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

基于皮尔逊积矩相关系数的基因共表达网络

——基因关系数据分析以及可视化方法探究

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

肺癌是对人类健康危害最严重的恶性肿瘤之一, 其发病率以及致死率均居恶性肿瘤之首. 非小细胞肺癌约占肺癌的 80%-85%, 治疗难度大, 传统治疗方案以手术切除为主. 而随着分子生物学技术的发展, 微阵列芯片技术尤其是基因芯片在生物学研究中发挥着越来越重要的作用, 也为从基因层面寻求非小细胞肺癌的治疗方式提供了契机.

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数据, 依照基因之间的关系规律, 利用皮尔逊积矩相关系数这一工具对数据进行分析, 最后使用 python 中的 matplotlib 库将数据可视化.

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 数据来源以及初步分析

2.1 数据来源

GEO 数据库 (Gene Expression Omnibus) 是由美国国家生物信息中心 (National Center for Biotechnology Information) 提供维护服务的公共基因表达数据资源. 该数据库储存实验采集的原始数据并将其公开, 我们通过对 GEO 数据库上某个实验相关样本数据集系列 (Series) 进行分析, 就可以获得对基因间关系的深入认知, 以实现对疾病诱导原因的深入研究, 以及疾病的辅助治疗.

由于队伍成员设备运算能力的限制, 在处理单个系列时需要耗费大量时间 (约 193 小时), 最后我们决定暂时只以 GSE81089 为数据来源, 计算非小细胞肺癌 (non-small cell lung cancer, NSCLC) 的基因关联网络.

2.2 疾病介绍

肺癌是世界上最常见的恶性肿瘤之一, 并且肺癌发生过程涉及的分子机制仍未被全部解开. 非小细胞肺癌[?] 包括鳞状细胞癌, 腺癌, 大细胞癌, 与小细胞癌相比其癌细胞生长分裂较慢, 扩散转移相对较晚, 对化疗相对不敏感. 非小细胞肺癌约占所有肺癌的 85%, 75% 的患者发现时已处于中晚期, 5 年生存率很低. 而在非小细胞肺癌中, 病变部位的组织学分类决定了患者治疗的方式. 传统治疗方案以针对病变部位的组织通过手术切除治疗为主, 可以说治标不治本. 如果我们可以得知诱导 NSCLC 出现的基因是什么, 那么通过无效化这个基因就可以避免 NSCLC 的出现, 这样就形成了一种新的治疗方法.

2.3 初步分析

我们下载的数据是 GEO 上储存的经过实验团队初步处理的 NSCLC 的基因表达数据, 这是实验团队将 199 个来自 seq 的患病样本和 142 个正常样本进行基因测序比较后的结果. 经过实验团队的初步处理, 我们已经可以利用比较便利的方法读取数据, 而不是对实验原始数据进行文本处理.

3 假设与符号

3.1 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. 认为正相关即相关系数 $\rho \in [\alpha, 1]$

ii. 认为负相关为相关系数 $\rho \in [-1, \alpha]$

iii. 认为无关为不属于上述二者的数据

3.2 符号

表 1: 符号说明

符号	意义
ρ	皮尔逊积矩相关系数
α	相关系数中的参数, 用于确定基因之间的关系
X/Y	互相比较的两列数据, 即两个基因所对应的的所有样本上的数值

4 问题分析

4.1 基因之间的联系表达

基因之间的复杂的表达关系往往无法直接用数据表达, 因为基因 A 可能会对基因 B 起正相关作用, 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.

由于基因之间的相对独立性, 我们可以在考虑 AB 之间的互相作用是忽略 C

4.2 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, 而是将等位基因看做普通基因以参与计算.

5 求解问题

5.1 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.

下列三式可以求的不同的皮尔逊积矩相关系数, 我们在这里采用第一个式子.

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

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

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

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 $[-1, 1]$, 假设参数 α, 又假设 IV 中的内容, 只要选定了 α 的值, 我们就可以将基因之间的关系区别为上述三种关系, 这样形成的矩阵无疑会对后续研究带来便利. α 的大小直接对结果造成影响, 根据文献[?], 我们选取 $\alpha = 0.9$, 这样可以得到数据量不是特别巨大而且又可靠的联系矩阵.

我们利用 python 计算了 GSE81089 的数据, 所得到的矩阵可以在支撑材料里找到.

