



组 别 本科生
题 目 B
队 号 020

北京航空航天大学
B E I H A N G U N I V E R S I T Y

2019 年数学建模竞赛

基于互信息的前列腺癌基因网络构建

姓名及学号	陈宇轩 17374384
姓名及学号	朱旭 17374376
姓名及学号	赵婉如 17373240
队伍联系电话	13001101899
队伍联系邮箱	446816968@buaa.edu.cn



摘要

本文使用一种基于互信息的方法对于前列腺癌这一疾病进行了基因谱表达分析（基因谱的数据集 GDS2545），根据基因间的关系强弱来预测潜在的患病基因。首先对靶点错误基因数据进行剔除、合并处理，再根据信息熵原则对不同的样本进行加权、正则处理，得到任意两两基因的相关边权，即成功构建基因表达网络。后处理中对所得到的结果进行了可视化处理，并采用了两种方法：直接邻居法和度分析法进行了基因网络分析，预测了可能的致病基因，为相关医学研究提供参考。

关键词：互信息 基因表达谱 前列腺癌 直接邻居法 度分析法



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

摘要.....	1
队伍声明.....	2
一、绪论.....	4
1.1 研究背景和意义.....	4
1.1.1 基因芯片的测序.....	4
1.1.2 信息熵.....	4
二、模型建立与求解.....	5
2.1 基本假设.....	5
2.2 相关符号解释及定义.....	5
2.3 基因表达各样本的权重确定.....	6
2.4 计算信息熵及相关系数.....	7
2.5 可视化及后处理.....	9
2.6 结果讨论与分析.....	10
三、模型评价与检验.....	10
3.1 直接邻居评价.....	10
3.2 度分析评价.....	13
四、模型评价与展望.....	14
4.1 模型的评价.....	14
4.1.1 模型的优点.....	14
4.1.2 模型的不足.....	14
4.2 模型的改进与展望.....	15
五、主要参考文献.....	16
附录.....	17



一、绪论

1.1 研究背景和意义

1.1.1 基因芯片的测序

随着人类基因组（测序）计划的逐步实施，越来越多的动植物、微生物基因组序列得以测定，基因序列的数据正在以前所未有的速度迅速增长。为此，建立新型杂交和测序方法以对大量的遗传信息进行高效、快速的检测、分析就显得格外重要了。应运而生的就是基因芯片技术，它可以通过微加工技术，一次性检测数以万计的 DNA 片段，测得相关基因的表达情况。这样得出的数据就是基因表达数据表，基因表达数据的分析和应用是现代生物学最重要和最丰富的研究领域之一，通常来说，基因表达数据中隐含着基因之间的相互作用关系，因而可以通过分析基因表达数据来构建基因调控网络，从基因表达数据中寻找分子间的直接关联或依赖关系对于重构生物网络意义重大。常用的处理基因表达数据的方法模型有：布尔网络模型^[1]，线性组合模型^[2]，加权矩阵模型^[3]，互信息关联网络^[4]，微分方程模型^[5]。本文将采用互信息关联网络来建基因关联网络并分析网络结构特征。

1.1.2 信息熵

熵是信息论中的一个非常重要的概念，它表示某一种能量在空间中分布的均匀程度，能量分布越均匀，越不确定，熵也越大。Shannon 将热力学中的概念运用于信息处理，提出了“信息熵”的概念[15]。信息熵是信息的量化度量，也表明了信息中随机变量取值的不确定性程度。因此对于某一个体而言，它能带来的信息越多，这个个体也就越重要。令 x 为随机变量的不同取值， $x_i (i=1,2,\dots)$ ，对应着不同的概率 $P(x_i) (i=1,2,\dots)$ ，那么 x 的信息熵定义为

$$H(X) = -\sum_i P(x_i) \log_2(p(x_i)) \quad (1-1)$$

通过式（2-1）可以看出随机变量的变化越多，通过它取得的信息量就越大。

信息熵具有如下的基本性质：

- 1、单调性，发生概率越高的事件，其携带的信息量越低；
- 2、非负性，信息熵可以看作作为一种广度量，非负性是一种合理的必然；
- 3、累加性，即多随机事件同时发生存在的总不确定性的量度是可以表示为各事件



