Phylo_reutils

Anna_Balan

2023-02-11

```
if (!("reutils" %in% installed.packages()))
  install.packages("reutils")
library(reutils)
options(reutils.email = "your_email@gmail.com")
```

1. searches PubMed for articles of interest to abstracts articles in plain text format;

```
esearch(db = "pubmed", term = "crustacyanin")
```

```
## Object of class 'esearch'
## List of UIDs from the 'pubmed' database.
##  [1] "35247793" "35010161" "34436301" "33919403" "33465290" "32851672"
##  [7] "32596057" "32236233" "31813041" "30860355" "29683674" "29178679"
## [13] "28851818" "26220698" "25797168" "25605312" "24782450" "23570752"
## [19] "23510436" "23441225" "22869108" "22428138" "22189778" "21391640"
## [25] "21169698" "19579223" "19416706" "19414522" "19317475" "19299880"
## [31] "19058530" "18667761" "17374944" "17188641" "17124125" "17124122"
## [37] "17028694" "16833638" "16407115" "15686376" "15644340" "14993674"
## [43] "14770227" "14646064" "12876374" "12832753" "12782314" "12777800"
## [49] "12123366" "12119396" "11526314" "11526313" "11341939" "10944355"
## [55] "10604288" "9761813 " "9200677 " "11540431" "8931133 " "15299714"
## [61] "7698348 " "11542700" "1548709 " "1935978 " "2026162 " "2001254 "
## [67] "2306227 " "4033433 " "6202261 " "6627105 " "7419516 " "760804  "
## [73] "830471  " "5644143 " "6078541 " "4959560 " "5971798 " "14234502"
## [79] "18933429"
```

```
ms <- esearch(db = "pubmed", term = "crustacyanin")
abstr <- efetch(ms, rettype = "abstract")
abstr
```

```
## Object of class 'efetch'
## 1. Comp Biochem Physiol Part D Genomics Proteomics. 2022 Jun;42:100977. doi:
## 10.1016/j.cbd.2022.100977. Epub 2022 Feb 16.
##
## Searching and identifying pigmentation genes from Neocaridina denticulate
## sinensis via comparison of transcriptome in different color strains.
##
## Lin S(1), Zhang L(2), Wang G(1), Huang S(1), Wang Y(1).
##
```

```
## Author information:
## (1)Key Laboratory of Healthy Mariculture for the East China Sea, Ministry of
## Agriculture, Fisheries College, Jimei University, Xiamen 361021, China.
## (2)Key Laboratory of Healthy Mariculture for the East China Sea, Ministry of
## ...
## EFetch query using the 'pubmed' database.
## Query url: 'https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?=efe...'
## Retrieval type: 'abstract', retrieval mode: 'text'
```

```
write(content(abstr), "abstracts.txt")
```

