Retrieving data from NCBI with E Utilities

Edit | New Page

Jump to bottom

AmyStonelake edited this page 2 hours ago · 4 revisions

Always remember to start the bioinformatics environment when working on Biostar class material.

```
conda activate bioinfo
```

Let's start by creating a directory for class data (if you do not already have one.)

```
mkdir biostar_class
```

Now let's go to that directory and create a directory to use for the GenBank data we will be retrieving.

```
cd biostar_class
mkdir genbank
cd genbank
```

Try an "ls" and "pwd" to see what's in the directory, and also the "path" to where you are.

```
ls
pwd
```

Okay, here is the command line we'll be working with. In this document, we will look at each of the parts of the command, and then run the command.

```
efetch -db nuccore -id NC_001501 -format gb > NC_001501.gb
```

EFetch is one of NCBI's "E-Utilities" that allows access to NCBI databases from the command line. Each utility (EInfo, ESearch, EPost, EFetch, ELink, EGQuery, ESpell, ECitMatch) has required parameters. They are the gateway to the "*Entrez system, which currently includes 38 databases covering a variety of biomedical data, including nucleotide and protein sequences, gene records, three-dimensional molecular structures, and the biomedical literature*".

Access to "EFetch" is provided by installation of the "Entrez-Direct" tools, which happened during the computer set-up phase of the course.

Your command line so far...

```
efetch
```

If you type "efetch" at the command line, you may get an error message. To find out more about how to use "efetch", you can look at the help documentation.

```
efetch --help
```

A required parameter for EFetch is "db", the database from which to retrieve records. The "nuccore" database is the "nucleotide" database.

Your command line so far...

```
efetch -db nuccore
```

Table 1

– Entrez Unique Identifiers (UIDs) for selected databases

Entrez Database	UID common name	E-utility Database Name
BioProject	BioProject ID	bioproject
BioSample	BioSample ID	biosample
Biosystems	BSID	biosystems
Books	Book ID	books
Conserved Domains	PSSM-ID	cdd
dbGaP	dbGaP ID	gap
dbVar	dbVar ID	dbvar
Epigenomics	Epigenomics ID	epigenomics
EST	GI number	nucest
Gene	Gene ID	gene
Genome	Genome ID	genome
GEO Datasets	GDS ID	gds
GEO Profiles	GEO ID	geoprofiles
GSS	GI number	nucgss
HomoloGene	HomoloGene ID	homologene

MeSH	MeSH ID	mesh
NCBI C++ Toolkit	Toolkit ID	toolkit
NCBI Web Site	Web Site ID	ncbisearch
NLM Catalog	NLM Catalog ID	nlmcatalog
Nucleotide	GI number	nuccore
OMIA	OMIA ID	omia
PopSet	PopSet ID	popset
Probe	Probe ID	probe
Protein	GI number	protein
Protein Clusters	Protein Cluster ID	proteinclusters
PubChem BioAssay	AID	pcassay

Each database entry has a UID, or "unique identifier" (-id). This is the second parameter that must be specified for the EFetch command.

```
-id NC_001501
```

```
-id NC_001501,NC_002549,NC_045512
```

In this example, the UID is "NC_001501", a NCBI RefSeq entry for "Moloney murine leukemia virus, complete genome".

Your command line so far...

```
efetch -db nuccore -id NC_001501
```

Sidebar: What is RefSeq?

RefSeq is the NCBI Reference Sequence database, "a comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein".

Moving on...what does "-format gb" mean? You can set the format of the returned information. Here, we are interested in GenBank format.

db = nlmcatalog		
Full record	*null*	text, *default*
XML	*null*	xml
db = nuccore, nucest, nucgss, protein or popset		
text ASN.1	*null*	text, *default*
binary ASN.1	*null*	asn.1
Full record in XML	native	xml
Accession number(s)	acc	text
FASTA	fasta	text
TinySeq XML	fasta	xml
SeqID string	seqid	text
Additional options for db = nuccore, nucest, nucgss or popset		
GenBank flat file	gb	text
GBSeq XML	gb	xml
INSDSeq XML	gbc	xml
Additional option for db = nuccore and protein		
Feature table	ft	text
Additional option for db = nuccore		
GenBank flat file with full sequence (contigs)	gbwithparts	text
CDS nucleotide FASTA	fasta_cds_na	text
CDS protein FASTA	fasta_cds_aa	text
Additional option for db = nucest		
EST report	est	text

Table 1 from NBK25499, a full list of allowed values for each database

For GenBank format,

```
-format gb
```

Your command line so far...

```
efetch -db nuccore -id NC_001501 -format gb
```

We're getting closer to the final version of the command line.

```
efetch -db nuccore -id NC_001501 -format gb > NC_001501.gb
```

So let's look at the last part of the command line, the output...

```
> NC_001501.gb
```

The right-facing caret ">", creates the output file, which is named "NC_001501.gb". You could name this file whatever you want, like this...

```
> file_whatever_you_want
```

But, it's better to choose a useful name that tells you something about the data. In this case, it is RefSeq NC_001501, in GenBank format. This is the only part of this command line where you can type in something other than what the example shows and it would still "work".

Your finished command line.

```
efetch -db nuccore -id NC_001501 -format gb > NC_001501.gb
```

Questions

1. How would you modify this command line to retrieve from a different GenBank database (not nucleotide)? Which part of the command line would you change?
2. Do "NC_001501", "NC001501" and "NC_oo15o1" all refer to the same GenBank entry? How would you test this? Why does/doesn't it work?
3. In what other formats can you retrieve data?

```
efetch -db nuccore -id NC_001501 -format fasta > NC_001501.fa
```

Clone this wiki locally

https://github.com/AmyStonelake/BTEP.wiki.git