不确定性的量度的和，这也是广度量的一种体现。

而两个基因表达模式的互信息公式为：

$$MI(A, B) = H(A) + H(B) - H(A, B) \quad (1-2)$$

互信息是两个变量之间的量度，直观上理解，若 $MI(A, B) = 0$ ，则表达两个基因不相关，反之说明二者的生物关系十分密切。

二、模型建立与求解

2.1 基本假设

1、本文所采用的数据来自 NCBI 网站，假设其数据真实可靠，不考虑由于技术提取层面带来的误差。

2.2 相关符号解释及定义

符号	解释
Cov	基因向量的协方差矩阵
sgn	符号函数，大于 0 为 1，小于 0 为 -1
H	信息熵
$MI(X_i, Y_j)$	X_i 与 Y_j 的互信息
b_i	样本 i 对应的权重

表 1

定义一、基因表达谱列表，设 (X, Y) 是二维离散定性指标变量，纵坐标为基因种类，横坐标为不同的样本，列表中 n_{ij} 是基因表达量，单位为 FPKM (Reads Per Kilobase of exon model per Million mapped reads)。



X \ Y	样本 1	...	样本 J	...	样本 S
基因 1	n_{11}	...	n_{1j}	...	n_{1s}
...
基因 I	n_{i1}	...	n_{ij}	...	n_{is}
...
基因 R	n_{r1}	...	n_{rj}	...	n_{rs}

表 2

定义二、对于某一样本基因谱 $A = (a_{ij})_{n \times m}$ ，利用适当方法把它正则化为 $R = (r_{ij})_{n \times m}$ ，则第 j 个属性的信息熵为：

$$E_j = \frac{1}{\log n} \sum_{i=1}^n r_{ij}' \log r_{ij}', \text{ 其中 } r_{ij}' = \frac{r_{ij}}{\sum_{i=1}^n r_{ij}}, \text{ 且规定, 当 } r_{ij}' = 0 \text{ 时有 } r_{ij}' \log r_{ij}' = 0。$$

定义三、样本基因谱 $A = (a_{ij})_{n \times m}$ 的第 j 个属性权重定义为：

$$b_i = \frac{1 - E_i}{\sum_{i=1}^m (1 - E_k)} \quad (2-1)$$

如上定义会有以下的性质：

当某个样本某一属性值相差越大，该属性的信息熵就越小，所占的权重也就越大，说明该属性可以更好的区分样本，所含的信息价值更大。而信息熵越大，所占的权重越小，说明该属性对于区分来说并不能提供非常有用的信息，对该属性考虑的必要也就越小。当某个样本各个属性值都相同时，该属性的信息熵达到最大值，此时权重为零，说明完全无法区分。因此属性的信息熵和权重与样本基因谱有直接关系，且

$$0 \leq E_i \leq 1, 0 \leq b_i \leq 1, \sum_{i=1}^m b_i = 1。$$

2.3 基因表达各样本的权重确定

引入权重是为了将各统计样本性状调整到同一个变异幅度、同一个重要值的变量水

平上。因此，权重必须能客观的反映出各统计性状的变异度或相对稳定度。信息熵是用来度量信息量的大小，也是信息有用程度的一种表现形式。信息熵又完全建立在原始数据的基础上，客观性比较强，所以本文采用了以信息熵确定属性权重的方法。在不同的样本乘上所对应的权重后，可以将每个加权个体的基因表达水平看作是均匀分布的[6]。

2.4 计算信息熵及相关系数

首先对基因表达谱的数据进行正则化处理：

$$x_{ij} = \frac{r_{ij}}{\max_i(r_{ij})} \quad (2-2)$$

处理之后的部分结果如下表所示：

IDENTIFIER	GSM153115	GSM153116	...	GSM153118	GSM153119	GSM153120
31397_at	0.000163463	0.000322291	...	0.002990433	0.003748265	0.002202712
31423_at	0.001082075	0.002032883	...	0.005779044	0.004730413	0.002458807
31709_at	0.003327377	0.003314229	...	0.00616808	0.002982214	0.003454057
...
31818_at	0.00160874	0.000985174	...	0.001863608	0.005007871	0.000637912
32010_at	0.003066562	0.004823931	...	0.008946031	0.006734101	0.006118309
32011_g_at	0.002912769	0.000876249	...	0.003462256	0.002184328	0.003948535