2. request to the nucleotide database sequences all in the footsteps by gene name for an organism by name view and returns a list of identities or information about their number;

```
esearch(db = "nucleotide", term = "crustacyanin")
```

```
## Object of class 'esearch'
## List of UIDs from the 'nucleotide' database.
##   [1] "2312022563" "2312022559" "2312022555" "2311876279" "2311876276"
##   [6] "2311876274" "2311876272" "2311863192" "2311863190" "2311863187"
##  [11] "2311863185" "2311848461" "2307929728" "2307929724" "2307929643"
##  [16] "2307928020" "2311226022" "2301310459" "2277724840" "2277724838"
##  [21] "2231603375" "2231601330" "2231597514" "2231595668" "2231593297"
##  [26] "2231593226" "2214798008" "2214798006" "2214797999" "2214797990"
##  [31] "2214797986" "2214797970" "2175922880" "2175922879" "2171713983"
##  [36] "2171713967" "2171713949" "2171713932" "2171713927" "2171713898"
##  [41] "2154992186" "2154992038" "2128399489" "2096064420" "2077536330"
##  [46] "2077535960" "2077535433" "2077534997" "2077534241" "2077530957"
##  [51] "2077530956" "2077530797" "2077525862" "2077525742" "2077524696"
##  [56] "2077524339" "2065206281" "2065193120" "2065190079" "2065188392"
##  [61] "2065186734" "2065171964" "2065159855" "2065028193" "2056514813"
##  [66] "2056498811" "2056493211" "2056492151" "2056488359" "2056483089"
##  [71] "2056467578" "2056465480" "2056439345" "2037090538" "2032923631"
##  [76] "1953685317" "1953635727" "1950720059" "1950720057" "1941200435"
##  [81] "1935954895" "1935954881" "1933347811" "1933347799" "1933347780"
##  [86] "1933347777" "1933347776" "1933347770" "1933329650" "1838624700"
##  [91] "1511198136" "1511198134" "1511198132" "1721459068" "1721459065"
##  [96] "1721459062" "1511198335" "1511198333" "1644883274" "1595304136"
```

```
esearch(db = "nucleotide", term = "crustacyanin AND human[orgn]") #human doesnt have this protein
```

```
## Object of class 'esearch'
## List of UIDs from the 'nucleotide' database.
## [1] "NA"
```

```
esearch(db = "nucleotide", term = "crustacyanin AND lobster[orgn]") #it doesnt know lobster
```

```
## Error(s):
##  PhraseNotFound  lobster[orgn]
```

```
## Warning(s):
##  OutputMessage   No items found.

## Object of class 'esearch'
##  PhraseNotFound
## "lobster[orgn]"
##     OutputMessage
## "No items found."
```

```
esearch(db = "nucleotide", term = "crustacyanin AND Homarus americanus[orgn]")
```

```
## Warning: HTTPS error: Status 429;

## Object of class 'esearch'
## [1] "HTTPS error: Status 429; "
```

```
crcn <- esearch(db = "nucleotide", term = "crustacyanin AND Homarus americanus[orgn]")
```

3. searches for an organism ID by name on the base;

```
esearch(db = "taxonomy", term = "Homarus americanus")
```

```
## Object of class 'esearch'
## List of UIDs from the 'taxonomy' database.
## [1] "6706"
```

```
esearch(db = "taxonomy", term = "Human")
```

```
## Object of class 'esearch'
## List of UIDs from the 'taxonomy' database.
## [1] "9606"
```

```
esearch(db = "taxonomy", term = "Homo sapiens")
```

```
## Warning: HTTPS error: Status 429;

## Object of class 'esearch'
## [1] "HTTPS error: Status 429; "
```

```
esearch(db = "taxonomy", term = "Mouse") #why two species??? we dont know
```

```
## Object of class 'esearch'
## List of UIDs from the 'taxonomy' database.
## [1] "10090" "10088"
```

```
esearch(db = "taxonomy", term = "Ape") #why two species??? we dont know
```

```
## Object of class 'esearch'
## List of UIDs from the 'taxonomy' database.
## [1] "314295" "4456  "
```

```
#efetch(db = "taxonomy", uid = apes) #doesnt work for some reason
```

4. requests to protein databases or nucleotide sequences by name of the gene, after which it returns table with UID (in XML this field is called Id), inventory number (in XML this field is called Caption), long in direction (Slen);

```
crcnp <- esearch(db = "protein", term = "crustacyanin AND Homarus americanus[orgn]")
su <- esummary(crcnp)
```

```
## Warning: HTTPS error: Status 429;
```

```
cosu <- content(su, "parsed")
```

```
## Warning: Errors parsing DocumentSummary
```

```
as.data.frame(cosu[c("Id", "Caption", "Slen")])
```

```
## data frame with 0 columns and 0 rows
```

