实验三：GO 富集分析

2021 年 5 月 27 日

学　　校:	华中农业大学
学院班级:	信息学院生信 1801 班
姓　　名:	邓启东
学　　号:	2018317220103
指导教师:	夏静波

目录

1 实验目的

使用 clusterProfiler 包进行 GO、KEGG 的富集分析方法，结果输出及内置的图形展示。

2 实验材料与方法

2.1 GO(Gene Ontology) 介绍

GO 是 Gene Ontology 的简称，是基因功能国际标准分类体系。它旨在建立一个适用于各种物种的，对基因和蛋白质功能进行限定和描述的，并能随着研究不断深入而更新的语言词汇标准。GO 分为分子功能（Molecular Function）、生物过程（Biological Process）、和细胞组成（Cellular Component）三个部分。

2.1.1 基因注释数据库 KEGG

京都基因与基因组百科全书 (Kyoto encyclopedia of genes and genomes, KEGG)，是系统分析基因功能与基因组信息的数据库，它整合了基因组学、生物化学和系统功能组学的信息，有助于研究者把基因及表达信息的过程作为一个网络进行整体研究。

2.2 获取差异表达基因数据

2.2.1 阿兹海默症 (AD) 简介

阿尔茨海默症（Alzheimer's disease, AD），俗称“老年痴呆症”，是当今世界范围内患病最广泛、病情最严重的神经退行性疾病，患者通常会出现以记忆力衰退、学习能力减弱为主的症状，并伴有情绪调节障碍以及运动能力丧失，极大地影响个人、家庭乃至社会的发展。

2.2.2 差异基因数据来源

我们的差异基因来自埃默里大学医学院发表的一篇文章,团队运用全蛋白质组关联研究 (proteome-wide association study，PWAS)，将阿尔茨海默症 (AD) 队列 GWAS 结果与人脑蛋白质组进行了整合，旨在鉴定通过影响脑蛋白丰度而导致 AD 风险的基因，深入了解这些基因座如何影响 AD 的发病机制 [1]。

该研究团队鉴定了 11 个与 AD 发病相关的基因，它们通过顺式调节的脑蛋白丰度发挥作用。9 个在 PWAS 验证队列中重现。我们的差异基因正是这已知的九种。

根据 map 的结果显示有 22.22% 的基因没有得到相应的 ENTREZID,分别是”CARHSP”,”STX” 这两个基因。最终的差异表达基数据输入为 7 个。

2.3 背景基因获取

如果 clusterProfiler 包没有所需要物种的内置数据库，可以通过自定义注释文件或者自建注释库的方法进行富集分析。待富集的背景基因是由函数：list data(geneList, package) 获取的，这里使用的是”DOSE” 这个库。

2.4 富集原理

根据挑选出的差异基因，计算这些差异基因同 GO 分类中某（几）个特定的分支的超几何分布关系，GO 分析会对每个有差异基因存在的 GO Term 返回一个假定值 p-value，小的 p 值表示差异基因在该 GO 中出现了富集。

3 结果及分析

3.1 差异表达基因 GO 富集分析

GO 富集分析结果主要由有向无环图 DAG 和柱状图两种形式所表示。

3.1.1 DAG

得到的有向无环图如下图??所示。结果中箭头"→" 代表 GO term 的上下层级关系，圆圈代表富集程度未在前十的 GO term，方格代表富集程度在前十的 GO term。