表 3

根据每个样本的属性权重，即可得出每个样本对于数据的贡献程度，拿权重乘上各自样本的表达水平之后，再进行一次标准化处理，标准化的公式为：

$$x'_{ij} = \frac{x_{ij} - \min_i x_{ij}}{\max_i x_{ij} - \min_i x_{ij}} \quad (2-3)$$

由于基因表达谱中的各个样本的基因表达水平均在实验中获得，所得到的为离散的分布频率，但考虑到数据的量足够大，可在等分处理后转化为概率，转换方式如下：

取 $X=0, 1, \dots, 10$. 对于

$$f(X) = \begin{cases} X & X > 0 \\ 0 & X \leq 0 \end{cases} \quad (2-4)$$

$$P_j(X=i) = \frac{\sum_j f(\text{sgn}(x_{ij}' - \frac{i-1}{10}) \times \text{sgn}(\frac{i}{10} - x_{ij}'))}{j} \quad i=1,2,\dots,10 \quad (2-5)$$

即将 0-1 分为十等分的区间，区间大小为 0.1，假设总共有 10 个值，3 个落在 0.3-0.4 之间，那么取值为 0.3 的概率即为 0.3，其余区间同理。等分后的各基因信息熵可用如下公式计算：

$$H_j = -\sum_{i=1}^{10} P_j(X=i) \log_2(P_j(X=i)) \quad (2-6)$$

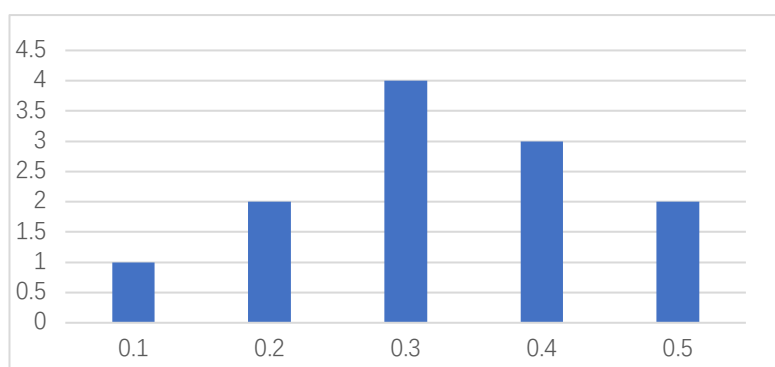


表 4 举例

为了求得两个基因向量之间的关系，我们使用互信息作为评定标准。互信息是一个随机变量中包含的关于另一个随机变量的信息量，或者说是一个随机变量由于已知另一个随机变量而减少的不确定性^[5]

互信息的计算公式为：

$$MI(X_i, Y_j) = H(X_i) + H(Y_j) - H(X_i, Y_j) \quad (2-7)$$

其中 X_i, Y_j 为对应的两个基因向量，事实上，根据高斯核概率密度函数^[7]，上式可以写为：

$$MI(X_i, Y_j) = \frac{1}{2} \log_2 \frac{|Cov(X_i)| \cdot |Cov(Y_j)|}{|Cov(X_i, Y_j)|} \quad (2-8)$$

其中 Cov 表示基因向量的协方差矩阵， $|Cov|$ 表示矩阵的行列式，从式中可以看出，若 X_i, Y_j 相互独立，那么 $MI(X_i, Y_j) = 0$ ，这大大简化了运算。最终采用的标准相关指标采用的计算公式为：

