

NCBI News, July 2010

Peter Cooper, Ph.D.¹ and Dawn Lipshultz, M.S.²

Created: July 31, 2010.

Featured Resource: Updated Entrez Sequence Database Interfaces

The Entrez sequence databases (Nucleotide, Protein, GSS, and EST) have recently completed migration to the streamlined discovery-oriented design that has been in service in PubMed for nearly a year, described fully in the [November 2009](#) issue of the NCBI News. The sequence database re-design includes new homepages, a simpler interface, and new options for downloading, displaying sequences, and connecting to related data.

New homepages

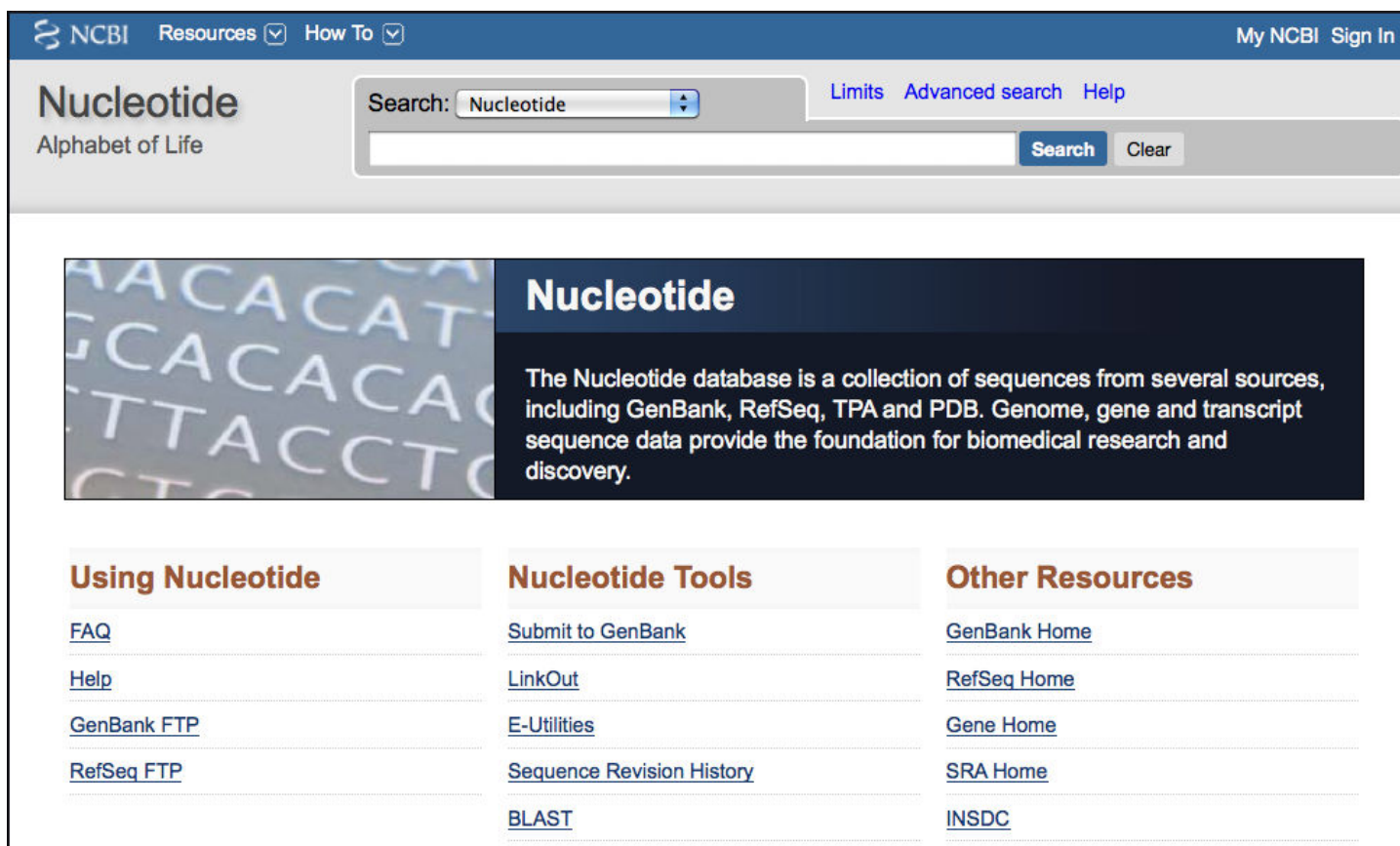
The sequence database homepages now have a simplified design with three columns of links with access to information about using the resource, tools for submitting data, searching and analysis, and other related resources at the NCBI site. As in the PubMed and the Site Guide (NCBI Homepage), the new sequence database pages have the new header including the search bar with access to all NCBI Entrez databases and the *Resources* and *How To* pull-down lists to aid navigation and to access to practical task-oriented help and the footer that provides rapid navigation to all major areas of the NCBI site. Figure 1 shows the new Entrez nucleotide homepage.

