Traineeship Part 1 NCBI-esummary+csv

April 23, 2020

Traineeship Part 1: Data collection using NCBI eUtils and esummary (+ generates csv files)

Author: Iris Raes

The University of Antwerp, Medical Biochemistry, Campus Drie Eiken

Loading required packages

[ ]:
```
# pip3 install --user eutils
from eutils import Client
from Bio import Entrez
import csv
```

Personal API-key

[ ]:
```
eclient = Client(api_key="8ecce891e7fa036ff84bccc7c74e5138dc09")
```

---

1) Entrez Nucleotide Search - mRNA Transcript Variants

[ ]:
```
### Creating query
mRNAtranscripts = []
transcriptmRNA_esearch = eclient.esearch(db='nucleotide',
            term='DPP8[gene] AND "Homo sapiens"[Primary Organism] AND␣
 ↪(biomol_mrna[PROP] AND refseq[filter])')
print("\nLoading currently available ids from Entrez nucleotide...")
print("="*50)
print("\nTranscript variant ids: ")
print(transcriptmRNA_esearch.ids)
for item in transcriptmRNA_esearch.ids:
    mRNAtranscripts.append(item)
print("\nSearch results: {}\n".format(transcriptmRNA_esearch.count))
```

[ ]:
```
### Esummary for retrieving information
Entrez.email = "iris.raes@hotmail.com"
### For each id in mRNAtranscripts
### Save data in csv file
```

```
with open('results-nucleotide.csv', mode='w') as result_nucleotide:
    result_writer = csv.writer(result_nucleotide,delimiter=';')
    result_writer.
 ↪writerow(["transcript_id","description","transcript_variant","accession","length_in_bp"])
    counter = 1
    for ids in mRNAtranscripts:
        handle = Entrez.esummary(db="nucleotide", id=ids)
        record = Entrez.read(handle)
        handle.close()
        ### Write info to csv file, row by row
        splittedtitle = record[0]["Title"].split(",")
        print(splittedtitle)
        result_writer.
 ↪writerow([record[0]["Id"],splittedtitle[0],splittedtitle[1],record[0]["AccessionVersion"],r
        ###
        counter += 1
### Close csv file
result_nucleotide.close()
```

2) dbVar Search - Pathogenic Copy Number Variation in Human

[ ]:
```
### Creating query
CNV = []
CNV_esearch = eclient.esearch(db='dbVar',
            term='DPP8[All Fields] AND ("Homo sapiens"[Organism] AND "copy␣
 ↪number variation"[Variant Type] AND "Pathogenic"[clinical_interpretation])')
print("\nLoading currently available ids from dbVar...")
print("="*50)
print("dbVar ids: ")
print(CNV_esearch.ids)
for item in CNV_esearch.ids:
    CNV.append(item)
print("\nSearch results: {}\n".format(CNV_esearch.count))
```

[ ]:
```
### Esummary for retrieving information
Entrez.email = "iris.raes@hotmail.com"
### For each id in CNV
### Save data in csv file
with open('results-CNV-dbVar.csv', mode='w') as result_CNV:
    result_writer = csv.writer(result_CNV,delimiter=';')
    result_writer.
 ↪writerow(["CNV_variant_id","variant_region_id","type","study_ID","clinical_assertion","asse
    counter = 1
    for ids in CNV:
        handle = Entrez.esummary(db="dbVar", id=ids)
        record = Entrez.read(handle)
```

```
        handle.close()
        varregid = record['DocumentSummarySet']['DocumentSummary'][0].get('SV')
        types = record['DocumentSummarySet']['DocumentSummary'][0].
↪get('dbVarVariantTypeList')
        studyid = record['DocumentSummarySet']['DocumentSummary'][0].get('ST')
        clinicalassertion = record['DocumentSummarySet']['DocumentSummary'][0].
↪get('dbVarClinicalSignificanceList')
        if␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'] !=␣
↪[]:
            assembly1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Assembly')
            start1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Chr_start')
            end1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Chr_end')
            assembly2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Assembly')
            start2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Chr_start')
            end2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Chr_end')
        ### Write info to csv file, row by row
        result_writer.
↪writerow([ids,varregid,types,studyid,clinicalassertion,assembly1+":
↪"+start1+"-"+end1,assembly2+":"+start2+"-"+end2])
        ###
        counter += 1
### Close csv file
result_CNV.close()
```

