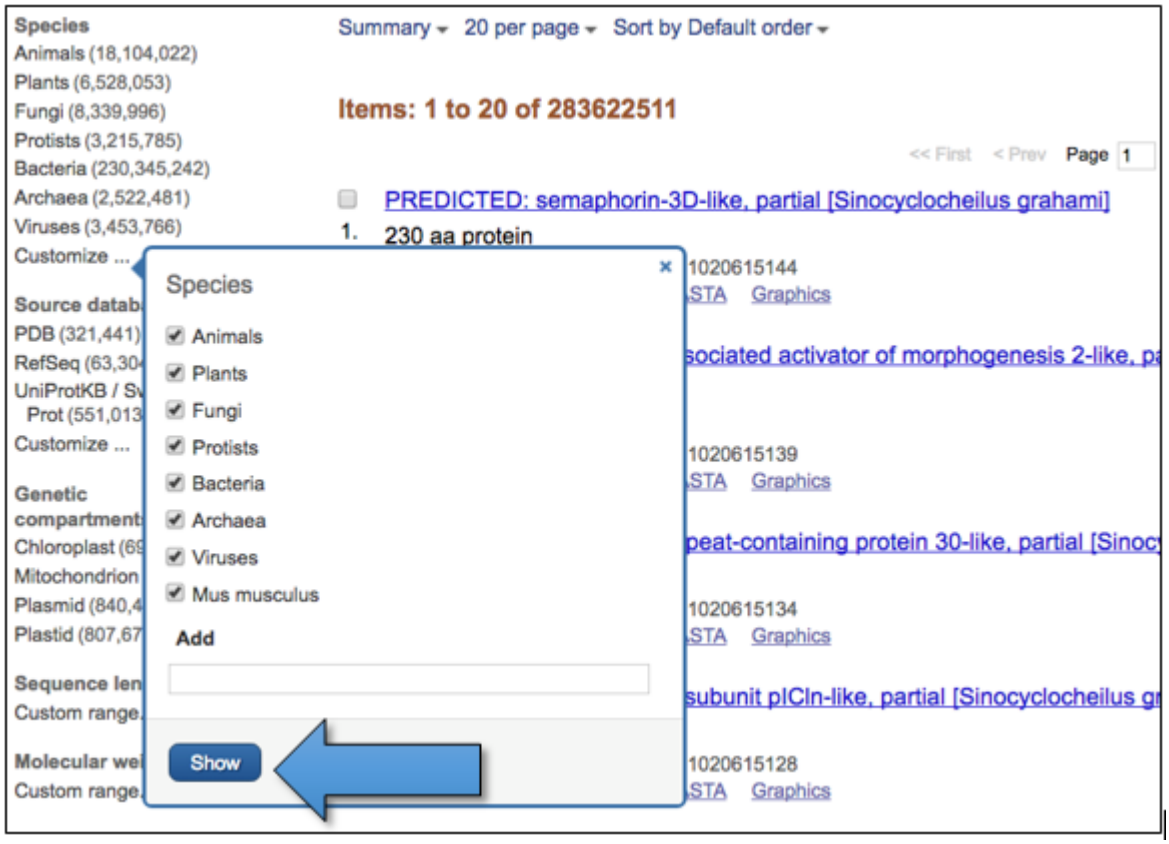
Use the facets on the left-hand side of any of the Protein, Nucleotide, GSS, or EST webpages to restrict the types of records shown.

