组　别: 本科生

题　目: B

队　号: 030

北京航空航天大學
BEIHANG UNIVERSITY

2019 年数学建模竞赛

biaoti

李亦龙 18373580

叶凡 18374449

栾帅 18373298

队伍联系电话:13718250032

队伍联系邮箱:18373580@buaa.edu.cn

摘要

本文基于 GEO 数据库公开的对于某些疾病的多个样本的基因测序结果,

队伍声明

我代表参赛队伍全体队员声明，本论文及其研究工作是由队伍成员独立完成的，在完成论文时所利用的一切资料均已在参考文献中列出

目录

1 问题重述

基因共表达网络 (Gene co-expression network) 是现代生物基因工程研究的重要方向, 也是研究基因与基因之间对于在某疾病上的表达影响的重要方法. 网络中节点代表基因, 节点与节点之间的连线代表基因之间的联系. 一张某疾病的基因共表达网络形如:

图 1: 基因共表达网络示例[?]

根据对表达出某疾病的许多样本的基因测序结果, 我们可以通过对同时出现的基因进行统计, 认为同时出现次数越多的基因可能会对该性状起阳性结果, 而出现次数较少的可以认为对该性状起阴性结果, 次数适中的认为是无关基因.

2 假设与符号

2.1 假设

I. 基因与基因之间相对独立, 即全体样本中基因频率为定值

II. 基因 A 与基因 B 之间的关系有三种: 正相关, 负相关和无关

III. 视在实验选定的样本里出现次数多的基因对该疾病有促进作用

IV. 关于基因之间的关系, 我们认为有正相关, 负相关, 无关 3 种情况:

i. 认为正相关即出现频率相近或者 p 值接近

ii. 认为负相关为出现频率相加接近 1

iii. 认为无关为出现频率接近其原基因频率

2.2 符号

3 问题分析

3.1 基因之间的联系表达

基因之间的复杂的表达关系无法直接用数据表达, 因为基因 A 可能会对基因 B 起正相关, 对基因 C 起负相关, 而基因 B 和 C 对疾病都可能呈现正相关效果. 为了解释清楚这种关系, 我们将 A,B,C 与疾病之间的复杂关系描述为 A 与 B, B 与 C, A 与 C 之间的关系, 即相当于描述为网络中节点之间的关系.

由于基因

3.2 等位基因

等位基因指位于一对同源染色体相同位置上控制同一性状不同形态的基因.

在 GEO 数据库中, 等位基因会使用多个编号来标记, 即若 A, B 互为等位基因, 则 A 与 B 的 ENSG 编号是不同的. 这样就代表着在数据中会包含互补的数据 (即等位基因). 这样的数据一般呈现 A 与 B 的出现频率相加约等于 100% 的情况, 与我们认为的负相关情况类似, 所以我们不单独考虑等位基因对图中联系造成的影响.

kanbudong[?]wosss

4 求解问题

4.1 皮尔逊积矩相关系数

对于两组数据, 我们可以利用皮尔逊积矩相关系数[?] 来求他们之间的相似程度.

皮尔逊积矩相关系数 (Pearson product-moment correlation coefficient，又称作 PPMCC 或 PCCs, 文章中常用 r 或 Pearson's r 表示) 是一种比较非中心化数据相似程度的一种方法, 该方法的特点在于得到的结果不受线性变换的影响, 即结果与数据的单位等无关.

由于基因的测量量单位 FPKM(Fragments Per Kilobase of exon model per Million mapped fragments, 每千个碱基的转录每百万映射读取的 fragments) 与基因本身的规模有关, 所以数据库中得到的数据的均值由于基因的变化会发生较大的升降. 为了抵消这种对数据本身的拉伸对数据特征造成的破坏, 我们选择使用皮尔逊积矩相关系数这种更加本质的特征, 它反映了两组数据在线性上的相似性, 而非数据大小.

皮尔逊积矩相关系数的取值范围在 $-1, 1$,

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X \sigma_Y} \tag{1}$$

$$r = \frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}} \tag{2}$$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma_X}\right)\left(\frac{Y_i-\overline{Y}}{\sigma_Y}\right) \tag{3}$$

结论

附录

支撑材料文件列表