

Zhang-Lab 生信小课堂 第十九期

Applied Bioinformatics Club (ABC)

和趣求真  秉实生信

(张建伟生物信息学课题组 <https://zhang.hzau.edu.cn>)

eggNOG-mapper 基因功能注释

2024.11.01 二综C102 15:00 欢迎大家交流学习!

主讲人: 陈盈

时间: 2024/11/01

eggNOG-mapper

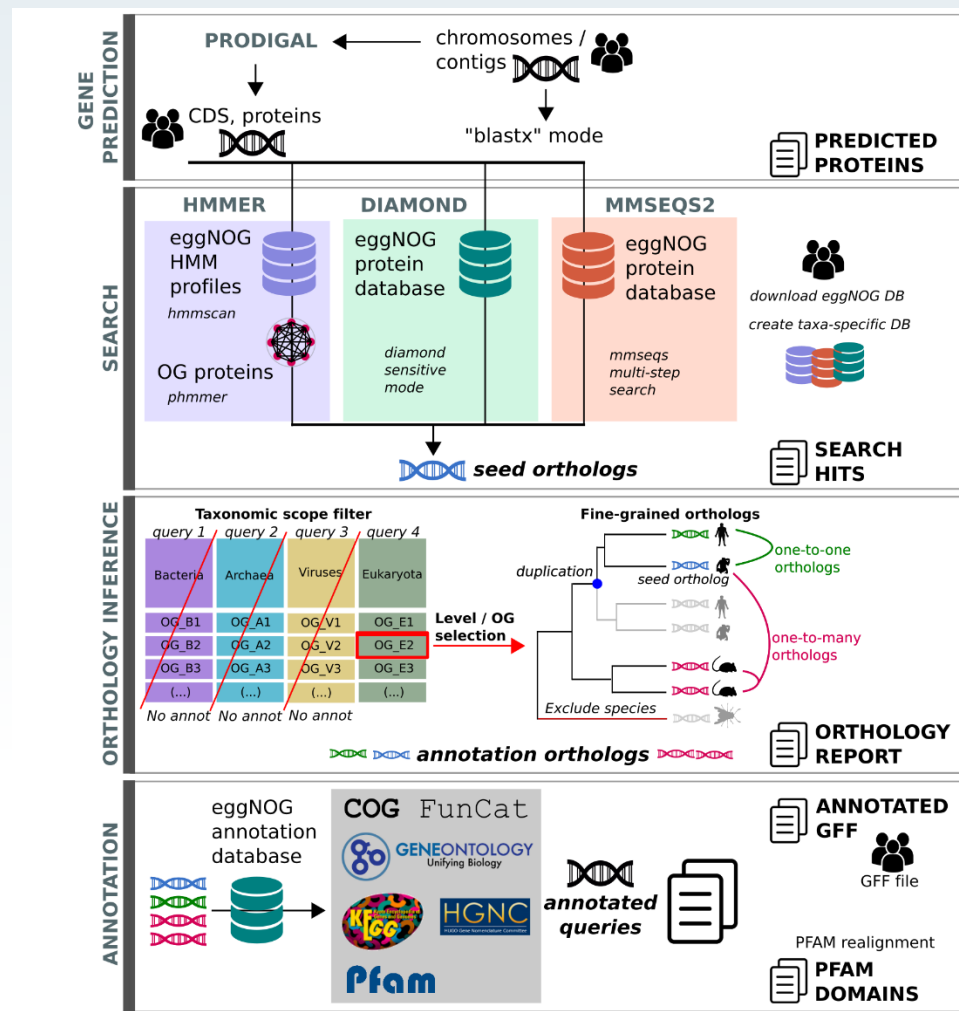
[eggNOG-mapper \(embl.de\)](http://eggNOG-mapper.embl.de): 是一款进行批量基因功能注释的工具。

A: 编码基因预测阶段，用Prodigal从组装的contigs中进行蛋白预测（或者blastx模式预测）。

B: 搜索阶段，选择HMMER或者DIAMOND或者MMESQS2将输入的蛋白序列和EggNOG数据库进行比对，生成seed orthologs。

C: 直系同源推断阶段，根据所需要的分类范围，生成直系同源报告。

D: 注释阶段，根据蛋白注释信息和蛋白结构域注释信息，生成表格和gff格式的报告。



01

使用

Proteins: 蛋白序列，上限是10万条序列。

CDS: CDS序列，上限也是10万条，会在搜索前倍翻译成蛋白序列。

Genomic: 基因组序列，支持最多1000条DNA序列，1000万个核苷酸。直接上传基因组序列会多一步编码蛋白预测，可以选择使用Prodigal或者Blastx-like这两者之一。

Metagenomic: Contig级别的基因组序列，本质上和上面是一样的，所以限制条件也相同。同样会进行编码蛋白预测。

Seeds: eggNOG-mapper跑的seed orthologs，支持最多上传10万条，这个主要用于重新注释。

Annotate a file

What kind of data?