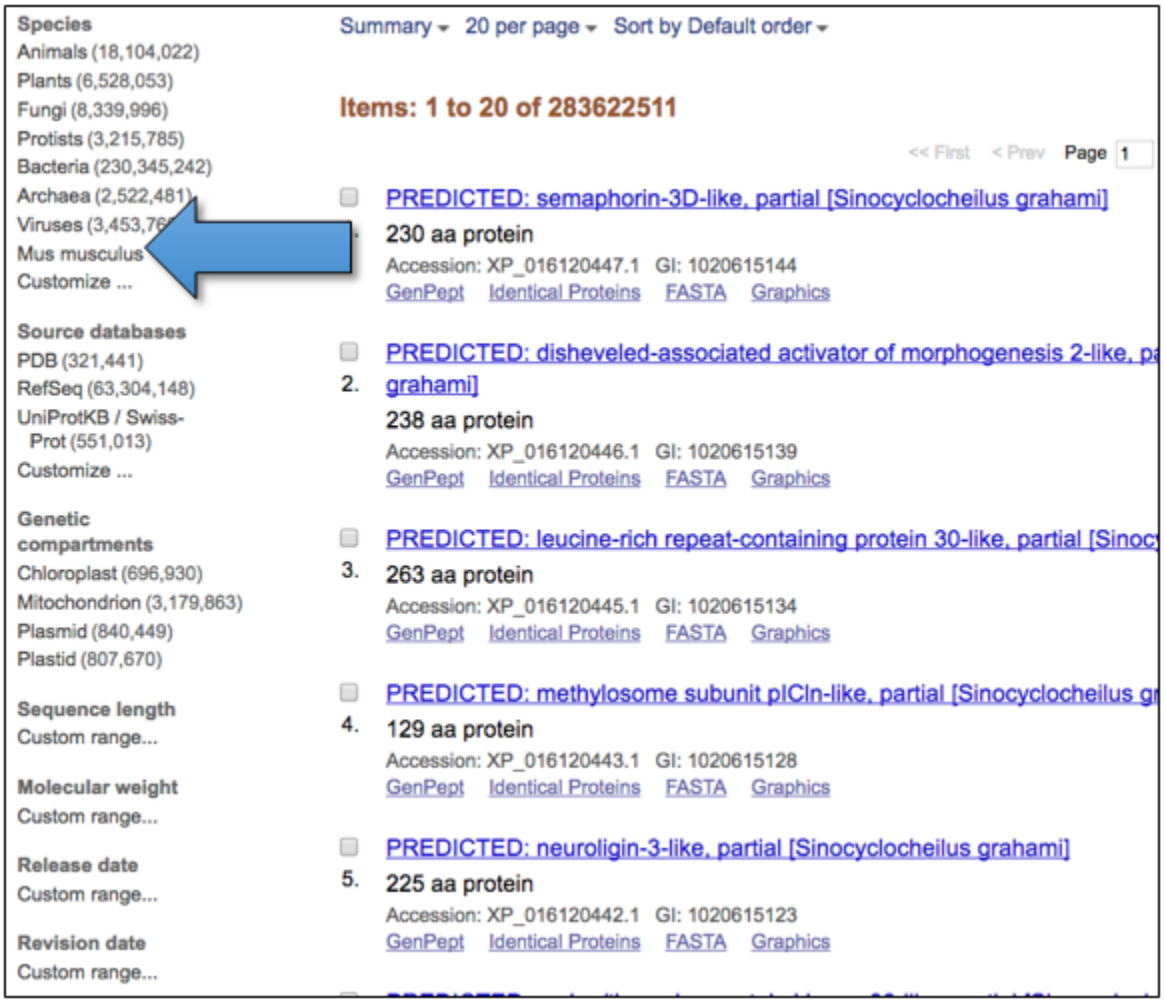


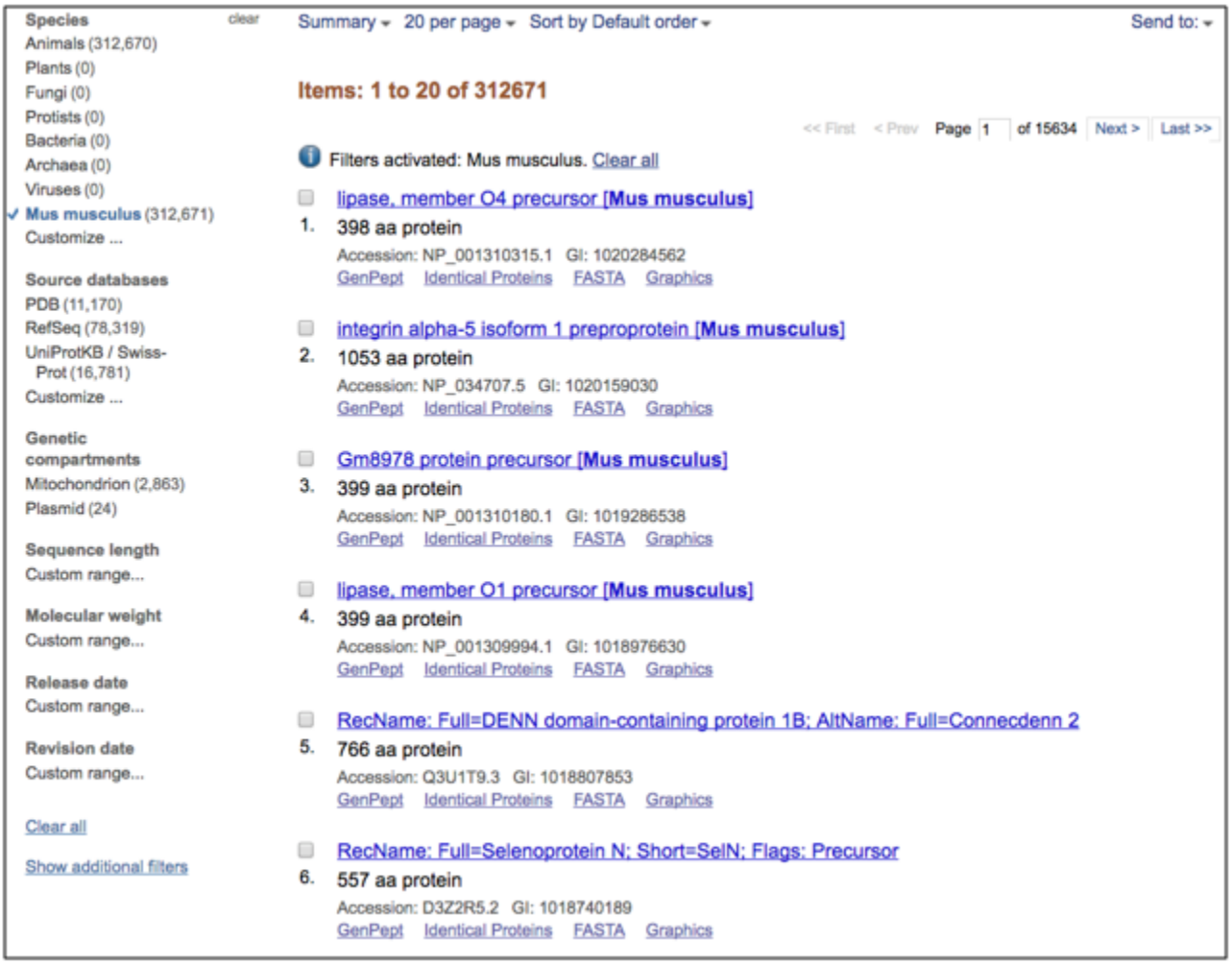
Organism

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.

