與生物資料庫的連結

Annotation Database

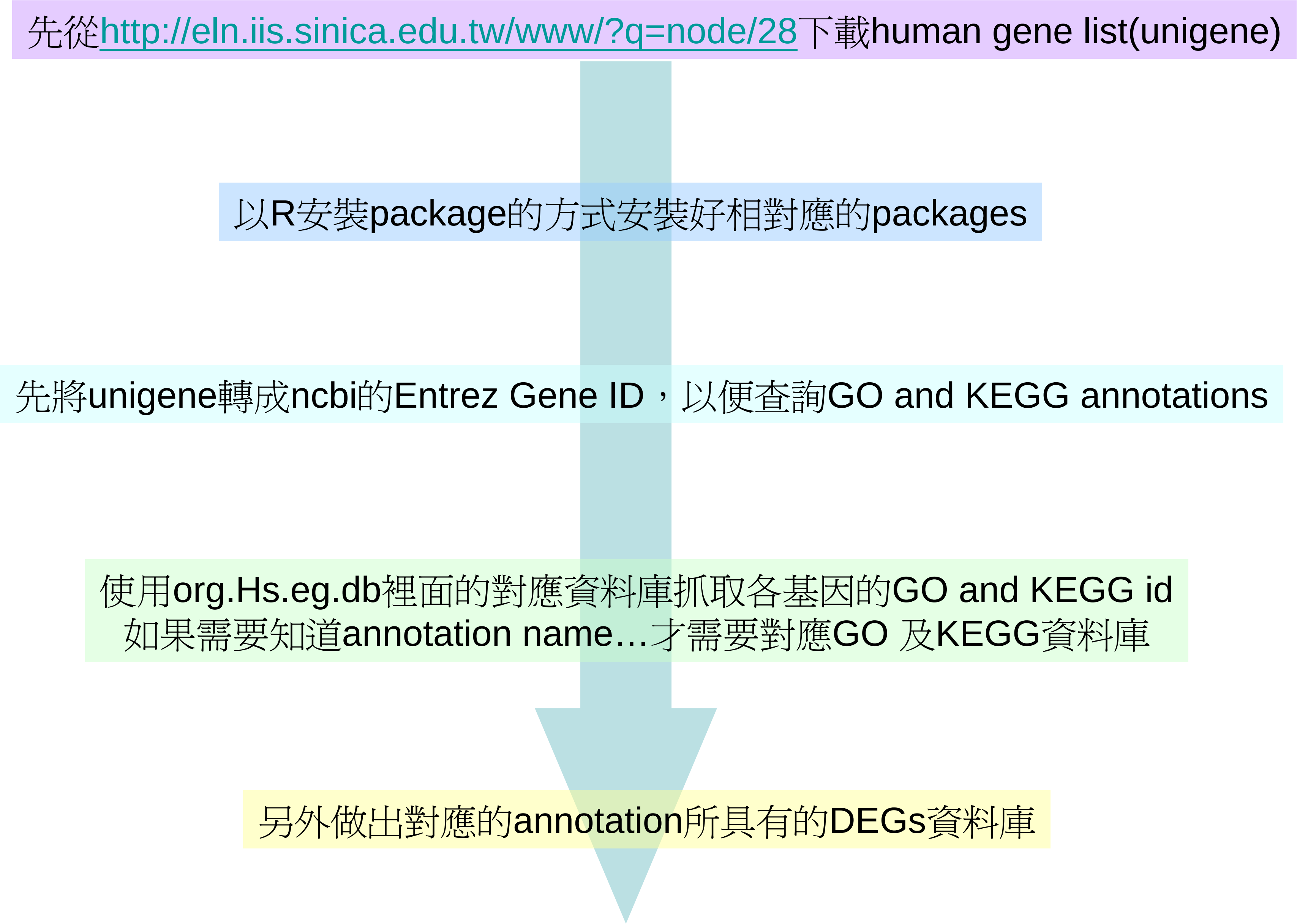
通常在已知的生物資料庫(GO or KEGG)，都會提供sql檔或純文字檔的資料下載，更新的動作通常是一個禮拜一次，部分特殊的格式(ex. cvs of GO)，則是一天更新一次。

大部分的使用方式，都是下載那些資料後，用字串處理的方式將database經由parser處理成自己想使用的格式，再串接到mysql及php去使用。

bioconductor爲了R的使用者，建了一些Model animal的annotation database提供使用者下載使用，但因爲是整合性資料，所以更新時間不定。

而現在R有提供的annotation database主要有以下幾種:
org.Hs.eg.db
org.Mm.eg.db
org.Rn.eg.db
org.Sc.sgd.db
org.Dm.eg.db
org.Pf.plasmo.db

Summary



先從http://eln.iis.sinica.edu.tw/www/?q=node/28下載human gene list(unigene)

以R安裝package的方式安裝好相對應的packages

先將unigene轉成ncbi的Entrez Gene ID，以便查詢GO and KEGG annotations

使用org.Hs.eg.db裡面的對應資料庫抓取各基因的GO and KEGG id
如果需要知道annotation name…才需要對應GO 及KEGG資料庫

另外做出對應的annotation所具有的DEGs資料庫

read a dataset

先從網路上下載事先準備好的dataset，來進行以下查詢GO與KEGG的資料

需要的package…” org.Hs.eg.db”（6/5/2009),” GO.db” ,” KEGG.db”

