How to Download Bacterial Genomes Using the Entrez API

Posted on **February 19, 2013**

Given the size of modern sequence databases, finding the complete genome sequence for a bacterium among the many other partial sequences can be a challenge. In addition, if you want to download sequences for many bacterial species, an automated solution might be preferable.

In this post we'll discuss how to download bacterial genomes programmatically for a list of species using the E-utilities, the application programming interface (API) to NCBI's Entrez system of databases. We'll also take advantage of NCBI's redesigned Genome database, which links all genome sequences for a given species to one record, making it easy to obtain the desired sequences once you find the right Genome record. In principle you can apply the procedure below to other simple genomes that are represented by a single sequence. Future posts will address additional considerations that apply to complex, eukaryotic genomes.

You'll find that several types of genome sequences are linked to a Genome record. There may be complete chromosomes and/or plasmids along with whole genome shotgun (WGS) sequences. There may be NCBI Reference Sequences (RefSeqs) and original submissions to GenBank. You can limit your download to any combination of these subsets, as you'll see below.

(Note: In this post, long e-utilities calls and lines of code are sometimes wrapped onto the next line for readability. This break is indicated by a backslash, '\', at the end of the line. Of course, these wrapped lines should be on one line when you use them.)

Procedure

1. Use esearch.fcgi to find the Genome record, using the bacterial species name as the query.

```
esearch.fcgi?db=genome&term=<species name>
```

[Parse out genome ID from XML output]

2. Use elink.fcgi to find the desired Nucleotide records linked to the Genome record.

```
elink.fcgi?dbfrom=genome&db=nuccore&id=<genome ID>&term=<sequence
&cmd=neighbor_history
```

[Parse out <query_key> and <WebEnv>]

3. Use efetch.fcgi to download the Nucleotide records in one of several formats.

```
efetch.fcgi?db=nuccore&query_key=<query_key>&WebEnv=<WebEnv>\
&rettype=<record type>&retmode=<record format>
```

Alternative for step 2.

If you remove the "&cmd" parameter from step 2, elink will return the nucleotide GI numbers for the linked sequences rather than a query_key and WebEnv. You will then need to parse each of the GI numbers from the XML output and pass them to efetch in step 3 using the "&id" parameter.

Now let's look at some tricks. Here are the Entrez search terms for <sequence type> in Step 2:

Desired Sequence	**&term value**
Completed chromosomes	gene+in+chromosome[prop]
Plasmids	gene+in+plasmid[prop]
RefSeqs	srcdb+refseq[prop]
INSDC (DDBJ, EMBL-Bank, GenBank)	srcdb+ddbj/embl/genbank[prop]
WGS	wgs[prop]
Other genomic sequences	gene+in+genomic[prop]

You can combine these with Boolean operators to retrieve, for example, all RefSeq genomic sequences:

```
&term=(gene+in+chromosome[prop]+OR+gene+in+genomic[prop])\
+AND+srcdb+refseq[prop]
```

Please see Table 1 in Chapter 4 of the efetch documentation for available values of &rettype and &retmode that will generate the format you want, such as FASTA, GenBank flat file, feature table or XML.

Example

For this example our goal will be to explore the genome data available for *Corynebacterium efficiens.*

1. esearch.fcgi?db=genome&term=corynebacterium+efficiens

   This call returns the genome ID 1076.

2. elink.fcgi?dbfrom=genome&db=nuccore&id=1076

   The results of the elink call reveal a total of eight sequences (at the time of writing). By using a series of the "&term" values listed in Table 1, you'll see that both RefSeq and WGS sequences are available. In this case we are using the alternative approach to step 2 above that does not use the "&cmd" parameter in the elink request. You might decide, for instance, to download the RefSeq sequence for the chromosome in FASTA format. As long as you have included the appropriate "&term" value in the elink call, the final step below will accomplish this.

3. efetch.fcgi?db=nuccore&id=25026556&rettype=fasta&retmode=text

Next steps

Now that you've seen the basic method, it's a relatively straightforward extension to produce a script that can read a file of species names and make the set of calls for each one. In this way you can download data for the entire set. We've included below a sample Perl script that downloads RefSeq chromosome sequences in FASTA format for a list of species provided as an array in the code. You can easily modify this script to, for example, read in species names from a file.

```
use strict;
use LWP::Simple;
my ($name, $outname, $url, $xml, $out, $count, $query_key,\
 $webenv, $ids);
my @genomeId;
my $base = 'http://eutils.ncbi.nlm.nih.gov/entrez/eutils/';
my $limit = 'srcdb+refseq[prop]+AND+gene+in+chromosome[prop])';
```

```
my @species = ('Corynebacterium efficiens',\
 'Acidimicrobium ferrooxidans', 'Fluviicola taffensis');

foreach my $s (@species) {
undef @genomeId;
$query_key = $webenv = '';
$s =~ s/ /+/g;
# ESearch
$url = $base . "esearch.fcgi?db=genome&term=$s";
$xml = get($url);
$count = $1 if ($xml =~ /<Count>(\d+)<\/Count>/);
if ($count > 20) {
$url = $base . "esearch.fcgi?db=genome&term=$s&retmax=$count";
$xml = get($url);
}
while ($xml =~ /<Id>(\d+?)<\/Id>/gs) {
push(@genomeId, $1);
}
$ids = join(',', @genomeId);
# ELink
$url = $base . "elink.fcgi?dbfrom=genome&db=nuccore\
&cmd=neighbor_history&id=$ids&term=$limit";
$xml = get($url);
$query_key = $1 if ($xml =~ /<QueryKey>(\d+)<\/QueryKey>/);
$webenv = $1 if ($xml =~ /<WebEnv>(\S+)<\/WebEnv>/);
# EFetch
$url = $base . "efetch.fcgi?db=nuccore&query_key=$query_key\
&WebEnv=$webenv&rettype=fasta&retmode=text";
$out = get($url);
open (OUT, ">$s.fna");
print OUT $out;
close OUT;
}
```

For more information:

- Genome database
- ESearch documentation
- ELink documentation
- EFetch documentation
- E-utilities documentation

Rate this: 2 Votes

Share this post:

Like this: ★Like

Be the first to like this.

This entry was posted in **Quick Tips** and tagged **API**, **e-utilities**, **genome**, **nucleotide**, **prokaryotes**, **sample script** by **ncbiinsights**. Bookmark the **permalink [http://ncbiinsights.ncbi.nlm.nih.gov/2013/02/19/how-to-download-bacterial-genomes-using-the-entrez-api/]** .