

# Entrez Help

Last Updated: May 31, 2016



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

NLM Citation: Entrez Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-.

This book contains information on Entrez, the indexing and data retrieval system developed by the National Center for Biotechnology Information (NCBI).

## Table of Contents

<b>Entrez Help</b> .....	1
The Entrez Databases .....	1
Access to the Entrez System .....	6
Entrez Searching Options .....	6
Displaying and Saving a Set of Records .....	13
Related data: Neighbors and Links .....	13
Creating Links to Web Pages in the Entrez System .....	15
Linking to Records in the Entrez System .....	16
Programmatic Access to the Entrez System .....	16

# Entrez Help

Created: January 20, 2006; Updated: May 31, 2016.

Entrez is NCBI's primary text search and retrieval system that integrates the PubMed database of biomedical literature with 38 other literature and molecular databases including DNA and protein sequence, structure, gene, genome, genetic variation and gene expression. This document is an overview of the Entrez databases, with general information on searching and displaying data. More detailed help is available for the individual Entrez databases in the [NCBI Help Manual](#) sections on the NCBI bookshelf.

The Entrez search interface features powerful options for constructing precise searches and managing results. Options include popular configurable preset facet filters to help focus on specific kinds of results, an Advanced Search interface that facilitates constructing more sophisticated queries. Specialized search fields are available for each database and can be browsed and selected in the Search Builder section of the Advanced Search interface. Other useful Entrez features include Search History with access to recent results and a Clipboard where search results can be saved temporarily. A [My NCBI](#) account increases the power of the system by providing even more flexibility. Most importantly Entrez integrates data with links within and between databases. Not only does this interconnectivity enhance navigation and allow search results to be quickly focused or expanded; but also, more importantly, these relationships often expose unexpected connections that can lead to scientific discoveries

## The Entrez Databases

The Entrez system comprises 39 molecular and literature databases. New databases are added as biomedical science advances and new kinds of data become available. An alphabetical list of the current databases with a brief description of each is given below.

### Assembly

The [Assembly](#) resource provides access to genome assemblies for both submitted data and NCBI RefSeq assemblies. Assembly provides versioned accession identifiers for submitted and RefSeq assemblies, links to the components in the Nucleotide system and direct access to the downloads on the NCBI FTP site.

### BioProject

The [BioProject](#) database is a searchable collection of complete and incomplete (in-progress) large-scale molecular projects including genome sequencing and assembly, transcriptome, metagenomic, annotation, expression and mapping projects. BioProject provides a central point to link to all data associated with a project in the NCBI molecular and literature databases.

### BioSample

[BioSample](#) contains descriptions of biological source materials used in studies that have data in other NCBI molecular databases such as Assembly, Nucleotide and SRA.

### BioSystems

The [BioSystems](#) database collects information on interacting sets of biomolecules involved in metabolic and signaling pathways, disease states, and other biological processes. BioSystems currently contains biological pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the EcoCyc (*Escherichia coli* K-12 MG1655) subset of the BioCyc databases and is designed to accommodate other data in the future. BioSystems records link to related literature, genes, protein sequences, structures, chemical data, to related BioSystems. When available each record links to detailed diagrams and annotations for individual pathways on the Web sites of the source databases.

## Bookshelf

The NCBI [Bookshelf](#) contains a collection of full-text books that can be searched online and that are linked to PubMed records through research paper citations within the text. The collection includes biomedical textbooks, other scientific titles, and NCBI help manuals.

## ClinVar

[ClinVar](#) is a public archive of submitted reports of clinically relevant human genetic variants and their relationships to phenotypes, with supporting evidence. ClinVar provides standardized nomenclature for variants and phenotypes, a review status for variants, and links to related NCBI literature and molecular databases.

## Conserved Domains

[Conserved Domains](#) is a database of protein domains represented by sequence alignments and profiles for protein domains conserved in molecular evolution. It also includes alignments of the domains to known three-dimensional protein structures in the MMDB database. The source databases for Conserved Domains are Pfam, Smart, and COG.

## dbGaP

[dbGaP](#) (Database of Genotypes and Phenotypes) provides the results of studies that have investigated the interaction of genotype and phenotype including genome-wide association studies, medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits.

## dbVAR

[dbVAR](#) (Database of Genomic Structural Variation) contains information about large-scale genomic variation, including large insertions, deletions, translocations and inversions. dbVar also provides associations of defined variants with phenotype information.

## EST

The [EST](#) database contains sequence records from the bulk EST (Expressed Sequence Tag) division of GenBank. These are typically short single-pass reads from cDNA libraries often generated as large survey project. Data from EST can be used to catalog expressed genes for a particular organ, tissue or cell type or general for a species, and compare expression levels of genes in various library sources.

## Gene

[Gene](#) is a searchable database of genes, focusing on genomes that have been completely sequenced and that have an active research community to contribute gene-specific data. Information in Gene records includes nomenclature, chromosomal localization, gene products and their attributes (e.g., protein interactions), associated markers, phenotypes, interactions, and links to citations, sequences, variation details, maps, expression reports, homologs, protein domain content, and external databases.

## Genome

The [Genome](#) database contains sequence and map data from the whole genomes of over 1000 species or strains. The genomes represent both completely sequenced genomes and those with sequencing in-progress. All three main domains of life (bacteria, archaea, and eukaryota) are represented, as well as many viruses, phages, viroids, plasmids, and organelles.

## GEO Datasets

[GEO Datasets](#) stores curated gene expression and molecular abundance data sets assembled by NCBI from the Gene Expression Omnibus (GEO) repository of microarray data.

## GEO Profiles

[GEO Profiles](#) is a database that stores individual gene expression and molecular abundance profiles assembled from the [Gene Expression Omnibus \(GEO\)](#) repository of microarray data.