5. gives nucleotide or protein bases text query sequences, and then writes the sequences to a file in fasta format (show the beginning of the file);

```
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")
head(fastaf)
```

```
## [1] ">XP_042236484.1 crustacyanin-C1 subunit-like [Homarus americanus]"
## [2] "MNSLSILLVFVASVAADKIPDFVVPGKCASVDRNKLWAEQTPNRNNYAGVWYQFALTNNPYQLIEKCVRN"
## [3] "EYSFDGEQFVITSTGIAYDGNLLKRNGKLYPNPFGEPHLSIDYENSFAAPLVILETDYSNYACLYSCIDY"
## [4] "NFGYHSDFSFIFSRSANLAEQYVKKCEAAFKNINVDTTRFVKTVQGSSCPYDTQKTL"
## [5] ""
## [6] ">XP_042236483.1 crustacyanin-A2 subunit-like [Homarus americanus]"
```

6. downloads a protein corresponding to a known nucleotide UID;

```
lnk1 <- elink(uid = "2065188392", dbFrom = "nucleotide", dbTo = "protein")
efetch(lnk1, rettype = "fasta", retmode = "text")
```

```
## Object of class 'efetch'
## >XP_042223242.1 crustacyanin-A2 subunit-like [Homarus americanus]
## MGVWYEIQAQPNIFQSIKSCLASSYKRVKTEIHVLSEGLDSSGASTTTKSILKIVDPQNPAHMVTDFVPG
## VEPPFDIVDTDYKTFSCAHSCLSIVGIKTEFVFIYSRNRTLRSNSTQHCLSIFEVSIIGIISFYTNANNY
##
##
## ...
## EFetch query using the 'protein' database.
## Query url: 'https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?=efe...'
## Retrieval type: 'fasta', retrieval mode: 'text'
```

7. downloads all sequences from work with PMID ... (for example, from the first task) and writes them to the fasta file.

```
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")
```

```
from Bio import Entrez
Entrez.email = 'annabalan267@gmail.com'
```

1. esearch searches articles in NCBI PubMed

```
handle = Entrez.esearch(db = "pubmed", term = "crustacyanin")
record = Entrez.read(handle)
print(record)
```

```
{'Count': '79', 'RetMax': '20', 'RetStart': '0', 'IdList': ['35247793', '35010161', '34436301', '33919403', '33465290',
```

2. efetch returns abstracts of 3 first articles

```
mshandle = Entrez.efetch(db="pubmed", id=record["IdList"][0:3], rettype="abstract", retmode="text")
print(mshandle.read())
```

```
1. Comp Biochem Physiol Part D Genomics Proteomics. 2022 Jun;42:100977. doi:
10.1016/j.cbd.2022.100977. Epub 2022 Feb 16.

Searching and identifying pigmentation genes from Neocaridina denticulate
sinensis via comparison of transcriptome in different color strains.

Lin S(1), Zhang L(2), Wang G(1), Huang S(1), Wang Y(1).

Author information:
(1)Key Laboratory of Healthy Mariculture for the East China Sea, Ministry of
Agriculture, Fisheries College, Jimei University, Xiamen 361021, China.
(2)Key Laboratory of Healthy Mariculture for the East China Sea, Ministry of
Agriculture, Fisheries College, Jimei University, Xiamen 361021, China.
Electronic address: llzhang@jmu.edu.cn.

Aquaria species are characterized by their amazing colors and patterns. Research
on the breeding molecular genetics of ornamental shrimps is surprisingly
limited. We conducted a transcriptome analysis to investigate the expression of
encoding genes in the integument of the strains Neocaridina denticulate
sinensis. After assembled and filtered, 19,992 unigenes were annotated by
aligning with public functional databases (NR, Swiss-Prot, KEGG, COG). 14,915
unigenes with significantly different expressions were found by comparing three
strains integument transcriptomes. Ribosomal protein genes, ABC transporter
families, calmodulin, carotenoid proteins and crustacyanin may play roles in the
cytological process of pigment migration and chromatophore maintenance. Numerous
color genes associated with multiple pathways including melanin, ommochrome and
pteridines pathways were identified. The expression patterns of 25 candidate
genes were analysis by qPCR in red, yellow, transparent and glass strains. The
qPCR results in red, yellow and transparent were consistent with the level of
RPKM values in the transcriptomes. The above results will advance our knowledge
of integument color varieties in N. denticulate sinensis and help the genetic
selection of crustaceans with consumer-favored colors. Furthermore, it also
provides some candidate pigmentation genes to investigate the correlation
between coloration and sympatric speciation in crustaceans.

Copyright © 2022. Published by Elsevier Inc.

DOI: 10.1016/j.cbd.2022.100977
PMID: 35247793 [Indexed for MEDLINE]


2. Foods. 2021 Dec 23;11(1):35. doi: 10.3390/foods11010035.

Purification and Characterisation of Two Novel Pigment Proteins from the
Carapace of Red Swamp Crayfish (Procambarus clarkii).

Chen H(1)(2), Ji H(1)(3)(4)(5)(6), Pan C(6)(7), Zhang D(1)(3)(4)(5), Su
W(1)(3)(4)(5), Liu S(1)(3)(4)(5)(6), Deng Y(1), Huang X(1).

Author information:
(1)Guangdong Provincial Key Laboratory of Aquatic Product Processing and Safety,
College of Food Science and Technology, Guangdong Ocean University, Zhanjiang
524088, China.
(2)Hunan Provincial Key Laboratory of Soybean Products Processing and Safety
Control, College of Food and Chemical Engineering, Shaoyang University, Shaoyang
422000, China.
(3)Guangdong Provincial Engineering Technology Research Center of Seafood,
College of Food Science and Technology, Guangdong Ocean University, Zhanjiang
```

