

Bioinformatics: Life Sciences on Your Computer

NCBI Relational Database

This is supplemental reading to the video on the NCBI Sequence Database.

Entrez

Entrez is the interface, or portal, to NCBI databases. It utilizes an index-based search system. Just like an index of a book has page numbers for terms, an index can be generated using GI numbers.

Exercise: Go to www.ncbi.nlm.nih.gov. Enter 17865153 into the search window of the Nucleotide database. View the record in GenBank (default) format. Then view in FASTA format and finally in graphics format.

Typical weekday traffic

NCBI has several million accesses daily.

Total IP addresses: 500,000

Total users: ~1.5M

Total hits: 75M

Total page views: 24M

Total bytes: 2.2 Terabytes

Entrez records

Each record has a unique ID (UID). The primary key is GI number for sequences. The GI number is an integer unique to that database. Each record has Document Summary (DocSum). Each record has links to biologically relevant UIDs. Each record is indexed by data fields, including [author], [title], [organism] and others.

GenBank divisions

There are two major databases of GenBank. The Traditional database includes direct submissions, accurate to better than ~1 error per 10,000 base pairs. They are well-characterized, but not necessarily curated, and organized by taxonomy. It includes primary, secondary databases from America, Europe and Japan. At one time, the Traditional database was called Core Nucleotide, but since many found that confusing, it is now just Nucleotide.

Discussion topic: What do these tags stand for? PRI, ROD, PLN, BCT, VRT, INV, VRL, MAM, SYN, PHG, UNA. What about the Bulk tags? EST, GSS, HTG, CON, PAT, STS, HTC, ENV.

Searching summary

Web-based databases rely on searches. Those searches are usually index-based, indexing all primary keys associated with that subject. Links represent primary keys. When searching a sequence database, most PubMed rules apply. AND, NOT, OR are Boolean operators that should always be capitalized.

All fields searched unless the search is limited to a specific field. By using brackets or by using the limits feature, we will carefully examine how to narrow searches to specific fields.

Nucleotide/protein fields

[ACCN] – Accession number
[SLEN] – Sequence length
[MOLWT] – Molecular weight (mass) of a protein
[MDAT] – Modification date
[PDAT] – Publication Date
[ORGN] – Organism
[GENE] – Gene name
[PROT] – Protein name
[WORD] – Text word
[TITLE] – Word in descriptive title of record

There is an complete table online at <http://www.ncbi.nlm.nih.gov/books/NBK49540/>

Phrase search

Double quotes around phrase will narrow search.

Exercise/Discussion topic: In the Nucleotide database, type in:

16S
RNA
16S RNA
"16S RNA"
16S AND RNA

What is the preferred phrase for 16S and RNA?

Range searches

The colon used in five searches to limit to range of values: [ACCN] [SLEN] [MOLWT] [MDAT] [PDAT]

Date searches use YYYY/MM/DD:YYYY/MM/DD

1995/02:2013/06/08 [MDAT] would get you February 1, 1995 to June 8, 2013.

Sequence length [SLEN]

Sequence length [SLEN] is the number of amino acids in the sequence in the protein database. 200:300 [SLEN] will hit sequences between 200 and 300 amino acids.

However, in the nucleotide sequence, [SLEN] refers to the number of base pairs in the sequence. 130000:140000 [SLEN] returns sequences between 130,000 and 140,000 base pairs.

Molecular Weight [MOLWT]

Molecular Weight field can be queried in the protein database as a single molecular weight (measured in daltons):

2002 [MOLWT]

Or range of weights:

2002:2009 [MOLWT]

It can be combined with other Entrez search terms:

2002:2009 [MOLWT] AND human [ORGN]

Organism [ORGN]

Even if you might not think so, the [ORGN] tag is the most important tag in sequence database searching.

Exercise/Discussion topic: Search the protein database:

Search Protein database

2002:2009 [MOLWT] AND human

2002:2009 [MOLWT] AND human [orgn]

Why are the results different? Where might human be found in the non-human sequences?

Revision history

Under Display Settings, one of the options is Revision History. When you access the Revision History, you will see sequence revisions and dates of the Accession Number. Remember that updates with sequence changes are assigned a new GI and updates without sequence changes keep the GI.

