3 GO功能注释

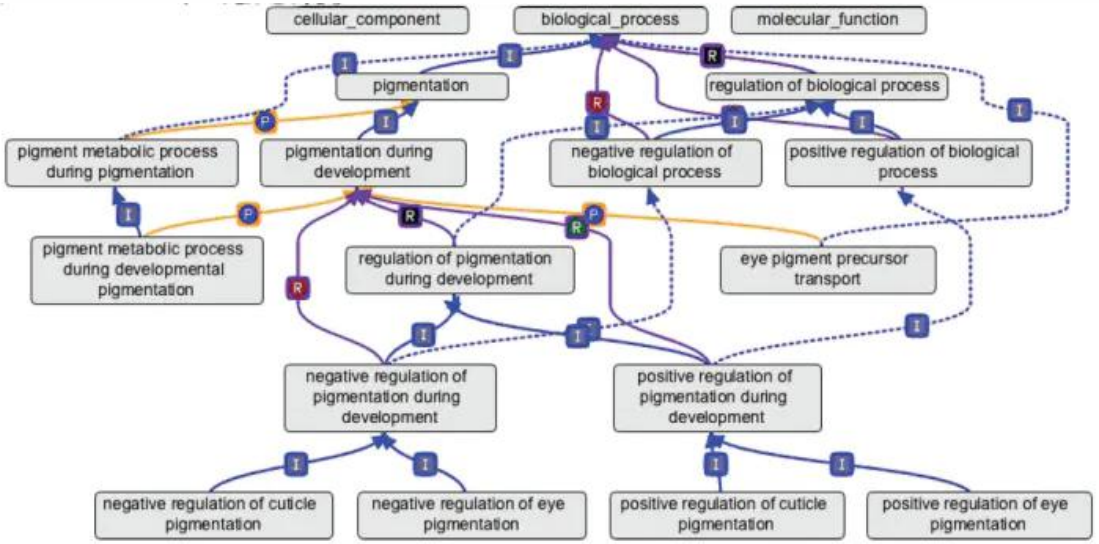
中国農業大學 China Agricultural University

Gene Ontology是一个 **描述这些跨物种的同源基因及其基因产物的功能** 的数据库，每个条目包含3类term和3类关系，构成了一个有向无环图

- 基因执行的分子功能（Molecular Function）
- 基因所处的细胞组分（Cellular Component）
- 基因以及参与的生物学过程（Biological Process）

GENEONTOLOGY Unifying Biology

- is_a
- part_of
- regulate



下载数据库后可以通过Scaffold软件对比完成

GO功能注释

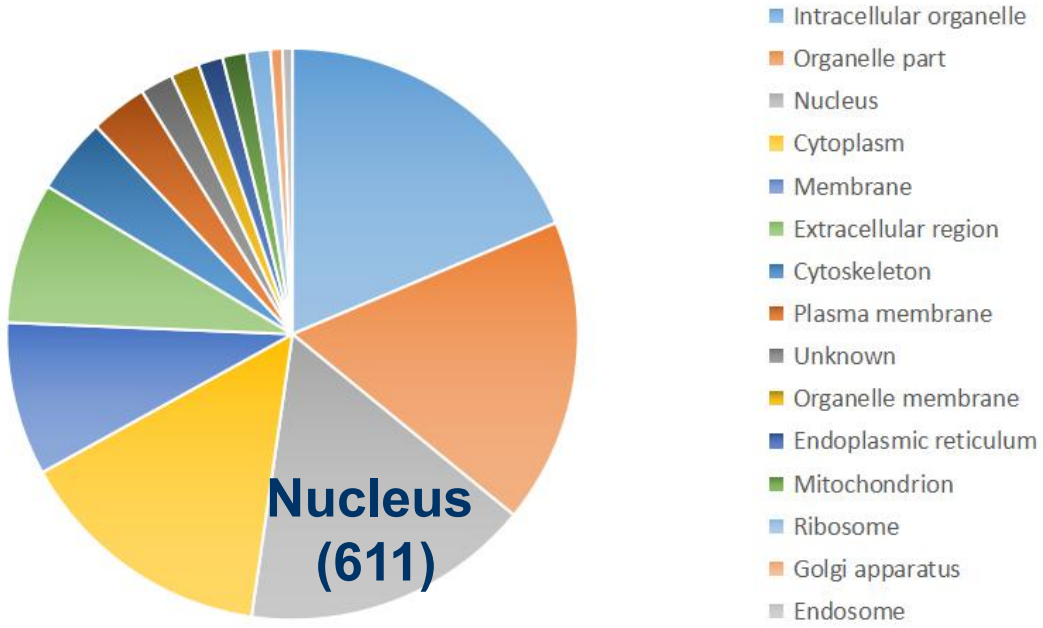
在Scafflod中，参考EBI数据库中下载的GO数据库，可以完成功能标注，输出有两个文件

GO标注的生物过程统计.xls	2020-03-25 11:42	XLS 工作表	350 KB
overview_protein_probabilities.xls	2020-03-26 17:56	XLS 工作表	1,195 KB

包含802个输出蛋白的清单（P=100%的部分）

这个输出把各个层次的term混在了一起，而且一个蛋白可能对应多条term，因而统计蛋白数的时候会出现重复

（即加起来总共超过了802，并且二级term的数量也明显多于三级term）



ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/goa_human.gaf.gz

4 GO富集分析

中國農業大學 China Agricultural University

GO Enrichment Analysis 功能的显著性分析

该分析对差异基因等按GO分类，并对分类结果进行基于离散分布的**显著性分析、错判率分析、富集度分析**，得到与实验目的有显著联系的、低误判率的、靶向性的基因功能分类，该分类即导致样本性状差异的最重要的功能差别。

