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

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 ...

Source datab
PDB (321,441)
RefSeq (63,30
UniProtKB / S
  Prot (551,013

Customize ...

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...

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   ✕   1020615144

☑ Animals   STA   Graphics

☑ Plants   sociated activator of morphogenesis 2-like, p

☑ Fungi

☑ Protists   1020615139

☑ Bacteria   STA   Graphics

☑ Archaea   peat-containing protein 30-like, partial [Sinoc

☑ Viruses

Add   1020615134

| Mus musculus   STA   Graphics

Mus musculus

Mus musculus domesticus   subunit pICln-like, partial [Sinocyclocheilus g

Mus musculus praetextus

Mus musculus castaneus   1020615128

Mus musculus hortulanus   FASTA   Graphics

Mus musculus musculus   3-like, partial [Sinocyclocheilus grahami]

Mus musculus spretus

See all results   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 datab
PDB (321,441)
RefSeq (63,30
UniProtKB / Sv
  Prot (551,013
Customize ...

**Genetic
compartment**
Chloroplast (69
Mitochondrion
Plasmid (840,4
Plastid (807,67

Sequence len
Custom range.

Molecular wei
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   sociated activator of morphogenesis 2-like, pa
  ☑ Fungi
  ☑ Protists   1020615139
  ☑ Bacteria   STA   Graphics
  ☑ Archaea
  ☑ Viruses   peat-containing protein 30-like, partial [Sinoc
  ☑ Mus musculus
Add   1020615134
                              STA   Graphics
_____
                              subunit pICln-like, partial [Sinocyclocheilus g
Show
                              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 [Sinoc
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

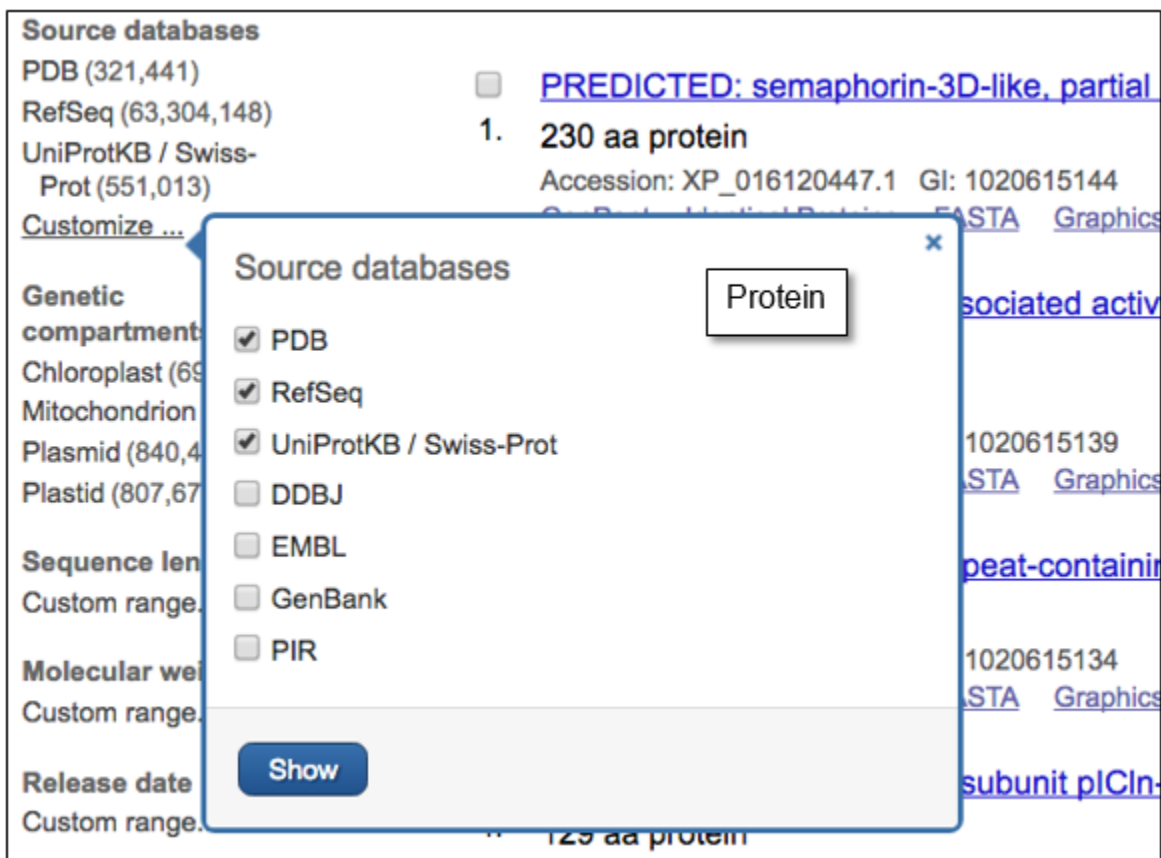In the Nucleotide database you can use the Molecule types facet to limit results to particular molecule type.

Molecule types

genomic
  DNA/RNA (107,209,960)
mRNA (38,091,526)
rRNA (176,537)
Customize ...

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.

