HW_2 (работа с NCBI)

In [61]:

```
from Bio import Entrez
```

In [71]:

```
Entrez.email = 'wwoskie@gmail.com'
```

In [72]:

```
%load_ext rpy2.ipython
```

```
The rpy2.ipython extension is already loaded. To reload it, use:
  %reload_ext rpy2.ipython
```

In [73]:

```
%%R
if (!("reutils" %in% installed.packages()))
  install.packages("reutils")
library(reutils)
```

1. Найдем в PubMed статьи по интересному для нас запросу и возвратим абстракты этих статей;

In [74]:

```
handle = Entrez.esearch(db = "pubmed", term = "cnv detection review")
record = Entrez.read(handle)
mshandle = Entrez.efetch(db="pubmed", id=record["IdList"][0:3],
rettype="abstract", retmode="text")
with open("abstracts_py.txt", "w") as f:
    for line in mshandle:
        f.write(line)
```

In [20]:

```
%%R
ms <- esearch(db = "pubmed", term = "cnv detection review")[1:3]
abstr <- efetch(ms, rettype = "abstract")
abstr
write(content(abstr), "abstracts_R.txt")
```

In [151…

```
!esearch -email wwoskie@gmail.com -db pubmed -query "cnv detection
review" | efetch -mode text -format abstract  > abstracts_bash.txt
```

2. Найдем ID организма по названию в базе taxonomy;

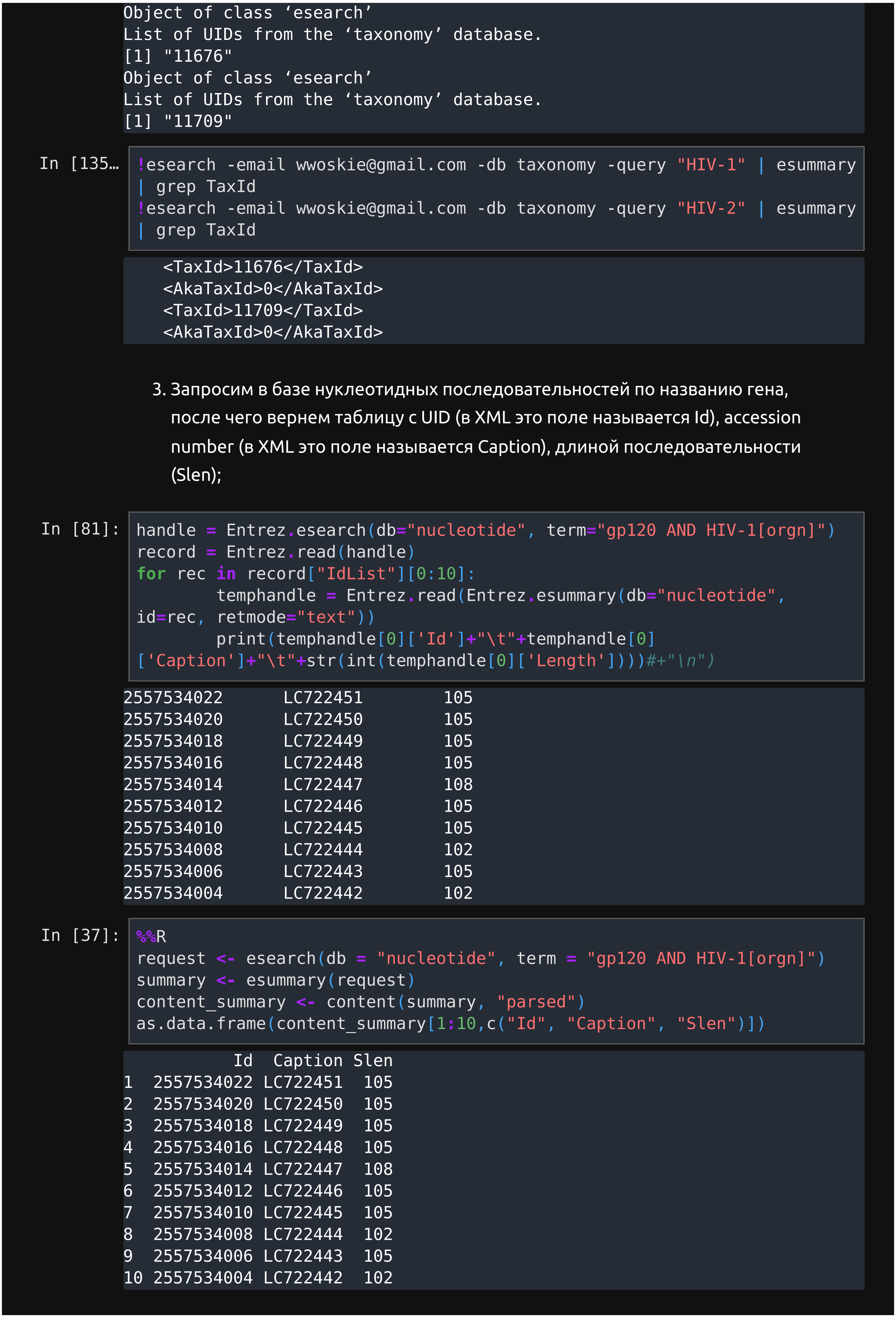
In [77]:

```
handle = Entrez.esearch(db = "taxonomy", term ="HIV-1")
record = Entrez.read(handle)
print(record['IdList'])
handle = Entrez.esearch(db = "taxonomy", term ="HIV-2")
record = Entrez.read(handle)
print(record['IdList'])
```

```
['11676']
['11709']
```