☒ Proteins **1.选择数据类型** ☐ CDS ☐ Genomic ☐ Metagenomic ☐ Seeds

Up to 100,000 proteins in FASTA format.

Upload sequences

Files may be compressed in gzip format (file name must end in '.gz')

未选择文件 **2.上传gz文件**

Email address *(Required for job scheduling and notifications)*

3.输入邮箱

Advanced Options

Database

Search against database:

☒ eggNOG 5 ☐ Novel families

Search filters

Annotation options

4.参数选择，默认

5.点击提交

02

邮箱

Job MM_6arem7do Submitted

[Click to manage your job](#)

← 点击，再点击start job

Job info

Job name	User	Date created	File	Sequences	Total length	Tax. scope	Orth. restrictions	GO evidence	Min. hit e-value	Min. hit bit-score	Min. % of query cov.	Min. % of subject cov.
MM_6arem7do	2959473298@qq.com	10/10/24	Soybean.all.maker.proteins.fasta.gz	46,690	18,699,895 proteins	auto	all	non-electronic	0.001	60	20	20

[Access your job files here](#)

Job MM_6arem7do Status: Done 完成

Job name	User	Date created	File	Sequences (type)
MM_6arem7do	2959473298@qq.com	10/10/24	Soybean.all.maker.proteins.fasta.gz	46690 (proteins)

```
emapper.py --cpu 20 --mp_start_method forserver --data_dir /dev/shm/ -o out --output_dir /emapper_web_jobs/emapper_jobs/user_data/MM_6arem7do/  
/emapper_web_jobs/emapper_jobs/user_data/MM_6arem7do/queries.fasta --evalue 0.001 --score 60 --pident 40 --query_cover 20 --subject_cover 20 --itype  
/emapper_web_jobs/emapper_jobs/user_data/MM_6arem7do/emapper.out 2> /emapper_web_jobs/emapper_jobs/user_data/MM_6arem7do/emapper.err
```

[Access your job files here](#) 也可以点击access your job

Important note: results are kept for a limited period of time. Please, download your results as soon as possible.

../

[emapper.err](#)

10-Oct-2024 08:26 6854

[emapper.out](#)

10-Oct-2024 08:26 3197

[info.txt](#)

10-Oct-2024 08:11 1285

[out.emapper.annotations](#)

10-Oct-2024 08:26 29M

[out.emapper.annotations.xlsx](#)

10-Oct-2024 08:26 6M

[out.emapper.decorated.gff](#)

10-Oct-2024 08:26 38M

[out.emapper.hits](#)

10-Oct-2024 08:24 11M

[out.emapper.orthologs](#)

10-Oct-2024 08:26 193M

[out.emapper.seed.orthologs](#)

10-Oct-2024 08:24 4M

[queries.fasta](#)

10-Oct-2024 08:11 20M

[queries.raw](#)

10-Oct-2024 08:11 22M

Explore annotations

Downloads

CSV

Excel

GFF

Orthologs

Seeds

可直接下载

03

完成

Thu Oct 10 08:22:52 2024## emapper-2.1.12## /data/shared/home/emapper/miniconda3/envs/eggno-mapper-2.1/bin/emapper.py --cpu 20 --mp_start_method forkserver --data_dir /dev/shm/ -o out --output_dir /emapper_web_jobs/emapper_jobs/user_data/MM_6arem7do --temp_dir /emapper_web_jo