## GSS

The [GSS](#) database contains sequence records from the bulk GSS (Genome Survey Sequence) division of GenBank. These are the genomic equivalent of EST records; short single pass reads from gDNA libraries. Insert end and other reads from BAC and other large insert genomic libraries used to identify and assemble candidates for genome sequencing are common examples of GSS records.

## GTR

The [Genetic Testing Registry \(GTR\)](#) is a repository for voluntary submissions of genetic test information by providers. The scope of GTR includes the purpose of the test, methodology, validity, evidence of the usefulness of the test, and laboratory contacts and credentials. GTR includes information from and links to NCBI resources such as Gene, ClinVar and MedGen as well as many resources from outside the NIH.

## HomoloGene

The [HomoloGene](#) database contains automatically generated sets of homologous genes and their corresponding mRNA, genomic, and protein sequence data from selected eukaryotic organisms. Potential homologs from other organisms are included through sequence similarity to UniGene clusters.

## MedGen

[MedGen](#) is NCBI's portal to information about human disorders and other phenotypes having a genetic component. MedGen is intended for health care professionals, the medical genetics community and provides centralized access to diverse types of content. MedGen aggregates the wide variety of terms used for particular disorders into a specific concept. Each concept may have associated clinical findings, causative genetic variants and the genes in which they occur, available clinical and research tests, molecular resources, professional guidelines, original and review literature, consumer resources, clinical trials, and links to other related NCBI molecular and literature databases as well as non-NCBI resources.

## MeSH

[MeSH \(Medical Subject Headings\)](#) is the National Library of Medicine's controlled vocabulary and classification system (ontology) used for indexing articles in PubMed. MeSH terminology provides a consistent way to retrieve information that may use different terminology for the same concepts. Searches in the Entrez MeSH database provide synonymous MeSH terms that can provide more useful results in PubMed. The MeSH database records show subheadings access the MeSH browser showing related concepts and hierarchical relationships among MeSH terms.

## NCBI Web Site Search

[NCBI Site Search](#) is database of static NCBI web pages, documentation, and online tools. Searching this database is a quick way to find specialized online sequence analysis tools, back issues of newsletters, legacy resource description pages, sample code, and other miscellaneous resources.

## NLM Catalog

The [NLM Catalog](#) contains records for books, journals, audiovisuals, computer software, electronic resources, and other materials in the National Library of Medicine (NLM) collections. The old Journals database was merged into the NLM Catalog database and the information once retrieved via Journals, is provided by the NLM Catalog. This includes data such as journal title, MEDLINE abbreviation, NLM ID, ISO abbreviation, or ISSN.

## Nucleotide

Apart from sequence data in the EST (Expressed Sequence Tag) and GSS (Genome Survey Sequence divisions of GenBank, the [Nucleotide](#) database contains all the sequence data from GenBank, EMBL, and DDBJ, the members of the [International Nucleotide Sequence Databases Collaboration \(INSDC\)](#). Nucleotide also includes NCBI-curated Reference Sequences (RefSeqs), submitted assemblies and annotations from the [Third Party Annotation \(TPA\)](#) database, and nucleotide sequences extracted from structure records from the [Protein Databank \(PDB\)](#).

## OMIM

The [OMIM \(Online Mendelian Inheritance in Man\)](#) database allows searches of OMIM articles about human genes, genetic disorders, and other inherited traits. OMIM articles provide links to associated literature references, sequence records, maps, and related databases. OMIM records are hosted and served by the independent OMIM site ([www.omim.org](http://www.omim.org)). The NCBI service provides searching capabilities.

## PopSet

The [PopSet](#) database contains related nucleotide sequences that originate from comparative studies: phylogenetic, population, environmental (ecosystem), and mutational. Each record in the database is a set of nucleotide sequences representing the same molecule from the same species (population, mutation), different identifiable species (phylogenetic), or anonymous species from the same biological community (ecosystem).

## Probe

[Probe](#) is a database of nucleic acid reagents designed for use in a wide variety of biomedical research applications including genotyping, gene expression studies, SNP discovery, genome mapping, and gene silencing. Probe records contain information on reagent distributors, probe effectiveness, and computed sequence similarities.

## Protein

The [Protein](#) database contains amino acid sequences created from the translations of coding regions provided on nucleotide records in GenBank, EMBL, and DDBJ, the members of the [International Nucleotide Sequence Databases Collaboration \(INSDC\)](#) as well as those from coding regions on NCBI Reference Sequences and the [Third Party Annotation \(TPA\)](#) database records. Protein records are also imported from the outside protein-only data sources [Protein Information Resource \(PIR\)](#), [UniProtKB/Swiss-Prot](#), [Protein Research Foundation \(PRF\)](#). Protein sequences are also extracted from structure records from the [Protein Data Bank \(PDB\)](#).



## Protein Clusters

[Protein Clusters](#) is a collection of related protein sequences (clusters) consisting of Reference Sequence proteins that are encoded by complete prokaryotic genomes as well those encoded eukaryotic organelle plasmids and genomes. The database provides easy access to annotation information, publications, domains, structures, external links, and analysis tools.

## PubChem BioAssay

[PubChem BioAssay](#) is a database that contains bioactivity screens of chemical substances described in PubChem Substance. It provides searchable descriptions of each bioassay, including descriptions of the conditions and readouts specific to that screening procedure.

## PubChem Compound

The [PubChem Compound](#) database contains unique, validated chemical structures (small molecules) that can be searched using names, synonyms or keywords. The compound records may link to more than one PubChem Substance record if different depositors supplied the same structure. Structures in PubChem Compounds are pre-clustered and cross-referenced by identity and similarity groups. Additionally, calculated properties and descriptors are available for searching and filtering of chemical structures. Compound records are linked to related PubChem Substance Records, PubMed citations, protein 3D structures, and biological screening results that are available in PubChem BioAssay.