In [39]:

```
%%R
print(esearch(db = "taxonomy", term = "HIV-1"))
print(esearch(db = "taxonomy", term = "HIV-2"))
```



```
Object of class ‘esearch’
List of UIDs from the ‘taxonomy’ database.
[1] "11676"
Object of class ‘esearch’
List of UIDs from the ‘taxonomy’ database.
[1] "11709"
```

In [135…

```
!esearch -email wwoskie@gmail.com -db taxonomy -query "HIV-1" | esummary
| grep TaxId
!esearch -email wwoskie@gmail.com -db taxonomy -query "HIV-2" | esummary
| grep TaxId
```

```
    <TaxId>11676</TaxId>
    <AkaTaxId>0</AkaTaxId>
    <TaxId>11709</TaxId>
    <AkaTaxId>0</AkaTaxId>
```

3. Запросим в базе нуклеотидных последовательностей по названию гена, после чего вернем таблицу с UID (в XML это поле называется Id), accession number (в XML это поле называется Caption), длиной последовательности (Slen);

In [81]:

```
handle = Entrez.esearch(db="nucleotide", term="gp120 AND HIV-1[orgn]")
record = Entrez.read(handle)
for rec in record["IdList"][0:10]:
        temphandle = Entrez.read(Entrez.esummary(db="nucleotide",
id=rec, retmode="text"))
        print(temphandle[0]['Id']+"\t"+temphandle[0]
['Caption']+"\t"+str(int(temphandle[0]['Length'])))#+"\n")
```

```
2557534022      LC722451        105
2557534020      LC722450        105
2557534018      LC722449        105
2557534016      LC722448        105
2557534014      LC722447        108
2557534012      LC722446        105
2557534010      LC722445        105
2557534008      LC722444        102
2557534006      LC722443        105
2557534004      LC722442        102
```

In [37]:

```
%%R
request <- esearch(db = "nucleotide", term = "gp120 AND HIV-1[orgn]")
summary <- esummary(request)
content_summary <- content(summary, "parsed")
as.data.frame(content_summary[1:10,c("Id", "Caption", "Slen")])
```

```
           Id  Caption Slen
1  2557534022 LC722451  105
2  2557534020 LC722450  105
3  2557534018 LC722449  105
4  2557534016 LC722448  105
5  2557534014 LC722447  108
6  2557534012 LC722446  105
7  2557534010 LC722445  105
8  2557534008 LC722444  102
9  2557534006 LC722443  105
10 2557534004 LC722442  102
```