$$r_{ij} = \frac{MI(X_i, Y_j)}{\max(H(X_i), H(Y_j))} \quad (2-9)$$

2.5 可视化及后处理

画出相关系数的热力图如图所示（255 灰度显示）：

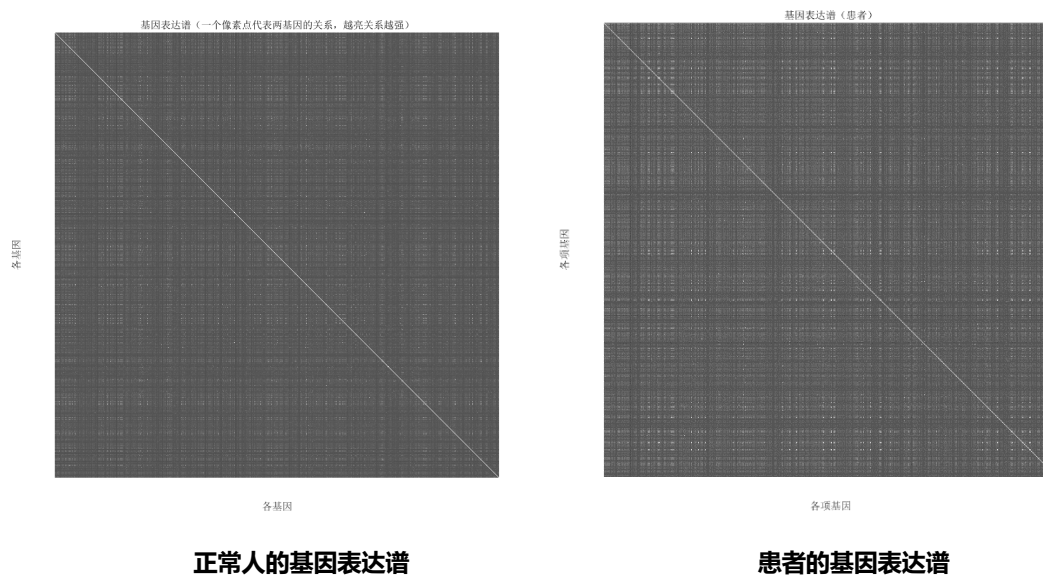


图 1

取小于 0.4 的相关性为 0，则得到如图所示：

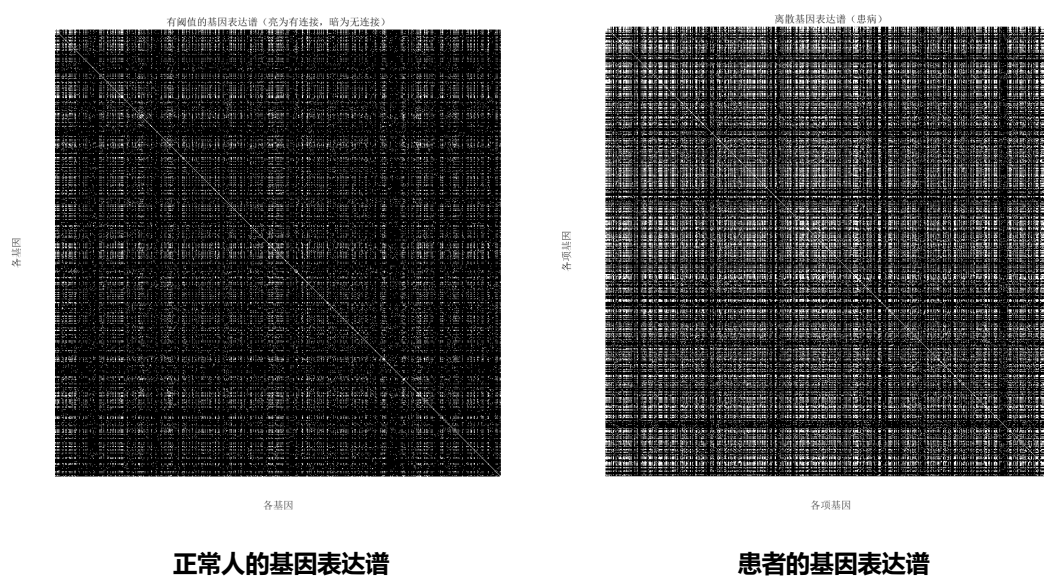


图 2

正常人的基因表达谱（阈值 0.4，即每两个节点之间的相关系数大于等于 0.4 则设为 1，视为有联系，小于 0.4 设为 0，视为无联系）

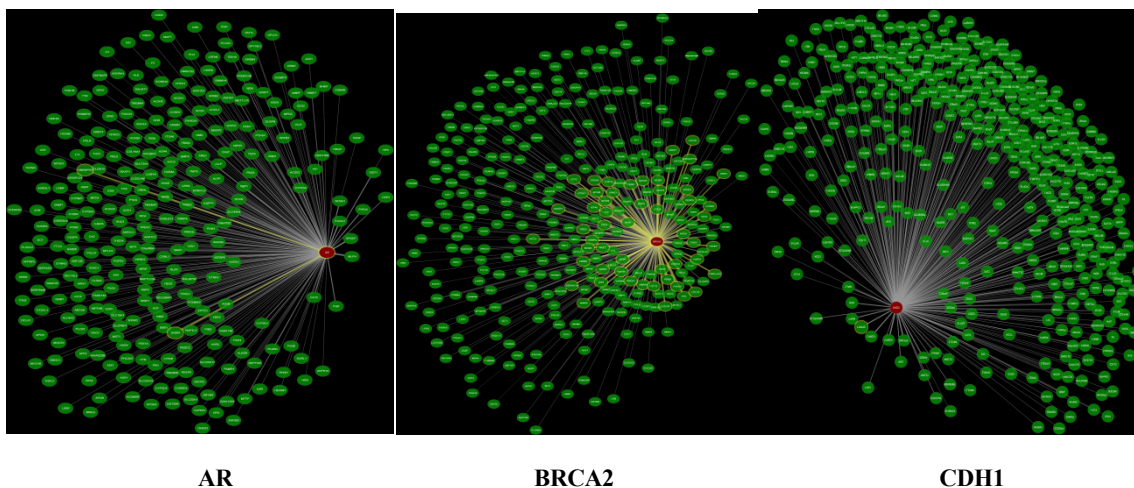
2.6 结果讨论与分析

该结果可以直观的表示任意两个基因的关联强弱关系，但我们也发现了一个问题，那就是存在一些纯色的横纵白纹和横纵黑纹，这说明有一些基因在无论样本状况如何时都表达或都不表达，这些基因对于我们揭示患病基因的内在联系并没有实际的帮助，应在预处理时予以剔除，但由于缺少生物相关知识，没能在建模的初期发现并处理。对于相关基因网络的分析，我们会在第三章模型评价部分介绍。

三、模型评价与检验

3.1 直接邻居评价

本文从 NCBI 的字库 GTR 中收集已知前列腺恶性肿瘤的致病基因： $seed = \{AR, BRCA2, CDH1, CHEK2, KLF6, MAD1L1, MXI1, PTEN, ZFH3\}$ 其在前列腺肿瘤可能基因的节点图为：