5.2 结果的可视化

仅仅获得相关系数的矩阵只是数据分析的一步, 要获得对于后续研究有所作用的直观图表, 我们需要将上述矩阵可视化. 基于假设 I 和 IV, 我们可以认为基因在图中是点, 基因与基因之间的关系用不同颜色的线来代替. 如图 2, 其中红色的线表示该线所连接的一对基因之间的关系是正相关, 蓝色的线表示负相关. 无关的基因之间没有连线.

图 2: 简化的基因共表达网络

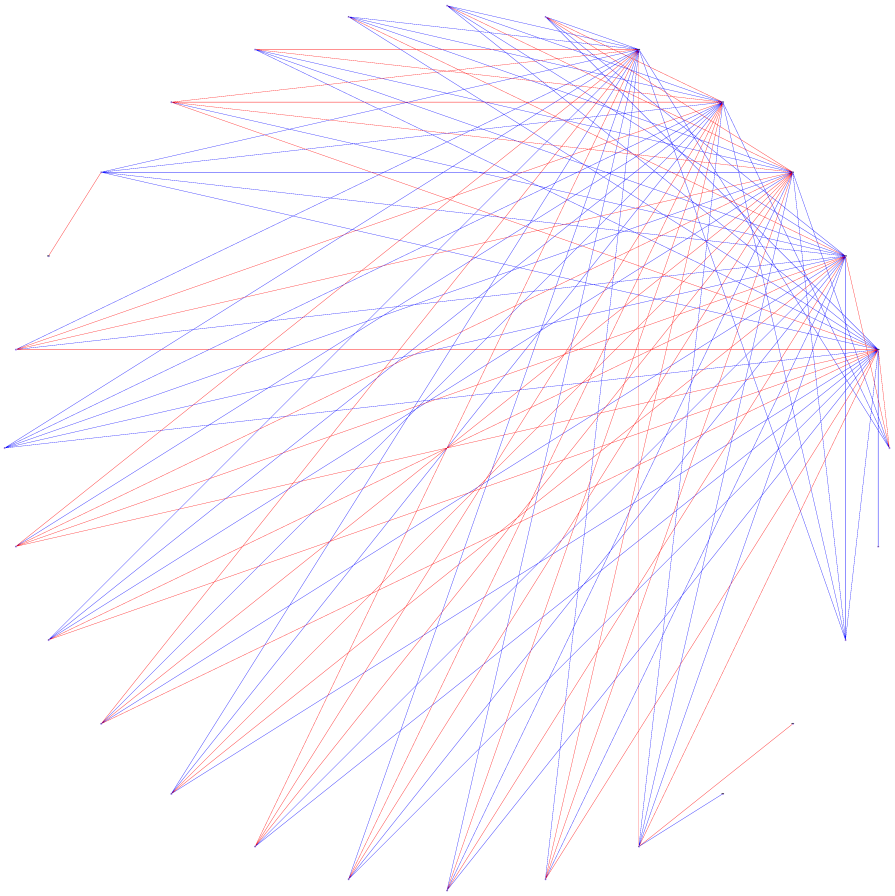

图 2只是整张图的一部分. 为了可以在论文里清晰的看出基因之间的联系, 我们舍弃了部分基因, 采取一部分基因来画出图 2. 完整的图可以在支撑材料里找到.

结论

人体基因数量有 60000+, 这样做出的两两关系在理论上不超过 3.6e9 个, 这样的数字仅凭人力是无法处理的, 而我们通过相关系数的方法将这个数量降到了 20000 以下, 即便这样, 可视化方法画出来的图依旧是一团乱麻, 难以分析.

实际上, 如果采用上述提取的数据辅以基因本身的名字以及特性, 研究人员可以由他们的专业知识对其中某几个或者某几十个基因提取出相关因子, 这样就可以辅助研究的进行.

综上, 我们提出了一种对于基因共表达关系分析的方法, 并基于这种方法对于某肺癌疾病的基因测序数据利用 python 进行了处理, 得到了我们想要的数据.

附录

支撑材料文件列表

```
/
├── data ..... 从 GEO 下载的实验数据
│   └── FPKM-cufflinks.tsv ..... GSE81089 的实验数据
├── module
│   ├── data-proceed.py ..... 处理数据得到相关系数矩阵脚本
│   ├── math-model.ipynb ..... 数据可视化的 jupyter 文件
│   ├── visualizeV1.py ..... 第一版本的数据可视化
│   └── visualizeV2.py ..... 第二版本的数据可视化
├── result
│   ├── plot100.png ..... 取其中 100 条关系所绘制的图像
│   ├── plotfull.png ..... 所有关系绘制的图像
│   └── plotmid.png ..... 其中一部分关系绘制的图像
└── readme ..... 文件列表
```