## PubChem Substance

The [PubChem Substance](#) database contains information on chemical substances including mixtures electronically submitted to PubChem by depositors. This includes any chemical structure information submitted, as well as chemical names, comments, and links to the depositor's web site.

## PubMed

[PubMed](#) is database of citations and abstracts for biomedical literature from MEDLINE and additional life science journals. Links are provided when full text versions of the articles are available through PubMed Central or other websites.

## PubMed Central

[PubMed Central \(PMC\)](#) is the U.S. National Library of Medicine's digital archive of life sciences journal literature. PMC contains full-text manuscripts deposited by authors or articles provided by the publisher.

## SNP

The [SNP \(Single Nucleotide Polymorphism\)](#) database is a central repository for single nucleotide polymorphisms, microsatellites, and small-scale insertions and deletions. Both submitted SNPs and NCBI-produced non-redundant reference records (RefSNPs) that cluster reports of the same polymorphism from different sources are available. SNP also contains population-specific frequency and genotype data, experimental conditions, molecular context, and mapping information for both neutral polymorphisms and clinical mutations.

## SRA

The [SRA \(Sequence Read Archive\)](#) contains sequencing data from the next generation sequencing platforms. SRA accepts and presents data from all current next-generation sequencing platforms including 454 (Roche),

Illumina, SOLiD (Applied Biosystems), HeliScope, and Complete Genomics. Data can include sequence, quality scores, color values, and intensity graphs depending on the platform involved.

## Structure

The **Structure** or Molecular Modeling Database (MMDB) contains experimental data from crystallographic and NMR structure determinations. The data for MMDB are obtained from the Protein Data Bank (PDB). Structure records link to bibliographic information, the sequence databases, and to the NCBI taxonomy. **Cn3D**, the NCBI 3D structure viewer, allows for easy interactive visualization of molecular structures from Entrez.

## Taxonomy

The **Taxonomy** database contains the names and phylogenetic lineages of the more than 350,000 species that have molecular data in the NCBI databases. New taxa are added to the Taxonomy database as data are deposited for them. The taxonomy records include links to all molecular data for the organism or group as well as links to outside classification resources. The taxonomy provides the major controlled vocabulary for classifying molecular data across the Entrez system.

## UniGene

**UniGene** is a database that provides automatically generated nonredundant sets (clusters) of transcript sequences, each cluster representing a distinct transcription locus (gene or expressed pseudogene). UniGene clusters also provide information on protein similarities, gene expression, cDNA clone reagents, and genomic location.

## Access to the Entrez System

Nearly all search boxes that appear on the NCBI site access the Entrez system. The search box at the top of the **NCBI homepage** is a convenient place to begin Entrez searches. With the default All Databases selection, the results are presented on the **Global Query** page shown in Figure 1. This page lists the Entrez databases and the corresponding number of records found by the query in each database. The databases are organized into six broad categories on the Global Query page: Literature, Health, Genomes, Genes, Proteins and Chemicals. Of course, the Global Query page itself can be used to search all database by entering a simple search term or phrase in the *Search across databases* query box. Clicking on the number or the adjacent database name in Global Query retrieves the results in that database.

The search box on the **NCBI homepage** also has a pull-down list that allows selection of any of the individual databases. Alternatively, searches can be launched from the individual Entrez database pages. Many of the database homepages are linked directly to the NCBI homepage from the Popular Resource box in the upper right or from the lists in the footer area. All Entrez homepages are linked from the **Resources list** on the NCBI homepage. It is also easy to access the database homepages directly using the simplified addresses that are formed by adding the database name to that of the NCBI homepage. For example, the address for the gene database homepage is simply [www.ncbi.nlm.nih.gov/gene](http://www.ncbi.nlm.nih.gov/gene). Searches launched from the database homepage allow for more precise search strategies tailored to the database. These can be constructed using Boolean operators and combinations of one or more search field limits as described below.

## Entrez Searching Options

Entrez queries can be single words, short phrases, sentences, database identifiers, gene symbols, or names ... just about anything. Often simple searches can result in overwhelming numbers of results or even no results at all. There are a number of built-in Entrez features that can help in creating more effective queries. These include

Search NCBI databases

all[filter] Search

Results found in 39 databases for "all[filter]"

Literature			Genes		
Books	509,092	books and reports	EST	76,176,725	expressed sequence tag sequences
MeSH	263,919	ontology used for PubMed indexing	Gene	23,647,902	collected information about gene loci
NLM Catalog	1,544,867	books, journals and more in the NLM Collections	GEO DataSets	1,895,235	functional genomics studies
PubMed	26,077,776	scientific & medical abstracts/citations	GEO Profiles	108,708,851	gene expression and molecular abundance profiles
PubMed Central	3,942,576	full-text journal articles	HomoloGene	141,268	homologous gene sets for selected organisms
Health			PopSet	250,954	sequence sets from phylogenetic and population studies
ClinVar	140,560	human variations of clinical significance	UniGene	6,473,284	clusters of expressed transcripts
dbGaP	216,812	genotype/phenotype interaction studies	Proteins		
GTR	34,338	genetic testing registry	Conserved Domains	50,648	conserved protein domains
MedGen	288,466	medical genetics literature and links	Protein	290,286,470	protein sequences
OMIM	24,437	online mendelian inheritance in man	Protein Clusters	820,546	sequence similarity-based protein clusters
PubMed Health	62,565	clinical effectiveness, disease and drug reports	Structure	118,410	experimentally-determined biomolecular structures
Genomes			Chemicals		
Assembly	85,786	genome assembly information	BioSystems	867,042	molecular pathways with links to genes, proteins and chemicals
BioProject	177,333	biological projects providing data to NCBI	PubChem BioAssay	1,218,630	bioactivity screening studies
BioSample	4,710,305	descriptions of biological source materials	PubChem Compound	89,162,175	chemical information with structures, information and links
Clone	37,486,015	genomic and cDNA clones	PubChem Substance	219,814,579	deposited substance and chemical information
dbVar	4,693,301	genome structural variation studies			
Genome	16,443	genome sequencing projects by organism			
GSS	39,518,073	genome survey sequences			
Nucleotide	204,712,631	DNA and RNA sequences			
Probe	32,392,659	sequence-based probes and primers			
SNP	774,407,218	short genetic variations			
SRA	2,560,356	high-throughput DNA and RNA sequence read archive			
Taxonomy	1,575,072	taxonomic classification and nomenclature catalog			