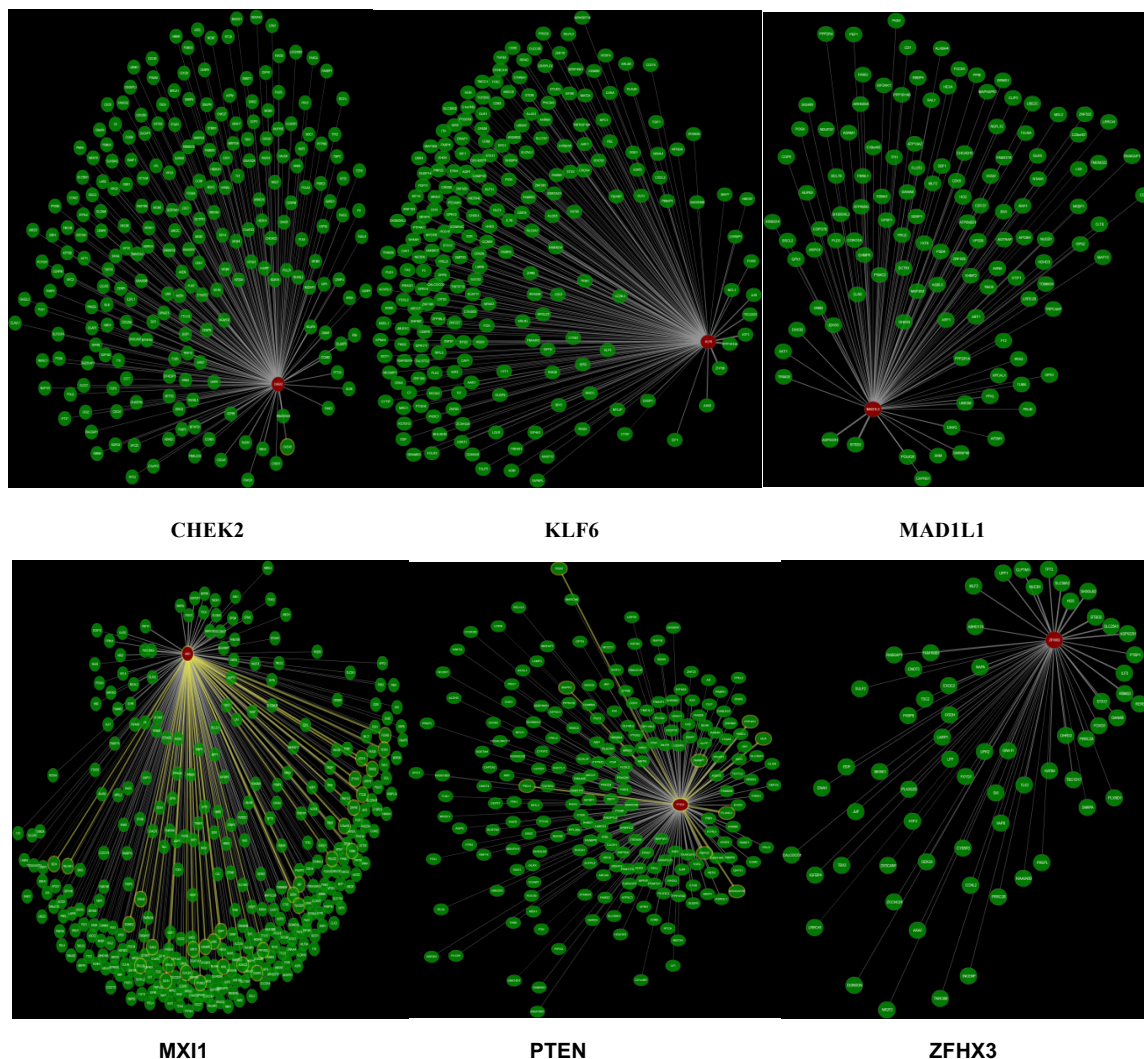


图 3

导入的数据文件包含 9462 个与前列腺癌可能相关的基因，每个基因有 63 个正常人样本点，90 个患者样本点，其中每一行的第一个数据表示 Affymetrix ID 号（Affymetrix 公司生产的基因芯片），每个 ID 都对应着一个基因符号（Gene Symbol）其后的数值表示基因在每个样本点的数值，其中正常人基因表达谱矩阵的部分数据为：

IDENTIFIER	GSM152931	GSM152932	GSM152933	GSM152934	GSM152935	GSM152936
1949_at	5.2	31.9	33.8	5.4	31.3	57.4
31397_at	1.6	3.1	5.5	28.4	35.8	21.1
31418_at	19.8	44.3	48.5	19.8	80.8	52.5
31420_at	10.6	13	8.7	19.6	18.3	27.5
31423_at	14.3	26.4	22.2	74.1	61	31.8
31430_at	20.1	6.9	8.1	41.9	90.7	27



31456_at	63.8	54.9	52.6	41	77.3	86.4
31588_at	42	68.7	57.8	81.1	11.9	39
31655_at	146.5	81.5	182.3	128.7	127.2	120.8
31709_at	56.6	55.4	50.7	101.8	49.5	57.5
31776_at	33.2	78.7	48.5	43.4	15.9	94.9
31818_at	482.4	290.3	1697.2	542.2	1465.3	187.2
32010_at	217.1	335.6	227.4	614.5	465.2	423.9

