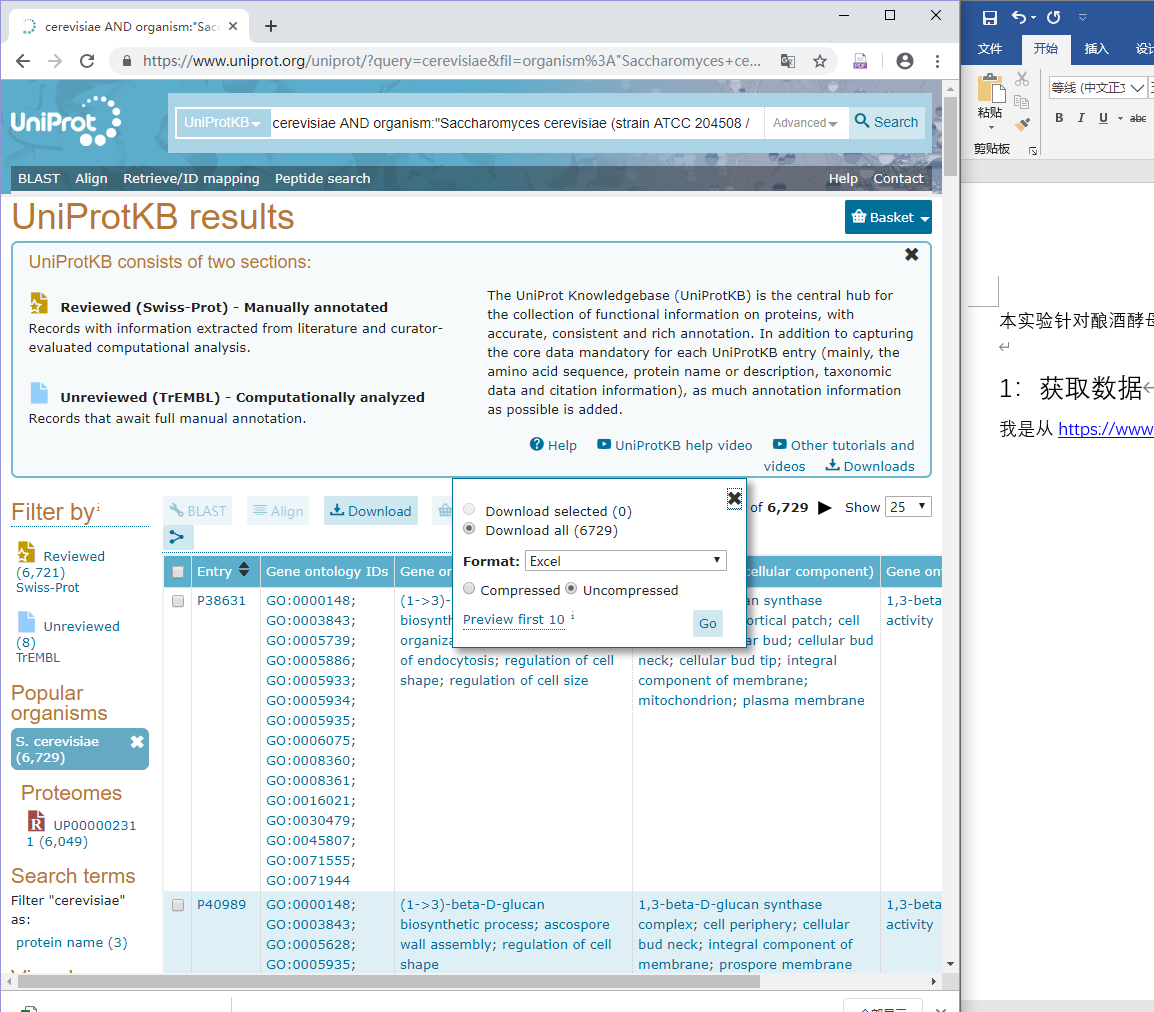
本实验针对酿酒酵母六千多种蛋白质进行逐一判断分析

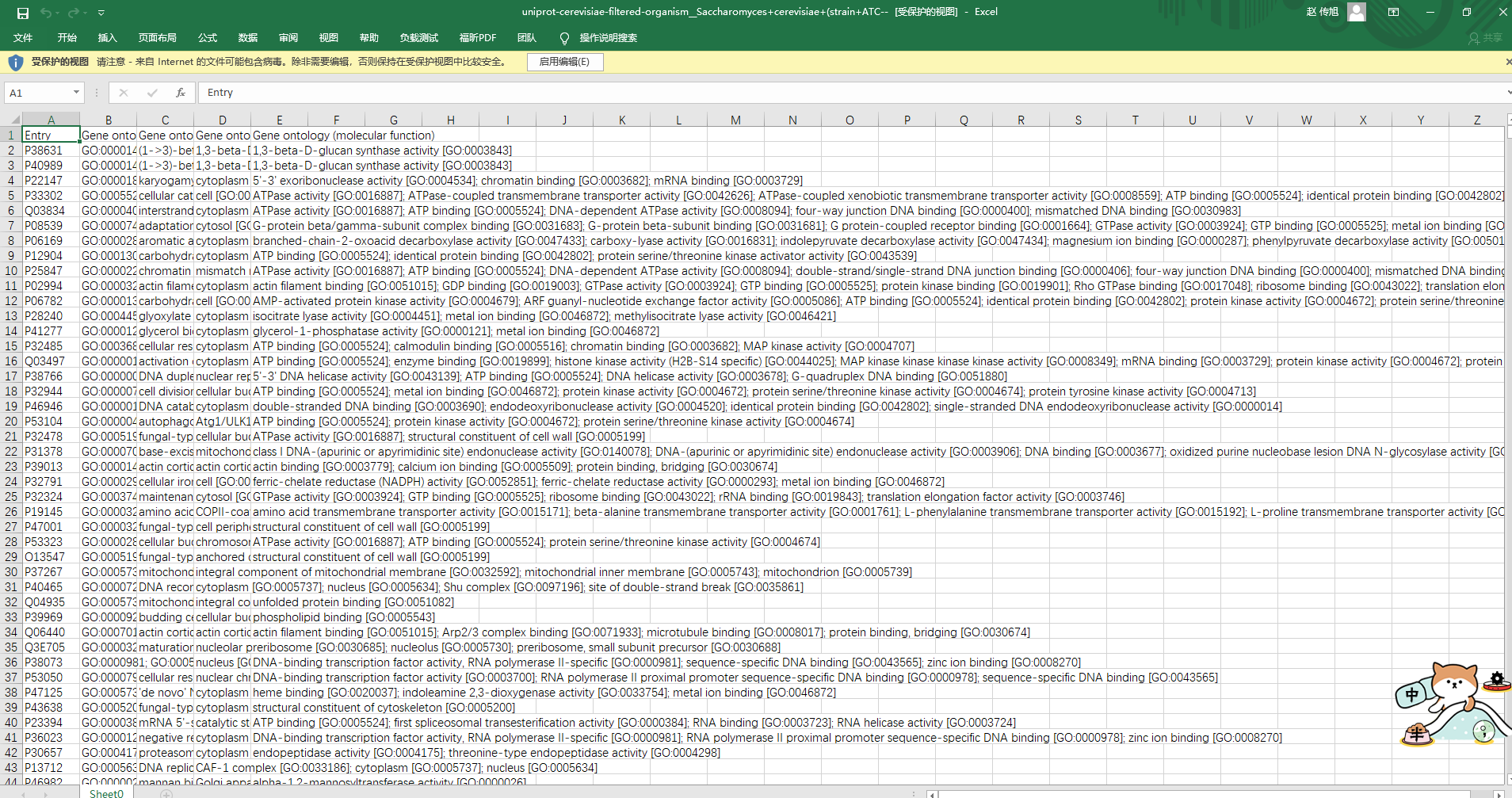
1：获取数据

我是从[https://www.uniprot.org](https://www.uniprot.org/)获取的数据 获取的酿酒酵母6729种蛋白质



获取的是一个execl文件

打开后如下：



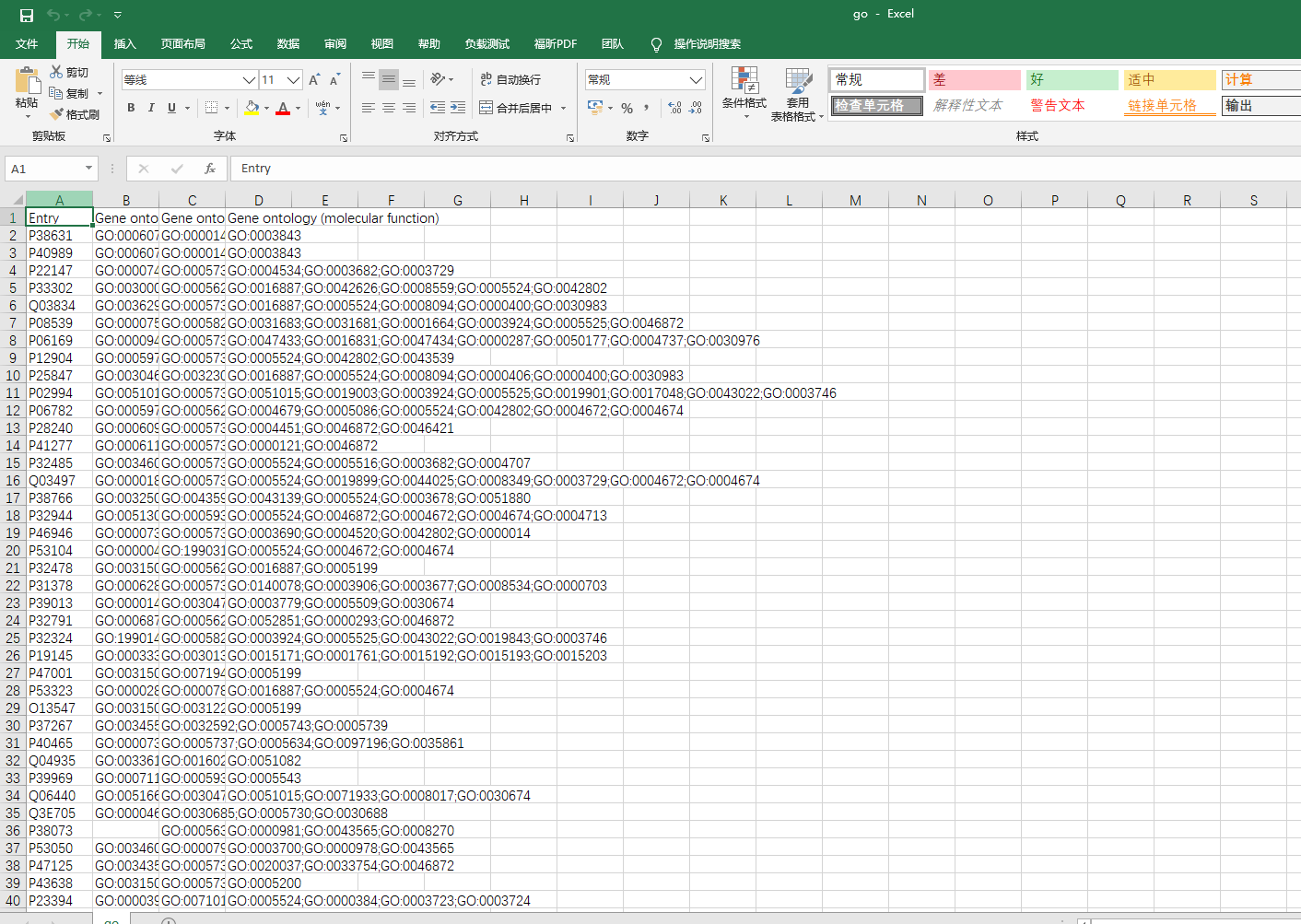
然后进行数据清洗 主要是将其中包含的基因本体洗出来

我用的python清洗

代码如下：

**import** pandas **as** pd  
**import** re  
  
*#t提取所有的*yeast = pd.read\_excel(**"./source.xlsx"**)  
*# print(yeast)  
  
#Entry = yeast['Entry']  
  
