Class 26: BLAST

April 25, 2019

The web interface to BLAST is available here: http://blast.ncbi.nlm.nih.gov/Blast.cgi (http://blast.ncbi.nlm.nih.gov/Blast.cgi)

Let's search for proteins related to the following query sequence, which is the glycoprotein of Machupo virus (causative agent of Bolivian hemorrhagic fever):

```
>GI:45825963|Machupo virus glycoprotein
MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLFLAGRSCS
DGTFKIGLHTEFQSVTFTMQRLLANHSNELPSLCMLNNSFYYMKGGANIFLIRVSDVS
VLMKEYDVSVYEPEDLGNCLNKSDSSWAIHWFSIALGHDWLMDPPMLCRNKTKKEGSN
IQFNISKADESRVYGKKIRNGMRHLFRGFYDPCEEGKVCYVTINQCGDPSSFEYCGTN
YLSKCQFDHVNTLHFLVRSKTHLNF
```

We can download the blast results from the NCBI website in XML format and store them as `Machupo_BLAST.xml`. This file is available here. (http://wilkelab.org/classes/SDS348/data_sets/Machupo_BLAST.xml)

Now we can process this file with Biopython.

In [1]:

```
from Bio.Blast import NCBIXML
from urllib.request import urlretrieve # to download xml file

# download file from course website and store locally
urlretrieve('http://wilkelab.org/classes/SDS348/data_sets/Machupo_BLAST.xml', '
Machupo_BLAST.xml')

# open the downloaded file and parse with NCBIXML.read()
blast_handle = open("Machupo_BLAST.xml")
blast_record = NCBIXML.read(blast_handle)
blast_handle.close()

imax = 30 # process the first 30 alignments
i = 0
for alignment in blast_record.alignments:
    i += 1
    if i > imax:
        break
    # we need a for loop here because in theory we could have
    # more than one hsp (High-scoring Segment Pair) per alignment
    for hsp in alignment.hsps:
        print('\n****Alignment****')
        print('sequence ID:', alignment.title)
        print('length:', alignment.length)
        print('score:', hsp.score)
        print('e value:', hsp.expect)
        print("Query:", hsp.query[0:100] + '...')
        print("Match:", hsp.match[0:100] + '...')
        print("  Hit:", hsp.sbjct[0:100] + '...')
```

```
****Alignment****
sequence ID: gi|45825964|gb|AAS77647.1| glycoprotein 1, partial [Machupo mammar
enavirus]
length: 257
score: 1381.0
e value: 0.0
Query: MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLFLAGRSCSDGTFKIGLHTEFQS
VTFTMQRLLANHSNELPSLCMLNNSFYY...
Match: MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLFLAGRSCSDGTFKIGLHTEFQS
VTFTMQRLLANHSNELPSLCMLNNSFYY...
  Hit: MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLFLAGRSCSDGTFKIGLHTEFQS
VTFTMQRLLANHSNELPSLCMLNNSFYY...

****Alignment****
sequence ID: gi|45826506|gb|AAS77879.1| glycoprotein precursor [Machupo mammare
navirus]
length: 496
score: 1379.0
e value: 0.0
Query: MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLFLAGRSCSDGTFKIGLHTEFQS
VTFTMQRLLANHSNELPSLCMLNNSFYY...
Match: MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLFLAGRSCSDGTFKIGLHTEFQS
VTFTMQRLLANHSNELPSLCMLNNSFYY...
  Hit: MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLFLAGRSCSDGTFKIGLHTEFQS
VTFTMQRLLANHSNELPSLCMLNNSFYY...

****Alignment****
sequence ID: gi|45825936|gb|AAS77633.1| glycoprotein 1, partial [Machupo mammar
enavirus]
length: 257
score: 1274.0
e value: 4.8109e-175
Query: MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLFLAGRSCSDGTFKIGLHTEFQS
VTFTMQRLLANHSNELPSLCMLNNSFYY...
Match: MGQL+SFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFL LAGRSCSDGTFKIGLHTEFQS
VT TMQRLLANHSNELPSLCMLNNSFYY...
  Hit: MGQLVSFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLLLAGRSCSDGTFKIGLHTEFQS
VTLTMQRLLANHSNELPSLCMLNNSFYY...

****Alignment****
sequence ID: gi|45825934|gb|AAS77632.1| glycoprotein 1, partial [Machupo mammar
enavirus]
length: 257
score: 1274.0
e value: 5.5461e-175
Query: MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLFLAGRSCSDGTFKIGLHTEFQS
VTFTMQRLLANHSNELPSLCMLNNSFYY...
Match: MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFL LAGRSCSDGTFKIGLHTEFQS
VT TMQRLLANHSNELPSLCMLNNSFYY...
  Hit: MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLLLAGRSCSDGTFKIGLHTEFQS
VTLTMQRLLANHSNELPSLCMLNNSFYY...

****Alignment****
sequence ID: gi|45825948|gb|AAS77639.1| glycoprotein 1, partial [Machupo mammar
enavirus] >gi|45825950|gb|AAS77640.1| glycoprotein 1, partial [Machupo mammaren
avirus]
length: 257
score: 1269.0
e value: 3.05564e-174
Query: MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLFLAGRSCSDGTFKIGLHTEFQS
VTFTMQRLLANHSNELPSLCMLNNSFYY...
Match: MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFL LAGRSCSDGTFKIGLHTEFQS
VT TMQRLLANHSNELPSLCMLNNSFYY...
```

Problems

Problem 1:

Count the number of hits with an E value of less than or equal to 1e-100.

