Task 4. Download Data from Pubtator and Extract Gene Annotation

Github link: https://github.com/Allen-ZKW/NLP_HZAU/tree/Task4

Abstract

In this Task, we try to use a script which is supported by teacher to download pubtator data. Then we classify and extract information for raw data inorder to learn from related papers.

Measure

1) Using Edirect to Get Papers' Abstract (Edirect on Linux)

```
esearch -db pubmed -query "Oryza sativa L"|efilter -pub english|efilter -pub
abstract|efetch -format json > abstract.json
```

2) Filter the Abstract and Get PMID (Python 3.8 64-bit)

```
def get_pmid(dirpath):
    datapath = dirpath + 'data/'
    raw_data = []
    with open(datapath + 'abstract.json','r',encoding = 'utf-8') as f:
        for row in f:
            if row[0:4]=='PMID':
                raw_data.append(row.strip())
    pmids = []
    for i in raw_data:
        pmids.append(i[6:14])
    with open(datapath + 'PMID.txt','w') as f:
        for i in pmids:
            f.write(i+"\n")
    return pmids

def main():
    dirpath = "D:/junior_n/NLP/term/"
    pmids = get_pmid(dirpath)

main()
```

2) Run Script in The Background (Linux)

Before runing the script, we learn and try to understand the script, then we also write some annotation to avoid forgetting

```
#!/bin/bash
F_OUT="result_MM_Pubtator.txt" #A file to write the result of this script
F_LIST="PMID.txt" #A file which record a list of PMID
```

```
echo -e "\n I am curating the result of Oryza sativa L.\n" #Alart information
echo -e "\n" >$F_OUT #Seperate results from different PMID

i=1 #Count the index of PMID
while IFS= read -r line #Do this loop untill all PMID being downloaded
do
        xx=`echo https://www.ncbi.nlm.nih.gov/research/pubtator-
api/publications/export/pubtator?pmids=$line` # API of pubtator
        curl $xx >>$F_OUT #Using curl to write the result to file
        printf "$i -th result out of xxxxx is processing...\n" #Tell using when
this job is complete
        i=$[i+1] # move to the next PMID
        sleep 5.8s #In order to avoid pubtator ban IP
done <"$F_LIST"
```

After we run this script in background because of a huge sum of time it will take.

```
nohup bash ./NCBI3.sh &
ps -ef|grep NCBI3.sh
exit
```

3) Extract Gene Annotation (Python 3.8 64-bit)

By reading the Pubtator result, It's easy to sepreate all information to three parts which have their own unique form.

$$Title \rightarrow PMID + |t| + Title \quad of \quad this \quad artile$$
$$Abstract \rightarrow PMID + |a| + Abstract \quad of \quad this \quad artile$$
$$Annotation \rightarrow PMID + Start + End + Ontology + Classification + Additional \quad information$$

Understanding the rules, we can filter the raw data and get information we want. In this task, we only need Gene Ontology , so I write a script to get annotation which is related about gene information.

```
import json

def pubtator_reader(dirpath):#function to read and classify pubtator infromation
    datapath = dirpath + 'data/'
    p_d = {}
    with open(datapath + 'result_MM_Pubtator.txt','r',encoding='utf-8') as f:
        for row in f:
            if '|t|' in row:
                key = row.split('|t|')[0]
                title = row.split('|t|')[1].strip()+' '
                p_d[key] = {}
                p_d[key]['paper'] = title
                p_d[key]['annotation'] = []
            elif '|a|' in row:
                key = row.split('|a|')[0]
                abstract = row.split('|a|')[1].strip()+' '
                p_d[key]['paper'] = p_d[key]['paper'] + abstract
                start = [0]
                stop = []
                for i in range(len(p_d[key]['paper'])):
                    if p_d[key]['paper'][i:i+2] == '. ':
                        stop.append(i+2)
```

```
                        start.append(i+2)
                del start[-1]
                sentence = []
                for i in range(len(start)):
                    sentence.append([start[i],stop[i]])
                p_d[key]['sentence'] = sentence
            elif row != '\n':
                note = row.strip().split('\t')
                if note[4] == 'Gene':
                    p_d[key]['annotation'].append(note)
    e = json.dumps(p_d)
    with open(datapath + 'pubtator_data.json','w') as f:
        f.write(e)
    return p_d

def main():
    dirpath = "D:/junior_n/NLP/term/"
    pubtator_data = pubtator_reader(dirpath)

main()
```