3. esearch searches in the base all the sequences for a certain gene and species, returning a list if IDs

```
handle = Entrez.esearch(db = "nucleotide", term = "crustacyanin AND Homarus[orgn]") #orgn=organism
record = Entrez.read(handle)
```

```
print(record)
Entrez.efetch(db = "nucleotide", id = record["IdList"])
```

```
{'Count': '19', 'RetMax': '19', 'RetStart': '0', 'IdList': ['2065206281', '2065193120', '2065190079', '2065188392', '206
<_io.TextIOWrapper encoding='UTF-8'>
```

searches taxon ID with a given name

```
handle = Entrez.esearch(db = "taxonomy", term = "Homarus americanus")
record = Entrez.read(handle)
print(record)
print(record['IdList'])
```

```
{'Count': '1', 'RetMax': '1', 'RetStart': '0', 'IdList': ['6706'], 'TranslationSet': [], 'TranslationStack': [{'Term': '
['6706']
```

4. esearch+summary searches in database of proteins and nucl sequences with a name of a protein. Returns UID. actually its XML parcing.

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

```
2068680993      XP_042236484    197
2068680990      XP_042236483    190
2068650119      XP_042225885    190
2068650116      XP_042225884    197
2068642615      XP_042227234    190
2068642613      XP_042227223    197
2068642611      XP_042227211    190
2068642608      XP_042227198    197
2068642605      XP_042227187    190
2068642602      XP_042227176    197
2056515232      KAG7177238      197
2056515231      KAG7177237      190
2056515230      KAG7177236      197
2056515229      KAG7177235      190
2056515228      KAG7177234      197
2056515227      KAG7177233      190
2056483091      KAG7166898      197
2056483090      KAG7166897      190
2056467580      KAG7160784      197
2056467579      KAG7160783      190
```