#	query	seed_orth	evaluate	score	eggNOG_OG	max_annot	COG_cat	Description	Preferred	GOs	EC	KEGG_ko	KEGG_Pat	KEGG_Mo	KEGG_Rea	KEGG_rcl	BRITE	KEGG_TC	CAZy	BIGG_Rea	PFAMs
	maker-Chr13-s	3847.GLYN	1.38e-26	111.0	COG0676@1	35493 Str	G	Belongs to	-	-	5.1.3.15	ko:K01792	ko00010,k	-	R02739	RC00563	ko00000,k	-	-	-	Aldose_epim
	maker-Chr13-s	3847.GLYN	1.63e-166	475.0	KOG1477@1	35493 Str	S	Ran-bindir	-	GO:00055	-	-	-	-	-	-	-	-	-	-	CLTH,SPRY
	maker-Chr13-s	3847.GLYN	0.0	1003.0	COG0143@1	35493 Str	J	Belongs to	-	GO:00055	6.1.1.10	ko:K01874	ko00450,k	M00359,N	R03659,R	RC00055,F	ko00000,k	-	-	-	Anticodon_1,tRNA-synt_1g,tRNA_bind
	maker-Chr13-s	3847.GLYN	1.53e-87	265.0	2CXQS@1	ro 35493 Str	S	Late embr	-	-	-	-	-	-	-	-	-	-	-	-	LEA_2
	augustus_mas	3847.GLYN	2.41e-60	194.0	2A7Z7@1	ro 35493 Str	S	Protein FA	-	-	-	-	-	-	-	-	-	-	-	-	FAR1,MULE,SWIM

•最重要的是annotations文件。此文件共有21列，

1. query: 蛋白序列名称
2. seed_ortholog: 搜索阶段比对上的seed ortholog编号
3. evaluate: evaluate值，越小结果越可靠
4. score: 比对的得分，越大越可靠
5. eggNOG_OGs: 为序列确定的以逗号分隔、按照进化分支深度排序的直系同源组（orthologs groups, OGs）列表。每个直系同源组以 OG@tax_id|tax_name方式展现。
6. max_annot_lvl: 用于检索注释的最宽的直系同源组，以tax_id|tax_name方式展现。
7. COG_category: 从最佳OG中推测的COG功能分类（一个字母），具体有哪些可以看这里NCBI的介绍[COG - NCBI \(nih.gov\)](https://www.ncbi.nlm.nih.gov/cog/)。
8. Description: 注释的基因功能描述
9. Preferred_name: 普遍使用的基因名缩写
10. GOs: 注释的GO编号，一个基因可能对映非常多的GO号

03

完成

Thu Oct 10 08:22:52 2024## emapper-2.1.12## /data/shared/home/emapper/miniconda3/envs/eggno-mapper-2.1/bin/emapper.py --cpu 20 --mp_start_method forkserver --data_dir /dev/shm/ -o out --output_dir /emapper_web_jobs/emapper_jobs/user_data/MM_6arem7do --temp_dir /emapper_web_jo

#	query	seed_orth	eval	score	eggNOG_OG	max_annc	COG_cat	Description	Preferred	GOs	EC	KEGG_ko	KEGG_Path	KEGG_Mo	KEGG_Rea	KEGG_rcla	BRITE	KEGG_TC	CAZy	BiGG_Rea	PFAMs
	maker-Chr13-s	3847.GLYN	1.38e-26	111.0	COG0676@1	35493	Str G	Belongs to	-	-	5.1.3.15	ko:K01792	ko00010,k-	-	R02739	RC00563	ko00000,k-	-	-	-	Aldose_epim
	maker-Chr13-s	3847.GLYN	1.63e-166	475.0	KOG1477@1	35493	Str S	Ran-bindir	-	GO:00055	-	-	-	-	-	-	-	-	-	-	CLTH,SPRY
	maker-Chr13-s	3847.GLYN	0.0	1003.0	COG0143@1	35493	Str J	Belongs to	-	GO:00055	6.1.1.10	ko:K01874	ko00450,k	M00359,N	R03659,R	RC00055,F	ko00000,k-	-	-	-	Anticodon_1,tRNA-synt_1g,tRNA_bind
	maker-Chr13-s	3847.GLYN	1.53e-87	265.0	2CXQS@1	ro 35493	Str S	Late embr	-	-	-	-	-	-	-	-	-	-	-	-	LEA_2
	augustus_mas1	3847.GLYN	2.41e-60	194.0	2A7Z7@1	ro 35493	Str S	Protein FA	-	-	-	-	-	-	-	-	-	-	-	-	FAR1,MULE,SWIM

11. EC: KEGG的EC通路编号，代表相关的酶

12.KEGG_ko: KEGG的KO编号，表示直系同源基因，代表一个具体的功能

13.KEGG_Pathway: 通常有ko编号和map编号，map编号代表reference pathway，是一种代谢通路类型，比较具体有一般参考意义

14.KEGG_Module: KEGG Module数据库编号，以M开头，实际上就是多个KO划分在一个共同发挥功能的单元里

15.KEGG_Reaction: KEGG Reaction数据库编号，以R开头，包含代谢通路上酶促反应相关信息

16.KEGG_rclass: KEGG RCLASS数据库编号，以RC编号开头，手动整理的反应数据集

17.BRITE: KEGG的Brite数据库编号，用的不是很多，是一个储存分类信息的数据库

18.KEGG_TC: ?