**Figure 1.** The Entrez Global Query results page showing the results of a search for all records in the databases (all[Filter]). The search was performed from the Search box on the NCBI homepage with the default All Databases selected.

Boolean operators, query translation, and fielded searching using any of the indexed fields available for the database. Any of these can be used in manually writing and editing queries but are also incorporated into various aspects of the interface so that precise results are available without the need to write complex query statements. These aspects of the interface include facets, and an Advanced Search page with a Search Builder and Search History that can be used to generate more sophisticated queries. More details on these features and some examples are given below.

## Using Boolean Operators

Boolean operators provide a way of generating precise queries that produce well-defined sets of results. The Boolean operators used in Entrez and how they work are as follows.

**AND:** Finds documents that contain terms on both sides of the operator terms, the intersection of both searches.

**OR:** Finds documents that contain either term, the union of both searches.

**NOT:** Finds documents that contain the term on the left but not the term on the right of the operator, the subtraction of the right hand search from the one on the left.

Entrez requires the Boolean operator AND to be entered in uppercase. This is not required in all databases for the other two operators, but it is simplest to enter all of them in uppercase:

```
promoters OR response elements NOT human AND mammals
```

Entrez processes all Boolean operators in a left-to-right sequence. Enclosing individual concepts in parentheses changes this priority. The terms inside the parentheses are processed first as a unit and then incorporated into the overall strategy. For example, in the following search statement, the union of response element and promoter results is generated first and then is intersected with the result of the g1p3 search.

```
g1p3 AND (response element OR promoter)
```

## Default Boolean Combinations and Phrase Searching

Individual search terms separated by spaces are normally automatically combined as if they were joined by AND operators. The query tp53 mouse always gives the intersection of a search for mouse and a search for tp53. Each Entrez database also has an indexed phrase list. If a multi-word search matches a phrase, then only the phrase is used. For instance, the query protein kinase c is treated as a complete phrase rather than as the intersection of the three terms. The phrase indices and behavior may be different for different databases. In some cases enclosing search terms in quotes can override the automatic intersection of terms and force a phrase search. The results for the phrase insulin dependent in most Entrez databases change depending on whether the phrase is in quotes or not. Although phrase searching is useful, it should be used with caution because enclosing search terms in quotes restricts the documents retrieved to only those documents with exact matches to the text string within the quotes. Quoting a phrase may also prevent automatic term mapping of the individual terms to controlled vocabularies such as Medical Subject Headings or Organism (Taxonomy).

## Indexed Fields, Query Translation, and Automatic Term Mapping

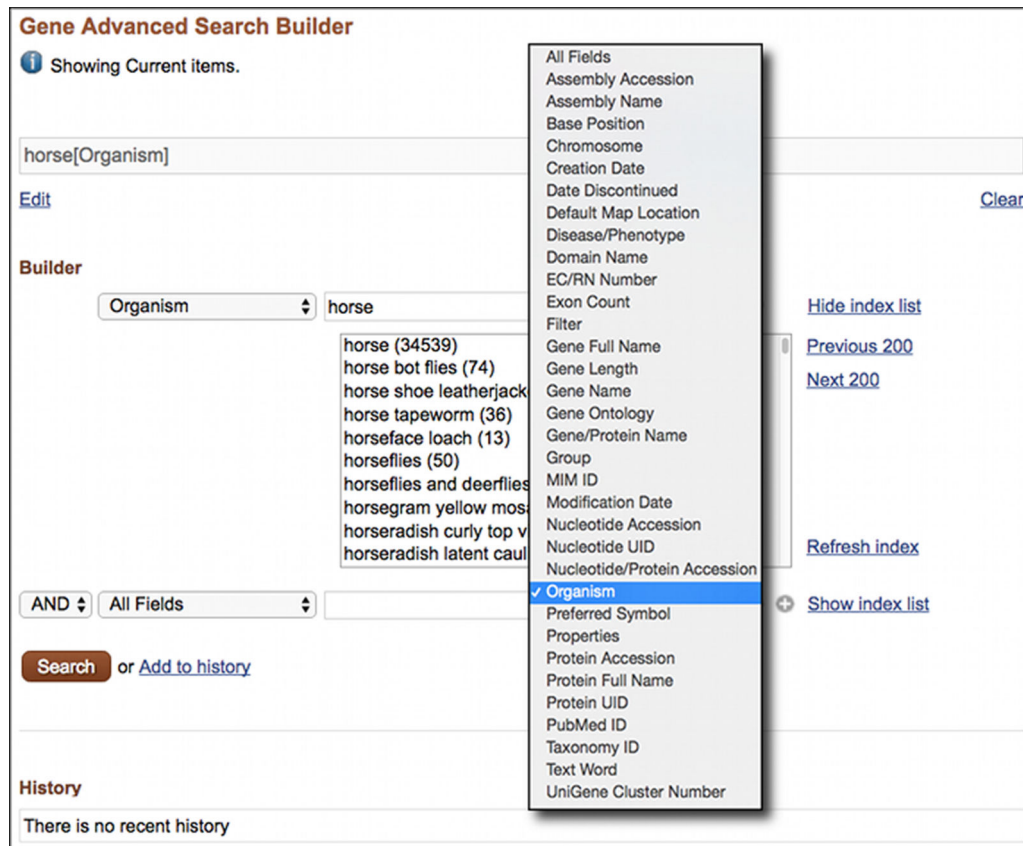
To facilitate searching, various indices are created for each Entrez database. These indices include information extracted from particular aspects of the record known as fields. Some of these fields contain essentially free text while other such as those for database identifiers (Accession, PMID), MeSH, and Organism are tightly controlled. Default searches in Entrez are All Fields searches. This usually results in the largest number of returned records but can produce unwanted results. For example, a search in any of the molecular databases with the term horse finds all records that contain this word in a variety of contexts, and many have nothing to do with that animal. If the goal is to find records specifically associated with that species, restricting the search to a particular field produces a more useful set of results. The available Fields and their indexed terms in any Entrez database can be explored on the Advanced Search page as part of the Search Builder. The Advanced Search interface, described in a separate section below, is linked beneath the search box of any page in an Entrez database as shown below for Nucleotide.