In [139…

```
!esearch -email wwoskie@gmail.com -db nucleotide -query "gp120 AND HIV-
1[orgn]" | esummary -mode xml | xtract -pattern DocumentSummary -element
Id Caption Slen | sed '11,$ d; s/"//g'
```

```
2557534022      LC722451        105
2557534020      LC722450        105
2557534018      LC722449        105
2557534016      LC722448        105
2557534014      LC722447        108
2557534012      LC722446        105
2557534010      LC722445        105
2557534008      LC722444        102
2557534006      LC722443        105
2557534004      LC722442        102
^C
```

4. Дадим в базу нуклеотидных или белковых последовательностей текстовый запрос, а затем вернем последовательности в формате fasta и запишем их в файл;

In [91]:

```
handle = Entrez.esearch(db="nucleotide", term="gp120 AND HIV-1[orgn]")
record = Entrez.read(handle)



with open("HIV-1_gp120_py.fa", "w") as ouf:
    for rec in record["IdList"][0:10]:
        lne = Entrez.efetch(db="nucleotide", id=rec, retmode="text",
rettype="fasta").read()
        ouf.write(lne+"\n")
```

In [140…

```
%%R
request <- esearch(db = "nucleotide", term = "gp120 AND HIV-1[orgn]")
fasta_nuc <- efetch(uid = request[1:10], db = "nucleotide", rettype =
"fasta", retmode = "text")
write(content(fasta_nuc), "HIV-1_gp120_R.fa")
```

In [145…

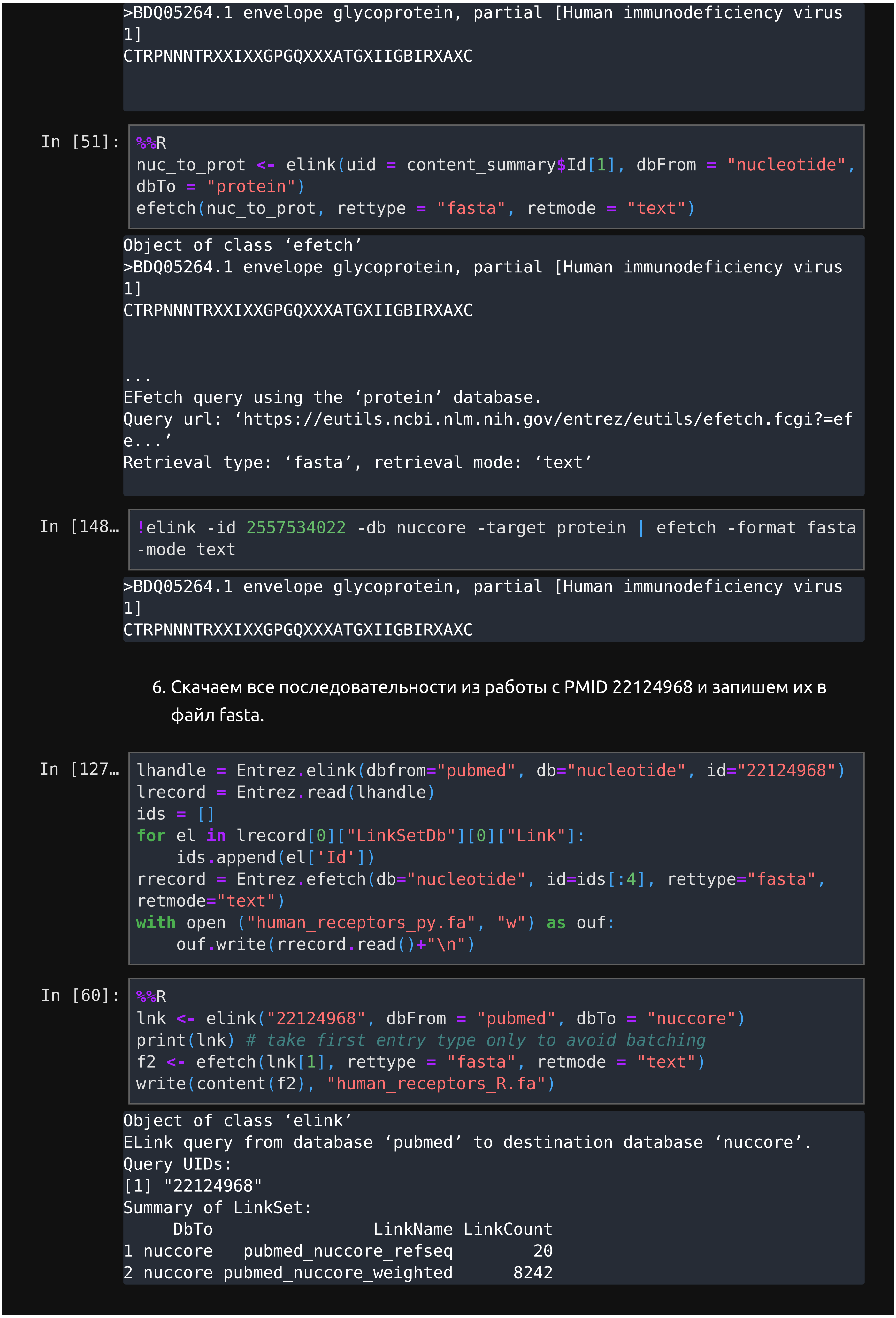
```
!esearch -email wwoskie@gmail.com -db nucleotide -query "gp120 AND HIV-
1[orgn][1:10]" | efetch -format fasta -mode text > HIV-1_gp120_bash.fa
```

```
^C
```