Improvements to the search interface

The new interface is simplified eliminating the four control tabs of the previous version: *Limits*, *Preview/Index*, *History*, *Clipboard*, and *Details*. These functions are still available but are now easier to access and use. A redesigned *Limits* page is linked at the upper-right of the *Search Box* that appears on all pages. The new *Advanced search* page also linked above the *Search Box* combines the functions of the old *Preview/Index* and *History* tabs. The *Search details* providing the translation of the query by the Entrez engine is linked on the *Advanced search* page but is also shown in the right-hand *Discovery* column as in the new PubMed. The *Clipboard*, when populated, is now accessible as a link at the top of the *Discovery* column as described below.

Using the *Advanced search* page

Figure 2 shows the *Advanced search* page for Entrez protein. The page functions as an independent search interface that allows formulation of complex queries. The *Search builder* and *Search history* sections replace the previous *Preview/Index* page and facilitate the construction of more precise queries. The pull-down list in the *Search Builder* shows all of the fields indexed for a particular database. The *Show Index* link opens an alphabetical list of terms for the selected field. When a term is entered in the *Search Builder*, the index will open to the closest match in the index. The *Add to Search Box* button puts the field-restricted queries into the *Search*



NCBI Resources How To My NCBI Sign In

Nucleotide
Alphabet of Life

Search: Nucleotide Limits Advanced search Help

Search Clear

Nucleotide

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

Using Nucleotide

- [FAQ](#)
- [Help](#)
- [GenBank FTP](#)
- [RefSeq FTP](#)

Nucleotide Tools

- [Submit to GenBank](#)
- [LinkOut](#)
- [E-Utilities](#)
- [Sequence Revision History](#)
- [BLAST](#)

Other Resources

- [GenBank Home](#)
- [RefSeq Home](#)
- [Gene Home](#)
- [SRA Home](#)
- [INSDC](#)

Figure 1. The new nucleotide homepage with access to related resources. The new sequence database homepages include the NCBI header and footer (not shown) that provide easy navigation to other parts of the site and links to task-oriented help documentation.

Box. These may be run using the *Search* button or may be added to the *Search History* using the *Preview* button. Entries in the *Search History* may be combined to give very precise results. The example in Figure 2 combines searches for frogs (#5), RefSeq proteins (#2), and prolactin (#3) to obtain the prolactin protein records for *Xenopus laevis*, NP_001086486, and *Xenopus (Siluriana) tropicalis*, NP_001093699).

Search results and other multiple record pages

Search result pages showing document summaries and all other multiple records views now incorporate fully the features of the PubMed redesign. Figure 3 shows the new search results for a protein search – prolactin[protein name] AND (Birds OR Mammals). The document summaries are presented in a new format with the record title first and hyperlinked to the record view. The summary now also shows the length of the sequence. The sequence identifiers, accession version and GI number, are listed below the summary. The old links menu has been removed from the document summaries in search results. However links to retrieve related sequences (similar by BLAST) are shown under the summary by default in the nucleotide and protein databases. The protein summaries also include a link to identical sequences. All links available for each record can be displayed if desired though individual settings in a My NCBI account. See the [My NCBI Help](#) manual for more information on customizing these preferences.