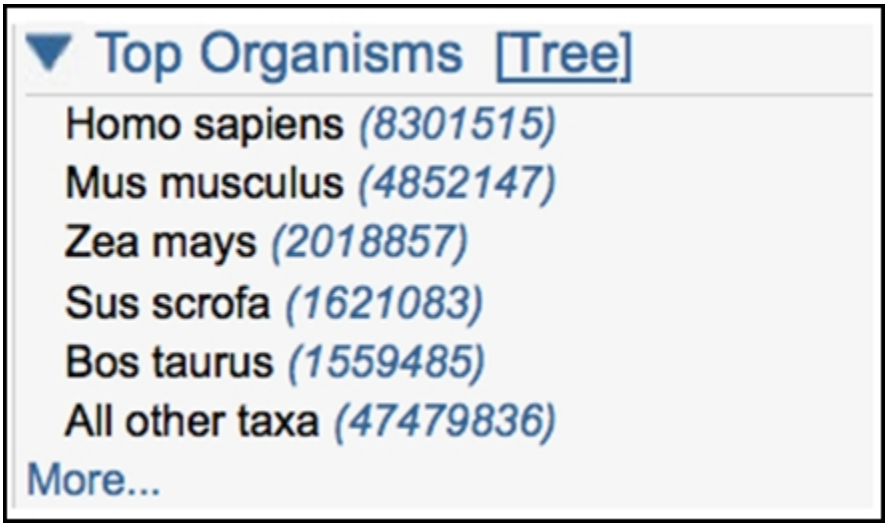

Species
Animals (18,104,022)
Plants (6,528,053)
Fungi (8,339,996)
Protists (3,215,785)
Bacteria (230,345,242)
Archaea (2,522,481)
Viruses (3,453,766)
Customize ...
Summary
20 per page
Sort by Default order
Items: 1 to 20 of 283622511
<< First
< Prev
Page 1
PREDICTED: semaphorin-3D-like, partial [Sinocyclocheilus grahami]
1. 230 aa protein
Species
Animals
Plants
Fungi
Protists
Bacteria
Archaea
Viruses
Add
Mus musculus
Mus musculus
Mus musculus domesticus
Mus musculus praetextus
Mus musculus castaneus
Mus musculus hortulanus
Mus musculus musculus
Mus musculus spretus
See all results
Source datab
PDB (321,441)
RefSeq (63,30
UniProtKB / Sv
Prot (551,013
Genetic compartment
Chloroplast (69
Mitochondrion
Plasmid (840,4
Plastid (807,67
Sequence len
Custom range.
Molecular wei
Custom range...
Release date
Custom range...
Revision date
Custom range...
1020615144
STA Graphics
sociated activator of morphogenesis 2-like, p
1020615139
STA Graphics
peat-containing protein 30-like, partial [Sinoc
1020615134
STA Graphics
subunit pICln-like, partial [Sinocyclocheilus g
1020615128
FASTA Graphics
3-like, partial [Sinocyclocheilus grahami]
GI: 1020615123
FASTA Graphics


Species
Animals (18,104,022)
Plants (6,528,053)
Fungi (8,339,996)
Protists (3,215,785)
Bacteria (230,345,242)
Archaea (2,522,481)
Viruses (3,453,766)
Customize ...
Source databases
PDB (321,441)
RefSeq (63,304,148)
UniProtKB / Swiss-Prot (551,013)
Customize ...
Genetic compartments
Chloroplast (696,930)
Mitochondrion (3,179,863)
Plasmid (840,449)
Plastid (807,670)
Sequence length
Custom range...
Molecular weight
Custom range...
Release date
Custom range...
Revision date
Custom range...
Summary
20 per page
Sort by Default order
Items: 1 to 20 of 283622511
<< First
< Prev
Page 1
PREDICTED: semaphorin-3D-like, partial [Sinocyclocheilus grahami]
1. 230 aa protein
Species
Animals
Plants
Fungi
Protists
Bacteria
Archaea
Viruses
Mus musculus
Add
Show
1020615144
STA
Graphics
sociated activator of morphogenesis 2-like, pa
1020615139
STA
Graphics
peat-containing protein 30-like, partial [Sinocy
1020615134
STA
Graphics
subunit pICln-like, partial [Sinocyclocheilus gr
1020615128
STA
Graphics


