利用shell编程实现NLP

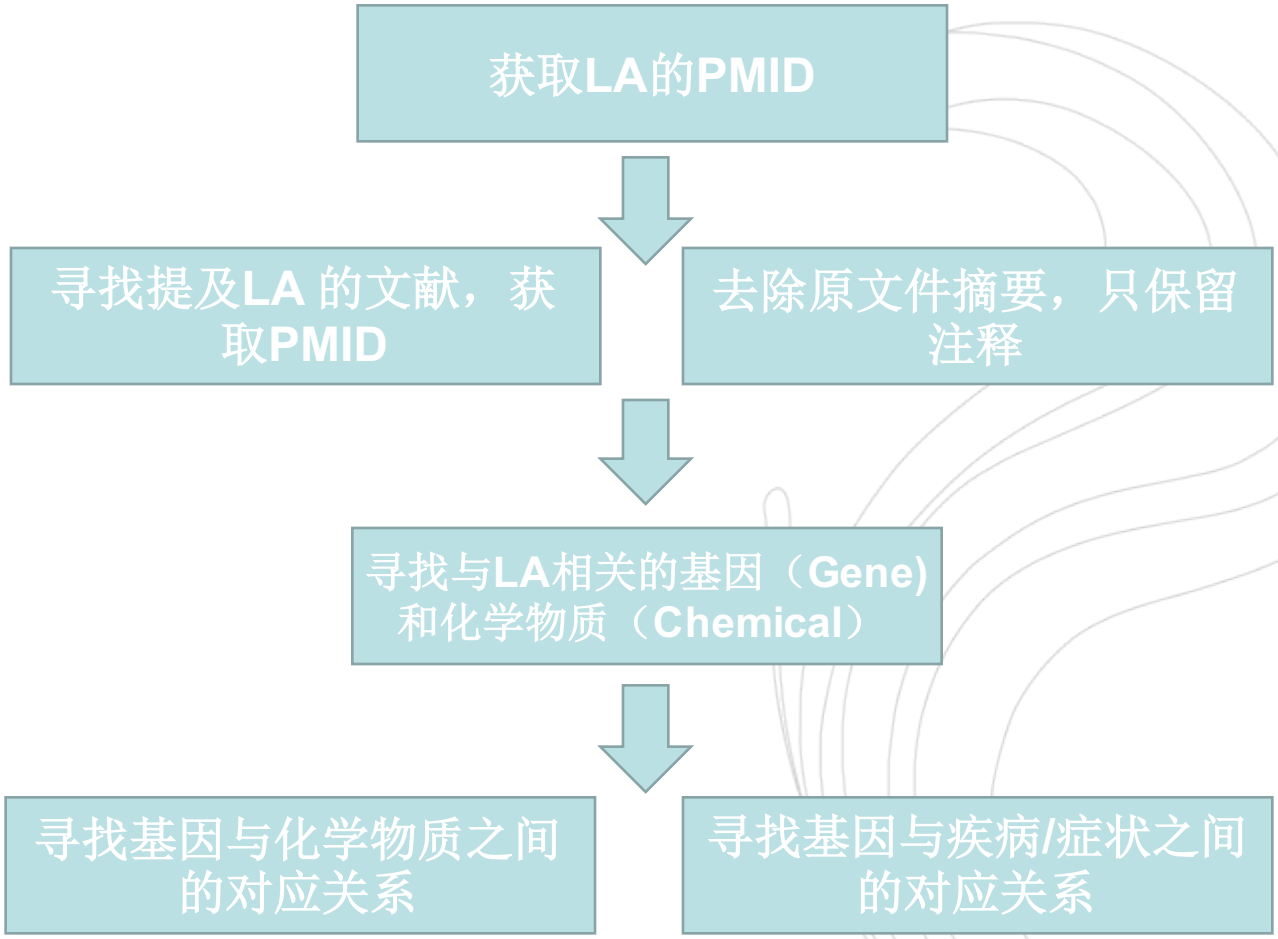
薛雅文

Lung Adenocarcinoma

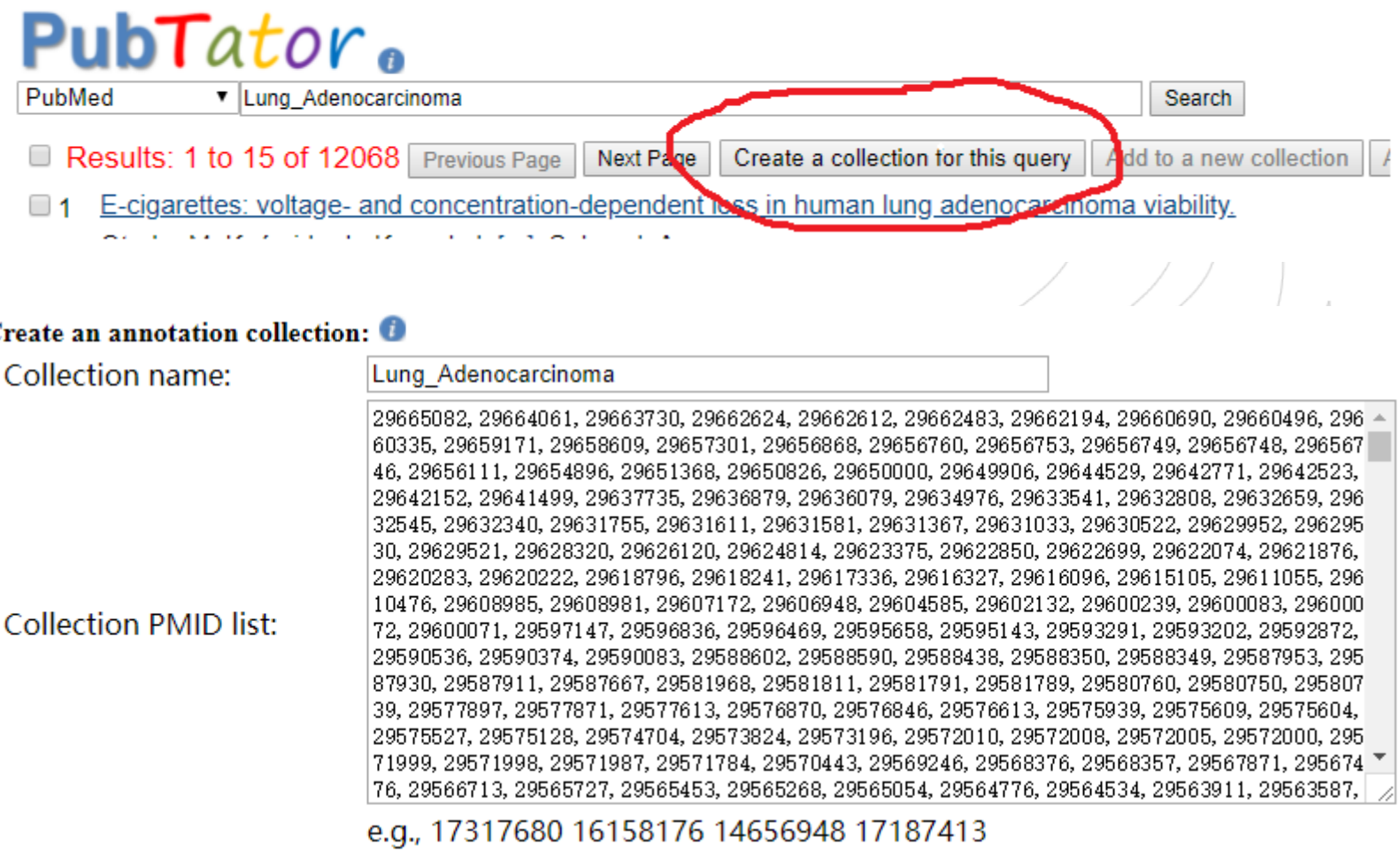
- 简介：肺腺癌：是肺癌的一种，属于非小细胞癌 。通常起源于支气管粘膜上皮，少数起源于大支气管的粘液腺，发病年龄较小，女性相对多见。

- 发病原因：吸烟
  大气污染
  职业因素长期接触铀镭等放射性物质
  肺部慢性疾病如肺结核、尘肺

Workflow


获取LA的PMID
寻找提及LA 的文献，获取PMID
去除原文件摘要，只保留注释
寻找与LA相关的基因（Gene)和化学物质（Chemical）
寻找基因与化学物质之间的对应关系
寻找基因与疾病/症状之间的对应关系