Discussion topic: Find the dates of all sequence changes of accession number NM_005806. Find what was changed in the sequence updates and the last few non-sequence revisions.

Structure database

[Molecular Modeling Database \(MMDB\)](#) at NCBI is derived from the Protein Data Bank (PDB). It's searchable like the sequence databases and most of the tags are the same. Neighbors are found with VAST and structures can be viewed with NCBI's structure-sequence viewer, Cn3D, a free download.

Genome database

The [NCBI Genome database](#) contains genome information from over 1000 sequenced organisms or viruses. They are listed in [table form](#). The prokaryotic table includes only archaea and bacteria. As of May 19, 2014, 167 archaea and 2,736 bacteria are completely sequenced with no gaps.

Discussion topic: What are archaea? Find some examples.

OMIM database

Online Mendelian Inheritance in Man ([OMIM](#)) is a Johns Hopkins database in the NCBI suite. It grew out of a catalog of human genes, genetic disorders, edited by [Victor McKusick](#) and others who mapped human genes for decades. It's very useful to physicians and other professionals concerned with genetic disorders. With all of its references, it would be really useful to a high school student writing a term paper.

The database contains information on allelic variants and SNPs and is very informative.

Taxonomy database

In the [Taxonomy database](#), TaxBrowser, taxid is a stable primary key that can be more stable than the name of the organism! Reclassifications and renamings of organisms are relatively common. There are entire journals dedicated to taxonomic revision. The Taxonomy database includes extinct organisms. You can enter a common name (human) or a Latin name (Homo sapiens).

RefSeq database

Try the following search in the Nucleotide database:

"thyroid peroxidase" [prot] AND human [orgn] AND biomol mrna [prop]
You should get seven hits.

Now, use the filter links on the right. Click RefSeq.

The RefSeq database is a secondary (curated) database that aims to be non-redundant. The goal is to get the best sequence for each transcript, linked to the protein product that is translated from that transcript. Curated links to nucleotide and protein are updated to reflect the current known sequence. Proteins and transcripts are validated by hand, and there is a distinct accession series. Notably, NM represents mRNA and NP represents proteins in a consistent format.

RefSeq accessions

Here is the naming convention for RefSeq accession numbers:

NM_123456 (mRNA) --> NP_123456 (protein)
NR_123456 (noncoding RNA)
Model transcripts, proteins (predicted)
XM_123456 (mRNA) --> XP_123456 (protein)
XR_123456 (noncoding RNA)
Assembled genomic regions (contigs)
NT_123456 (BAC clones)
NW_123456 (whole genome shotgun)
NC_123456 (complex regions, pseudogenes)

RefSeq example

Human myoglobin has 3 transcripts through alternative splicing:

[NM_005368.2](#) has 1,078 base pairs. Its protein product is [NP_005359.1](#) with 154 amino acids.

[NM_203377.1](#) has 1,170 base pairs. Its protein product is [NP_976311.1](#) with 154 amino acids.

[NM_203378.1](#) has 1,153 base pairs. Its protein product is [NP_976312.1](#) with 154 amino acids.

All three of the proteins are identical. The transcripts differ in untranslated regions only--the coding regions are the same.

Discussion topic: Why make three different mRNAs that all produce the same protein??

Entrez Gene (Gene) database

The [NCBI Gene database](#) is essentially the outlet for RefSeq data. It is a great starting point for many searches with extensive links to other databases displayed prominently on the right menu. Splice variants, gene expression, oligos for microarrays are also part of the Gene record. Try finding human myoglobin and the variants listed in the RefSeq example section.

GEO database

The [Gene Expression Omnibus database](#) links to experimental data on gene expression, predominantly expression experiments based on microarrays. Microarrays will be discussed at the end of the course.

SNP database

The [Single Nucleotide Polymorphism \(SNP\) database](#) contains data on single nucleotide polymorphisms within populations. A SNP is defined as a genetic variant in which at least 1% of the population has the

variant allele. DNA markers and DNA fingerprinting are based on SNPs, which are important in police work, prenatal testing particularly in epidemiology.

LinkOut

LinkOut is an underutilized feature at NCBI but it's getting better, particularly in the Taxonomy database. It allows outside publishers to display links on a record. The idea is to include relevant outside links to researchers, web pages, full-text publications, biological databases, consumer health information and research tools.