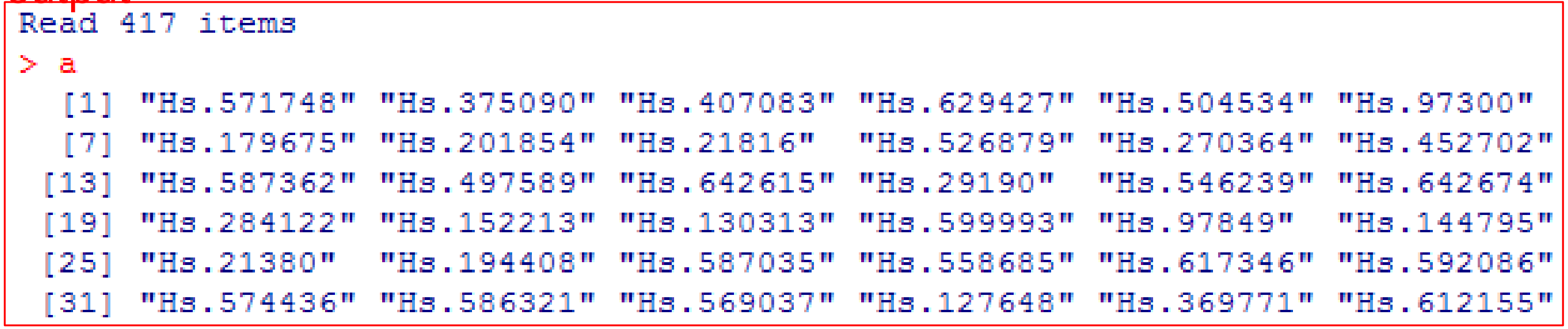
```
library('org.Hs.eg.db')
library('GO.db')
library('KEGG.db')
library('annotate')
```

先將需要的package讀入記憶體

```
DEGs <- scan("c:/unigene_gh_0001_5",what = character(0),sep='\n')
```

將已經下載的gene list讀入R

output



```
Read 417 items
> a
  [1] "Hs.571748" "Hs.375090" "Hs.407083" "Hs.629427" "Hs.504534" "Hs.97300"
  [7] "Hs.179675" "Hs.201854" "Hs.21816"  "Hs.526879" "Hs.270364" "Hs.452702"
 [13] "Hs.587362" "Hs.497589" "Hs.642615" "Hs.29190"  "Hs.546239" "Hs.642674"
 [19] "Hs.284122" "Hs.152213" "Hs.130313" "Hs.599993" "Hs.97849"  "Hs.144795"
 [25] "Hs.21380"  "Hs.194408" "Hs.587035" "Hs.558685" "Hs.617346" "Hs.592086"
 [31] "Hs.574436" "Hs.586321" "Hs.569037" "Hs.127648" "Hs.369771" "Hs.612155"
```

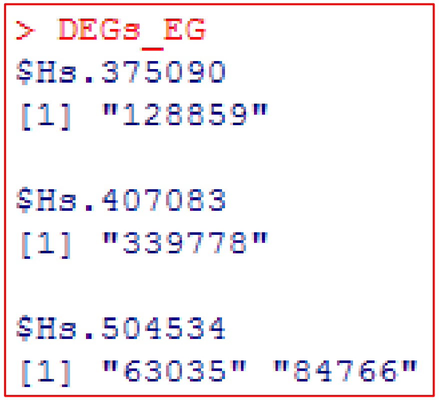
org.Hs.eg.db

是一個由NCBI提供，用來做id mapping及biological annotation查詢的工具

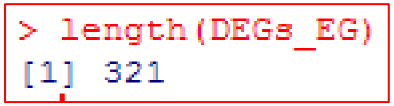
裡面包括許多mapping的環境，我們會使用到的
有”org.Hs.egUNIGENE2EG”,”org.Hs.egGO”跟”org.Hs.egPATH”

org.Hs.egUNIGENE2EG…將Unigene id對應回去Entrez Gene ID
org.Hs.egGO…查詢Entrez Gene ID所標記的Gene Ontology ID
org.Hs.egPATH…查詢Entrez Gene ID所標記的KEGG pathway ID

```
DEGs_EG =list()  #建立空的list
#分別將各個DEGs對應出其EG ID
for(i in 1:length(DEGs)){
            EG=org.Hs.egUNIGENE2EG[[DEGs[i]]]
            if(!is.null(EG)){
                        DEGs_EG[[DEGs[i]]]=EG
            }
}
```



```
> DEGs_EG
$Hs.375090
[1] "128859"

$Hs.407083
[1] "339778"

$Hs.504534
[1] "63035" "84766"
```