Summary ▾ 20 per page ▾ Sort by Default order ▾

Format
- ◉ Summary
- ○ GenPept
- ○ GenPept (full)
- ○ FASTA
- ○ FASTA (text)
- ○ ASN.1
- ○ Revision History
- ○ Accession List
- ○ GI List

83933920

<< First   < Prev   Page 1

transporter 9 isoform b [Homo sapiens]

10906.1   GI: 1021643673
Proteins   FASTA   Graphics

☐   jmjC domain-containing protein 8 isoform 3 precursor [Homo sapiens]

2.   **240 aa protein**
Accession: NP_001310848.1   GI: 1021643671
GenPept   Identical Proteins   FASTA   Graphics

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.

Send to: ⌄

Choose Destination

◉ File   ○ Clipboard
○ Collections

Download 207 items.

Format
| GenBank | ⬍ |

(Create File)

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.
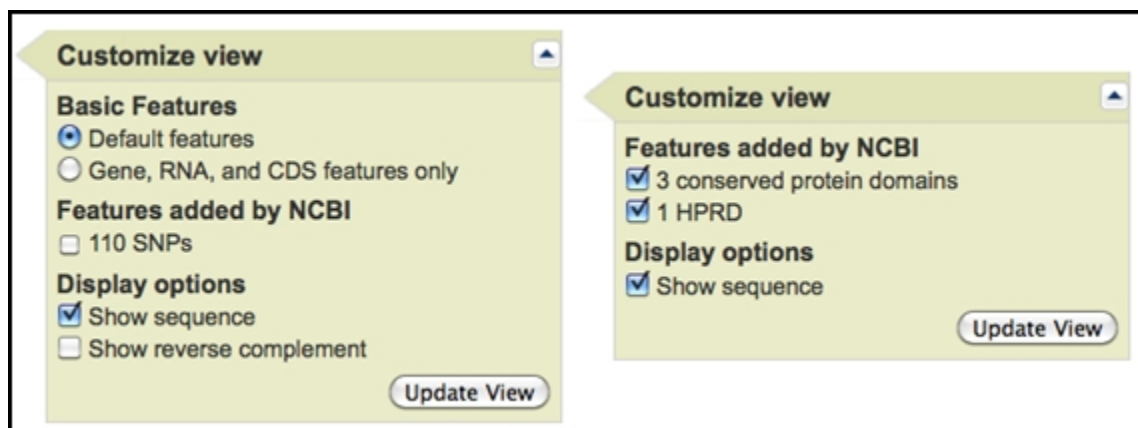
How can I display a portion of the sequence?

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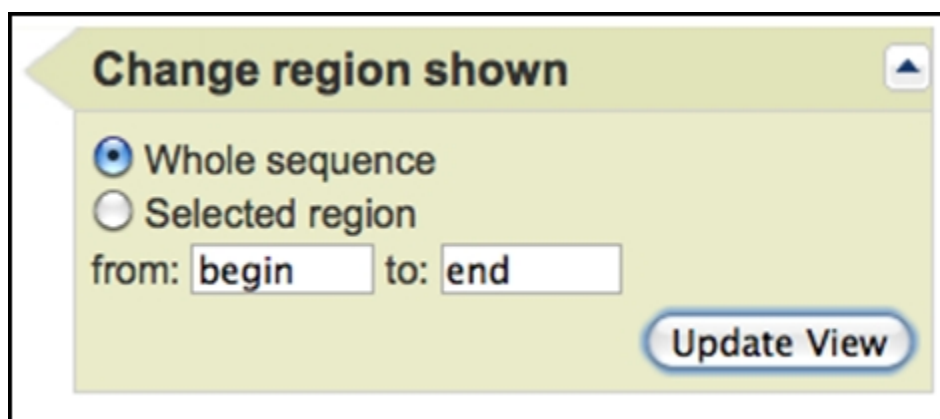
How do I analyze the sequence data directly or find additional related data?

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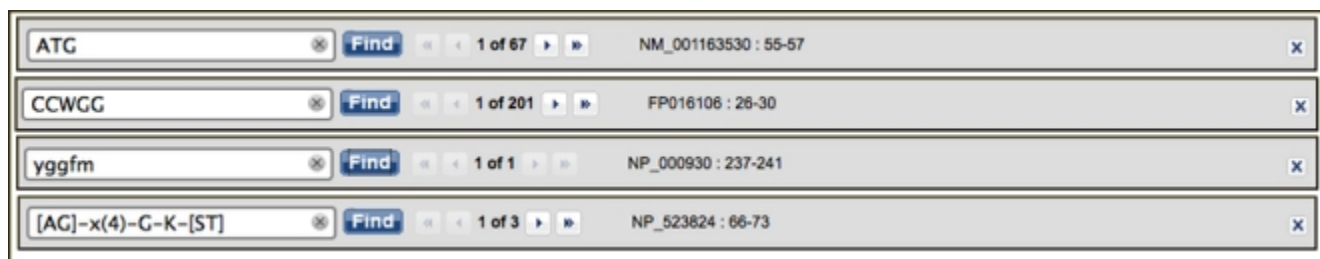
There are also links to related data in the right-hand column that may provide additional information and pre-computed analyses for the displayed records.

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.

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
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).

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).