从Pubtator上下载LA的PMID



寻找提及LA的文献，提取PMID

```
29660690|t|New 1,2,4-triazole-Chalcone hybrids induce Caspase-3 dependent apoptosis in A549 human lung aden
29660690|a|A series of novel 1, 2, 4-triazole/chalcone hybrids was prepared and identified with different s
repared compounds showed remarkable cytotoxic activity against different cancer cell lines. Compounds 24, 2
 highest cytotoxicity among the tested compounds against human lung adenocarcinoma A549   cells with IC<sub
16.04    M compared to cisplatin with IC<sub>50</sub> of 15.3    M. Flow cytometric analysis of the tested
n the number of apoptotic cells in a dose-dependent manner. The further mechanistic study demonstrated that
rids induced apoptosis via increased level of proapoptotic protein Bax, release of cytochrome c from mitoch
se-3/8/9 proteins. However, general caspase inhibition by the pan-caspase inhibitor, z-VAD-fmk, significant
uced by the tested hybrids, suggesting dependency of apoptosis on activation of the caspase-3 pathway.
29660690        4       27      1,2,4-triazole-Chalcone Chemical
29660690        43      52      Caspase-3       Gene    836
29660690        81      86      human   Species 9606
29660690        87      106     lung adenocarcinoma     Disease C538231
29660690        132     148     1, 2, 4-triazole        Chemical        C045575
29660690        149     157     chalcone        Chemical        D002599
29660690        313     319     cancer  Disease D009369
29660690        4       27      1,2,4-triazole-Chalcone Chemical
29660690        43      52      Caspase-3       Gene    836
29660690        81      86      human   Species 9606
29660690        87      106     lung adenocarcinoma     Disease C538231
29660690        132     148     1, 2, 4-triazole        Chemical        C045575
29660690        149     157     chalcone        Chemical        D002599
29660690        313     319     cancer  Disease D009369
```

```
81307
139205
140093
155940
278851
294062
373915
422569
479125
636378
667821
722727
903615
943561
947267
954007
991153
1159836
1185808
1248011
1280979
1287141
1294457
```

Command：*grep -i 'lung adenocarcinoma' laresult.txt | awk '{print $1}' | sort -n | uniq | awk -F '|' '{print $1}' | uniq > laID.txt*

去除原文件摘要，只保留注释

```
29660690        4       27      1,2,4-triazole-Chalcone Chemical
29660690        43      52      Caspase-3       Gene    836
29660690        81      86      human   Species 9606
29660690        87      106     lung adenocarcinoma     Disease C538231
29660690        132     148     1, 2, 4-triazole        Chemical        C045575
29660690        149     157     chalcone        Chemical        D002599
29660690        313     319     cancer  Disease D009369
29660690        386     398     cytotoxicity    Disease D064420
29660690        434     439     human   Species 9606
29660690        440     459     lung adenocarcinoma     Disease C538231
29660690        537     546     cisplatin       Chemical        D002945
29660690        759     784     1, 2, 4-triazole-chalcone       Chemical
29660690        855     858     Bax     Gene    581
29660690        871     883     cytochrome c    Gene    54205
29660690        920     929     caspase-3       Gene    836
29660690        1010    1019    z-VAD-fmk       Chemical
29660690        1146    1155    caspase-3       Gene    836
29660496        65      73      tyrosine        Chemical        D014443
29660496        112     116     EGFR    Gene    1956
29660496        125     144     lung adenocarcinoma     Disease C538231
29660496        146     157     Lung cancer     Disease D008175
29660496        197     204     cancers Disease D009369
29660496        206     214     Patients        Species 9606
29660496        256     288     epidermal growth factor receptor        Gene    1956
29660496        290     294     EGFR    Gene    1956
29660496        307     311     EGFR    Gene    1956
29660496        312     320     tyrosine        Chemical        D014443
29660496        413     421     patients        Species 9606
```

Command：*grep '^[0-9]\{8\}[^\|]' laresult.txt > ref.txt*

寻找与LA有关的基因和化学物质

```
#!/bin/bash
# find gene about LA

echo -e "data pre-processing"

grep -i 'lung adenocarcinoma' laresult.txt | awk '{print $1}' | sort -n | uniq | awk -F '|' '{print $1}' | uniq > laID.txt
grep '^[0-9]\{8\}[^|]' laresult.txt > ref.txt

echo -e "data pre-processing done!"

F_id="laID.txt"
F_ref="ref.txt"

echo -e "\n start finding gene and chemical\n"

while IFS= read -r line
do
    grep $line $F_ref | grep 'Gene' | awk -F '\t'  '{print $4}' >> gene.txt
    grep $line $F_ref | grep 'Chemical' | awk -F '\t'  '{print $4}' >> chemical.txt
done < "$F_id"

echo -e "finish finding!"
echo -e "start remove duplicates"

cat gene.txt | sort | uniq > uniq_gene.txt
cat chemical.txt | sort | uniq > uniq_chemical.txt

echo "finish remove!"
"findGeneandChemical.sh" 28L, 737C written                                    3,0-1
```