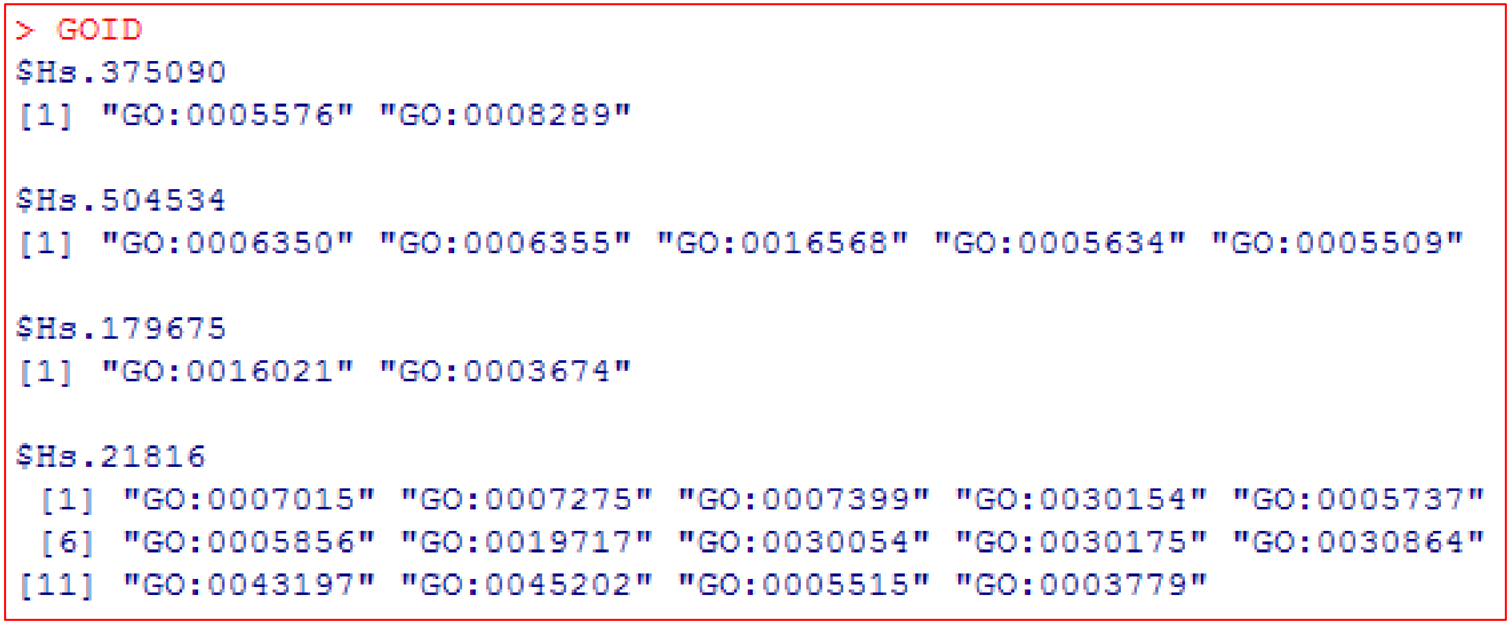


```
> length(DEGs_EG)
[1] 321
```

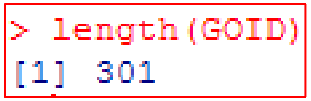
find GO annotation

```
GOID=list()
for(i in 1:length(DEGs_EG)){
        for(j in 1:length(DEGs_EG[[i]])){
                GOs=org.Hs.egGO[[DEGs_EG[[i]][[j]]]]
                if(!is.na(GOs[1])){
                        for(k in 1:length(GOs)){
                                GOID[[names(DEGs_EG[i])]]=append(GOID[[names(DEGs_EG[i])]],GOs[[k]][['GOID']])
                        }
                }
        }
}
```

以上則是用for loop將DEGs_EG所對應出來的Entrez Gene ID，利用org.Hs.egGO查詢出所標記的GO ID，再將其歸屬給原來的Unigene ID