#读取Gene ontology (biological process) 生化过程*bp = yeast[**'Gene ontology (biological process)'**]  
bp = bp.fillna(value=**''**)  
  
*#正则表达式提取 用；分隔开***for** i **in** range(len(bp)):  
 temp = re.findall(**r"GO:\d{7}"**,bp[i])  
 bp[i]=**';'**.join(temp)  
 *#print(bp[i])  
  
#读取Gene ontology (cellular component) 细胞结构*cc = yeast[**'Gene ontology (cellular component)'**]  
cc = cc.fillna(value=**''**)  
  
*#用正则表达式提取 用；分隔开***for** i **in** range(len(cc)):  
 temp = re.findall(**r"GO:\d{7}"**,cc[i])  
 cc[i]=**';'**.join(temp)  
  
  
 *#print(i,cc[i])  
  
  
###操作字符串***""" 忽略了字符串定地质  
for j in cc:  
 for k in range(len(j)):  
 if j[k]=='","' and j[k+1]=='"':  
 m = k  
 for n in range(m,len(j)):  
 temp = j[n]  
 j[n+1]=temp  
 j[m]=','  
print(j)  
"""***# # #  
  
  
  
#读取Gene ontology (molecular function) 分子作用*mf = yeast[**"Gene ontology (molecular function)"**]  
mf = mf.fillna(value=**""**)  
  
**for** i **in** range(len(mf)):  
 temp = re.findall(**r"GO:\d{7}"**,mf[i])  
 mf[i]=**';'**.join(temp)  
 *#print(i,mf[i])  
  
