Question: Fetching genbank entries for list of accession numbers.

I have a looong list of accession numbers, for which I need to fetch genbank entries. Usually, I used to use epost-efetch workflow for long lists. But now I can not, because epost doesn't accept accession numbers as IDs.

So the question is:

a) is there a way of batch downloading using accession numbers? if not b) is there a way of mapping Accession number to normal ids?

ncbi **entrez** **efetch** **accession**

created 10 months ago by Sanjarbek Hudaiberdiev **10** • 2

updated 7 months ago by Sanjarbek Hudaiberdiev

can you give some example of your accession number? where is it from?

log in to reply • written 10 months ago by Leszek ♦ **2,780** • 5 • 19

Ex: A22237,A22239,A32021,A32022,A33397 Those are accessions from NCBI. When you post them using epost, it gives this error: "IDs contain invalid characters which was treated as delimiters." So it appears to me that epost doesn't accept non-numeric characters for ID field. I tried to change the letters to their ascii codes, didn't help.

log in to reply • written 10 months ago by Sanjarbek Hudaiberdiev **10** • 2

One more thing that I noticed today: All the BioXXX libraries just stop when they get error from epost. But I noticed that along with error, epost still returns WebEnv and query_key. But what it does is that, it takes the accession number, trims out the non-numeric characters, and searches for resultant GID. So, A22237 turns to 22237. Don't know what to do. Such a tiny problem taking up lot's of time.

log in to reply • written 10 months ago by Sanjarbek Hudaiberdiev **10** • 2

6 answers

You can try this:

```
#!/usr/bin/env python
"""Fetch GenBank entries for given accessions.

USAGE:
  python acc2gb.py A22237 A22239 A32021 A32022 A33397 > out.gb
or
```

created 10 months ago by

```
  cat ids | python acc2gb.py > out.gb

DEPENDENCIES:
Biopython
"""

import sys
from Bio import Entrez

#define email for entrez login
db           = "nuccore"
Entrez.email = "some_email@somedomain.com"

#load accessions from arguments
if len(sys.argv[1:]) > 1:
  accs = sys.argv[1:]
else: #load accesions from stdin
  accs = [ l.strip() for l in sys.stdin if l.strip() ]
#fetch
sys.stderr.write( "Fetching %s entries from GenBank: %s\n" % (len(accs)
, ", ".join(accs[:10])))
for i,acc in enumerate(accs):
  try:
    sys.stderr.write( " %9i %s          \r" % (i+1,acc))
    handle = Entrez.efetch(db=db, rettype="gb", id=acc)
    #print output to stdout
    sys.stdout.write(handle.read())
  except:
    sys.stderr.write( "Error! Cannot fetch: %s        \n" % acc)
```

EDIT
The same using epost:

```
import sys
from Bio import Entrez

#define email for entrez login
db           = "nuccore"
Entrez.email = "some_email@somedomain.com"
batchSize    = 100
retmax       = 10**9

#load accessions from arguments
if len(sys.argv[1:]) > 1:
  accs = sys.argv[1:]
else: #load accesions from stdin
  accs = [ l.strip() for l in sys.stdin if l.strip() ]
#first get GI for query accesions
sys.stderr.write( "Fetching %s entries from GenBank: %s\n" % (len(accs)
, ", ".join(accs[:10])))
query  = " ".join(accs)
handle = Entrez.esearch( db=db,term=query,retmax=retmax )
giList = Entrez.read(handle)['IdList']
sys.stderr.write( "Found %s GI: %s\n" % (len(giList), ", ".join(giList[
```

Leszek ♦
2,780 • 5 • 19
updated 10 months ago
by Leszek ♦

```
:10])))
#post NCBI query
search_handle     = Entrez.epost(db=db, id=",".join(giList))
search_results    = Entrez.read(search_handle)
webenv,query_key  = search_results["WebEnv"], search_results["QueryKey"
]
#fecth all results in batch of batchSize entries at once
for start in range( 0,len(giList),batchSize ):
  sys.stderr.write( " %9i" % (start+1,))
  #fetch entries in batch
  handle = Entrez.efetch(db=db, rettype="gb", retstart=start, retmax=ba
tchSize, webenv=webenv, query_key=query_key)
  #print output to stdout
  sys.stdout.write(handle.read())
```

I usually use this way. But it's extremely slow when you want to download thousands of entries and sometimes falls to timeout. Using epost would be much better. But, apparently there's no way of doing that.

