

GO 富集分析说明

基因本体(Gene Ontology ,GO)是一个在生物信息学领域中广泛使用的本体。1998年由研究三种模式(果蝇、小鼠和酵母)基因组的研究者共同发起组织了一个称为基因本体联盟的专业团队。创建基因本体的初衷是希望提供一个可具代表性的规范化的基因和基因产物特性的术语描绘或词义解释的工作平台。现在已包含数十个动物、植物、微生物的数据库。基因本体涉及的基因和基因产物词汇分为三大类,涵盖生物学的三个方面:

细胞组分 (cellular component): 细胞的每个部分和细胞外环境。

分子功能 (molecular function): 可以描述为分子水平的活性 (activity), 如催化 (catalytic) 或结合 (binding) 活性。

生物过程 (biological process): 生物过程系指由一个或多个分子功能有序组合而产生的系列事件。其定义有广义和狭义之分,在词义上可以区分为泛指和特指。一般规律是,一个过程是由多个不同的步骤组成。**注:生物过程与途径或通路 (pathway) 不是同一回事。**

通过将差异基因做 GO 富集分析,可以把基因按照不同的功能进行归类,达到对基因进行注释和分类的目的。

富集分析方法说明:

采取的方法是 fisher 精确检验,数据包是 clusterProfiler, 来自 R/bioconductor;

挑选的标准是落在某个 term/GO 上差异的基因数目 ≥ 2 , $p_value < 0.05$,

画图中取得 term/GO 是按照 enrich factor 的值从大小降序排列,取前 30 个结果。

enrich_factor 定义 = (某个 term 中的差异基因数目/总的差异基因数目)/(数据库 term 中总的基因数目/数据库中总的基因数目)

字段说明:

ID	对应 GO 数据库中的 ID
Description	GO 的描述
GeneRatio	对应 GO 差异基因数 / 能够对应到 GO 数据库中同类型的差异基因数
BgRatio	对应 GO 包含对应物种的基因数 / GO 数据库中包含对应物种的基因数
pvalue	富集分析得到的 p-value

p.adjust	校正后的 p-value
qvalue	富集分析得到的 qvalue
geneID	富集基因列表
Count	富集基因数目
TYPE	所属三大分类类别
enrich_factor	富集程度

上海伯豪生物技术有限公司

技术支持部

2016 年 01 月