Species
Animals (18,104,022)
Plants (6,528,053)
Fungi (8,339,996)
Protists (3,215,785)
Bacteria (230,345,242)
Archaea (2,522,481)
Viruses (3,453,76
Mus musculus
Customize ...
Source databases
PDB (321,441)
RefSeq (63,304,148)
UniProtKB / Swiss-Prot (551,013)
Customize ...
Genetic compartments
Chloroplast (696,930)
Mitochondrion (3,179,863)
Plasmid (840,449)
Plastid (807,670)
Sequence length
Custom range...
Molecular weight
Custom range...
Release date
Custom range...
Revision date
Custom range...
Summary
20 per page
Sort by Default order
Items: 1 to 20 of 283622511
<< First
< Prev
Page 1
PREDICTED: semaphorin-3D-like, partial [Sinocyclocheilus grahami]
230 aa protein
Accession: XP_016120447.1 GI: 1020615144
GenPept Identical Proteins FASTA Graphics
PREDICTED: disheveled-associated activator of morphogenesis 2-like, pa
2. grahami]
238 aa protein
Accession: XP_016120446.1 GI: 1020615139
GenPept Identical Proteins FASTA Graphics
PREDICTED: leucine-rich repeat-containing protein 30-like, partial [Sinocy
3. 263 aa protein
Accession: XP_016120445.1 GI: 1020615134
GenPept Identical Proteins FASTA Graphics
PREDICTED: methylosome subunit pICln-like, partial [Sinocyclocheilus gr
4. 129 aa protein
Accession: XP_016120443.1 GI: 1020615128
GenPept Identical Proteins FASTA Graphics
PREDICTED: neuroligin-3-like, partial [Sinocyclocheilus grahami]
5. 225 aa protein
Accession: XP_016120442.1 GI: 1020615123
GenPept Identical Proteins FASTA Graphics

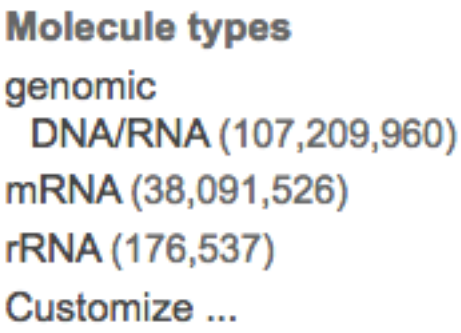


You can also use the linked numbers in the Top Organisms list in the right-hand column of search results to filter select records from specific organisms from your results.