In [2]:

```
E_cutoff = 1e-100
count = 0
for alignment in blast_record.alignments:
    for hsp in alignment.hsps:
        if hsp.expect <= E_cutoff:
            count += 1

print("There are", count, "hits with E <=", E_cutoff)
```

```
There are 28 hits with E <= 1e-100
```

Problem 2:

Extract the genbank identifiers (written as `gb|`*`string`*`|`, where *`string`* is the actual identifier, consisting of letters, numbers, and the period symbol) for all matches with an E value of less than or equal to 1e-100, and store them in a python list. For matches that list multiple genbank identifiers, only extract the first one.

In [3]:

```
import re

E_cutoff = 1e-100
gb_list = []
for alignment in blast_record.alignments:
    for hsp in alignment.hsps:
        if hsp.expect <= E_cutoff:
            match = re.search(r'gb\|([\w\d\.]+)\|', alignment.title)
            if match:
                gb_id = match.group(1)
                gb_list.append(gb_id)
            else:
                print("could not find genbank identifier in ", alignment.title)

print(gb_list)
```

```
could not find genbank identifier in  gi|240104274|pdb|2WFO|A Chain A, Crystal
Structure Of Machupo Virus Envelope Glycoprotein Gp1
could not find genbank identifier in  gi|290790109|pdb|3KAS|B Chain B, Machupo
Virus Gp1 Bound To Human Transferrin Receptor 1
['AAS77647.1', 'AAS77879.1', 'AAS77633.1', 'AAS77632.1', 'AAS77639.1', 'AAS7764
1.1', 'AAS77637.1', 'AAS77645.1', 'AAS77631.1', 'AAS77621.1', 'AAS77636.1', 'AA
X99337.1', 'AAS77634.1', 'AAS77635.1', 'AAN09942.1', 'AAX99339.1', 'AAX99329.1'
, 'AAS77646.1', 'AAT45081.1', 'AAX99333.1', 'AAT40455.1', 'AAX99331.1', 'AEX083
76.1', 'ACU24736.1', 'ACU24728.1', 'ACU24734.1']
```

If this was easy

Problem 3:

Using the list of genbank identifiers obtained in the previous exercise, download the corresponding sequences from genbank and print them out in FASTA format. Hint: You will have to specify the database as "protein" for this to work, since the previous exercise generated identifiers for protein sequences.

Hint: Use the function `SeqIO.write()` to output your results in FASTA format, and use `sys.stdout` from the `sys` module as your output handle.