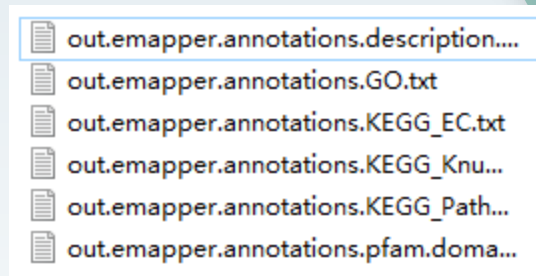
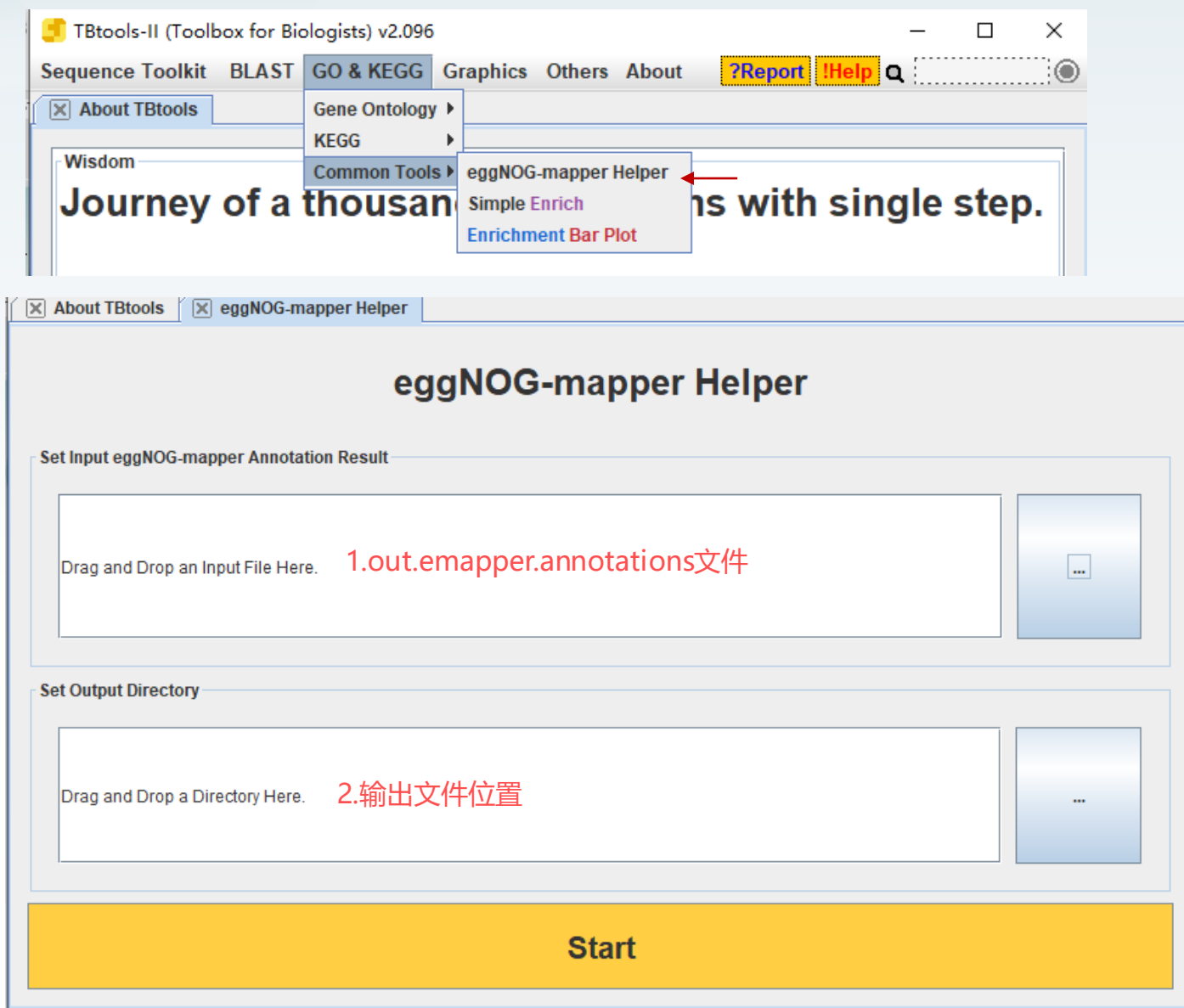
19.CAZy: 碳水化合物酶相关的专业数据库，可以看官网[CAZy – Home](https://www.cazy.org/)