Result

Using PMID: 33871646 as an example

1) Edirect Result (.txt)

```
16. Plant Physiol. 2021 Apr 19. pii: kiab175. doi: 10.1093/plphys/kiab175. [Epub
ahead of print]
Post-Golgi Trafficking of Rice Storage Proteins Requires the Small GTPase Rab7
Activation Complex MON1-CCZ1.
Pan T(1), Wang Y(1), Jing R(1), Wang Y(1), Wei Z(2), Zhang B(2), Lei C(2), Qi
Y(2), Wang F(1), Bao X(1), Yan M(2), Zhang Y(1), Zhang P(1), Yu M(1), Wan G(2),
Chen Y(2), Yang W(2), Zhu J(1), Zhu Y(2), Zhu S(2), Cheng Z(2), Zhang X(2),
Jiang
L(1), Ren Y(2), Wan J(1)(2).
Author information:
(1)State Key Laboratory for Crop Genetics and Germplasm Enhancement, Jiangsu
Plant Gene Engineering Research Center, Nanjing Agricultural University, Nanjing

210095, China.
(2)National Key Facility for Crop Gene Resources and Genetic Improvement,
Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing
100081, China.
Protein storage vacuoles (PSVs) are unique organelles that accumulate storage
proteins in plant seeds. Although morphological evidence points to the existence

of multiple PSV-trafficking pathways for storage protein targeting, the
molecular
mechanisms that regulate these processes remain mostly unknown. Here, we report
the functional characterization of the rice (Oryza sativa) glutelin precursor
accumulation7 (gpa7) mutant, which over-accumulates 57-kD glutelin precursors in

dry seeds. Cytological and immunocytochemistry studies revealed that the gpa7
mutant exhibits abnormal accumulation of storage pre-vacuolar compartment-like
structures, accompanied by the partial mistargeting of glutelins to the
```

extracellular space. The gpa7 mutant was altered in the CCZ1 locus, which encodes
the rice homolog of Arabidopsis (Arabidopsis thaliana) CALCIUM CAFFEINE ZINC SENSITIVITY1a (CCZ1a) and CCZ1b. Biochemical evidence showed that rice CCZ1 interacts with MONENSIN SENSITIVITY1 (MON1) and that these proteins function together as the Rat brain 5 (Rab5) effector and the Rab7 guanine nucleotide exchange factor (GEF). Notably, loss of CCZ1 function promoted the endosomal localization of Vacuolar Protein Sorting-associated protein 9 (VPS9), which is the GEF for Rab5 in plants. Together, our results indicate that the MON1-CCZ1 complex is involved in post-Golgi trafficking of rice storage protein through a Rab5 and Rab7-dependent pathway.

DOI: 10.1093/plphys/kiab175
PMID: 33871646

2) PMID Information (.txt)

```
33871646
```

3) Pubtator Raw Data (.txt)

```
33871646|t|Post-Golgi Trafficking of Rice Storage Proteins Requires the Small GTPase Rab7 Activation Complex MON1-CCZ1.
33871646|a|Protein storage vacuoles (PSVs) are unique organelles that accumulate storage proteins in plant seeds. Although morphological evidence points to the existence of multiple PSV-trafficking pathways for storage protein targeting, the molecular mechanisms that regulate these processes remain mostly unknown. Here, we report the functional characterization of the rice (Oryza sativa) glutelin precursor accumulation7 (gpa7) mutant, which over-accumulates 57-kD glutelin precursors in dry seeds. Cytological and immunocytochemistry studies revealed that the gpa7 mutant exhibits abnormal accumulation of storage pre-vacuolar compartment-like structures, accompanied by the partial mistargeting of glutelins to the extracellular space. The gpa7 mutant was altered in the CCZ1 locus, which encodes the rice homolog of Arabidopsis (Arabidopsis thaliana) CALCIUM CAFFEINE ZINC SENSITIVITY1a (CCZ1a) and CCZ1b. Biochemical evidence showed that rice CCZ1 interacts with MONENSIN SENSITIVITY1 (MON1) and that these proteins function together as the Rat brain 5 (Rab5) effector and the Rab7 guanine nucleotide exchange factor (GEF). Notably, loss of CCZ1 function promoted the endosomal localization of Vacuolar Protein Sorting-associated protein 9 (VPS9), which is the GEF for Rab5 in plants. Together, our results indicate that the MON1-CCZ1 complex is involved in post-Golgi trafficking of rice storage protein through a Rab5 and Rab7-dependent pathway.
33871646    26  30  Rice    Species 4530
33871646    74  78  Rab7    Gene    29448
33871646    103 107 CCZ1    Gene    360768
33871646    469 473 rice    Species 4530
33871646    475 487 Oryza sativa    Species 4530
33871646    874 878 CCZ1    Gene    360768
33871646    904 908 rice    Species 4530
33871646    920 931 Arabidopsis Species 3702
33871646    933 953 Arabidopsis thaliana    Species 3702
33871646    955 990 CALCIUM CAFFEINE ZINC SENSITIVITY1a Gene    838172
```

```
33871646    992 997 CCZ1a   Gene    838172
33871646    1003    1008    CCZ1b   Gene    844431
33871646    1043    1047    rice    Species 4530
33871646    1048    1052    CCZ1    Gene    360768
33871646    1182    1186    Rab7    Gene    29448
33871646    1187    1221    guanine nucleotide exchange factor  Gene    362799
33871646    1223    1226    GEF Gene    362799
33871646    1246    1250    CCZ1    Gene    360768
33871646    1366    1369    GEF Gene    362799
33871646    1435    1439    CCZ1    Gene    360768
33871646    1489    1493    rice    Species 4530
33871646    1529    1533    Rab7    Gene    29448
```