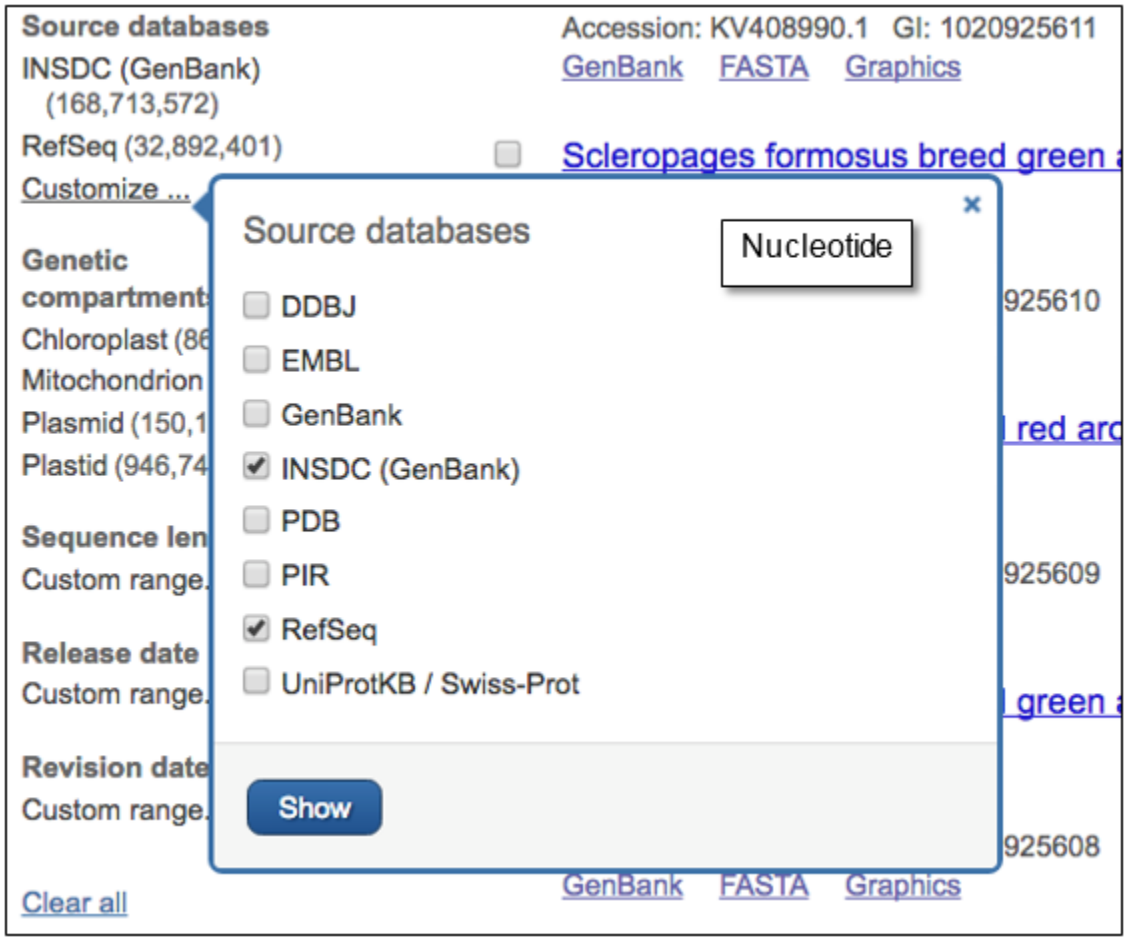
Molecule type

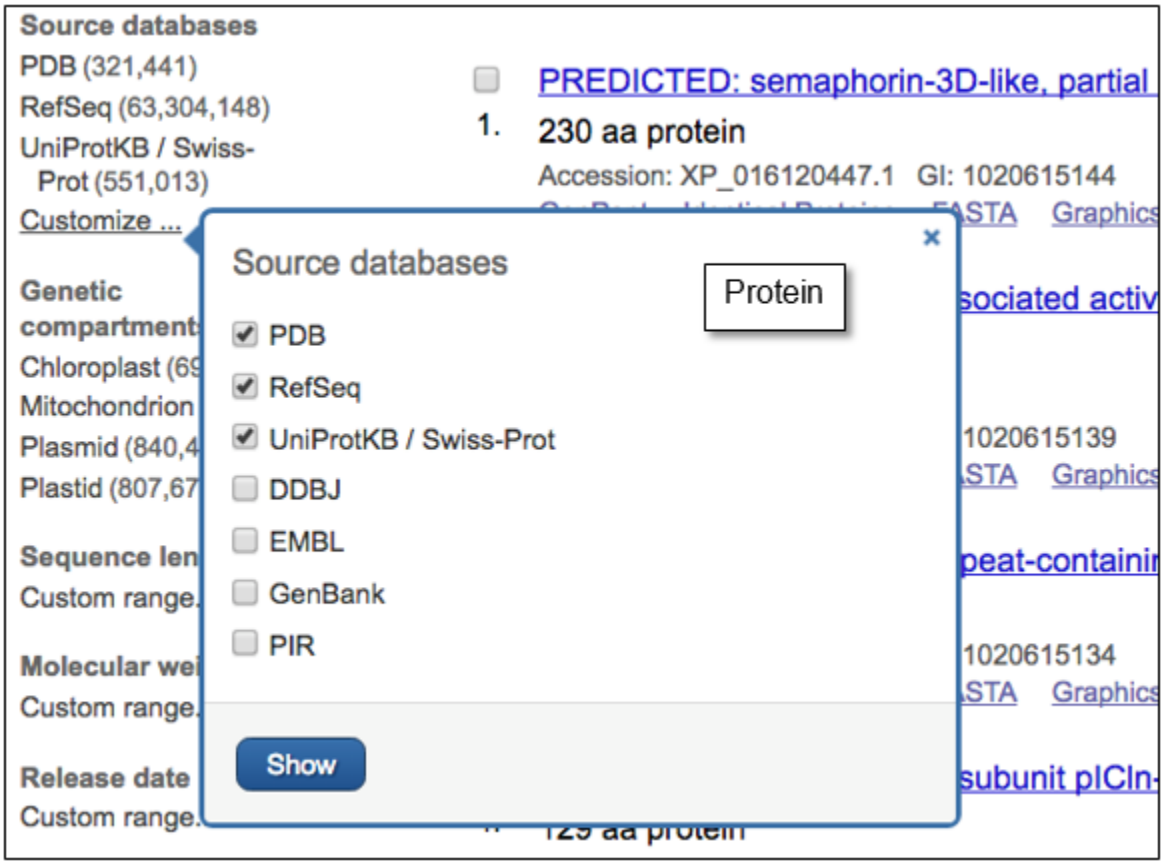
In the Nucleotide database you can use the Molecule types facet to limit results to particular molecule type.



Source database

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





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

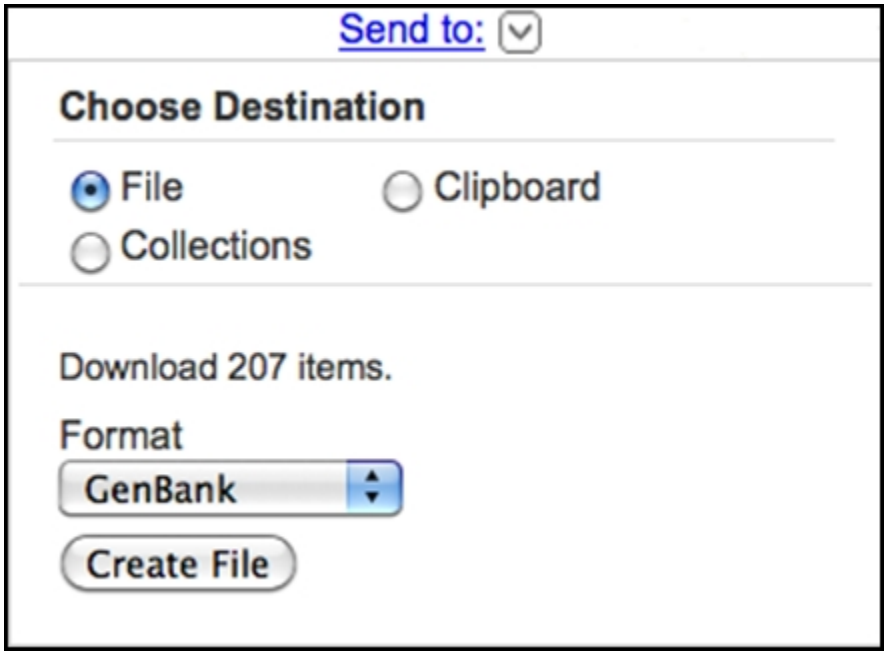
- **Protein:** SwissProt and PIR components of UniProt; Protein Research Foundation (PRF); Protein Data Bank (PDB); and translations of coding regions on sequences in Entrez Nucleotide (RefSeq, International Sequence Database Collaboration – DDBJ / EMBL / GenBank.
- **Nucleotide:** International Sequence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

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.