提取Gene和Chemical

排序，去重

结果文件

uniq_gene.txt

```
14-3-3 isoforms (theta, epsilon, and sigma) and annexin A5
14-3-3 sigma
14-3-3zeta
15-hydroxyprostaglandin dehydrogenase
15-Hydroxyprostaglandin dehydrogenase
15-lipoxygenase
15-PGDH
37-kDa laminin receptor
37-kDa laminin receptor precursor
37LRP
3-hydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase
3'-phosphoadenosine 5'-phosphosulfate synthase 1
-3-phosphoinositide-dependent protein kinase-1
4EBP1
4E-BP1
53BP1
5-lipoxygenase
78 kDa glucose-regulated protein
78-kDa glucose-regulated protein
8-oxo-guanine DNA glycosylase
8-oxoguanine DNA glycosylase
8-Oxoguanine DNA glycosylase
8-oxo-guanine glycosylase-1
```

uniq_chemical.txt

```
10-epi-7,10-epoxy-ar-bisabol-11-ol
10-epi-kalihinol X
10-formyltetrahydrofolate
10-hydroxy-aplysin
10-hydroxycamptothecin
10-hydroxy-debromoepiaplysin
10-hydroxy-epiaplysin
10-methoxy-7-methyl-2H-benzo[g]chromen-2-one
10<sup>-6</sup>
1,10-phenanthroline
1,10-Phenanthroline
11(13)-triene-6,12-olide
1,12-epoxybenz[a]anthracene
1,1'-(3,3'-dimethoxybiphenyl-4,4'-diyl)bis(thiourea)
(11)C
(11)C-acetate
11C-acetate
11C-Choline
(11)C-PDT
1,1' dioactadecyl-3-3-3',3'-tetramethylindotricarbocyanine iodide
```

寻找基因与化学物质之间的关系

```
#!/bin/bash

echo -e "start finding"
while IFS= read -r line1
do
    echo "gene: $line1" >> gc_relationship.txt
    grep "$line1" ref.txt | awk '{print $1}' | sort | uniq > tmp.txt
    while IFS= read -r line2
    do
        grep "$line2" ref.txt | grep 'Chemical' | awk -F '\t'  '{print $4}' >> gc_relationship.txt
    done < "tmp.txt"
done < "uniq_gene.txt"

echo -e "done!"
~
~
~
~
~
```

```
gene: 14-3-3 isoforms (theta, epsilon, and sigma) and annexin A5
nitrosamine
nitrosamine 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone
NNK
1,4-phenylenebis(methylene)selenocyanate
alpha-1-antitrypsin
NNK
gene: 14-3-3 sigma
manganese superoxide
gene: 14-3-3zeta
BH3
gene: 15-hydroxyprostaglandin dehydrogenase
NAD
phorbol ester
phorbol 12-myristate 13-acetate
PMA
PGE2
sodium butyrate
apicidin
oxamflatin
NAD
prostaglandins
Flurbiprofen
indomethacin
(R)-flurbiprofen
```

- Gene和Chemical唯一的联系是PMID
- tmp.txt：全部提及该gene的文献的PMID

```
#!/bin/bash

echo -e "start finding!"
while IFS= read -r line1
do
    echo "chemical  $line1" >> dc_relationship.txt
    grep "$line1" ref.txt | awk '{print $1}' | sort | uniq > tmp.txt
    while IFS= read -r line2
    do
        grep "$line2" ref.txt | grep 'Disease' | awk -F '\t' '{print $4}' >> dc_relationship.txt
    done < "tmp.txt"
done < "uniq_chemical.txt"

echo -e "done!"

~
~
~
~
~
```

```
chemical  10-epi-7,10-epoxy-ar-bisabol-11-ol
cancer
lung adenocarcinoma
stomach cancer
hepatoma
colon cancer
cytotoxicity
chemical  10-epi-kalihinol X
Hainan sponge Acanthella sp
lung adenocarcinoma
chemical  10-formyltetrahydrofolate
Cancers
tumors
cancer
cancers
cancer
lung adenocarcinomas
tumor
hepatocellular carcinoma
tumor
chemical  10-hydroxy-aplysin
cancer
lung adenocarcinoma
stomach cancer
hepatoma
```

Thanks !