表 5

基于网络的疾病基因预测方法将已知的疾病基因作为先验信息组成种子集，根据候选基因与种子基因在网络上的拓扑关系，预测候选基因是疾病基因的可能性。我们运用直接邻居法，把与已知疾病致病基因相连的基因作为疾病的可能致病基因，给予全网络对没有个基因进行打分，得出其与已知基因直接相连的总得分 S_i ，即与致病基因直接相连的权重总和。其模型为：

$$S_i = \sum_{j \in seed} r_{ij} \quad (3-1)$$

式中， S_i 表示得分 r_{ij} 表示基因 i 和致病基因 j 的连边权重； $seed$ 表示已知致病基因集。

由此可得全网络中的每一个基因的得分值，再将所有基因依据其分值由大到小进行排序。本文选出 $seed=\{AR, BRCA2, CDH1, CHEK2, KLF6\}$ ，并用 $test=\{MAD1L1, MXI1, PTEN, ZFHX3\}$ 作为测试集，由此可得全网络中每一个基因的得分，再将所有基因依据其分值由大到小进行排序。同时找出 $test$ 中的基因得分，观察其排名：

我们列出分数排名前 23 的基因：

rank	IDENTIFIER	score
1	BASP1	3.277156718
2	IL6	3.063775791
3	EIF1AX	3.054699786
4	ASNS	3.005404971
5	PGK1	2.996436712
6	MVD	2.974546776
7	UBE2C	2.941512551
8	SLC2A3	2.928898529



9	CDKN3	2.926912219
10	HBB	2.923854993
11	TK1	2.923798466
12	CEL	2.907788144
13	AR	2.900100362
14	HS3ST1	2.89153261
15	KLF6	2.884829952
16	KCNJ9	2.866673321
17	LAMA1	2.85744212
18	ITGAM	2.839709327
19	PTH1R	2.837585591
20	GAP43	2.832413326
21	SLC26A3	2.832370971
22	CHEK2	2.831097453

表 6

完整数据在附件中给出。

test={MAD1L1, MXI1, PTEN, ZFH3}4 个致病基因都在靠前的位置 (2.5%, 4%, 17%, 50%), 说明基因网络比较成功。但注意到由于没有进行预处理, 可以发现位于前列基因得分都相差很少 (影响的噪音太多)。而且我们所处理的数据样本容量太大, 而已知的致病基因又太少 (只有 9 个), 对于 GDS2545 表中的九千多个基因来说, 能起到的筛选作用太小, 若采用网上通用的标准数据集进行数据训练, 会对预测结果有非常大的改善。

3.2 度分析评价

对基因网络结构差异的分析, 要用到复杂网络的分析方法。度是复杂网络中常用的一个指标, 一个基因在正常组的度与癌症组的度差别越大, 说明此基因在正常组合癌症组的作用发生了较大的变化, 它在癌症发病中起的作用便越重要, 是癌症的关键基因。

提取了 10 个度差最大的基因, 其度差见表 3:

IDENTIFIER	正常组的度	患病组的度	度差 (绝对值)
RALA	47	5293	5246



BPTF	157	5348	5191
AURKA	25	5134	5109
ATXN2L	95	5118	5023
DIAPH1	99	5113	5014
DDX39A	6	5001	4995
MAPK8IP3	36	5028	4992
AF009267	204	5178	4974
ZMYND8	89	5042	4953

表 7

完整度差表见附件。

而致病基因 AR, BRCA2, CDH1, CHEK2, KLF6, MAD1L1, MXI1, PTEN, ZFH3 都比较靠前, 且上表中的 10 个基因度差也很大, 很有可能是潜在的致病基因。

四、模型评价与展望

4.1 模型的评价

4.1.1 模型的优点

(1) 将医学领域的问题转化到数学领域上进行分析和讨论, 可以通过对大量生物医学数据的分析, 寻找分子间的直接关联关系, 有助于阐明复杂生物过程的分子机制, 最终实现辅助癌症等疾病的治疗诊断目的。具有很强的理论性和可靠性。

(2) 模型中涉及到的参变量都有相应的数据来源, 结合一定的数据可以很方便计算出, 而且各变量间关系明确, 易于模型的求解。

(3) 由于在真实生物分子网络中基因间的非线性相关性扮演者很重要的角色, 基于互信息的建模方法不仅可以从高维低样本基因表达数据中发现基因变量间的线性关系, 也可以发现基因变量间的非线性关系。