In [4]:

```
from Bio import Entrez, SeqIO
import sys

Entrez.email = "wilke@austin.utexas.edu" # put your email here

handle = Entrez.efetch(db="protein", id=gb_list, rettype="gb", retmode="text")
records = SeqIO.parse(handle, "genbank")

for record in records:
    SeqIO.write(record, sys.stdout, "fasta")

handle.close() # important, close the handle only after you have iterated over
the records. Otherwise you will get an error!
```

```
>AAS77647.1 glycoprotein 1, partial [Machupo mammarenavirus]
MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLFLAGRSCSDG
TFKIGLHTEFQSVTFTMQRLLANHSNELPSLCMLNNSFYYMKGGANIFLIRVSDVSVLMK
EYDVSVYEPEDLGNCLNKSDSSWAIHWFSIALGHDWLMDPPMLCRNKTKKEGSNIQFNIS
KADESRVYGKKIRNGMRHLFRGFYDPCEEGKVCYVTINQCGDPSSFEYCGTNYLSKCQFD
HVNTLHFLVRSKTHLNF
>AAS77879.1 glycoprotein precursor [Machupo mammarenavirus]
MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLFLAGRSCSDG
TFKIGLHTEFQSVTFTMQRLLANHSNELPSLCMLNNSFYYMKGGANIFLIRVSDVSVLMK
EYDVSVYEPEDLGNCLNKSDSSWAIHWFSIALGHDWLMDPPMLCRNKTKKEGSNIQFNIS
KADESRVYGKKIRNGMRHLFRGFYDPCEEGKVCYVTINQCGDPSSFEYCGTNYLSKCQFD
HVNTLHFLVRSKTHLNFERSLKAFFSWSLTDSSGKDMPGGYCLEEWMLIAAKMKCFGNTA
VAKCNQNHDSEFCDMLRLFDYNKNAIKTLNDESKKEINFLSQTVNALISDNLLMKNKIRE
LMSVPYCNYTKFWYVNHTLTGQHTLPRCWLIRNGSYLNISEFRNDWILESDHLISEMLSK
EYAERQSKTPITLVDICFWSTIFFTASLFLHLVGIPTHRHLKGEACPLPHRLDSLGGCRC
GKYPRLKKPTVWHRRH
>AAS77633.1 glycoprotein 1, partial [Machupo mammarenavirus]
MGQLVSFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLLLAGRSCSDG
TFKIGLHTEFQSVTLTMQRLLANHSNELPSLCMLNNSFYYMKGGVNTFLIRVSDVSVLMK
EYDVSVYEPEDLGNCLNKSDSSWAIHWFSNALGHDWLMDPPMLCRNRTKKEGSNIQFNIS
KADDVRVYGKKIRNGMRHLFRGFHDPCEEGKVCYLTINQCGDPSSFDYCGTNYLSKCQFD
HVNTLHFLVRSKTHLNF
>AAS77632.1 glycoprotein 1, partial [Machupo mammarenavirus]
MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLLLAGRSCSDG
TFKIGLHTEFQSVTLTMQRLLANHSNELPSLCMLNNSFYYMKGGVNTFLIRVSDISVLMK
EYDVSVYEPEDLGNCLNKSDSSWAIHWFSNALGHDWLMDPPMLCRNRTKKEGSNIQFNIS
KADDVRVYGKKIRNGMRHLFRGFHDPCEEGKVCYLTINQCGDPSSFDYCGTNYLSKCQFD
HVNTLHFLVRSKTHLNF
>AAS77639.1 glycoprotein 1, partial [Machupo mammarenavirus]
MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLLLAGRSCSDG
TFKIGLHTEFQSVTLTMQRLLANHSNELPSLCMLNNSFYYMKGGVNTFLIRVSDISVLMK
EYDVSIYEPEDLGNCLNKSDSSWAIHWFSNALGHDWLMDPPMLCRNKTKKEGSNIQFNIS
KADDARVYGKKIRNGMRHLFRGFHDPCEEGKVCYLTINQCGDPSSFDYCGVNHLSKCQFD
HVNTLHFLVRSKTHLNF
>AAS77641.1 glycoprotein 1, partial [Machupo mammarenavirus]
MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLLLAGRSCSDG
TFKIGLHTEFQSVTLTMQRLLANHSNELPSLCMLNNSFYYMRGGVNTFLIRVSDISVLMK
EYDVSIYEPEDLGNCLNKSDSSWAIHWFSNALGHDWLMDPPMLCRNKTKKEGSNIQFNIS
KADDARVYGKKIRNGMRHLFRGFHDPCEEGKVCYLTINQCGDPSSFDYCGVNHLSKCQFD
HVNTLHFLVRSKTHLNF
>AAS77637.1 glycoprotein 1, partial [Machupo mammarenavirus]
MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLLLAGRSCSDG
TFKIGLHTEFQSVTLTMQRLLANHSNELPSLCMLNNSFYYMKGGVNTFLIRVSDISVLMK
EHDVSIYEPEDLGNCLNKSDSSWAIHWFSNALGHDWLMDPPMLCRNKTKREGSNIQFNIS
KADDARVYGKKIRNGMRHLFRGFHDPCEEGKVCYLTINQCGDPSSFDYCGVNHLSKCQFD
HVNTLHFLVRSKTHLNF
>AAS77645.1 glycoprotein 1, partial [Machupo mammarenavirus]
MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLLLAGRSCSDG
TFKIGLHTEFQSVTLTMQRLLANHSNELPSLCILNNNFYYMKGGVNTFLIRVSDISVLMK
EYDVSIYEPEDLGNCLNKSDSSWAVHWFSNALGHDWLMDPPMLCRNKTKKEGSNIQFNIS
KADDTRVYGKKIRNGMRHLFRGFHDPCEEGKVCYLTINQCGDPSSFDYCGVNHLSKCQFD
HVNTLHFLVRSKTHLNF
>AAS77631.1 glycoprotein 1, partial [Machupo mammarenavirus]
MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGIINLYKSGLFQFIFFLLLAGRSCSDG
TFKIGLHTEFQSVTLTMQRLLANHSNELPSLCMLNNSFYYMRGGVNTFLIRVSDVSVLMK
EYDVSIYEPEDLGNCLNKSDSSWAVHWFSNALGHDWLMDPPMLCRNKTKKEGSNIQFNIS
KADDTKVYGKKIRNGMRHLFRGFHDLCEEGKVCYLTINQCGDPSSFDYCNTNYLSKCQFD
HVNTLHFLVRSKTHLNF
>AAS77621.1 glycoprotein 1, partial [Machupo mammarenavirus]
MGQLISFFQEIPVFLQEALNIALVAVSLIAVIKGVINLYKSGLFQFIFFLLLAGRSCSDG
TFKIGLHTEFQSVTLTMQRLLANHSSELPSLCMLNNSFYYMKGGVNTFLIRVSDVSVLMK
EYDVSIYEPEDLGNCLNKSDSSWAVHWFSNALGHDWLMDPPMLCRNKTKKEGSNIQFNIS
KADDTKVYGKKIRNGMRHLFRGFHDLCEEGKVCYLTINQCGDPSSFDYCNTNYLSKCQFD
```