Question: retrieving FASTA sequences from ncbi using biopython

I have a file with GI numbers and would like to get `FASTA` sequences from ncbi.

```
from Bio import Entrez
import time
Entrez.email ="eigtw59tyjrt403@gmail.com"
f = open("C:\\bioinformatics\\gilist.txt")
for line in iter(f):
    handle = Entrez.efetch(db="nucleotide", id=line, retmode="xml")
    records = Entrez.read(handle)
    print ">GI "+line.rstrip()+" "+records[0]["GBSeq_primary-accession"
]+" "+records[0]["GBSeq_definition"]+"\n"+records[0]["GBSeq_sequence"]
    time.sleep(1) # to make sure not many requests go per second to ncb
i
f.close()
```

This script runs fine but I suddenly get this error message after a few sequences.

```
Traceback (most recent call last):
  File "C:/Users/Ankur/PycharmProjects/ncbiseq/getncbiSeq.py", line 7,
in <module>
    handle = Entrez.efetch(db="nucleotide", id=line, retmode="xml")
  File "C:\Python27\lib\site-packages\Bio\Entrez\__init__.py", line 139
, in efetch
    return _open(cgi, variables)
  File "C:\Python27\lib\site-packages\Bio\Entrez\__init__.py", line 455
, in _open
    raise exception
urllib2.HTTPError: HTTP Error 500: Internal Server Error
```

Of course I can use `http://www.ncbi.nlm.nih.gov/sites/batchentrez` but I am trying to create a pipeline and would like something automated.

How can I prevent ncbi from "kicking me out"

python | **biopython** | **ncbi** | **entrez**

created 12 months ago by Ankur
30 • 2

updated 12 months ago by Martombo

3 answers

Are you following the Entrez Usage Guidlines? Specifically:

- For any series of more than 100 requests, do this at weekends or outside USA peak times. This is up to you to obey.
- Make no more than three requests every seconds (relaxed from at most one request every three seconds in early 2009). This is automatically enforced by Biopython.

Also, read the paragraph about Minimizing the Number of Requests:

created 12 months ago by Jeroen

> If a task requires searching for and/or downloading a large number of records, it is much more efficient to use the Entrez History to upload and/or retrieve these records in batches rather than using separate requests for each record. [...] Many thousands of IDs can be uploaded using a single EPost request, and several hundred records can be downloaded using one EFetch request.

Van Goey ♦
1,800 • 4 • 14
updated 12 months ago by Jeroen Van Goey ♦

Also some unsolicited Python advice:

Use the with statement to open a file, then you don't have to explicitly close it. And a file object is already an iterable, so you don't need to wrap it in `iter` :

```
with open("C:\\bioinformatics\\gilist.txt") as f:
    for line in f:
        # do your stuff
```

I second Jeroen's suggestions, here's some code to get you started with the batch mode:

created 12 months ago by Hayssam
200 • 4
updated 12 months ago by Hayssam

```
from Bio import Entrez
import time
Entrez.email ="eigtw59tyjrt403@gmail.com"
## We instead upload the list of ID beforehand
gis=[166706892,431822405,431822402]
request = Entrez.epost("nucleotide",id=",".join(map(str,gis)))
result = Entrez.read(request)
webEnv = result["WebEnv"]
queryKey = result["QueryKey"]
handle = Entrez.efetch(db="nucleotide",retmode="xml", webenv=webEnv, qu
ery_key=queryKey)
for r in Entrez.parse(handle):
    # Grab the GI
    try:
        gi=int([x for x in r['GBSeq_other-seqids'] if "gi" in x][0].spl
it("|")[1])
    except ValueError:
        gi=None
    print ">GI ",gi," "+r["GBSeq_primary-accession"]+" "+r["GBSeq_defin
ition"]+"\n"+r["GBSeq_sequence"][0:20]
```

Note that there might be a limit on the number of sequences you can retrieve in a batch. You should circumvent this by splitting your input set of ids in subsets of e.g. 200 ids.

---

my suggestion is to directly handle the exception and wait a longer time if you get "kicked out":

```
from Bio import Entrez
from urllib2 import HTTPError
import time
Entrez.email ="eigtw59tyjrt403@gmail.com"
f = open("C:\\bioinformatics\\gilist.txt")
for line in iter(f):
    try:
        handle = Entrez.efetch(db="nucleotide", id=line, retmode="xml")
    except HTTPError:
        time.sleep(20)
        handle = Entrez.efetch(db="nucleotide", id=line, retmode="xml")
    records = Entrez.read(handle)
    print ">GI "+line.rstrip()+" "+records[0]["GBSeq_primary-accession"
]+" "+records[0]["GBSeq_definition"]+"\n"+records[0]["GBSeq_sequence"]
    time.sleep(1) # to make sure not many requests go per second to ncb
i
f.close()
```

I tried with 20 seconds and it's working for me, if such a long time is not suitable for you, you can try to minimize it

created 12 months ago by Martombo
140 • 3