Multiple-record displays in one of the full-record formats such as GenBank or FASTA have the same set of controls in the new style as the summaries. The only additional feature is the *Customize view* control, described for single sequence views in the [March 2009 NCBI News](#). In the case of multiple records, this control allows toggling the reverse complement of the displayed sequences.

Protein Advanced Search [« Back to Protein](#)

Search Box [Limits](#) [Details](#) [Help](#)
 ((#2) AND #5) AND #3 [Search](#) [Preview](#) [Clear](#)

Search Builder
 All Fields [AND](#) [Add to Search Box](#)
[Show Index](#)

[Search Builder Instructions](#)

Search History

Search	Most Recent Queries	Time	Result
#16	Search ((#2) AND #5) AND #3	17:02:27	2
#15	Search creatine kinase[Protein Name]	14:19:27	149
#14	Search creatine kinase	14:13:08	986
#5	Search frogs	14:01:43	106353
#4	Search (#2) AND #3	13:58:27	10
#3	Search "srcdb refseq"[Properties]	13:57:38	10929822
#2	Search "prolactin"[Protein Name]	13:56:37	360
#1	Search guinea pig[organism]	13:54:22	1692
#0	protein clipboard	13:58:26	10

[Less History](#) [Clear History](#)

Protein Name: creatine kinase [AND](#) [Add to Search Box](#)

creatine kinase (149)
 creatine kinase 1 (5)
 creatine kinase 1 2 (1)
 creatine kinase 2 (4)
 creatine kinase 2 2 (1)
 creatine kinase 28aa (1)
 creatine kinase 3 (1)
 creatine kinase 4 (1)
 creatine kinase a (2)
 creatine kinase b (11)

[Show Index](#)
[Previous 200](#)
[Next 200](#)
[Close Index List](#)

Figure 2. Entrez protein Advanced search page. The Search Builder function and Search History replace the former Preview/Index page. The pull-down list shows all indexed fields for the database. The Show Index link (green arrow) expands the display to show all terms indexed for a particular field. The bottom panel shows the index for the Protein Name field matching the search term creatine kinase. Entries from the Search Builder and the Search History can be combined in the Search Box to construct complex queries. Clicking on the numbered entries in the Search History provides Options for combining searches, removing History entries, loading results, showing queries, and saving the search in My NCBI. Combining #2, #5 and #3 finds the two *Xenopus* RefSeq proteins for prolactin.

New items in the right-hand Discovery column

The search filters that formerly were a series of tabs are now implemented as a set of links at the upper part of the right-hand Discovery column. In the case of the results shown in Figure 3, the RefSeq filter has been clicked filtering the output to show only the six Reference Sequences. Clicking the plus sign (+) at the right of the selected filter will add the filter as a term to the search. The Search details, previously a control tab that shows how the query is interpreted or mapped to Entrez controlled vocabularies is now exposed in the Discovery column. The Search details at the bottom right of Figure 3 show how the terms Birds and Mammals were expanded, translated, and mapped to the organism field (NCBI Taxonomy) to generate more accurate results. The former Clipboard tab now appears as a link above the right-hand column only if the NCBI Clipboard contains items (Figure 3, upper right).

Display Settings: Summary, 20 per page, Sorted by Default order Send to: Clipboard: 265 items

Results: 6

[prolactin precursor \[Oryctolagus cuniculus\]](#)
1. 227 aa protein
NP_001076144.1 GI:130504953
[Related Sequences](#) [Identical Proteins](#) [Item in clipboard](#)

[prolactin \[Ovis aries\]](#)
2. 240 aa protein
NP_001009306.1 GI:57164329
[Related Sequences](#) [Identical Proteins](#) [Item in clipboard](#)

[prolactin precursor \[Equus caballus\]](#)
3. 229 aa protein
NP_001075365.1 GI:126352562
[Related Sequences](#) [Identical Proteins](#) [Item in clipboard](#)

