Отчет по работе с NCBI

Жерко Никита

Ниже будут представлены два скрипта (на *Python* и *R*), а также последовательность команд *bash*.

Python

```
# Поиск статей в PubMed по запросу 'zika' и вывод абстракта первой статьи
handle_2 = Entrez.esearch(db = "pubmed", term = "zika")
record_2 = Entrez.read(handle_2)
mshandle = Entrez.efetch(db="pubmed", id=record_2["IdList"][0:2], rettype="abstract",
retmode="text")
print(mshandle.read())
# 1. Nat Prod Res. 2024 Feb 11:1-7. doi: 10.1080/14786419.2024.2312418. Online
# ahead  of print.
#
# Larvicidal activity of diterpenes from Xylopia langsdorfiana St. Hilaire &
# Tulasne (Annonaceae) against Aedes aegypti linn. (diptera: culicidae).
#
# Ribeiro-Júnior KAL(1), Souza da Silva SA(2), Araújo-Júnior JX(2), Costa JGD(3),
# Fonseca Goulart H(4), Bernardo VB(1), Silva MSD(5), Tavares JF(5), Santana
# AEG(1)(4).
#
# Author information:
# (1)Institute of Chemistry and Biotechnology, Federal University of Alagoas A. C.
# Simões Campus, Maceió, AL, Brazil.
# (2)Laboratory of Medicinal Chemistry, Pharmaceutical Sciences Institute, Federal
# University of Alagoas A. C. Simões Campus, Maceió, AL, Brazil.
# (3)Embrapa Food and Territories, Embrapa, Maceió, AL, Brazil.
# (4)Agricultural Sciences Centre, Federal University of Alagoas, Rio Largo-AL,
# Brazil.
# (5)Postgraduate Program in Natural and Synthetic Bioactive Products, Federal
# University of Paraiba, João Pessoa, Brazil.
#
# Mosquitoes of the Aedes genus are responsible for transmitting many vector-borne
# viral diseases worldwide. Hundreds of thousands of people die annually from
# vector-borne diseases, including West Nile fever, dengue, tick-borne diseases,
# yellow fever, chikungunya, Rift Valley fever, and Zika. Billions of people are
# at the risk of infection on all continents, which is a cause of international
# concern. Therefore, new vector-control methods are essential for mitigating
# these illnesses. The bioactive hydrocarbons isolated from Xylopia langsdorfiana
# St. Hilaire & Tulasne are trachylobanes, a rare class of diterpenes found in the
# n-hexane fraction of the stem and leaf ethanolic extracts. These were tested
# against Ae. aegypti fourth-instar larvae over 48 h of exposure, with LC50 values
# ranging from 19.84 to 72.9 µg/mL, comparable to that of the positive control.
# The findings highlight the potential of Xylopia langsdorfiana St. Hilaire &
# Tulasne metabolites for controlling the main vectors of arthropod-borne viruses.
#
# DOI: 10.1080/14786419.2024.2312418
# PMID: 38343284


# Ищем ID организмов по названию в базе taxonomy
handle = Entrez.esearch(db = "taxonomy", term = "Mustela lutreola")
record = Entrez.read(handle)
print(record['IdList'])
# ['9666']
```

```
# Запрашиваем в базе белковых последовательностей по названию белка, после чего возвр
ащает таблицу
handle = Entrez.esearch(db="protein", term="crustacyanin AND Homarus americanus[org
n]")
record = Entrez.read(handle)
for rec in record["IdList"]:
        temphandle = Entrez.read(Entrez.esummary(db="protein", id=rec, retmode="tex
t"))
        print(temphandle[0]['Id']+"\t"+temphandle[0]['Caption']+"\t"+str(int(temphand