5. Скачаем белок, соответствующий известному UID нуклеотида;

In [126…

```
lhandle = Entrez.elink(dbfrom="nucleotide", id="2557534022",
linkname='nuccore_protein') # без linkname не работало
lrecord = Entrez.read(lhandle)
prothandle = lrecord[0]["LinkSetDb"][0]['Link'][0]['Id']
rrecord = Entrez.efetch(db="protein", id=prothandle, rettype="fasta",
retmode="text")
print(rrecord.read())
```



```
>BDQ05264.1 envelope glycoprotein, partial [Human immunodeficiency virus
1]
CTRPNNNTRXXIXXGPGQXXXATGXIIGBIRXAXC
```

```
In [51]: %%R
         nuc_to_prot <- elink(uid = content_summary$Id[1], dbFrom = "nucleotide",
         dbTo = "protein")
         efetch(nuc_to_prot, rettype = "fasta", retmode = "text")
```

```
Object of class ‘efetch’
>BDQ05264.1 envelope glycoprotein, partial [Human immunodeficiency virus
1]
CTRPNNNTRXXIXXGPGQXXXATGXIIGBIRXAXC



...
EFetch query using the ‘protein’ database.
Query url: ‘https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?=ef
e...’
Retrieval type: ‘fasta’, retrieval mode: ‘text’
```

```
In [148… !elink -id 2557534022 -db nuccore -target protein | efetch -format fasta
         -mode text
```