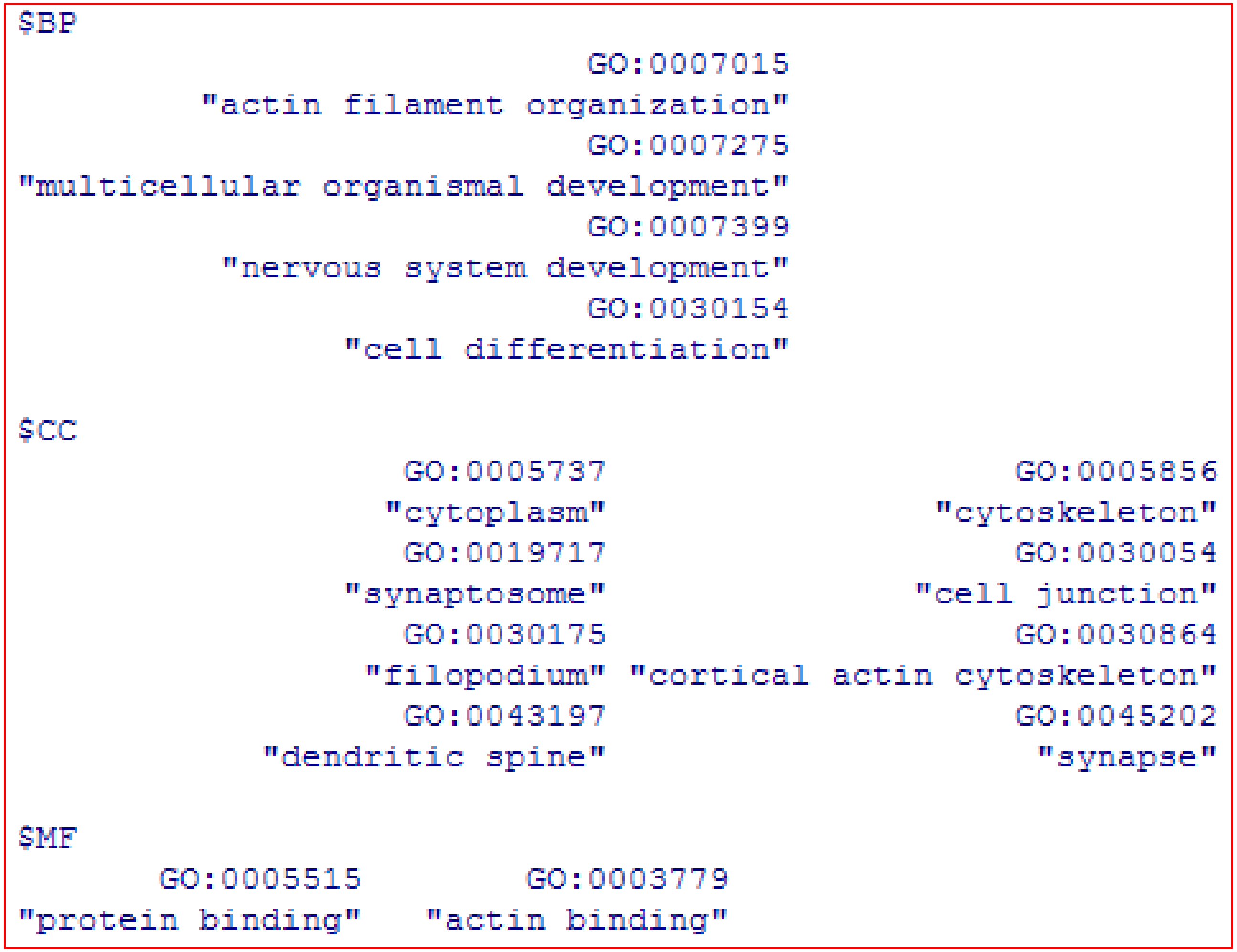
```
> GOID
$Hs.375090
[1] "GO:0005576" "GO:0008289"

$Hs.504534
[1] "GO:0006350" "GO:0006355" "GO:0016568" "GO:0005634" "GO:0005509"

$Hs.179675
[1] "GO:0016021" "GO:0003674"

$Hs.21816
 [1] "GO:0007015" "GO:0007275" "GO:0007399" "GO:0030154" "GO:0005737"
 [6] "GO:0005856" "GO:0019717" "GO:0030054" "GO:0030175" "GO:0030864"
[11] "GO:0043197" "GO:0045202" "GO:0005515" "GO:0003779"
```



```
> length(GOID)
[1] 301
```

How to get Gene Ontology names

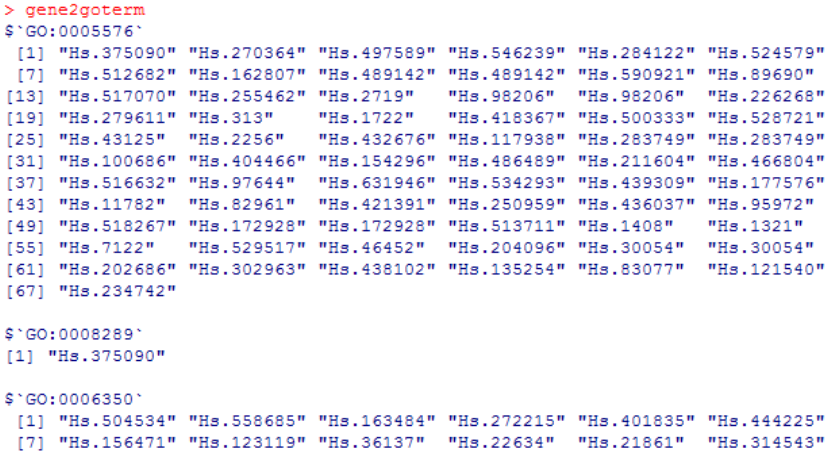
```
getGOTerm(GOID[[4]]) #直接以第幾個gene做查詢

getGOTerm(GOID[[‘Hs.21816’]]) #直接以基因名作查詢
```



```
$BP
                             GO:0007015
          "actin filament organization"
                             GO:0007275
"multicellular organismal development"
                             GO:0007399
           "nervous system development"
                             GO:0030154
                 "cell differentiation"

$CC
                   GO:0005737                    GO:0005856
                  "cytoplasm"                "cytoskeleton"
                   GO:0019717                    GO:0030054
                "synaptosome"               "cell junction"
                   GO:0030175                    GO:0030864
                 "filopodium" "cortical actin cytoskeleton"
                   GO:0043197                    GO:0045202
            "dendritic spine"                     "synapse"

$MF
       GO:0005515        GO:0003779
"protein binding"   "actin binding"
```