20.BiGG_Reaction: BiGG是一个整合基因组尺度代谢网络模型的数据库，官网[BiGG Models \(ucsd.edu\)](https://biogg.ucsd.edu/)

21.PFAMs: PFAM数据库，根据多序列比对结果和隐马尔可夫模型，将蛋白分为不同家族的一个数据库，官网[InterPro \(ebi.ac.uk\)](https://www.ebi.ac.uk/interpro/)。没错官网是InterPro，pfam数据库已经被InterPro合并，并且2023年1月之后原网页就失效了。

04

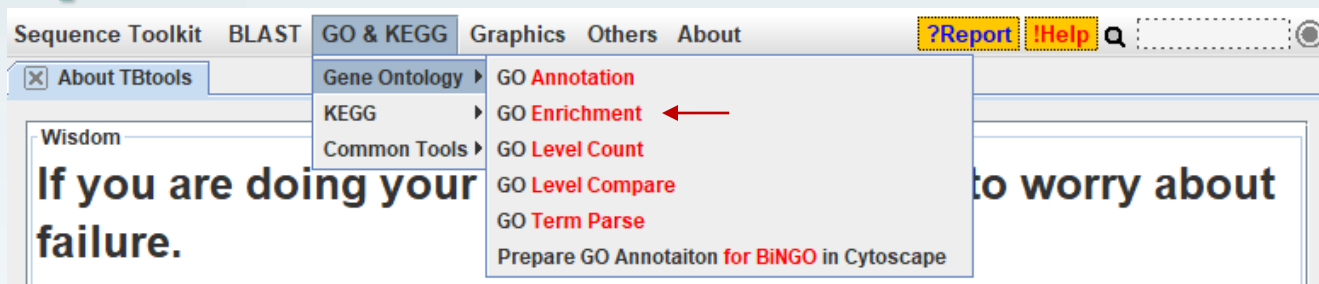
Tbtools数据处理



- out.emapper.annotations.description.txt, 对应的功能文本描述
- out.emapper.annotations.GO.txt, 对应的是GO注释结果, 可直接用于 TBtools GO富集分析, 当注释背景文件
- out.emapper.annotations.KEGG_Knum.txt, 对应的是KEGG注释结果, 可直接用于 TBtools KEGG富集分析, 当背景注释文件
- out.emapper.annotations.pfam.domain.txt, 对应的是PFAM结构域注释, 注意, 这个注释结果是定性的, 即有无某结构域, 如果一个序列有多个相同结构域, 只会显示一个

04

Tbtools GO分析



得到xxx.final.xls, 可用于可视化

04

Tbtools GO分析 可视化

Sequence Toolkit BLAST **GO & KEGG** Graphics Others About ?Report !Help

About TBtools

Wisdom

If you are doing your failure.

Gene Ontology
KEGG
Common Tools
eggNOG-mapper Helper
Simple Enrich
Enrichment Bar Plot

not have to worry about

Enrichment Bar Plot

Set Resultant File from Enrichment Analysis

Drag and Drop an Enrichment Result File Here xxx.final.xls

Preset Parameters

☒ GO Enrichment(Gene Ontology) ☐ KEGG Pathway Enrichment ☐ Other

Graph Parameters

Term Col Name: GO_Name

Class Col Name: Class

pValue/qValue Col Name: P_value

Max Term to Show: 0

X lab Label: -log10(P-value)