Nucleotide Nucleotide Frogs AND 2016/04[Publication Date] Search  
Create alert Advanced Help

After a search there is also a Create alert option above the Search box that allows saving a search strategy in a [My NCBI](#) account. My NCBI provides the ability to schedule saved searches to be run automatically. The [My NCBI help manual](#) has more information on Saved searches and the other features of My NCBI mentioned in this document.

The Entrez Gene Advanced Search Page is shown in Figure 2 with the index for the Organism field expanded. The term horse is in the Organism index for Gene. Selecting the Organism field before the search finds only Gene records for the horse (*Equus caballus*) while the default All Fields search finds records for other species as well. Field restricted searching can be performed using the Search Builder. Restricted searches may also be



**Figure 2.** The Entrez Gene Advanced Search page showing the Search Builder with the Index for the Organism Field expanded.

entered manually by following the term with the name of the field in square brackets “[ ]”, as shown in the following examples:

```
horse[Organism]
neoplasms[MeSH Terms]
prolactin[Protein Name]
srcdb_refseq[Properties]
2010/06[Publication Date]
```

## Dates and Other Ranges

Certain fields can accept ranges of values. Common examples are Publication Date, Modification Date, Accession, Molecular Weight, and Sequence Length. In these cases the low and high numbers of the range are entered with a colon “:” as the range operator between them followed by the field:

```
110:500[Sequence Length]
2015/3/1:2016/4/30[Publication Date]
```

## Facet Filters

Many of useful restrictions including some of the search terms described above can be applied to searches through the faceted filter links on the left-hand column of the Entrez search page. Figure 3 shows the a PubMed search page with several facets selected. Selecting any of the facet filters intersects the current search with the corresponding filter terms. Facet filters may restrict to certain types of records or exclude undesired ones.

PubMed.gov  
US National Library of Medicine  
National Institutes of Health

PubMed ebolavirus  
Create RSS Create alert Advanced

Article types clear Summary 20 per page Sort by Most Recent Send to: ▾  
Clinical Trial

Review  
Customize ...

Text availability clear  
Abstract

Free full text  
Full text

PubMed Commons  
Reader comments  
Trending articles

Publication dates clear  
5 years  
10 years

From 2015/01/01 to 2016/12/31

Species clear  
Humans  
Other Animals

Clear all  
Show additional filters

**Search results**  
Items: 18

Filters activated: Review, Free full text, Publication date from 2015/01/01 to 2016/12/31, Humans. [Clear all](#) to show 1910 items.

[A Review of the Role of Food and the Food System in the Transmission and Spread of Ebolavirus.](#)  
1. Mann E, Streng S, Bergeron J, Kircher A.  
PLoS Negl Trop Dis. 2015 Dec 3;9(12):e0004160. doi: 10.1371/journal.pntd.0004160. eCollection 2015 Dec. **Review.**  
PMID: 26633305 **Free PMC Article**  
[Similar articles](#)

[Ebola and blood transfusion: existing challenges and emerging opportunities.](#)  
2. de La Vega MA, Stein D, Kobinger GP.  
PLoS Pathog. 2015 Nov 12;11(11):e1005221. doi: 10.1371/journal.ppat.1005221. eCollection 2015. **Review.**  
PMID: 26562671 **Free PMC Article**  
[Similar articles](#)

[Ebola and blood transfusion: existing challenges and emerging opportunities.](#)  
3. Abdullah S, Karunamoorthi K.  
Eur Rev Med Pharmacol Sci. 2015 Aug;19(16):2983-96. **Review.**  
PMID: 26367717 **Free Article**  
[Similar articles](#)

**Figure 3.** A PubMed search results page with several facet filters selected: Review, Free full text, Publication date from 2015/01/01 to 2016/12/31, Humans. The results page reports in the line with the “I” icon the fact that the search was filtered with a link (Clear all) to clear the filters.