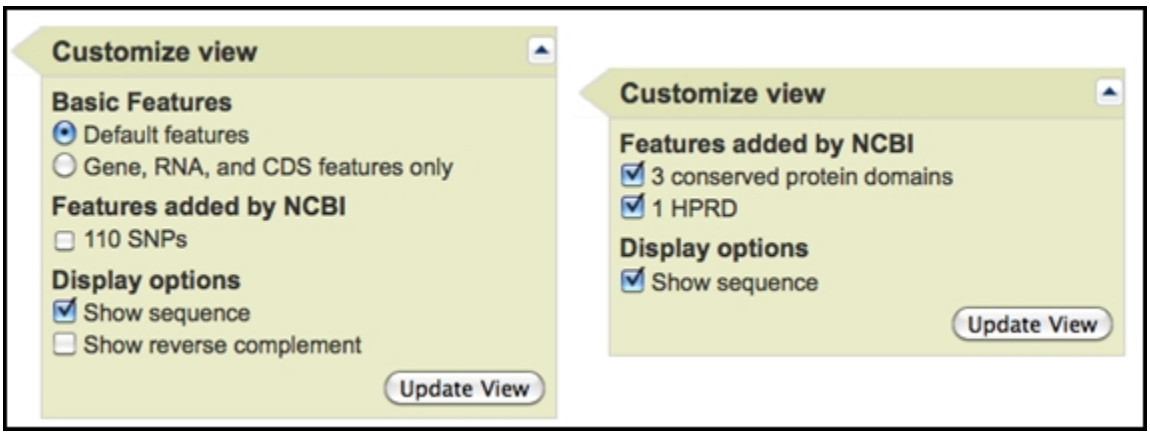
How can I download sequence records to a file on my computer?

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.



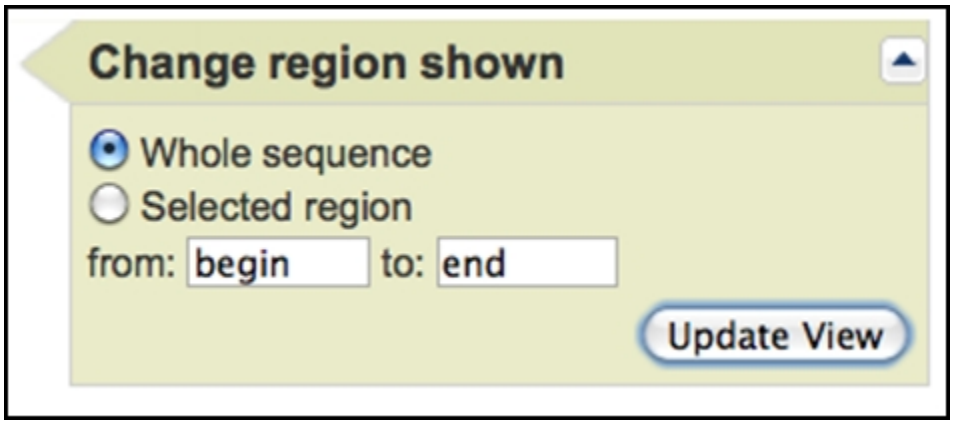
How do I change the information that is shown such as optional biological features or sequence?

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?

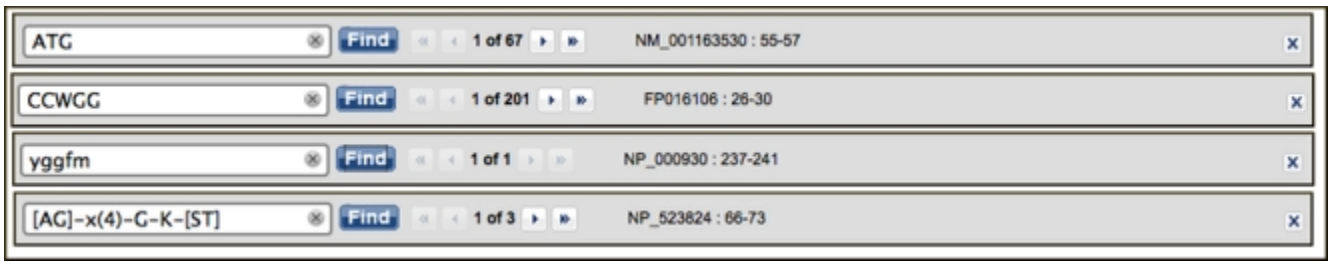
There are direct links to analysis tools including BLAST, Primer-BLAST (Nucleotide, GSS, and EST), and Conserved Domain Database Search (Protein) in the right-hand column of displayed records.

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 Discovery 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.