Y lab Label: GO Term

Value Format: 0.00

Bar Plot Type: Normal

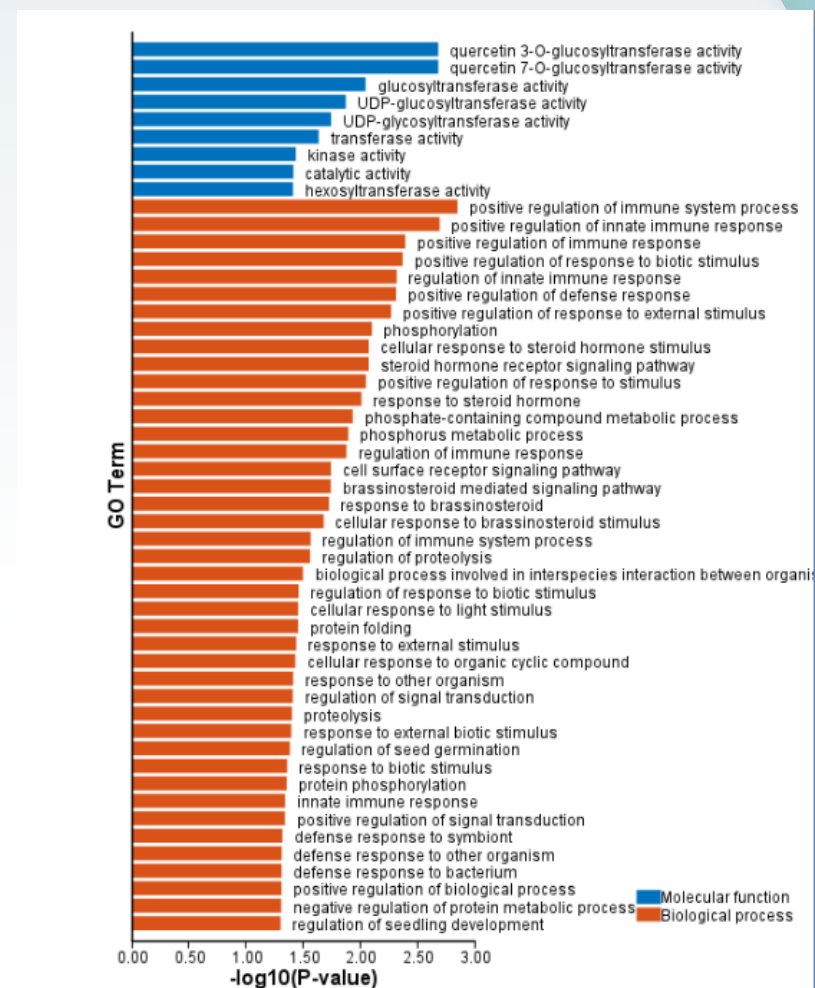
Note:

File for bar plot would be generated from any enrichment analysis software, including TBtools. It should possess at least two columns:

Term Col
pValue/qValue/FDR Col

To facilitate the visualization of enrichment analysis conducted on TBtools, two preset modes have been set.

Start



04

Tbtools KEGG分析

The screenshot shows the Tbtools KEGG Enrichment Analysis window. The menu bar includes 'Sequence Toolkit', 'BLAST', 'GO & KEGG', 'Graphics', 'Others', and 'About'. The 'GO & KEGG' menu is open, showing 'Gene Ontology', 'KEGG', and 'Common Tools'. The 'KEGG' submenu is selected, showing 'KEGG Enrichment Analyse' (highlighted with a red arrow) and 'KEGG Pathway Map Illustrate'. The main window has tabs for 'About TBtools' and 'KEGG Enrichment Analyse'. The title is 'KEGG Enrichment Analysis'. The interface includes four main sections with file selection fields and buttons, each with a red annotation:

- Download and Prepare KEGG Backend for Enrichment (Several Minutes Depending on Your Network Situation)**: A dropdown menu shows 'Plants' with the annotation '1. .KEGG.BackEnd 文件'. A button 'Make One Backend File From Web' is next to it. Below is a text field 'Drag and Drop a Backend File Which Has Been Made Last Time or Just Make a New One' with a button '....'.
- Set BackGround File (QueryID\tKnum SepByComma)**: A text field 'Drag and Drop a Query2Knum File Here' has the annotation '2. out.emapper.annotations.KEGG_Knum.txt文件'. A button '....' is next to it.
- Set Selection Set (One QueryId Per Line)**: A text field 'Drag and Drop a Query Set File Here' has the annotation '3.基因列表文件'. A button '....' is next to it.
- Set Output Directory**: A text field 'Drag and Drop an Output Directory Here' has the annotation '4.输出文件路径'. A button '....' is next to it.