## Controlled Vocabulary Fields and Query Mapping

The indexed MeSH (Medical Subject Headings) and Organism fields have special roles in the PubMed and the traditional biomolecular databases respectively. Both MeSH and the Organism fields are tightly controlled vocabularies that are also hierarchical classification systems for the database records in PubMed and molecular databases. Every PubMed record is assigned sets of MeSH terms that add important information about the subject matter of the original paper. Records in the small molecule databases that are component PubChem are also associated with MeSH terms for the chemical components. The PubMed help document gives more details on the importance of the MeSH system. In a similar manner nearly all of the biomolecular database records are attached to the source organism and its phylogenetic classification in the [NCBI Taxonomy database](#). The [MeSH Browser](#) and the [Taxonomy Browser](#) are useful ways of exploring these systems and their associated records. Because of the importance of these two systems, queries are automatically mapped to these vocabularies whenever possible. The search terms may be expanded and translated as well. In certain databases other fields may involved in this mapping as well. The query horse dopamine receptor D2 becomes the more complex search statements in the PubMed and Protein Entrez search systems as shown in the box immediately below.

**PubMed:** ("horses"[MeSH Terms] OR "horses"[All Fields] OR "horse"[All Fields] OR "equidae"[MeSH Terms] OR "equidae"[All Fields]) AND ("receptors, dopamine d2"[MeSH Terms] OR ("receptors"[All Fields] AND "dopamine"[All Fields] AND "d2"[All Fields]) OR "dopamine d2 receptors"[All Fields] OR ("dopamine"[All Fields] AND "receptor"[All Fields] AND "d2"[All Fields]) OR "dopamine receptor d2"[All Fields])

**Protein:** ("Equus caballus"[Organism] OR horse[All Fields]) AND (dopamine

```
receptor D2[Protein Name] OR (dopamine[All Fields] AND receptor[All Fields]
AND D2[All Fields]))
```

## Special cases: Author names, Database Identifiers, and Stopwords

There are some other special cases of query interpretation in Entrez. Entering an author name in the form last name, first initials without punctuation, such as Lipman DJ, automatically maps to an author field search. Entering a recognized identifier for certain databases bypasses the general indices and directly retrieves the record. Identifiers that behave this way include accessions and gi numbers for sequence records, PubMed identifiers (PMIDs), and gene identifiers. Another special case is that certain words are ignored in Entrez searches. These words, known as stopwords, occur frequently in text on records but are not informative. Simple examples are definite and indefinite articles, conjunctions, and prepositions. A complete list of Entrez (PubMed) stopwords is given in the [PubMed Help](#) book. Punctuation in search terms is also typically ignored by Entrez and can cause certain strings to be missed. Enclosing problematic terms in quotes can help.

## Using Wild Cards or Query Truncation

Entrez allows searching with single word stems where the ending of the term is replaced by an asterisk “\*” to represent any character. This is often called truncation searching. For example, the search term hors\* in the Protein database finds records with the terms hors, hors4, horse, horse's, horseradish, horst, and many more. Truncation is supported in fielded as well as All Fields searches, and is helpful if the spelling of a word is uncertain. It can also help gather together ranges of identifiers. For example the following search statement in the Entrez Nucleotide database will find records for all human chromosomes:

```
NC_0000*[Accession] AND Human[Organism]
```

Because the truncation searches use the only first 600 variations of a search term indexed for a particular field, poorly determined terms, for example cat\* in PubMed, will give incomplete results.

## Search Details Shows Query Interpretation

In some cases it is useful to see how Entrez interpreted, expanded, or mapped the query as described above. This information is provided in the Search Details in the *Search details* box in the right-hand or Discovery column of search results. The Search details box below shows how the following search is interpreted in PubMed.

```
Smith T about the life of a horse
```

**Search details**

(Smith T[Author] OR Smith T[Investigator]) AND ("life"[MeSH Terms] OR "life"[All Fields]) AND ("horses"[MeSH Terms] OR "horses"[All Fields] OR "horse"[All Fields] OR "equidae"[MeSH Terms] OR "equidae"[All Fields])

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

The words “about”, “the”, “of” and “a” are stopwords ignored by Entrez; “Smith T” is mapped and expanded to an author search; and the terms “life” and “horses” are expanded and mapped to the MeSH vocabulary

**Protein Advanced Search Builder**

(("prolactin"[Protein Name]) AND frogs) AND srcdb refseq[Properties]

[Edit](#) [Clear](#)

**Builder**

Protein Name  [Hide index list](#)

- prolactin (641)
- prolactin 1 (19)
- prolactin 1 like (9)
- prolactin 2 (28)
- prolactin 2 like (2)
- prolactin 2 precursor (1)
- prolactin 2a1 (4)
- prolactin 2a1 like (9)
- prolactin 2a1 like protein (2)
- prolactin 2a1 precursor (2)

[Previous 200](#)  
[Next 200](#)  
[Refresh index](#)

AND   [Show index list](#)

AND   [Show index list](#)

AND   [Show index list](#)

or [Add to history](#)

---

**History** [Download history](#) [Clear history](#)

Search	Add to builder	Query	Items found	Time
#30	<a href="#">Add</a>	Search ("prolactin"[Protein Name]) AND frogs) AND srcdb refseq[Properties]	1	16:18:17
#29	<a href="#">Add</a>	Search prolactin	9897	16:14:08
#28	<a href="#">Add</a>	Search frogs	196522	16:13:49
#27	<a href="#">Add</a>	Search creatine kinase	3273	16:13:33
#26	<a href="#">Add</a>	Search srcdb refseq[Properties]	64565536	16:13:10

**Figure 4.** The Protein Advanced Search interface showing the Search Builder and Search History. 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 #28 and #26 in the Search History with the protein name search for prolactin in the Search Builder finds the *Xenopus tropicalis* RefSeq protein for prolactin.

## Using the Advanced Search Page to Construct Complex Search Statements

The Advanced Search page for each Entrez database is useful for constructing complex and highly precise queries. Figure 4 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 in combination with Search History facilitates 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 Box. These may be run using the *Search* button or may be added to the Search History using the *Preview* button.