Then we calculate the enrichment ratio of GO/Interpro term t in group S, and with the equation of the hypergeometric distribution, we can also calculate its P-value:

$$\text{Enrichment\_ratio} = \frac{\frac{m}{M}}{\frac{n}{N}},$$

$$p\text{-value} = \sum_{m'=m}^{n} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} \quad (\text{Enrichment\_ratio} \geqslant 1)$$

超几何分布

or

$$p\text{-value} = \sum_{m'=0}^{m} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} \quad (\text{Enrichment\_ratio} < 1).$$

N	total number of proteins in group W annotated by GO/Interpro
n	number of proteins in group W annotated by GO/Interpro term t
M	total number of proteins in group S annotated by GO/Interpro
m	number of proteins in group S annotated by GO/Interpro term t

再具体的计算方法还需要再看看资料

GO富集分析

中国農業大學 China Agricultural University

将Scoffold中的**802个蛋白**的Unipot Accession Number转换为基因的EntrezID，得到**711个基因**（有部分蛋白编号无法识别，有部分蛋白来自同一个gene）

参考org.Hs.eg.db，带入R的clusterProfile，作GOenrich analyse（包括BP、CC、MF）并将前30个结果绘图可视化

这里相当于是重新作了标注，直观感觉上和前面标注的有些不一样

MF_bar_Rplot.pdf	2020-03-27 1:41	Chrome HTML D...	7 KB
MF_cnet_Rplot.pdf	2020-03-27 1:41	Chrome HTML D...	155 KB
BP_bar_Rplot.pdf	2020-03-27 1:41	Chrome HTML D...	7 KB
CC_dot_Rplot.pdf	2020-03-27 1:31	Chrome HTML D...	9 KB
CC_bar_Rplot.pdf	2020-03-27 1:29	Chrome HTML D...	7 KB
CC_cnet_Rplot.pdf	2020-03-27 1:29	Chrome HTML D...	194 KB
test_CC.txt	2020-03-27 0:48	文本文档	262 KB
BP_dot_Rplot.pdf	2020-03-27 0:48	Chrome HTML D...	9 KB
BP_cnet_Rplot.pdf	2020-03-27 0:48	Chrome HTML D...	563 KB

分析结果的数据表

中国农业大学 China Agricultural University

	ONTOLOGY ID	Description	GeneRati	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
BP	GO:0008380	RNA splicing	138/698	469/18670	3.52E-86	1.95E-82	1.47E-82	HSPA8/HNRNPM/HNRNPK/DHX9/HNRNPA1/	138
BP	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.41E-79	[illegible]	[illegible]
BP	GO:0000377	RNA splicing, via trans	123/698	379/18670	[illegible]	[illegible]	3.54E-79	HSPA8/HNRNPM/HNRNPK/DHX9/HNRNPA1/	123
BP	GO:0000398	mRNA splicing, via spli	123/698	379/18670	3.40E-82	4.72E-79	3.54E-79	HSPA8/HNRNPM/HNRNPK/DHX9/HNRNPA1/	123
BP	GO:0006401	RNA catabolic process	[illegible]	[illegible]	7.94E-57	8.82E-54	6.63E-54	VIM/HSPA8/HNRNPM/DHX9/HNRNPU/DDX5	102
BP	GO:0006402	mRNA catabolic process	98/698	364/18670	[illegible]	[illegible]	[illegible]	[illegible]	98
BP	GO:1903311	regulation of mRNA meta	91/698	324/18670	[illegible]	[illegible]	1.66E-51	[illegible]	[illegible]
BP	GO:0019080	viral gene expression	62/698	191/18670	3.09E-41	2.14E-38	1.61E-38	[illegible]	[illegible]
BP	GO:0019083	viral transcription	[illegible]	[illegible]	[illegible]	5.80E-35	4.36E-35	[illegible]	57
BP	GO:0000184	nuclear-transcribed mRN	[illegible]	[illegible]	[illegible]	4.53E-33	3.40E-33	[illegible]	[illegible]
BP	GO:0006413	translational initiatio	56/698	193/18670	2.24E-34	1.13E-31	8.49E-32	TPR/RPS3/EIF4A1/HSPB1/RPS4X/NPM1/	56

GO编号

GO term的内容

样本比率

背景比率

从背景中获取x个基因,有y个属于该term的概率

在该类别下的GeneName

基因的数量

在本次样本中出现该term的比例

p.adjust是对一组p值进行的矫正 这里使用的矫正方法是“BH”

p = 从18670个球（有469个是红色）中不放回挑选出698个，而至少有138个球为红色的概率

整个基因组中含有该term的比例大小

q=在该概率下的分位数?（有待验证）

Exp03.py × Test_H.py

```python
from scipy import stats

p = stats.hypergeom.sf(137,18670,469,698)
print(p)
```

```
3.518731359892522e-86
[Finished in 1.7s]
```

$$P(X=k)=\frac{C_k^M C_{N-M}^{n-k}}{C_N^n}$$

N个球中有M个红球，先不放回抽取n个，有k个为红色的概率

$$p=P(X\geq m)=\sum_{m\leq k\leq n}^{k} P(X=k)=1-CDF$$

http://geneontology.org/docs/go-enrichment-analysis/