3) dbVar Search - Insertions in Human

[ ]:
```
### Creating query
insertion_esearch = eclient.esearch(db='dbVar',
            term='DPP8[All Fields] AND ("Homo sapiens"[Organism] AND␣
↪"insertion"[Variant Type])')
print("\nLoading currently available ids from dbVar...")
print("="*50)
print("dbVar ids: ")
```

```
print(insertion_esearch.ids)
print("\nSearch results: {}\n".format(insertion_esearch.count))
```

[ ]:
```
### Esummary for retrieving information
Entrez.email = "iris.raes@hotmail.com"
### For each id in insertion
### Save data in csv file
with open('results-insertion-dbVar.csv', mode='w') as result_insertion:
    result_writer = csv.writer(result_insertion,delimiter=';')
    result_writer.
↪writerow(["insertion_variant_id","variant_region_id","type","study_ID","assembly1","assembl
    counter = 1
    for ids in insertion:
        handle = Entrez.esummary(db="dbVar", id=ids)
        record = Entrez.read(handle)
        handle.close()
        varregid = record['DocumentSummarySet']['DocumentSummary'][0].get('SV')
        types = record['DocumentSummarySet']['DocumentSummary'][0].
↪get('dbVarVariantTypeList')
        studyid = record['DocumentSummarySet']['DocumentSummary'][0].get('ST')
        if␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'] !=␣
↪[]:
            assembly1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Assembly')
            start1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Chr_start')
            end1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Chr_end')
            assembly2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Assembly')
            start2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Chr_start')
            end2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Chr_end')
        ### Write info to csv file, row by row
        result_writer.writerow([ids,varregid,types,studyid,assembly1+":
↪"+start1+"-"+end1,assembly2+":"+start2+"-"+end2])
        ###
        counter += 1
```

```
### Close csv file
result_insertion.close()
```

4) dbVar Search - Inversions in Human

[ ]:
```
### Creating query
inversion_esearch = eclient.esearch(db='dbVar',
            term='DPP8[All Fields] AND ("Homo sapiens"[Organism] AND␣
 ↪"inversion"[Variant Type])')
print("\nLoading currently available ids from dbVar...")
print("="*50)
print("dbVar ids: ")
print(inversion_esearch.ids)
print("\nSearch results: {}\n".format(inversion_esearch.count))
```

[ ]:
```
### Esummary for retrieving information
Entrez.email = "iris.raes@hotmail.com"
### For each id in inversion
### Save data in csv file
with open('results-inversion-dbVar.csv', mode='w') as result_inversion:
    result_writer = csv.writer(result_inversion,delimiter=';')
    result_writer.
 ↪writerow(["inversion_variant_id","variant_region_id","type","study_ID","assembly1","assembl
    counter = 1
    for ids in inversion:
        handle = Entrez.esummary(db="dbVar", id=ids)
        record = Entrez.read(handle)
        handle.close()
        varregid = record['DocumentSummarySet']['DocumentSummary'][0].get('SV')
        types = record['DocumentSummarySet']['DocumentSummary'][0].
 ↪get('dbVarVariantTypeList')
        studyid = record['DocumentSummarySet']['DocumentSummary'][0].get('ST')
        if␣
 ↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'] !=␣
 ↪[]:
            assembly1 =␣
 ↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
 ↪get('Assembly')
            start1 =␣
 ↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
 ↪get('Chr_start')
            end1 =␣
 ↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
 ↪get('Chr_end')
```

```
            assembly2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Assembly')
            start2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Chr_start')
            end2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Chr_end')
        ### Write info to csv file, row by row
        result_writer.writerow([ids,varregid,types,studyid,assembly1+":
↪"+start1+"-"+end1,assembly2+":"+start2+"-"+end2])
        ###
        counter += 1
### Close csv file
result_inversion.close()
```

5) dbVar Search - Short Tandem Repeats in Human (seems to be less important)

[ ]:
```
### Creating query
STR_esearch = eclient.esearch(db='dbVar',
            term='DPP8[All Fields] AND ("Homo sapiens"[Organism] AND "short␣
↪tandem repeat"[Variant Type])')
print("\nLoading currently available ids from dbVar...")
print("="*50)
print("dbVar ids: ")
print(STR_esearch.ids)
print("\nSearch results: {}\n".format(STR_esearch.count))
```

