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, source database, genomic or cDNA library?
- 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.
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, source database, genomic or cDNA library name or properties?

You can use the Limits page to limit your search to only certain kinds of records. You can also use the Filter your results to select categories of records after a search. Follow these links to jump to the limit of interest: organism, molecule type, source database, library name or properties.

Limits

Use the Limits page linked to top of any of the Protein, Nucleotide, GSS, or EST webpages to select the appropriate limit from the various pull-down lists.

Organism

To get records from a specific organism or group of organisms use the Search Field Tags pull-down list and select Organism.
You can then use the common or scientific name of a species, strain, or higher taxon as a search term. Examples: human, mouse, *Drosophila simulans*, green plants, bacteria.

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

Use the Molecule pull-down list on the Limits page to restrict results to particular molecule type.

You can find more information about using the field restriction to refine your query in Entrez Help.

Source database

Use the Source database pull-down list on the Limits page to restrict results to particular molecule type.
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.

**Library name or properties**

Molecular library names or properties are most useful for Genome Survey Sequences (GSS) or Expressed Sequence Tags (EST). To limit your search to these fields choose Library Name or Library Class (GSS only) from the Search Field Tags pull-down list on the Limits page.
The following are some example queries for Library Name.

**GSS:** mus musculus 129sv/ev, Biston betularia BAC library.

**EST:** acorn worm normalized neurula pexpess1 library, human liver regeneration after partial hepatectomy.

Examples of common Library Class terms indexed for GSS are bac ends, bac subclone, fosmid ends, gene trap, and methylation filtered.

**Filter your results**

You can also use the *Filter your results* links that appear in the right-hand column to select certain categories of records from your results after a search.
How do I change the format, number, or sorting order of records displayed?

Click the Display Settings menu that appears at the upper left of document summaries or record views and select the desired format, items per page, or sorting order from the listed radio buttons. Click the Apply button to activate the new settings.

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.

<table>
<thead>
<tr>
<th>Nucleotide Codes</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A adenine</td>
<td>Y T or C</td>
</tr>
<tr>
<td>C cytidine</td>
<td>M A or C</td>
</tr>
<tr>
<td>G guanine</td>
<td>W A or T</td>
</tr>
<tr>
<td>T thymidine</td>
<td>R G or A</td>
</tr>
<tr>
<td>N A, G, C, or T</td>
<td>B G, T, or C</td>
</tr>
<tr>
<td>U uridine (matches T)</td>
<td>D G, A, or T</td>
</tr>
<tr>
<td>K G or T</td>
<td>H A, C, or T</td>
</tr>
<tr>
<td>S G or C</td>
<td>V G, C, or A</td>
</tr>
</tbody>
</table>
Amino Acid Codes

<table>
<thead>
<tr>
<th>Letter</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>alanine</td>
</tr>
<tr>
<td>B</td>
<td>aspartate/asparagine</td>
</tr>
<tr>
<td>C</td>
<td>cysteine</td>
</tr>
<tr>
<td>D</td>
<td>aspartate</td>
</tr>
<tr>
<td>E</td>
<td>glutamate</td>
</tr>
<tr>
<td>F</td>
<td>phenylalanine</td>
</tr>
<tr>
<td>G</td>
<td>glycine</td>
</tr>
<tr>
<td>H</td>
<td>histidine</td>
</tr>
<tr>
<td>I</td>
<td>isoleucine</td>
</tr>
<tr>
<td>K</td>
<td>lysine</td>
</tr>
<tr>
<td>L</td>
<td>leucine</td>
</tr>
<tr>
<td>M</td>
<td>methionine</td>
</tr>
<tr>
<td>N</td>
<td>asparagine</td>
</tr>
<tr>
<td>P</td>
<td>proline</td>
</tr>
<tr>
<td>Q</td>
<td>glutamine</td>
</tr>
<tr>
<td>R</td>
<td>arginine</td>
</tr>
<tr>
<td>S</td>
<td>serine</td>
</tr>
<tr>
<td>T</td>
<td>threonine</td>
</tr>
<tr>
<td>V</td>
<td>valine</td>
</tr>
<tr>
<td>W</td>
<td>tryptophan</td>
</tr>
<tr>
<td>Y</td>
<td>tyrosine</td>
</tr>
<tr>
<td>Z</td>
<td>glutamate/glutamine</td>
</tr>
</tbody>
</table>

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.

Discontiguous 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).