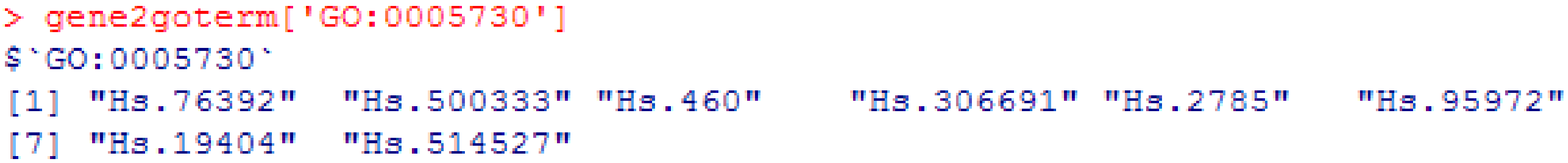
Search DEGs of a GO term

```
gene2goterm=list()
for(i in 1:length(GOID)){
                for(j in 1:length(GOID[[i]])){
                                gene2goterm[[GOID[[i]][j]]]=append(gene2goterm[[GOID[[i]][j]]],names(GOID[i]))
                }
}
```



```
> gene2goterm
$`GO:0005576`
 [1] "Hs.375090" "Hs.270364" "Hs.497589" "Hs.546239" "Hs.284122" "Hs.524579"
 [7] "Hs.512682" "Hs.162807" "Hs.489142" "Hs.489142" "Hs.590921" "Hs.89690"
[13] "Hs.517070" "Hs.255462" "Hs.2719"   "Hs.98206"  "Hs.98206"  "Hs.226268"
[19] "Hs.279611" "Hs.313"    "Hs.1722"   "Hs.418367" "Hs.500333" "Hs.528721"
[25] "Hs.43125"  "Hs.2256"   "Hs.432676" "Hs.117938" "Hs.283749" "Hs.283749"
[31] "Hs.100686" "Hs.404466" "Hs.154296" "Hs.486489" "Hs.211604" "Hs.466804"
[37] "Hs.516632" "Hs.97644"  "Hs.631946" "Hs.534293" "Hs.439309" "Hs.177576"
[43] "Hs.11782"  "Hs.82961"  "Hs.421391" "Hs.250959" "Hs.436037" "Hs.95972"
[49] "Hs.518267" "Hs.172928" "Hs.172928" "Hs.513711" "Hs.1408"   "Hs.1321"
[55] "Hs.7122"   "Hs.529517" "Hs.46452"  "Hs.204096" "Hs.30054"  "Hs.30054"
[61] "Hs.202686" "Hs.302963" "Hs.438102" "Hs.135254" "Hs.83077"  "Hs.121540"
[67] "Hs.234742"

$`GO:0008289`
[1] "Hs.375090"

$`GO:0006350`
 [1] "Hs.504534" "Hs.558685" "Hs.163484" "Hs.272215" "Hs.401835" "Hs.444225"
 [7] "Hs.156471" "Hs.123119" "Hs.36137"  "Hs.22634"  "Hs.21861"  "Hs.314543"
```

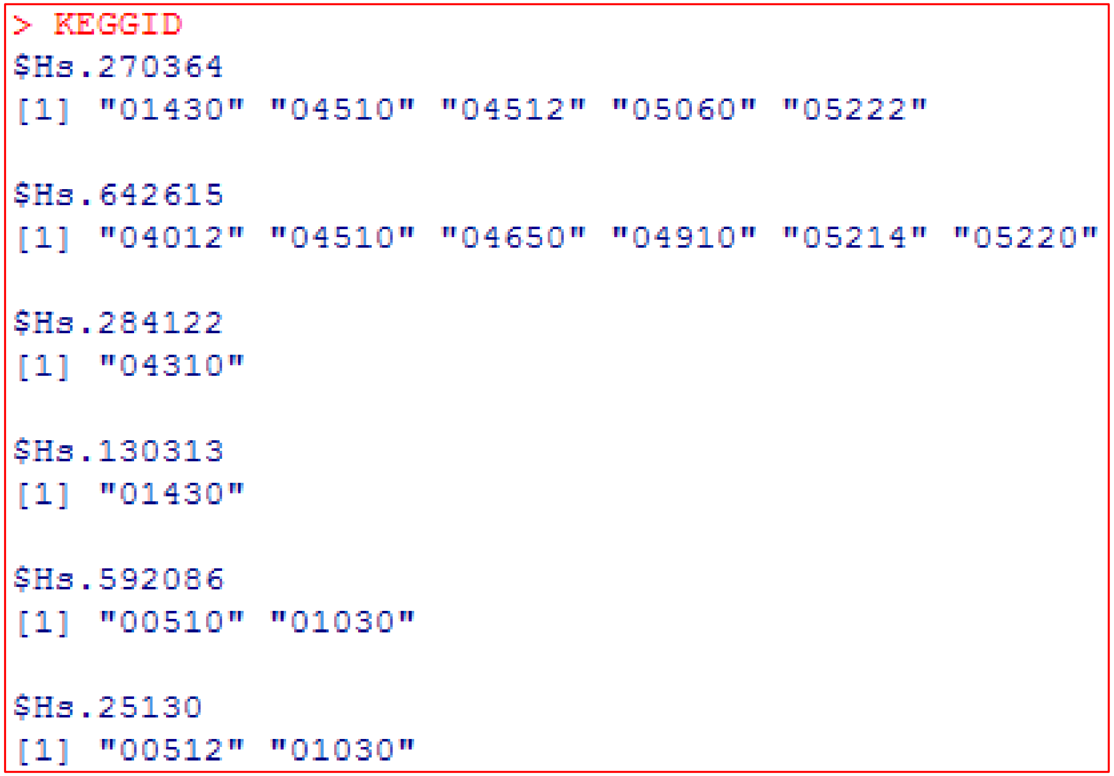
```
gene2goterm['GO:0005730']
```



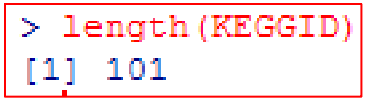
```
> gene2goterm['GO:0005730']
$`GO:0005730`
[1] "Hs.76392"  "Hs.500333" "Hs.460"    "Hs.306691" "Hs.2785"   "Hs.95972"
[7] "Hs.19404"  "Hs.514527"
```