[prolactin precursor \[Bos taurus\]](#)
4. 229 aa protein
NP_776378.2 GI:46810277
[Related Sequences](#) [Identical Proteins](#) [Item in clipboard](#)

[prolactin precursor \[Homo sapiens\]](#)
5. 227 aa protein
NP_001157030.1 GI:254675133
[Related Sequences](#) [Identical Proteins](#)

[prolactin precursor \[Gallus gallus\]](#)
6. 229 aa protein
NP_990797.1 GI:49169789
[Related Sequences](#) [Identical Proteins](#)

Filter your results:

[All \(265\)](#)

Bacteria (0)

[Related Structures \(198\)](#)

RefSeq (6) +

[Manage Filters](#)

Top Organisms [Tree]

- Oryctolagus cuniculus (1)
- Homo sapiens (1)
- Bos taurus (1)
- Ovis aries (1)
- Equus caballus (1)
- All other taxa (1)

[More...](#)

Analyze these sequences ▲

[Run BLAST](#)

[Align sequences with COBALT](#)

Find related data ▲

Database: [Select](#)

[Find items](#)

Search details ▲

```
prolactin[protein name] AND
(("Aves"[Organism] OR
Birds[All Fields]) OR
("Mammalia"[Organism] OR
Mammals[All Fields]))
```

[Search](#) [See more...](#)

Display Settings: Summary, 20 per page, Sorted by Default order

Send to:

Figure 3. Entrez protein search results for the query `Prolactin[protein name] AND (Birds OR Mammals)` showing the new summary style. The Discovery column now has the search filters, *Search details*, the analysis tool BLAST and COBALT, and the *Find related data* device that displays linked records related to the present set of results. The inset shows the Encoding mRNA selection that retrieves the corresponding RefSeq mRNA records for these proteins. The *Search details* show how the terms Birds and Mammals were mapped to the controlled vocabulary of the indexed Organism field.

Also in the Discovery column are two new items, *Analyze these sequences* and *Find related data* that provide access to analysis tools and pre-computed relationships – true Discovery components. *Analyze these sequences*, available for displays containing 20 or fewer records, provides direct access to the NCBI BLAST service from all sequence databases and, for multiple protein records, the NCBI multiple-alignment tool COBALT, described in the [May 2009 NCBI News](#). This is a convenient interface to COBALT since it allows direct submission after using Entrez to collect the desired input set. For example the set of homologous prolactin proteins from mammals and birds in Figure 3 can now be easily aligned using COBALT.

The links shown in the *Find related data* list previously were presented on the Display pull-down list at the top of the results. The new location in the Discovery column makes these links easier to find and use. These provide access to the wealth of pre-computed and pre-compiled relationships that make Entrez a powerful discovery system. In many cases there may be more than one kind of link to another database. The inset in Figure 3 shows the multiple links to the nucleotide database available from the protein data. Selecting “Encoding mRNA” and clicking *Find items* links to the six corresponding RefSeq mRNA records.

Display Settings and Send to menus

The *Display settings* and *Send to* menus are now accessed through links at the left (top and bottom) and right (top and bottom) respectively of single- and multiple-record displays (Figure 3). The upper panel of Figure 4 shows the expanded *Display settings* menu that provides the ability to select any of the record formats available for the database, to modify the number of records displayed, and to sort by publication, release date, accession, or organism. As in PubMed, My NCBI now allows the default settings for these options in the sequence database to be set to different values if desired. This new option in My NCBI is described in more detail in the Updates and Enhancements section of the current NCBI News. The *Send to* menu provides options for saving items to the NCBI clipboard, Collections in My NCBI, or to download sequence records in various formats to a local file (Figure 4, lower left). With the new mechanism for saving records it is no longer necessary to display the record in the desired format before downloading.