[ ]:
```
### Esummary for retrieving information
Entrez.email = "iris.raes@hotmail.com"
### For each id in STR
### Save data in csv file
with open('results-STR-dbVar.csv', mode='w') as result_STR:
    result_writer = csv.writer(result_STR,delimiter=';')
    result_writer.
↪writerow(["STR_variant_id","variant_region_id","type","study_ID","assembly1","assembly2"])
    counter = 1
    for ids in STR:
        handle = Entrez.esummary(db="dbVar", id=ids)
        record = Entrez.read(handle)
        handle.close()
        varregid = record['DocumentSummarySet']['DocumentSummary'][0].get('SV')
        types = record['DocumentSummarySet']['DocumentSummary'][0].
↪get('dbVarVariantTypeList')
        studyid = record['DocumentSummarySet']['DocumentSummary'][0].get('ST')
```

```
        if␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'] !=␣
↪[]:
            assembly1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Assembly')
            start1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Chr_start')
            end1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][0].
↪get('Chr_end')
            assembly2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Assembly')
            start2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Chr_start')
            end2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['dbVarPlacementList'][1].
↪get('Chr_end')
        ### Write info to csv file, row by row
        result_writer.writerow([ids,varregid,types,studyid,assembly1+":
↪"+start1+"-"+end1,assembly2+":"+start2+"-"+end2])
        ###
        counter += 1
### Close csv file
result_STR.close()
```

6) ClinVar Search - Genetic Variations in Human

[ ]:
```
### Creating query
ClinVar_esearch = eclient.esearch(db='ClinVar',
            term='DPP8[gene] AND "Single gene"')
print("\nLoading currently available ids from ClinVar...")
print("="*50)
print("\nClinVar ids: ")
print(ClinVar_esearch.ids)
print("\nSearch results: {}\n".format(ClinVar_esearch.count))
```

[ ]:
```
### Esummary for retrieving information
Entrez.email = "iris.raes@hotmail.com"
### For each id in ClinVar
### Save data in csv file
with open('results-ClinVar.csv', mode='w') as result_ClinVar:
    result_writer = csv.writer(result_ClinVar,delimiter=';')
```

```
   result_writer.
↪writerow(["ClinVar_variant_id","title","accession","type","description","protein_change","a
   counter = 1
   for ids in ClinVar:
       handle = Entrez.esummary(db="ClinVar", id=ids)
       record = Entrez.read(handle)
       handle.close()
       title = record['DocumentSummarySet']['DocumentSummary'][0].get('title')
       accession = record['DocumentSummarySet']['DocumentSummary'][0].
↪get('accession_version')
       types = record['DocumentSummarySet']['DocumentSummary'][0].
↪get('obj_type')
       description =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['clinical_significance'].
↪get('description')
       protein_change = record['DocumentSummarySet']['DocumentSummary'][0].
↪get('protein_change')
       if␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['variation_set'][0]['variation_loc']␣
↪!= []:
           assembly1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['variation_set'][0]['variation_loc'][0].
↪get('assembly_name')
           start1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['variation_set'][0]['variation_loc'][0].
↪get('start')
           end1 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['variation_set'][0]['variation_loc'][0].
↪get('stop')
           if␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['variation_set'][0].
↪get('variation_loc') != []:
               try:
                   assembly2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['variation_set'][0]['variation_loc'][1].
↪get('assembly_name')
                   start2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['variation_set'][0]['variation_loc'][1].
↪get('start')
                   end2 =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['variation_set'][0]['variation_loc'][1].
↪get('stop')
               except:
                   assembly2 = "not applicable"
                   start2 = "not applicable"
                   end2 = "not applicable"
```

```
        if record['DocumentSummarySet']['DocumentSummary'][0]['trait_set'] !=␣
↪[]:
            dbsource =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['trait_set'][0]['trait_xrefs'][0]['db_so
            dbid =␣
↪record['DocumentSummarySet']['DocumentSummary'][0]['trait_set'][0]['trait_xrefs'][0]['db_id
        ### Write info to csv file, row by row
        result_writer.
↪writerow([ids,title,accession,types,description,protein_change,assembly1+":
↪"+start1+"-"+end1,assembly2+":"+start2+"-"+end2,dbsource+" ("+dbid+")"])
        ###
        counter += 1
### Close csv file
result_ClinVar.close()
```