find KEGG pathway ID

```
KEGGID=list()
for(i in 1:length(DEGs_EG)){
            for(j in 1:length(DEGs_EG[[i]])){
                        KEGGs=org.Hs.egPATH[[DEGs_EG[[i]][[j]]]]
                        if(!is.na(KEGGs[1])){
                                    KEGGID[[names(DEGs_EG[i])]]=append(KEGGID[[names(DEGs_EG[i])]],KEGGs)
                        }
            }
}
```



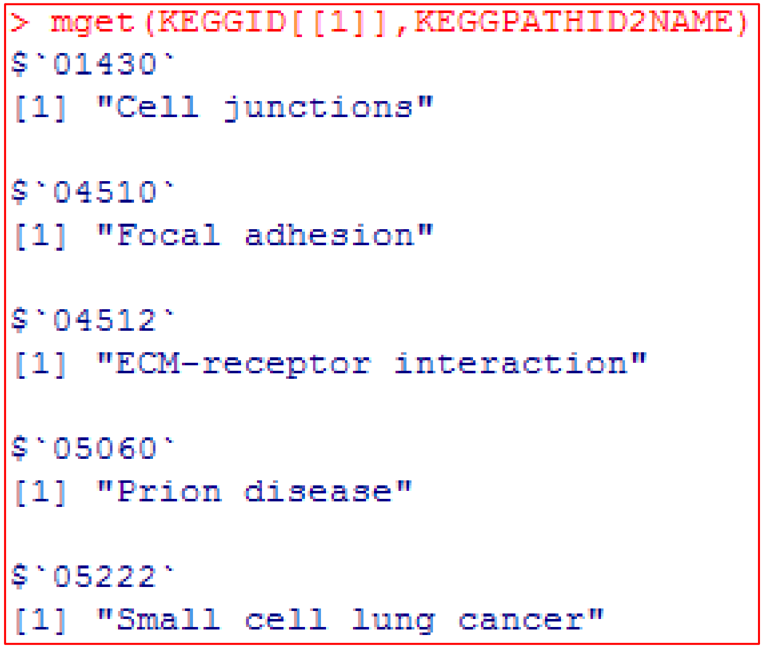
```
> KEGGID
$Hs.270364
[1] "01430" "04510" "04512" "05060" "05222"

$Hs.642615
[1] "04012" "04510" "04650" "04910" "05214" "05220"

$Hs.284122
[1] "04310"

$Hs.130313
[1] "01430"

$Hs.592086
[1] "00510" "01030"

$Hs.25130
[1] "00512" "01030"
```



```
> length(KEGGID)
[1] 101
```

How to get KEGG pathway names

```
mget(KEGGID[[1]],KEGGPATHID2NAME)

mget(KEGGID[[‘Hs.270364']],KEGGPATHID2NAME)
```



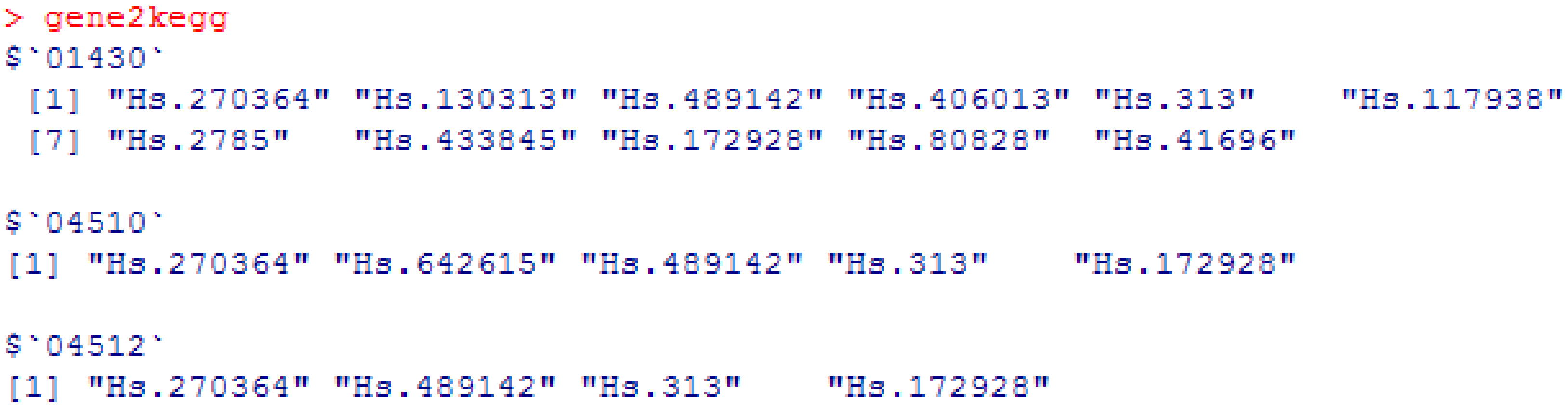
```
> mget(KEGGID[[1]],KEGGPATHID2NAME)
$`01430`
[1] "Cell junctions"

$`04510`
[1] "Focal adhesion"

$`04512`
[1] "ECM-receptor interaction"

$`05060`
[1] "Prion disease"

$`05222`
[1] "Small cell lung cancer"
```