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
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		

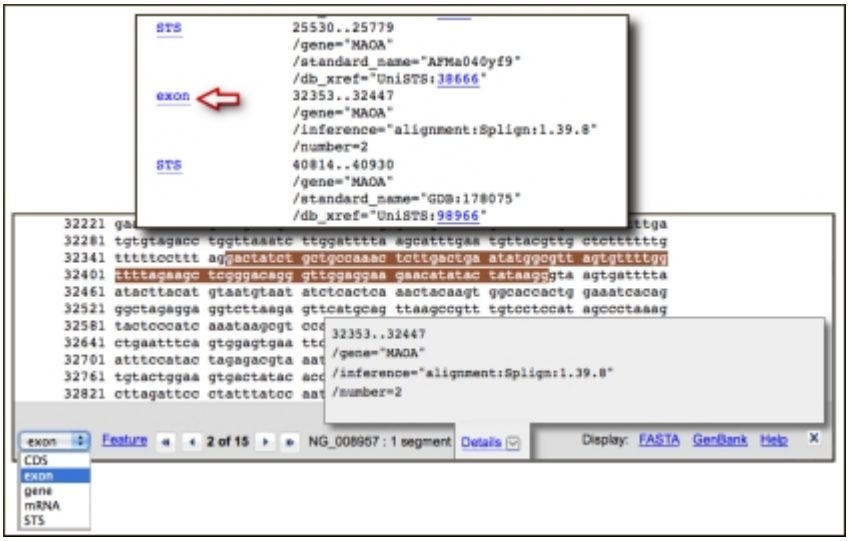
Find matches by clicking the find button. The first 500 matches are highlighted for each displayed sequence. The first or current match is highlighted in white text on a dark background in the sequence, and its position is shown in the search bar. The other matches are highlighted with a light blue background. 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).

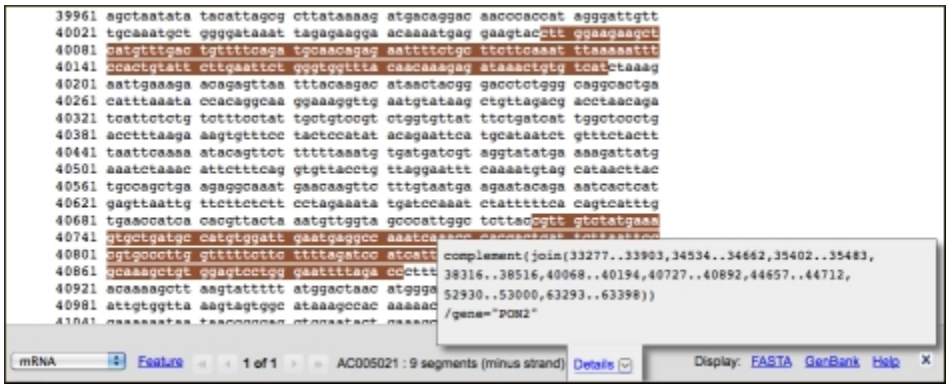
```
FEATURES             Location/Qualifiers
     source          1..97660
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="X"
                     /map="Xp11.3"
     STS             3835..4158
                     /standard_name="PMC130047P4"
                     /db_xref="UniSTS:270611"
     gene            5001..95660
                     /gene="MAOA"
                     /note="monoamine oxidase A"
                     /db_xref="GeneID:4128"
                     /db_xref="HGNC:6833"
                     /db_xref="MIM:309850"
     mRNA            join(5001..5254,32353..32447,42130..42267,60711..60815,
                     61544..61635,77012..77153,80080..80229,80533..80692,
                     81538..81634,85066..85119,89520..89577,90789..90886,
                     92633..92744,92948..93010,93206..95660)
                     /gene="MAOA"
                     /product="monoamine oxidase A"
                     /transcript_id="NM_000240.2"
                     /db_xref="GI:33469954"
                     /db_xref="GeneID:4128"
                     /db_xref="HGNC:6833"
                     /db_xref="MIM:309850"
```

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).



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).



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]

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 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?

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

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.

Search Field Descriptions for Sequence Database

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

Created: December 3, 2010; Updated: February 9, 2011.

Table 1. Fields available for all Sequence Databases (Nucleotide, Protein, EST, GSS). Fields only available for the EST and GSS databases are given in Table 2.

Search Field	Short Field SpecifierSequences_help_appe.TF.2	Definition
[Accession]	**[ACCN]**	The accession number assigned by NCBI. *Examples:* AF123456[ACCN] Nucleotide NP_000240[ACCN] Protein
[All Fields]	**[ALL]**	All terms from all search fields in the database. *Example:* human[All Fields] Nucleotide Protein EST GSS (Compare with human[Organism], see [Organism] entry in this table.)
[Author]	**[AU]** **[AUTH]**	All authors from all references in the records. The format is last name [space] first initial(s), without punctuation. *Example:* venter jc[AUTH] Nucleotide Protein
[EC/RN Number]	**[ECNO]**	Enzyme Commission (EC) number for an enzyme activity. *Example:* 5.3.1.9[ECNO]) Protein Nucleotide (glucose-6-phosphate isomerase)
[Feature Key] (Nucleotide, Protein, GSS)	**[FKEY]**	Biological features listed in the Feature Table of the sequence records. *Examples:* polya signal[FKEY] Nucleotide nonstdres[FKEY] Protein gene[FKEY] GSS 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 SpecifierSequences_help_appe.TF.2	Definition
[Filter]	**[FILT]** **[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 *Examples:* *Links* nucleotide_protein[Filter] Nucleotide protein_structure[Filter] Protein nucest_unigene[Filter] EST nucgss_unists[Filter] GSS *Organism or properties subsets* all[filter] Nucleotide Protein EST GSS mrna[filter] Nucleotide refseq[filter] Nucleotide Protein mammals[filter] Nucleotide Protein EST GSS
[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. *Example:* BRCA1[GENE] Nucleotide Protein
[Genome Project]	-	The numeric unique identifier for the genome project that produced the sequence records. *Examples:* 13139[Genome Project] Nucleotide Protein (Oryza sativa Japonica) 21117[Genome Project] Nucleotide EST GSS (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). *Examples:* proceedings of the national academy of sciences of the united states of america[Journal] Nucleotide Protein EST GSS Proc Natl Acad Sci U S A[Journal] Nucleotide Protein EST GSS 0027-8424[Journal] Nucleotide Protein EST GSS

Table 1. continued from previous page.