At the bottom, a yellow bar contains the text 'Start' and the annotation '5.点击start'.

得到xxx.final.xls, 可用于可视化,
参考GO可视化

其它

egglog-mapper服务器使用（不建议）：

下载：

```
git clone https://github.com/jhcepas/egglog-mapper.git
```

下载数据库：

```
mkdir -p egglog
```

```
python egglog-mapper/download_egglog_data.py --data_dir egglog -y -f euk bact arch viruses
```

使用：

```
python emapper.py -i test/polb.fa --output polb_bact -d bact --data_dir ~/data/db/egglog
```

-i输入、-output输出文件前缀、-d指定数据库数据、-data_dir指定数据库位置

diamond方法，默认为hmmer方法：

```
python emapper.py -i test/polb.fa --output diamond_bact_ -d bact --data_dir ~/data/db/egglog -m diamond
```

得到3个文件：

polb_bact.emapper.annotations、 polb_bact.emapper.hmm_hits、 polb_bact.emapper.seed_orthologs

基因功能的注释：从基因组上提取翻译后的蛋白序列和主流的数据库进行比对，完成功能注释。

其它基因组功能注释数据库：

- 1.Nr/Nt数据库：NR(Non-Redundant Protein Sequence Database)非冗余蛋白库，包括GenBank+EMBL+DDBJ+PDB中的非冗余蛋白序列，NT(Nucleotide Sequence Database),核酸序列数据库，是NR库的子集。NR和NT库都可以通过NCBI进行在线BLAST。
2. UniProt数据库：包括人工校正的高质量数据库Swiss-Prot和软件自动注释的TrEMBL数据库。
- 3.GO/KEGG数据库
4. pfam数据库：蛋白质家族数据库，根据多序列比对结果和隐马尔科夫模型，将蛋白分成不同家族。
5. InterPro：EBI开发的一个整合的蛋白家族功能注释数据库，包括Gene3D、CDD、Pfam等10几个数据库。

其它

通用数据库注释方法，以uniprot数据库为例：

1.下载数据库：

wget

ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/uniprot_sprot_plants.dat.gz

wget

ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/uniprot_trembl_plants.dat.gz

zcat uniprot_sprot_plants.dat.gz uniprot_trembl_plants.dat.gz > uniprot_plants.dat

2.ID转换

```
from Bio import SeqIO
```

```
count = SeqIO.convert("uniprot_plants.dat","swiss","uniprot_plants.fa","fasta")
```

```
print("Converted %i records" % count)
```

其它

通用数据库注释方法，以uniprot数据库为例：

3.构建blastp索引：

```
makeblastdb -in uniprot_plants.fasta -dbtype prot -title uniprot_plants -parse_seqids
```

或者快速的：diamond makedb --in uniprot_plants.fasta -d uniprot_plants

4.蛋白质序列与数据库进行比对：

```
blastp -query protein.fa -out uniprot_plants.xml -db uniprot_plants.fasta -evaluate 1e-5 -outfmt 6 -num_threads 24
```

或者：diamond blastp -d uniprot_plants.dmnd -q proteins.fasta --evaluate 1e-5 > blastp.outfmt6

5.比对结果中筛选每个query的最佳subject：

```
python -m jcvl.formats.blast best -n 1 blastp.outfmt6
```

6.使用add_annotation_from_dat.py根据blastp输出从dat提取GO/KEGG/同源基因

其它

可参考学习的网站

[基因组注释 \(6\) ——在线版eggNOG-mapper注释功能基因 - 我的小破站 \(phantom-aria.github.io\)](https://phantom-aria.github.io)

[基因组注释——适合初学者的流程分解版 | TianzhenWu' Blog \(wu-tz.github.io\)](https://wu-tz.github.io)

[基因组注释 \(三\) : 基因功能注释 | 生信技工 \(yanzhongsino.github.io\)](https://yanzhongsino.github.io)

[零基础快速完成基因功能注释 / GO / KEGG / PFAM... - 简书 \(jianshu.com\)](https://jianshu.com)

[零基础-绝对完成GO/KEGG pathway富集分析-和-绘图 - 简书 \(jianshu.com\)](https://jianshu.com)

[应该是最好的eggno-mapper功能注释教程 - 简书 \(jianshu.com\)](https://jianshu.com)

谢谢观看，欢迎讨论！

2024.11.01 二综C102 15:00