Find the map of KEGG pathway

http://www.genome.jp/dbget-bin/www_bfind?pathway

Search DEGs of a KEGG pathway

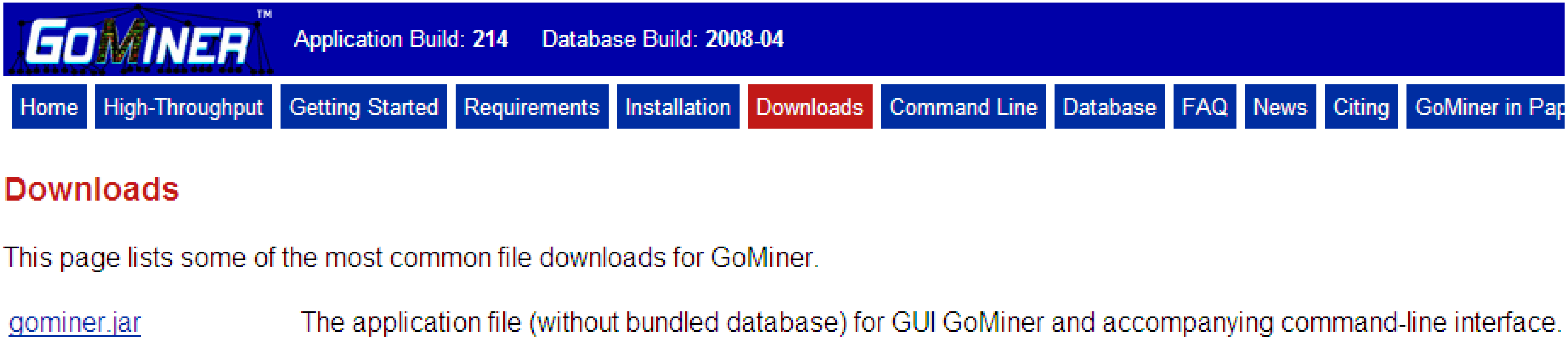
```
gene2kegg=list()
for(i in 1:length(KEGGID)){
        for(j in 1:length(KEGGID[[i]])){
                gene2kegg[[KEGGID[[i]][j]]]=append(gene2kegg[[KEGGID[[i]][j]]],names(KEGGID[i]))
        }
}
```



```
> gene2kegg
$`01430`
 [1] "Hs.270364" "Hs.130313" "Hs.489142" "Hs.406013" "Hs.313"    "Hs.117938"
 [7] "Hs.2785"   "Hs.433845" "Hs.172928" "Hs.80828"  "Hs.41696"

$`04510`
[1] "Hs.270364" "Hs.642615" "Hs.489142" "Hs.313"    "Hs.172928"

$`04512`
[1] "Hs.270364" "Hs.489142" "Hs.313"    "Hs.172928"
```