The Search History is maintained separately for each Entrez database and keeps track of all searches until the Web browser is closed or the history is deleted. Histories are automatically deleted after eight hours of inactivity. Entries in the Search History may be combined to create new searches that give precise results. The example in



Figure 4 combines searches for frogs (#28), RefSeq proteins (#26), with a protein name search for prolactin to obtain the prolactin protein record for *Xenopus tropicalis*, NP\_001093699.

## Displaying and Saving a Set of Records

The *Display Settings* and *Send to* menus at the upper left and upper right of Entrez pages manage how records are displayed and stored or downloaded. The *Display Settings* menus have options for format, number of results per page, and sorting order. The available formats and sorting options vary depending on the database. The default format for multiple search results in Entrez is the Summary format that is consistent across databases. Single record default formats depend on the database. The default number of records displayed is 20 per page presented in the default sorting order for the database. These default settings may be modified by setting personal Preferences in a My NCBI account as described in the [My NCBI Help](#) book.

The *Send to* menu has options for sending results to online storage in Collections in My NCBI, the NCBI Clipboard for the database, or to a local file. Additional options may be available depending on the database. When choosing the file option, the record format and sorting order can be specified. By default all *Display Settings* and *Send to* menu operations affect all records unless individual items are selected using the checkboxes at the left of the record title.

## About the Clipboard and My Collections

The Clipboard is a temporary place on the NCBI website to save records. Each Entrez database has its own independent Clipboard that is limited to 500 items. Items saved to the clipboard are lost after eight hours of inactivity. When there are items in the clipboard, a link to access the clipboard appears in the upper right of any Entrez page for that database. The Clipboard behaves in the same way as any other page view in the database with equivalent *Display Settings*, *Send to* menus, and other features of that database. Records can be removed from the Clipboard through the link next to each item or by selecting the items using the check box and clicking the Remove Selected Items link at the top of the page. This link also functions to clear the Clipboard when no records are selected.

My Collections that is a part of the My NCBI service is a more permanent place to save records.

## Related data: Neighbors and Links

One of the most useful and powerful features of the Entrez system is the integration of the data so that relationships between records can easily be explored. Many of these relationships may be unanticipated making the Entrez system an engine for scientific discovery. There are two major kinds of relationships established in the Entrez system: computationally derived associations within a database – items connected in this way are often called neighbors, and relationships based on information present on the records themselves, sometimes called hard links.

Neighbors are established automatically for many Entrez databases by computing on the data in the records. For example, related sequences are identified through similarity searches with the BLAST algorithm; related structures are determined using Vector Alignment Search Tool (VAST), a structure similarity algorithm; related PubMed citations are determined by an algorithm that compares information rich words and phrases in the abstracts.

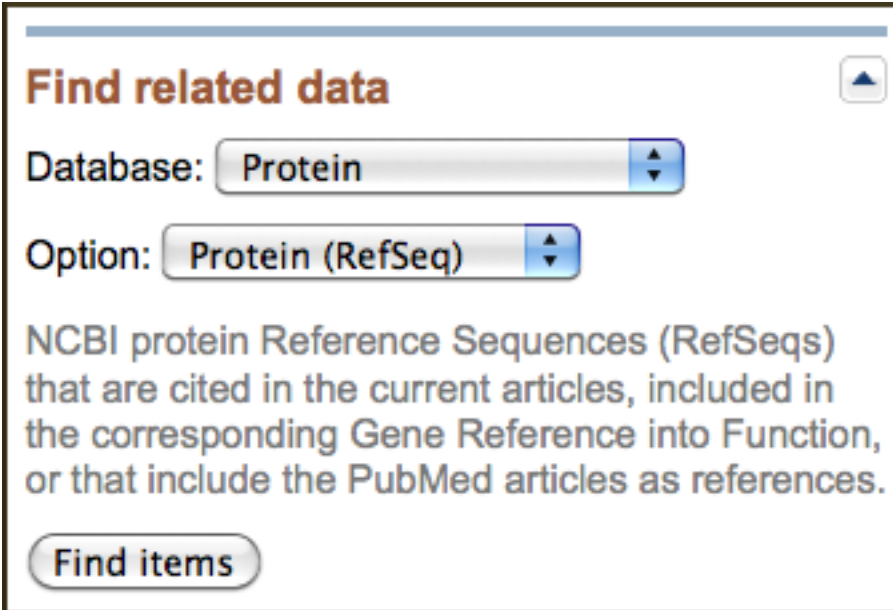
Reciprocal links between databases are established from the records themselves. For example, molecular records such as sequences or structures are linked to the PubMed citations where these data were reported. The literature citation has the reciprocal links back to the molecular databases. PubMed citations are linked to the full-text article in PubMed Central. Databases such as UniGene or Gene that gather together many records from different sources have links back to their source records.

Combining neighbors and hard links can be an especially effective method for navigating across data and finding the most useful information. Moreover, this can be accomplished without performing additional searches. For example going from an mRNA sequence in the nucleotide database to the three-dimensional structure for the corresponding or closely related protein can be a simple matter of following the link to protein, linking to related proteins, then using the Find related data component (described below) to link to the structure database from these similar proteins. A separate *Related Structure* link present on many protein records can make this even faster.

## Access to Related Data

### The Discovery Column

Easy access to related data is available on the individual records through the Links menu, the list of hypertext links available in the *All links from this record* section in the right-hand column of the full-record display. This right-hand column known as the Discovery Column contains additional access to related data and, in some cases, analysis tools. Many components in the Discovery column function as advertisements for related data providing more details about what information is available. The Discovery Column also functions in the search results providing additional options for searching and navigating including alternative search strategies as well as the *Search details* and *Recent activity* components described in the Entrez Searching section of this document. The Discovery Column on search results and other multiple record displays also provides access to related data through the *Find related data* component shown immediately below set for PubMed to Protein RefSeq. This follows the corresponding link for all records in the display when the *Find items* button is clicked.