le[0]['Length'])))#+"\n")
# 400260701 4ALO_B  181
# 400260700 4ALO_A  181


# Даём в базу белковых последовательностей текстовый запрос,
# а затем возвращает последовательности в формате fasta, которые записывает в файл
Entrez.efetch(db="protein", id=record["IdList"], retmode="text", rettype="fasta").rea
d()
with open("crcn.fasta", "w") as ouf:
    for rec in record["IdList"]:
        lne = Entrez.efetch(db="protein", id=rec, retmode="text", rettype="fasta").re
ad()
        ouf.write(lne+"\n")
with open("crcn.fasta", "r") as fastaf:
    snippet = [next(fastaf) for x in range(5)]
    print(snippet)
# crcn.fasta
# >pdb|4ALO|B Chain B, H1 APOCRUSTACYANIN
# DKIPDFVVPGKCASVTRNKLWAEQTPNRNMYAGVWYQFALTNNPYQLIEKCVRNEYSFDGEQFVITSTGI
# AYDGNLLKRNGKLYPNPFGEPHLSIDYEMSFAAPLVILETDYSNYACLYSCIDYNFGYHSDFSFIFSRSA
# NLADKYVKKCEAAFKNINVDTTRFVKTVQGSSCPYDTQKTV
#
#
# >pdb|4ALO|A Chain A, H1 APOCRUSTACYANIN
# DKIPDFVVPGKCASVTRNKLWAEQTPNRNMYAGVWYQFALTNNPYQLIEKCVRNEYSFDGEQFVITSTGI
# AYDGNLLKRNGKLYPNPFGEPHLSIDYEMSFAAPLVILETDYSNYACLYSCIDYNFGYHSDFSFIFSRSA
# NLADKYVKKCEAAFKNINVDTTRFVKTVQGSSCPYDTQKTV


# Скачиваем белок по нуклеотидному айдишнику
lhandle = Entrez.elink(dbfrom="nucleotide", db="protein", id="2065188392")
lrecord = Entrez.read(lhandle)
prothandle = lrecord[0]["LinkSetDb"][0]['Link'][0]['Id']
rrecord = Entrez.efetch(db="protein", id=prothandle, rettype="fasta", retmode="text")
with open ("prot_from_nt.fasta", "w") as ouf:
    ouf.write(rrecord.read()+"\n")
# prot_from_nt.fasta
# >XP_042223242.1 crustacyanin-A2 subunit-like [Homarus americanus]
# MGVWYEIQAQPNIFQSIKSCLASSYKRVKTEIHVLSEGLDSSGASTTTKSILKIVDPQNPAHMVTDFVPG
# VEPPFDIVDTDYKTFSCAHSCLSIVGIKTEFVFIYSRNRTLRSNSTQHCLSIFEVSIIGIISFYTNANNY


# Скачиваем все последовательности по айдишнику
lhandle = Entrez.elink(dbfrom="pubmed", db="nucleotide", id="20558169")
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 ("py_fasta_pmid.fasta", "w") as ouf:
    ouf.write(rrecord.read()+"\n")
# >HM140499.1 Myospora metanephrops isolate NZ6C small subunit ribosomal RNA gene, pa
rtial sequence
# GACGGCTACCAGGTCCAAGGACAGCAGCAGGCGCGAAAATTACCGAAGCCTACAACAGGGCGGTAGTAAT
# GAGACGTGAAAACTAGACACGAATAAAATACGTGTTAGCAACTGGAGGTCAAGTCTGGTGCCAGCATCCG
# CGGTAATACCAGCTCCAGGGGTGTCTATGATGATTGCTGCGATTAAAAGGTCCGTAGTCTTATGTCAGAA
# CCGATGTGTAAGATGCTCGATCTAAGAGCAAAAAGGATTGGTACAGACATACATATATAGTGGTGTGTAT
# ATAGAGATGTTATATTTGTAATGTTGATATGTATATGGTGCAATATATTGAAATGAGGAGCGACCGGGGG
# CTAGATTATCGAGCAACGAGAGGTGAAATTTGATGACTTGCTTGGGAGTAACAGAGGCGAAAGCGCTAGT
# CAAGGGCGAATCCGATGATCAAGGACGTAGGCTGGAGGATCGAACACGATTAGATACCGTAGTAGTTCCA
# GCAGTAAACTATGCCGACGCCGTGGGTTGTTTGACCCGCGGAAGAGAAATCTAGTAGGGCTTTGGGGAGA
# GTACGCGCGCAAGCGATAAATTTAAAGGAAATTGACGGAAGAACACCACAAGGAGTGGAGTGTGCGGCTT
# AATTTGACTCAACGCGGGACAGCTTACCAGGCCCGAGGATTGCACGAGCGAATACGCGATAGATCTGAAA
# GTGGTGCATGGCCGTTATCGACGAATGGAGTGATCTTTTGGTTAAATCCGTCAATTCGTGAGACCCTTTT
# AATTTGATTAATGTCAGTGGTTGATACAGGTATGAAAATACAGGGGGGAAAGGACAAGAACAGGTCAGTG
# ATGCCCTTAGATGGCCTGGGCTGCACGCGCACTACAGTGGTTACTTTAAACTGAAGAGAGGAATAAATGT
# AATCGAGAGGGAATGAGCGCTGCAAGGCCCACAGGAACGAGGAATTGCTAGTAATCGTAGGCTCAGTAAG
# ATACGATGAATGTGTCCCTGTTC
# ...
```