GOminer

http://discover.nci.nih.gov/gominer/htgm.jsp

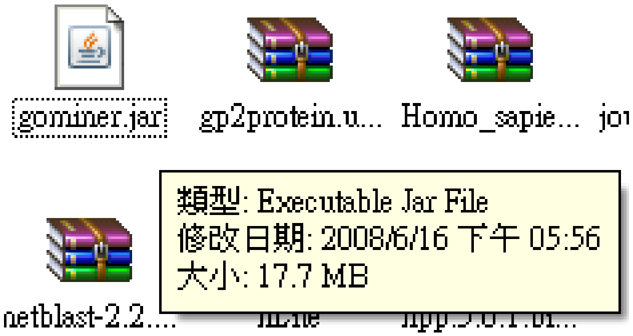


Application Build: **214** Database Build: **2008-04**

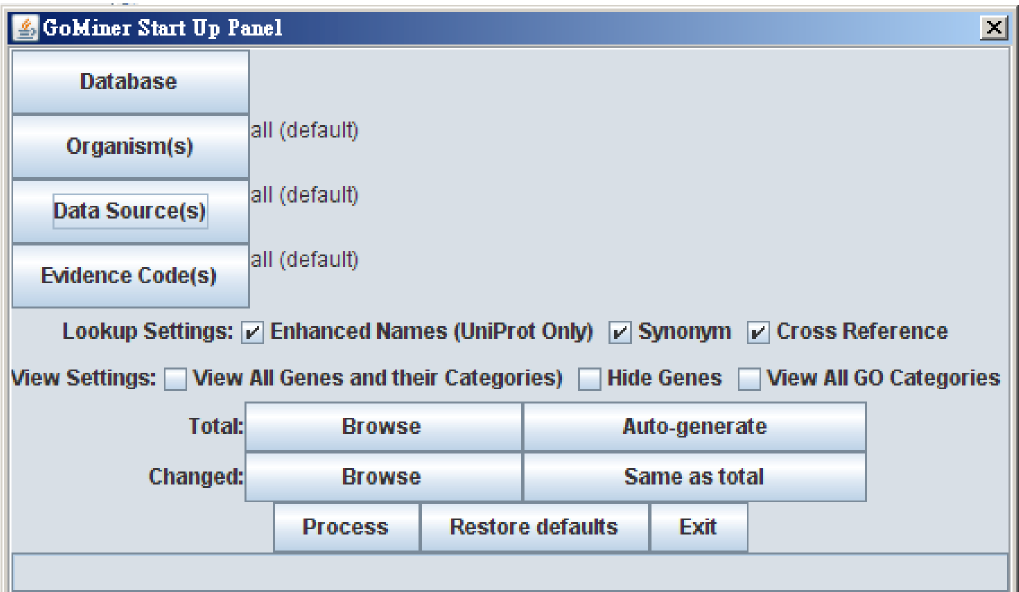
Downloads

This page lists some of the most common file downloads for GoMiner.

gominer.jar The application file (without bundled database) for GUI GoMiner and accompanying command-line interface.



類型: Executable Jar File
修改日期: 2008/6/16 下午 05:56
大小: 17.7 MB



Organism(s): all (default)
Data Source(s): all (default)
Evidence Code(s): all (default)

Lookup Settings: ☑ Enhanced Names (UniProt Only) ☑ Synonym ☑ Cross Reference

View Settings: □ View All Genes and their Categories) □ Hide Genes □ View All GO Categories

Total:

Changed:

BiNGO

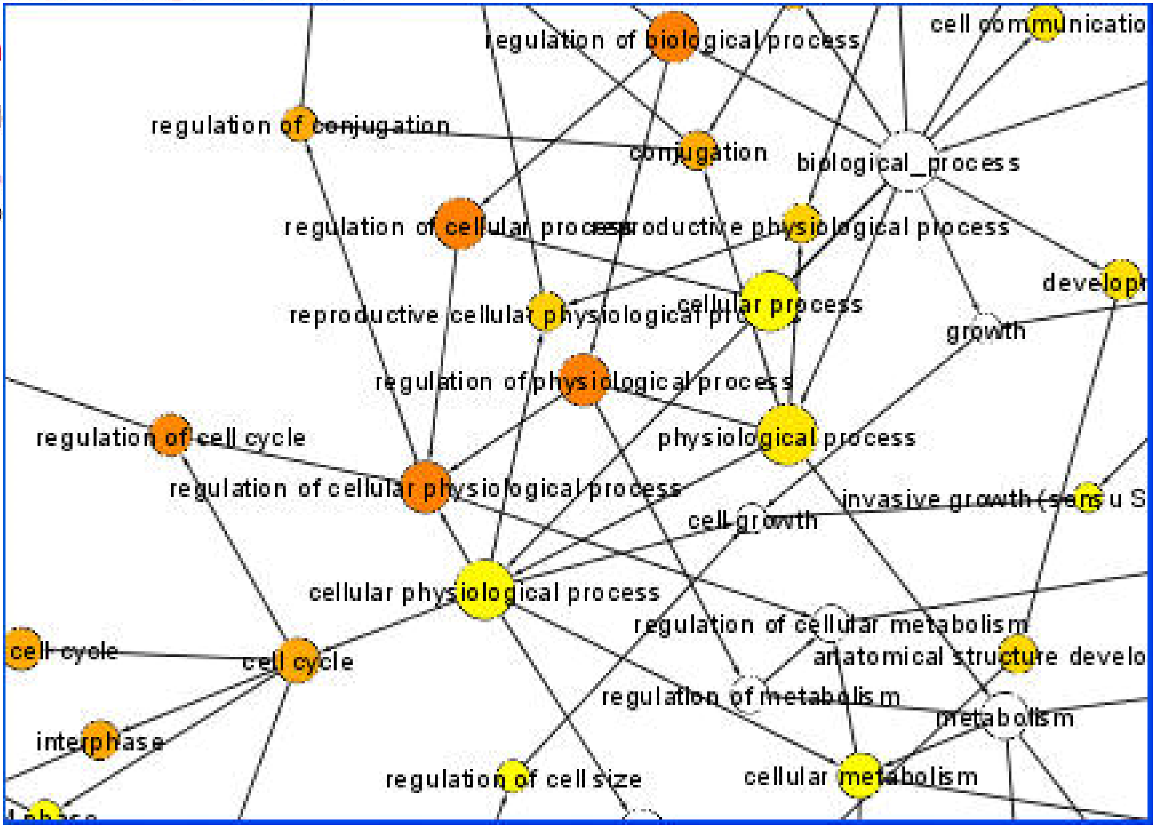
BiNGO

Latest update of Ontologies/Annotations : August 5th 2008

Current version : BiNGO 2.3 (com

If you use BiNGO in your research, pl

Maere, S., Heymans, K. and Kuiper, M. (20
overrepresentation of Gene Ontology catego
3448-3449. (PubMed)


regulation of biological process
cell communicatio
regulation of conjugation
conjugation
biological_process
regulation of cellular process
reproductive physiological process
develop
reproductive cellular physiological pr
cellular process
growth
regulation of physiological process
regulation of cell cycle
physiological process
regulation of cellular physiological process
invasive growth (sensu S
cell growth
cellular physiological process
regulation of cellular metabolism
cell cycle
cell cycle
anatomical structure develo
regulation of metabolism
metabolism
interphase
regulation of cell size
cellular metabolism
phase