log in to reply • written 10 months ago by Sanjarbek Hudaiberdiev **10** • 2

you can do epost easily with biopython - you just have to be sure to provide valid accessions, otherwise the script will crash... have a look here how to do it: http://biopython.org/DIST/docs/tutorial/Tutorial.html#htoc114

log in to reply • written 10 months ago by Leszek ♦ **2,780** • 5 • 19

I'm having the same issue with epost not taking accession numbers but integer UIDs instead...

log in to reply • written 7 months ago by junyinglim **0**

Look at edited option: you can provide any id that is accepted by NCBI and the script will get UIDs of these automatically and print fasta...

log in to reply • written 7 months ago by Leszek ♦ **2,780** • 5 • 19

This works fantastic! Just couldn't think myself of searching by accessions and getting ['IdList'] back. Thanks Leszek.

log in to reply • written 7 months ago by Sanjarbek Hudaiberdiev **10** • 2

1 you can find more tricks in biopython tutorial: http://biopython.org/DIST/docs/tutorial/Tutorial.html#htoc110

log in to reply • written 7 months ago by Leszek ♦ **2,780** • 5 • 19

The hit-it-on-the-head-with-a-hammer alternative is to batch download a whole section of genbank, then filter out the ones you want using bioperl which will take accessions.

i.e. something along the lines of

```
   foreach my $acc (@acc){
           if ($seq->accession eq $acc){
                   $fileout->write_seq($seq);          ##etc. etc.
etc.
           }
```

created 10 months ago by Mabeuf **900** • 1 • 8

Any suggestions on how to download the whole data of genbank?

log in to reply • written 10 months ago by Sanjarbek Hudaiberdiev **10** • 2

There is the ftp if you want to batch DL ftp://ftp.ncbi.nlm.nih.gov/genbank/ But if you're only concerned about a certain taxa (prok, euk, human, whatever) you can just browser download it via http://www.ncbi.nlm.nih.gov/nuccore

log in to reply • written 10 months ago by Mabeuf **900** • 1 • 8

I'm not too sure which accession number you mean and what normal ID you are referring to but http://david.abcc.ncifcrf.gov/conversion.jsp has the ability to map between various accessions (e.g. genbank, uniprot) and IDs (e.g. entrez, ensembl)

created 10 months ago by secretjess **130** • 4

I'm working with GenBank. So, I need to map Genbank accession numbers to Genbank GI number. DAVID can't find the majority of the accessions that I submit. Moreover, it doesn't provide API, afaik.

log in to reply • written 10 months ago by Sanjarbek Hudaiberdiev **10** • 2

Ah, sorry that wasn't helpful. This looks like a similar question to yours (in reverse) so maybe the solution will be more useful: http://www.biostars.org/p/50383/

log in to reply • written 10 months ago by secretjess **130** • 4

1 I had a look on that thread. The problem is not solved there. If I had GIDs first, and wanted to convert them to Accessions, then that would be viable via epost-efetch:). But can't do the other way around. Thank you for your suggestions.

log in to reply • written 10 months ago by Sanjarbek Hudaiberdiev **10** • 2

---

two not so specific cents:

1) NCBI has a well documented but under-used collection of E-utilities you may want to try

2) Bio-perl has modules to fetch GenBank entries using different IDs.

created 10 months ago by Wen.Huang **860** • 2 • 6

1) Yes, I like very much working with E-utils. It's the first time I got stuck with a problem. 2) True. There's no problem in fetching by accession number in one-request-at-a-time manner. But it's slow, due to 1/3 seconds limitation of NCBI. The problem comes when you want to try epost for batch downloading. Thank you for your suggestions!

log in to reply • written 10 months ago by Sanjarbek Hudaiberdiev **10** • 2

Depends on how long "long" is, but up to some length, you can even use the Entrez web interface to get these e.g. querying with "A22237 A22239 A32021 A32022 A33397" here http://www.ncbi.nlm.nih.gov gives you a link to this page:

http://www.ncbi.nlm.nih.gov/nuccore/A22237A22239A32021A32022A33397

and you can get these in a range of formats e.g. full genbank, fasta, etc., using the "Send to" functionality

created 10 months ago by aidan-budd ♦ **1,720** • 6

The script provided by Leszek solves the problem.

created 7 months ago by Sanjarbek Hudaiberdiev **10** • 2