4) Pubtator Filtered Data (.json)


```
"33871646": {"paper": "Post-Golgi Trafficking of Rice Storage Proteins Requires the Small GTPase Rab7 Activation Complex MON1-CCZ1. Protein storage vacuoles (PSVs) are unique organelles that accumulate storage proteins in plant seeds. Although morphological evidence points to the existence of multiple PSV-trafficking pathways for storage protein targeting, the molecular mechanisms that regulate these processes remain mostly unknown. Here, we report the functional characterization of the rice (Oryza sativa) glutelin precursor accumulation7 (gpa7) mutant, which over-accumulates 57-kD glutelin precursors in dry seeds. Cytological and immunocytochemistry studies revealed that the gpa7 mutant exhibits abnormal accumulation of storage pre-vacuolar compartment-like structures, accompanied by the partial mistargeting of glutelins to the extracellular space. The gpa7 mutant was altered in the CCZ1 locus, which encodes the rice homolog of Arabidopsis (Arabidopsis thaliana) CALCIUM CAFFEINE ZINC SENSITIVITY1a (CCZ1a) and CCZ1b. Biochemical evidence showed that rice CCZ1 interacts with MONENSIN SENSITIVITY1 (MON1) and that these proteins function together as the Rat brain 5 (Rab5) effector and the Rab7 guanine nucleotide exchange factor (GEF). Notably, loss of CCZ1 function promoted the endosomal localization of Vacuolar Protein Sorting-associated protein 9 (VPS9), which is the GEF for Rab5 in plants. Together, our results indicate that the MON1-CCZ1 complex is involved in post-Golgi trafficking of rice storage protein through a Rab5 and Rab7-dependent pathway. ", "annotation": [["33871646", "74", "78", "Rab7", "Gene", "29448"], ["33871646", "103", "107", "CCZ1", "Gene", "360768"], ["33871646", "874", "878", "CCZ1", "Gene", "360768"], ["33871646", "955", "990", "CALCIUM CAFFEINE ZINC SENSITIVITY1a", "Gene", "838172"], ["33871646", "992", "997", "CCZ1a", "Gene", "838172"], ["33871646", "1003", "1008", "CCZ1b", "Gene", "844431"], ["33871646", "1048", "1052", "CCZ1", "Gene", "360768"], ["33871646", "1182", "1186", "Rab7", "Gene", "29448"], ["33871646", "1187", "1221", "guanine nucleotide exchange factor", "Gene", "362799"], ["33871646", "1223", "1226", "GEF", "Gene", "362799"], ["33871646", "1246", "1250", "CCZ1", "Gene", "360768"], ["33871646", "1366", "1369", "GEF", "Gene", "362799"], ["33871646", "1435", "1439", "CCZ1", "Gene", "360768"], ["33871646", "1529", "1533", "Rab7", "Gene", "29448"]], "sentence": [[0, 109], [109, 212], [212, 414], [414, 600], [600, 839], [839, 1010], [1010, 1229], [1229, 1390], [1390, 1553]]}
```


Discussion

Because this task is realted with our term in this class, so actually, all codes above is a part of my term code, so if your are interested in this task, read the branch:term can be a very good choice.

Althouiugh this measure is good enough, there still exist some limits or disadvantages: First, it takes too much time to downlod paper about Oryza sativa L, nearly three days. Additionally, there are some ontologys which are not covered by Pubtator.

To solve this problem, we think this way may be useful. First, we wirte annotation for articals very very detailed. Then, we use these data to train CRF model and use this model to predict the tags of every words. The final result combines both predict result and Pubtator annotation. Because of no spare time, maybe in further study, we will have choice to try the is measure.