5. returns fasta and writes it in a file

```
handle = Entrez.esearch(db="protein", term="crustacyanin AND Homarus americanus[orgn]")
record = Entrez.read(handle)

Entrez.efetch(db="protein", id=record["IdList"], retmode="text", rettype="fasta").read()
with open("crcn.fasta", "w") as ouf:
    for rec in record["IdList"]:
        lne = Entrez.efetch(db="protein", id=rec, retmode="text", rettype="fasta").read()
        ouf.write(lne+"\n")
with open("crcn.fasta", "r") as fastaf:
    snippet = [next(fastaf) for x in range(5)]
    print(snippet)
```

```
['>XP_042236484.1 crustacyanin-C1 subunit-like [Homarus americanus]\n', 'MNSLSILLVFVASVAADKIPDFVVPGKCASVDRNKLWAEQTPNRNN
```

6. downloads a protein, takes a UID of a nucleotide

```
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")
```

7. Downloads fasta sequences from a work with PMID (e.g. fron the first task)

```
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")
```

SSH

1. ищет в PubMed статьи по интересному для вас запросу и возвращает абстракты этих статей (можно N первых статей в списке) в простом текстовом формате (можно записать в файл);

```
esearch -email your@email.com -db pubmed -query "crustacyanin"
esearch -email your@email.com -db pubmed -query "crustacyanin AND lobster[orgn]" | efetch -mode text -format abstract
```

2. запрашивает в базе нуклеотидных последовательностей все последовательности по названию гена для организма по названию вида и возвращает список ID или информацию об их количестве;

```
esearch -email your@email.com -db nucleotide -query "crustacyanin AND Homarus americanus[orgn]" | esummary
```

3. ищет ID организма по названию в базе;

```
esearch -email your@email.com -db taxonomy -query "Homarus gammarus" | esummary | grep TaxId
```

4. запрашивает в базе белковых или нуклеотидных последовательностей по названию гена, после чего возвращает таблицу с UID (в XML это поле называется Id), accession number (в XML это поле называется Caption), длиной последовательности (Slen);

```
esearch -email your@email.com -db protein -query "crustacyanin AND Homarus americanus[orgn]" | esummary -mode xml -format doc
```

5. даёт в базу нуклеотидных или белковых последовательностей текстовый запрос, а затем пишет последовательности в файл в формате fasta (покажите начало файла);

```
esearch -email your@email.com -db protein -query "crustacyanin AND Homarus americanus[orgn]" | efetch -format fasta -mode tex
head crcn.fa
```

> XP_042236484.1 crustacyanin-C1 subunit-like [Homarus americanus]
> MNSLSILLVFVASVAADKIPDFVVPGKCASVDRNKLWAEQTPNRNNYAGVWYQFALTNNPYQLIEKCVRN
> EYSFDGEQFVITSTGIAYDGNLLKRNGKLYPNPFGEPHLSIDYENSFAAPLVILETDYSNYACLYSCIDY
> NFGYHSDFSFIFSRSANLAEQYVKKCEAAFKNINVDTTRFVKTVQGSSCPYDTQKTL XP_042236483.1 crustacyanin-A2 subunit-like
> [Homarus americanus] MFRTVIVAALVACVAADGIPSFVTAGKCASVANQDNFDLRRYAGRWYQTHIIENAYQPVTRCINSNYEYS
> GNDYGFKVTTAGFNPNDEYLKIDFKVYPTKEFPAAHMLIDAPSVFAAPYEVIETDYDTYSCVYSCITTDN
> YKSEFAFVFSRTPQTSGPAVEKCAAVFNKNGVEFSKFVPVSHTAECVYRA XP_042225885.1 crustacyanin-A2 subunit-like [Homarus
> americanus] MFRTVIVAALVACVAADGIPSFVTAGKCASVANQDNFDLPRYAGRWYQTHIIENAYQPVTRCINSNYEYS

6. скачивает белок, соответствующий известному UID нуклеотида;

```
link -id 2065188392 -db nuccore -target protein | efetch -db protein -format fasta -mode text
```

7. скачивает все последовательности из работы с PMID … (например, из первого задания) и пишет их в файл fasta.

```
elink -db pubmed -target nucleotide -id 20558169 | efetch -format fasta -mode text > lobster_msp.fasta
```