Nucleotide Send menu and Coding Sequence Download

For nucleotide records the *Send* menu allows downloading either the complete record or all annotated coding sequence regions as FASTA formatted sequences directly from the parent records. The feature allows downloading either the nucleotide sequences or the corresponding protein translation. This often-requested feature works with annotated CDS features on any nucleotide record from simple open reading frames on mRNA sequences to complex multi-exon genes on mammalian genomic regions. Each downloaded CDS has its own structured title that includes a unique identifier incorporating the parent sequence accessions, gene symbol, protein product, reading frame, protein identifier, and location on parent sequence. This CDS download function can be used to quickly create a local database of sequences for analysis.

Summary

The updated Entrez interface now implemented in the sequence databases provides a streamlined and less complex search interface as well as improved consistency of form and function across the molecular and literature resources making the NCBI site easier to use. New options for displaying and downloading records especially the ability to download coding regions are important improvements to the utility and flexibility of the molecular biology Web services at the NCBI. The presence of analysis tools and improved access to related data on search results and record displays increase the power of Entrez as a system for scientific discovery.

New Databases and Tools

Bibliography Management

My NCBI will replace eRA Commons for grantee bibliography management. The *My Bibliography* function in My NCBI will link with Commons to allow scientists to maintain and manage a list of their publications including journal articles, book chapters, meeting abstracts, talks and presentations, patents, and other materials. Journal articles can be those found in PubMed, journals not indexed or not yet appearing in PubMed, or manuscripts submitted to NIHMS. Citations may not be entered manually into eRA Commons at this point, and users must use My Bibliography to manage their bibliographies. For more information, please see the [NIH Announcement](#).

The image displays the 'Display Settings' and 'Send to' menus for the NCBI Protein database. The 'Display Settings' menu is divided into three columns: 'Format', 'Items per page', and 'Sort by'. The 'Format' column includes options like Summary, GenPept, FASTA, and ASN.1. The 'Items per page' column has radio buttons for 5, 10, 20, 50, 100, and 200. The 'Sort by' column includes Default order, Accession, Date Modified, Date Released, Organism Name, and Taxonomy ID. Below this is an 'Apply' button. The 'Send to' menu is a dropdown menu with options like File, Clipboard, and Collections. The 'Send' menu is a dropdown menu with options like Complete Record and Coding Sequences. The 'Choose Destination' menu is a dropdown menu with options like File, Clipboard, and Collections. The 'Download features' menu is a dropdown menu with options like FASTA Nucleotide and FASTA Protein.

Display Settings: [v]

Format	Items per page	Sort by
<input checked="" type="radio"/> Summary	<input type="radio"/> 5	<input type="radio"/> Default order
<input type="radio"/> GenPept	<input type="radio"/> 10	<input type="radio"/> Accession
<input type="radio"/> GenPept (full)	<input checked="" type="radio"/> 20	<input checked="" type="radio"/> Date Modified
<input type="radio"/> FASTA	<input type="radio"/> 50	<input type="radio"/> Date Released
<input type="radio"/> FASTA (text)	<input type="radio"/> 100	<input type="radio"/> Organism Name
<input type="radio"/> ASN.1	<input type="radio"/> 200	<input type="radio"/> Taxonomy ID
<input type="radio"/> Revision History		
<input type="radio"/> Accession List		
<input type="radio"/> GI List		

Send to: [v]

Send: [v]

Choose Destination

File Clipboard

Collections

Download 6 items.

Format

GenPept [v]

Create File

GenPept
GenPept (full)
FASTA
ASN.1
XML
INSDSeq XML
TinySeq XML
Feature Table
Accession List
GI List

Complete Record
 Coding Sequences

Choose Destination

File Clipboard

Collections

Complete Record
 Coding Sequences

Download features.

Format

FASTA Nucleotide [v]

FASTA Nucleotide
FASTA Protein

Create File

Figure 4. Display settings and Send to menu for the Protein database. These menus have equivalent options in all sequence databases for multiple and single record displays. The *Send* menu that replaces the *Send to* menu for nucleotide records with annotated coding regions has an option to download either the complete record or coding regions in FASTA nucleotide or protein.