  
#write*yeast[**'Gene ontology (biological process)'**] =bp  
yeast[**'Gene ontology (cellular component)'**]=cc  
yeast[**"Gene ontology (molecular function)"**]=mf  
  
yeast.to\_csv(**'go.csv'**,index=**False**,columns=[**'Entry'**,  
 **'Gene ontology (biological process)'**,  
 **'Gene ontology (cellular component)'**,  
 **"Gene ontology (molecular function)"**]  
  
  
 )

清洗的目的是获取出对应每个蛋白质基因本体的biological process（生化过程）、cellular component（细胞结构）、molecular function（分子作用）对应的基因提取出来提取结果如下



然后就用R语言进行操作

2：进行蛋白质判断GoSemSim

首先去<http://bioconductor.org/>

1. 安装GoSemSim <http://bioconductor.org/packages/2.6/bioc/html/GOSemSim.html>

两种安装方式 1：下载压缩包

2：source("http://bioconductor.org/biocLite.R")

biocLite("GOSemSim")

上边是官网提供的 但是不安全直接用?BiocManager::install（package）进行下载 但是前提是现在?BiocManager

1. 获取数据库

<http://bioconductor.org/packages/3.10/data/annotation/html/org.Sc.sgd.db.html>

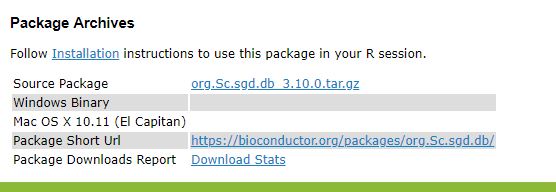
可以直接安装这个package 代码：

if (!requireNamespace("BiocManager", quietly = TRUE))

install.packages("BiocManager")

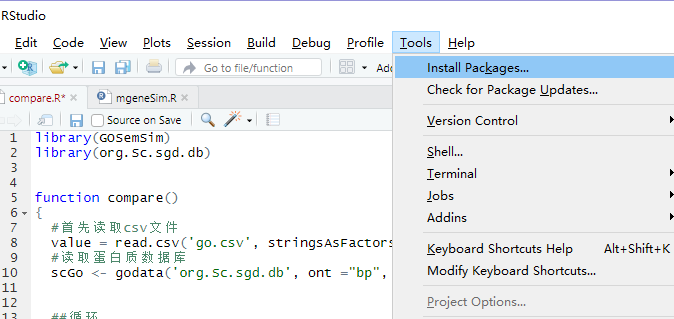
BiocManager::install("org.Sc.sgd.db")

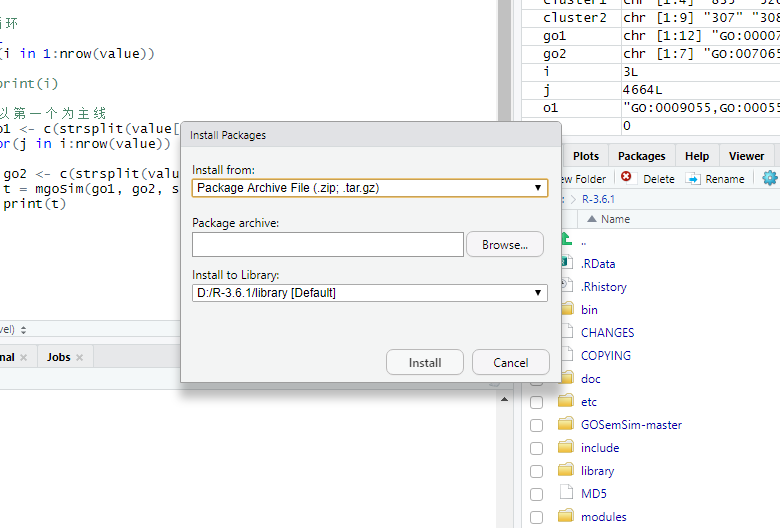
第二种方法是下载这个源文件



最下边

然后再RProject安装这个包





然后，载入GoSemSim各个函数，用于对蛋白质序列集的比较 我用的是mgoSim

然后我写了个compare脚本进行双重循环遍历比较

library(GOSemSim)

library(org.Sc.sgd.db)

function compare()

{

#首先读取csv文件

value = read.csv('go.csv', stringsAsFactors = F)

#读取蛋白质数据库

scGo <- godata('org.Sc.sgd.db', ont ="bp", computeIC = F)

##循环

i<-1

for(i in 1:nrow(value))

{

#print(i)

#以第一个为主线

go1 <- c(strsplit(value[i,2], split =";")[[1]])

for(j in i:nrow(value))

{

go2 <- c(strsplit(value[j,2], split =";")[[1]])

t = mgoSim(go1, go2, semData = scGo, measure ="Wang")

print(t)

}

}

}

还没导出数据 蛋白质比较的一部分