Search Field	Short Field SpecifierSequences_help_appe.TF.2	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. *Examples:* BARCODE[KYWD] Nucleotide Protein HTG[KYWD] Nucleotide RefSeqGene[KYWD] Nucleotide 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. *Examples:* 2009/01/08[MDAT] Nucleotide Protein EST GSS 1995/09[MDAT] Nucleotide Protein EST GSS 2010/01:2010/12/31[MDAT] Nucleotide Protein EST GSS
[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. *Examples:* 3039[MOLWT] Protein 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. *Examples:* cellular organisms[ORGN] Nucleotide Protein EST GSS firmicutes[ORGN] Nucleotide Protein human[ORGN] Nucleotide Protein EST GSS 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.

Table 1. continued from previous page.

Search Field	Short Field SpecifierSequences_help_appe.TF.2	Definition
[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. *Examples:* U01317[PACC] Nucleotide M18047[PACC] Nucleotide (Compare: M18047[ACCN] Nucleotide, see [Accession] entry in this table.)
[Primary Organism]	**[PORGN]**	The primary organism when there is more than one source organism. *Examples:* human[PORGN] Nucleotide (Compare with human[ORGN], 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. *Examples:* *Molecule type* biomol crna[PROP] Nucleotide biomol_genomic[PROP] Nucleotide biomol_mrna[PROP] Nucleotide *Cellular location* gene_in_genomic[PROP] Nucleotide Protein gene_in_mitochondrion[PROP] Nucleotide Protein *GenBank division* gbdiv_htg[PROP] Nucleotide gbdiv_vrt[PROP] Nucleotide Protein (These GenBank division queries must be combined with srcdb_genbank[PROP] to retrieve only GenBank records.) *Database source* srcdb_genbank[PROP] Nucleotide Protein EST GSS srcdb_ddbj/embl/genbank[PROP] Nucleotide Protein EST GSS srcdb_refseq_known[PROP] Nucleotide Protein srcdb_refseq_predicted[PROP] Nucleotide Protein srcdb_swiss-prot[PROP] Protein srcdb_pdb[PROP] Nucleotide Protein

Table 1. continued from previous page.

Search Field	Short Field SpecifierSequences_help_appe.TF.2	Definition
[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. *Examples:* 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. *Examples:* 2009/01/08[PDAT] Nucleotide EST GSS 2009/01/10[PDAT] Protein 1995/09[PDAT] Nucleotide Protein EST GSS 2010/01:2010/12/31[PDAT] Nucleotide Protein EST GSS
[SeqID String]	**[SQID]**	The NCBI identifier string for the sequence record. This is a brief structured format used by NCBI software. *Example:* 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. *Examples:* 755[SLEN] Nucleotide Protein EST GSS 100:1000[SLEN] Nucleotide Protein EST GSS
[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. *Examples:* mg, 1010[Substance Name] Nucleotide 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. complete cds[TI] Nucleotide kinesin[TI] Nucleotide Protein liver[TI] Nucleotide Protein EST uncultured[TI] Nucleotide Protein EST GSS

Table 1. continued from previous page.

Search Field	Short Field SpecifierSequences_help_appe.TF.2	Definition
[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.

Table 2. Fields available only for EST and GSS databases.

Index Search Field	Description
[Clone ID]	The clone identifier provided by the submitter of the EST or GSS records. *Example:* image 1000232[Clone ID] EST ZMMBBb0001G04f[Clone ID] GSS
[EST Name] **[GSS Name]**	The name given to the EST or GSS record by the submitter. *Examples:* R-OVA-119[EST Name] EST DKFZP761J17121[GSS Name] GSS
[EST ID] **[GSS ID]**	Legacy dbEST or dbGSS unique identifier provided by NCBI. *Examples:* 2081316[EST ID] EST 14283478[GSS ID] GSS
[Library Class] (GSS Only)	Information about the kind of genomic DNA library that was the source of the clone. Examples: bac ends[Library Class] GSS methylation filtered [Library Class] GSS cosmid ends[Library Class] GSS shotgun[Library Class] GSS
[Library Name] (EST Only)	The name given to the cDNA library that is the source of the clone, provided by the submitter and taken verbatim from the record. May contain useful information about the cell, tissue, or organ source. *Examples:* soares fetal liver spleen 1nfls[Library Name] EST full length enriched swine cdna library, adult adrenal gland[Library Name] EST
[Submitter Name]	Submitter name of EST and GSS records. Unlike [Author Name], the Submitter Name content is not controlled and is verbatim from the EST or GSS record *Examples:* smith tpl[Submitter Name] EST GSS david severson[Submitter Name] EST da lightfoot and chris town[Submitter Name] GSS