**Find related data**

Database: Protein

Option: Protein (RefSeq)

NCBI protein Reference Sequences (RefSeqs) that are cited in the current articles, included in the corresponding Gene Reference into Function, or that include the PubMed articles as references.


Find items

In some databases, certain links to related data are also available from the individual summaries in search results, for example Similar Article and Free article in PubMed and Identical Proteins in the Protein databases. Optionally, a preferences setting in My NCBI allows all links to be displayed in the summaries.


### Recent Activity


An additional component that appears at the bottom of the Discovery Column is the Recent Activity list. While the Search History described previously in this document as part of the Advanced Search page, can be used to navigate to previous search results in the same database, it does not allow cross-database access. The Recent Activity component, present on all Entrez searches and record views, provides navigation to searches and record views in other databases. Only links to the last five searches and record views are listed by default, but activity for


the past six months is available through a [My NCBI](#) account, where these searches can be saved and records can be added to collections.


**Recent activity** 


[Turn Off](#) [Clear](#)

 [tp53 AND human\[organism\] \(423\)](#) Gene

 [Lipman DJ \(60\)](#) PubMed

 [kinase \(314243\)](#) Nucleotide

 [DNA mismatch repair protein Mlh1 isoform 1 \[Homo sapiens\]](#) Protein

 [A study of somatolactin actions by ectopic expression in transgenic zebrafish la...](#) PubMed

[See more...](#)

## Creating Links to Web Pages in the Entrez System

The Entrez system now uses a standardized structure for web page addresses (URLs) that makes it easy to construct HTML links to address page displays and perform single searches. The standard URL format consists of the base URL for the database followed by options that can specify the record to be displayed, display options, and search terms. The Entrez Programming Utilities ([E-Utilities](#)) described in the next section of this document should be used to send frequent queries or retrieve large numbers of records from Entrez.

### Database Homepages, Advanced Search, Limits

The base URL alone retrieves the homepage for the resource. The Advanced Search and Limits pages may also be addressed directly.

#### Examples:

Nucleotide homepage: [www.ncbi.nlm.nih.gov/nucleotide](http://www.ncbi.nlm.nih.gov/nucleotide)

PubMed homepage: [www.ncbi.nlm.nih.gov/pubmed](http://www.ncbi.nlm.nih.gov/pubmed)

Gene Advanced Search: [www.ncbi.nlm.nih.gov/gene/advanced](http://www.ncbi.nlm.nih.gov/gene/advanced)

## Linking to Records in the Entrez System

### Linking Directly to Records Using an Identifier

The base URL for the database followed by a valid unique identifier for the database retrieves the record in the default single record view format. The summary format is displayed by default for more than one record. Most valid unique identifiers are purely numeric such as PubMed IDs, gene ids, and gi numbers for sequences. For the sequence databases accession numbers or accession.version numbers may also be used to address specific records. Click here for a list of [Accession Number Prefixes](#) for nucleotide and protein records. The report and format display options may be specified following the identifiers. The options available for report vary depending on the database and are listed on the Display settings menu for record views in that database. The [NCBI Help Manual](#) sections on individual Entrez databases have more information on available report formats. The format option may be used to retrieve text rather than the default HTML format

#### Examples:

Protein gi 4557757, GenPept format (default)

[www.ncbi.nlm.nih.gov/protein/4557757](http://www.ncbi.nlm.nih.gov/protein/4557757)

Nucleotide accessions NM\_000240 and NM\_000041 in GenBank format

[www.ncbi.nlm.nih.gov/nucleotide/NM\\_000240,NM\\_000041&report=genbank](http://www.ncbi.nlm.nih.gov/nucleotide/NM_000240,NM_000041&report=genbank)

Gene ID 348 full report (default)

[www.ncbi.nlm.nih.gov/gene/348](http://www.ncbi.nlm.nih.gov/gene/348)

Gene ID 348 in XML format

[www.ncbi.nlm.nih.gov/gene/348?report=XML](http://www.ncbi.nlm.nih.gov/gene/348?report=XML)

PubMed IDs 9705509 and 19745054 in abstract format, text

[www.ncbi.nlm.nih.gov/pubmed/9705509,19745054?report=abstract&format=text](http://www.ncbi.nlm.nih.gov/pubmed/9705509,19745054?report=abstract&format=text)

### Linking to Search Results

Links to search results in Entrez databases may be created by adding a term to the URL. Report options and the maximum number of records to display per page may also be specified.

#### Examples:

Search in nucleotide for APOE with gene field restriction and 200 records displayed

[www.ncbi.nlm.nih.gov/nucleotide/?term=APOE\[gene\]&dispmax=200](http://www.ncbi.nlm.nih.gov/nucleotide/?term=APOE[gene]&dispmax=200)

Search in PubMed for Lipman DJ with PMID display format

[www.ncbi.nlm.nih.gov/pubmed/?term=Lipman+DJ&report=ulist](http://www.ncbi.nlm.nih.gov/pubmed/?term=Lipman+DJ&report=ulist)

## Programmatic Access to the Entrez System

The Entrez system may be accessed programmatically for high volume non-interactive search and retrieval through the Entrez Programming Utilities (E-utilities). These are a set of eight server-side programs that provide

a stable interface to the Entrez query and database system. More information E-utilities is available in the [Entrez Programming Utilities Help](#) manual on the NCBI Bookshelf.