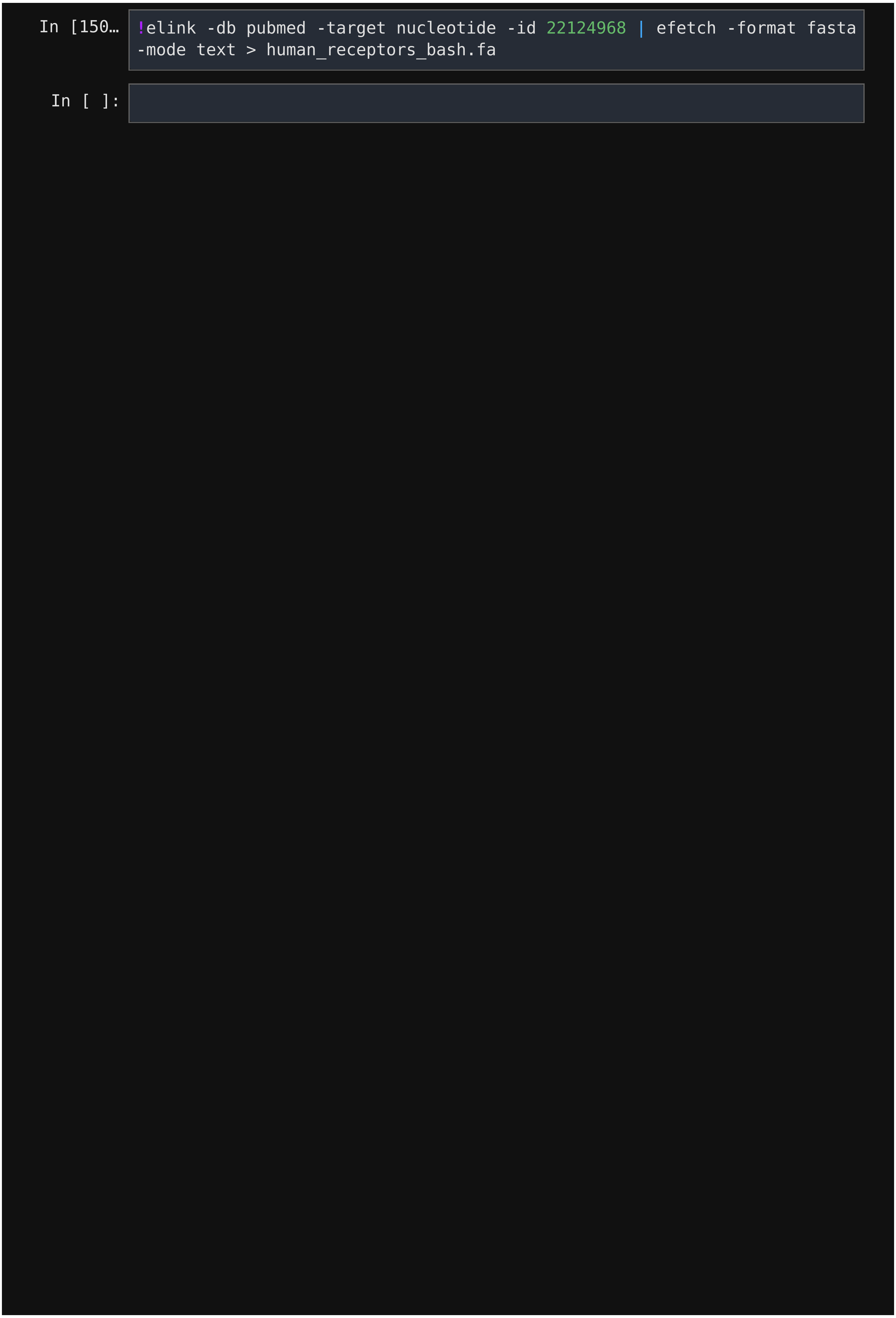
```
>BDQ05264.1 envelope glycoprotein, partial [Human immunodeficiency virus
1]
CTRPNNNTRXXIXXGPGQXXXATGXIIGBIRXAXC
```

6. Скачаем все последовательности из работы с PMID 22124968 и запишем их в файл fasta.

```
In [127… lhandle = Entrez.elink(dbfrom="pubmed", db="nucleotide", id="22124968")
         lrecord = Entrez.read(lhandle)
         ids = []
         for el in lrecord[0]["LinkSetDb"][0]["Link"]:
             ids.append(el['Id'])
         rrecord = Entrez.efetch(db="nucleotide", id=ids[:4], rettype="fasta",
         retmode="text")
         with open ("human_receptors_py.fa", "w") as ouf:
             ouf.write(rrecord.read()+"\n")
```

```
In [60]: %%R
         lnk <- elink("22124968", dbFrom = "pubmed", dbTo = "nuccore")
         print(lnk) # take first entry type only to avoid batching
         f2 <- efetch(lnk[1], rettype = "fasta", retmode = "text")
         write(content(f2), "human_receptors_R.fa")
```

```
Object of class ‘elink’
ELink query from database ‘pubmed’ to destination database ‘nuccore’.
Query UIDs:
[1] "22124968"
Summary of LinkSet:
     DbTo                LinkName LinkCount
1 nuccore   pubmed_nuccore_refseq        20
2 nuccore pubmed_nuccore_weighted      8242
```



```
In [150…  !elink -db pubmed -target nucleotide -id 22124968 | efetch -format fasta
          -mode text > human_receptors_bash.fa

In [ ]:
```