图形的颜色反应了基因在某一个 GO Term 上的富集程度，颜色越深代表富集程度越显著。GO Term 的层级越低，功能描述越具体。

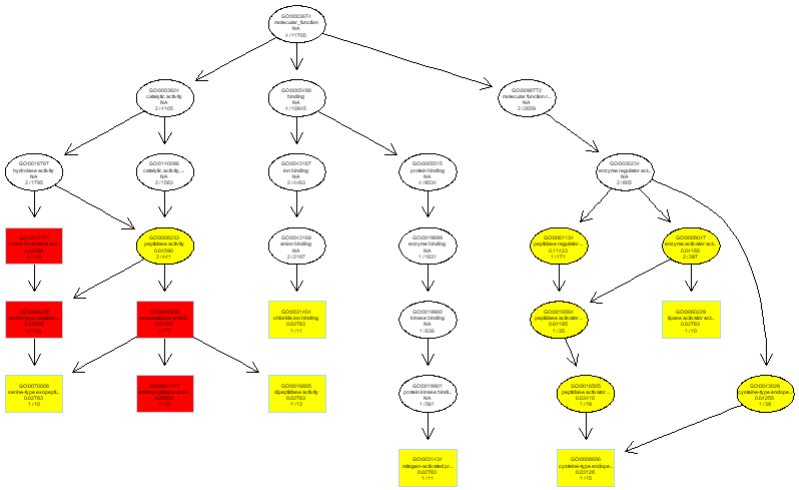

图 1: 差异表达基因富集分析 DAG 图

3.1.2 柱状图

柱状图结果如下图 (图2) 所示：

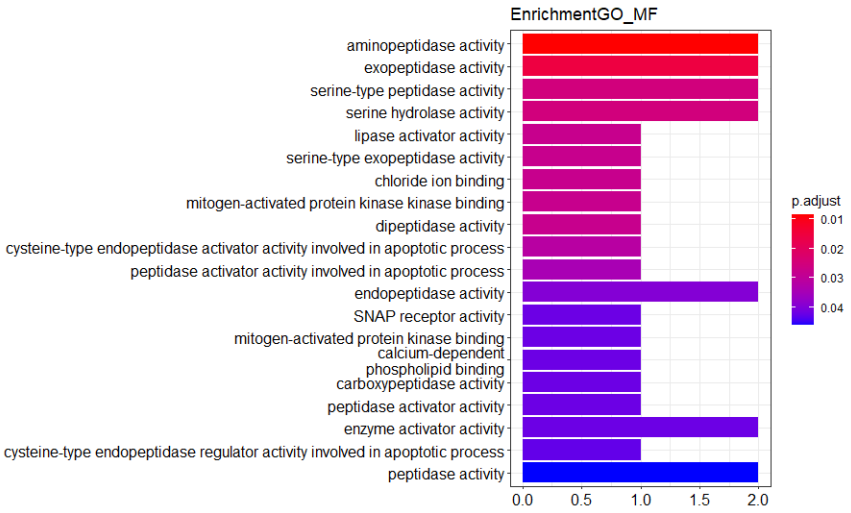

图 2: 差异表达基因富集分析柱状图

可以看到富集结果按照按照 p 值排名前十的分别是：

aminopeptidase activity（氨肽酶活性）、exopeptidase activity（外肽酶的活动）、serine type peptidase activity（丝氨酸型肽酶活性）、lipase activator activity（脂肪酶催化剂活性）、serine type exopeptidase activity（丝氨酸型外肽酶活性）、chloride ion binding（氯离子结合）、mitogen activated protein kinase kinase binding（丝裂原激活蛋白激酶激酶结合）、dipeptidase activity（二肽酶的活动）、cysteine-type endopeptidase activator activity involved in apoptotic process（半胱氨酸型内肽酶激活物活性参与凋亡过程）、endopeptidase activity（肽链内切酶活性）。

可以观察发现这个当中基本上每一个都涉及到酶的活性。通过相关文献检索发现，事实上阿兹海默症的确和酶的活性、结合有关。例如施一公团队发现了阿尔兹海默重要蛋白　分泌酶 [2]。该酶便能水解淀粉样蛋白。而阿兹海默症的症状之一就是淀粉样沉积。

关于出现的第六条氯离子结合，我们也在一个电分析化学学术会议文章中找到证据，通过微传感用于 AD 及缺血模型下鼠脑中三个脑区 (海马, 纹状体, 皮层)Cl -的测定，发现氯离子 (Cl -) 作为人体常见的阴离子在 AD 的病理过程起着重要作用(链接)

其中的肽链内切酶活性，也可以找到相应证据：天冬酰胺内肽酶（asparagine endopeptidase，AEP）也被称为豆荚蛋白，它是一个溶酶体半胱氨酸蛋白酶，能够将蛋白 c 端的天冬氨酸剪切掉。编码 AEP 基因缺失的 tau P301s 转基因小鼠，其 tau 的过度磷酸化被降低，突触损害降低以及认知障碍得到改善。这些结果表明，AEP 在 AD 中发挥了关键作用，抑制 AEP 或许能够成为治疗 AD 的有效方法 [3]。

4　结论与展望

本次实验，从文献当中找到了 GWAS、PWAS 等方法通过测序得到的一些阿兹海默的关联基因，但是由于 GWAS 的假阳性等以及缺乏证据等原因，这些基因还有待于进一步证实。通过基因富集分析，我们能够了解到这些突变位点具体和哪些分子层面的性状有关，这些性状再经过人工的文献检索，发现的确已经有相关工作作为证据表明其确实与疾病相关联。也印证了这一套基因富集分析的流程有利于我们发现疾病的一些潜在机理。

当然我们现在是拿着结果在验证。如果发现了新的知识，光靠富集分析还不足以作为可信的证据，有待后续进一步的分析和确认。

5 Github 链接

github 链接：https://github.com/LianzePuppet/nlphomework3

参考文献

[1] A. P. Wingo, Y. Liu, J. Gockley, B. A. Logsdon, and T. S. Wingo. Integrating human brain proteomes and genome©wide association results implicates new genes in alzheimer's disease: Functionalizing genetic variants in alzheimer's disease. *Alzheimer's and Dementia*, 16(S3), 2020.

[2] G. Yang, R. Zhou, X. Guo, C. Yan, and Y. Shi. Structural basis of ¦ -secretase inhibition and modulation by small molecule drugs. *Cell*, 184(2), 2020.

[3] J. Wang, H. J. Hu, Z. K. Liu, J. J. Liu, and M. Song. Pharmacological inhibition of asparaginyl endopeptidase by ¦ -secretase inhibitor 11 mitigates alzheimer's disease-related pathologies in a senescence-accelerated mouse model. *Translational Neurodegeneration*, 10(1), 2021.