NCBI Discovery Workshops

NCBI will present a two-day workshop on September 29-30, on the NIH campus in Bethesda, MD. The workshops provide hands-on experience exploring practical examples using tools and databases on the NCBI

website. The four workshops are Sequences, Genomes, and Maps; Proteins, Domains and Structures; NCBI BLAST Services; and Human Variation and Disease Genes. For more information see the [Discovery Workshop](#) page, which also includes a registration link.

Microbial Genomes

Twenty-one finished microbial genomes were released in July 2010. The original sequence data files submitted to GenBank/EMBL/DDBJ are available on the FTP site: <ftp.ncbi.nih.gov/genbank/genomes/Bacteria/>. The RefSeq provisional versions of these genomes are also available: <ftp.ncbi.nih.gov/genomes/Bacteria/>.

GenBank News

GenBank release 178.0 is available through the NCBI web and FTP sites. The current release includes information available as of June 11, 2010. Release notes are available on the on the NCBI ftp site: <ftp.ncbi.nih.gov/genbank/gbrel.txt>

Updates and Enhancements

My NCBI now offers sequence database preferences

My NCBI allows Preferences to be set for record format and result display settings for the Entrez sequence databases (Nucleotide, Protein, EST and GSS). Record formats can be changed from the default (GenBank, GenPept, EST, GSS) to FASTA or graphics for Nucleotide and Protein or to GenBank or FASTA for EST or GSS records. All available options on the *Display settings* menu in the sequence databases can also be changed from the default settings. The relevant preferences dialogs from My NCBI are shown below.

The screenshot shows the NCBI My NCBI interface. The top navigation bar includes links for Home, PubMed, GenBank, and BLAST, along with a user profile for 'pscooper' and options to Sign Out or My NCBI. The main content area is divided into sections for Nucleotide Preferences, Protein Preferences, GSS Preferences, and EST Preferences. Each section has links for 'Record Display Format' and 'Result Display Settings'. Two pop-up windows are overlaid on the Protein Preferences section. The first pop-up, titled 'Result Display Settings for Protein', allows users to set default values for sorting order and number of items per page. The second pop-up, titled 'Record Display Settings for Protein', allows users to set the default record format.

Result Display Settings for Protein

Set the default values for sorting order and number of items to be displayed per page in Protein:

Default items per page:

5 10
 20 50
 100 200

Default sort by:

Default Order Accession
 Date Modified Date Released
 Organism Name Taxonomy ID

Record Display Settings for Protein

Set the default values for displaying record in Protein:

Default record format:

GenPept
 FASTA
 Graphics

Or cancel and return to the [preferences page](#)

RefSeq

RefSeq Release 42 is available through the Entrez system and can be downloaded from the FTP site (<ftp.ncbi.nlm.nih.gov/refseq/release>). This full release incorporates genomic, transcript, and protein data available as of July 21, 2010. It includes 15,038,858 records from 10,728 different organisms. Changes since the last release can be found in the release notes (<ftp.ncbi.nih.gov/refseq/release/release-notes/RefSeq-release42.txt>)

New FTP file for Gene

Entrez Gene calculates matches between NCBI and Ensembl annotations and reports the matches in the Entrez gene Full Report display, the "matches Ensembl" index property, and the gene2ensembl FTP file. A new FTP file, README_ensembl, will soon be added to provide a summary of species whose annotations have been compared, including release and assembly information, and the date when the comparison was last performed. A complete description of the file is on the ftp site: <ftp.ncbi.nih.gov/gene/DATA/README>

Announce Lists and RSS Feeds

Eighteen topic-specific mailing lists are available which provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on

the Announcement List summary page: www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html. To receive updates on the *NCBI News*, please see: www.ncbi.nlm.nih.gov/About/news/announce_submit.html.

Twelve RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: www.ncbi.nlm.nih.gov/feed/.

Users can also stay updated on NCBI's resources on Facebook and Twitter: twitter.com/NCBI.

Send comments and questions about NCBI resources to: info@ncbi.nlm.nih.gov, or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.