R

```
# Поиск статей в PubMed по запросу 'zika' и вывод абстрактов статей в файл *abstract
s.txt*
esearch(db = "pubmed", term = "zika")
ms <- esearch(db = "pubmed", term = "zika")
abstr <- efetch(ms, rettype = "abstract")
write(content(abstr), "abstracts.txt")
# вывод, как и в питоне


# Ищем ID организма по названию в базе taxonomy
esearch(db = "taxonomy", term = "Mustela lutreola")
# "9666"


# Запрашиваем в базе белковых последовательностей по названию белка, после чего возвр
ащает таблицу
crcnp <- esearch(db = "protein", term = "crustacyanin AND Homarus americanus[orgn]")
su <- esummary(crcnp)
cosu <- content(su, "parsed")
as.data.frame(cosu[,c("Id", "Caption", "Slen")])
#          Id Caption Slen
# 1 400260701  4ALO_B  181
# 2 400260700  4ALO_A  181


# Даём в базу белковых последовательностей текстовый запрос,
# а затем возвращает последовательности в формате fasta, которые записывает в файл
s <- esearch(db = "protein", term = "crustacyanin AND Homarus americanus[orgn]")
f <- efetch(uid = s[1:10], db = "protein", rettype = "fasta", retmode = "text")
write(content(f), "Ham_crcn.fa")
fastaf <- readLines("Ham_crcn.fa")
# вывод, как и в питоне


# Находим белок по нуклеотидному айдишнику
lnk1 <- elink(uid = "2065188392", dbFrom = "nucleotide", dbTo = "protein")
efetch(lnk1, rettype = "fasta", retmode = "text")
# >XP_042223242.1 crustacyanin-A2 subunit-like [Homarus americanus]
# MGVWYEIQAQPNIFQSIKSCLASSYKRVKTEIHVLSEGLDSSGASTTTKSILKIVDPQNPAHMVTDFVPG
# VEPPFDIVDTDYKTFSCAHSCLSIVGIKTEFVFIYSRNRTLRSNSTQHCLSIFEVSIIGIISFYTNANNY


# Скачиваем все последовательности по айдишнику
ms2 <- esearch(term = "lobster microsporidia", db = "pubmed")
lnk <- elink(ms2[4], dbFrom = "pubmed", dbTo = "nuccore")
f2 <- efetch(lnk, rettype = "fasta", retmode = "text")
write(content(f2), "lobster_microsporidia.fa")
# вывод, как и в питоне
```

bash

```
# Поиск статей в PubMed по запросу 'zika' и вывод абстрактов
esearch -email nik.zherko@bk.ru -db pubmed -query "zika"
# <ENTREZ_DIRECT>
#   <Db>pubmed</Db>
#   <Count>12299</Count>
#   <Query>zika</Query>
#   <Step>1</Step>
#   <Elapsed>3</Elapsed>
# </ENTREZ_DIRECT>
esearch -email nik.zherko@bk.ru -db pubmed -query "zika" | efetch -mode text -format
abstract
# вывод абстрактов


# Ищем ID организмов по названию в базе taxonomy
esearch -email nik.zherko@bk.ru -db taxonomy -query "Mustela lutreola" | esummary | g
rep TaxId
# <TaxId>9666</TaxId>
# <AkaTaxId>0</AkaTaxId>


# Запрашиваем в базе белковых последовательностей по названию белка, после чего возвр
ащает таблицу
esearch -email nik.zherko@bk.ru -db protein -query "crustacyanin AND Homarus american
us[orgn]" | esummary -mode xml -format docsum | xtract -pattern DocumentSummary -elem
ent Id Caption Slen
# 400260701 4ALO_B  181
# 400260700 4ALO_A  181

# Даём в базу белковых последовательностей текстовый запрос,
# а затем возвращает последовательности в формате fasta, которые записывает в файл
esearch -email your@email.com -db protein -query "crustacyanin AND Homarus americanus
[orgn]" | efetch -format fasta -mode text >crcn.fa

# Находим белок по нуклеотидному айдишнику
link -id 2065188392 -db nuccore -target protein | efetch -db protein -format fasta -m
ode text

# Скачиваем все последовательности по айдишнику
elink -db pubmed -target nucleotide -id 20558169 | efetch -format fasta -mode text >
lobster_msp.fasta
```