4.1.2 模型的不足

(1) 基因网络是一个非常复杂的非线性系统, 数据库中样本数量太大, 而已知信息太少, 且没有预处理去除完全无关基因, 导致构建较大规模的基因关联网络所需计算时

间较长。

(2) 所使用方法存在过高估计的系统误差。假设存在基因变量 X 和 Y 彼此是独立的(即二者相关性为零),它们和变量 Z 都有一定的相关性,运用互信息公式(1-2)计算结点 X 和 Y 之间的相关性,其互信息值大于零。而实际上 X 和 Y 没有相互作用关系,即它间的相关性应该为零,可见互信息法过高地估计了变量之间的关系。

(3) 通过基因芯片数据推测并构建基因关联网络时,由于基因芯片的数据精度有限,且所构建的基因关联网络中的边是无向(双向)的,无法确定基因间调控关系的方向,不能够精确确定关联网络结构。

4.2 模型的改进与展望

本文从基因表达数据出发,首先构建基因完全图,然后采用部分互信息构建稀疏基因相关网络。为了避免互信息存在过高估计系统误差的问题,可以采用条件互信息(CMI)和部分互信息(PMI)[8][9][10]的方法,高精度地构建基因相关网络,提高网络构建的效率和可靠性。

在未来的研究中,为了确定基因间的调控关系,基因间调控方向的确定通过打分策略来实现。可以构建互信息测试打分函数,通过函数值得到可能的有向基因关联网络图。在搜索空间中找出一个得分最高的网络结构,从而确定基因调控方向,有效过滤冗余信息,使基因关联关系更加准确,有助于实现疾病的精准诊断与治疗。

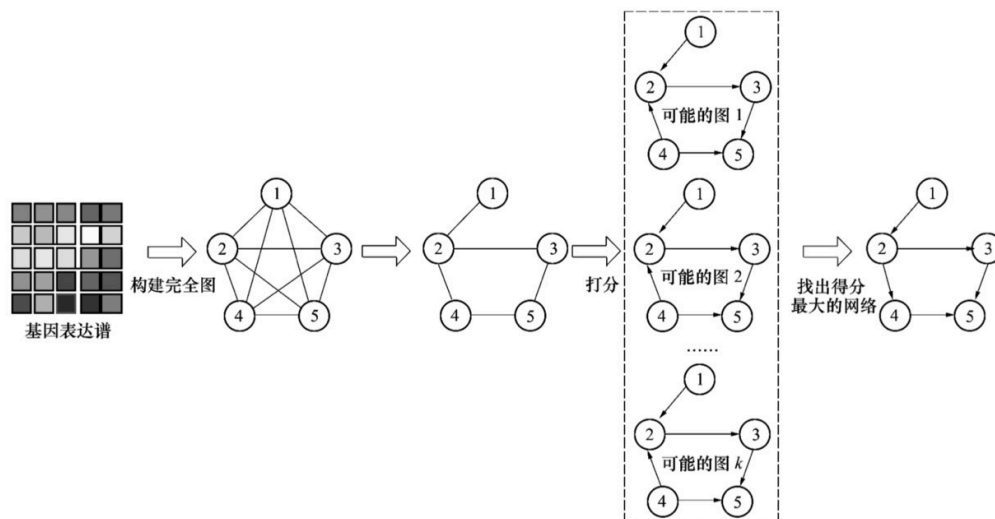


图 4



五、主要参考文献

- [1] Bellenson JL. Expression data and Bioinformatics challenges. In Mark Schemed DNA Microarrays-a practical approach. Oxford: Oxford University Press,1999,97. Andreas Wagner.Bioinformatics,2001,17:s270-s278
- [2] WeaverDC, Workman CT,Stormo GD. Modeling regulatory networks with weight matrices. Pac Symp Biocomput,1999,4:112-123.
- [3] Murphy K,Weiss Y. The Factored Frontier Algorithm for Approximate Inference in DBNs,2001.
- [4] Chen T,He HL,Church GM. Modeling gene expression with differential equations.PacSymp Biocomput,1999,4:29-40.
- [5] 刘青芳. 基于改进互信息的医学图像配准方法研究[D].山西大学,2010.
- [6] 王锐. 信息熵在基因调控网络构建中的应用研究[D].第三军医大学,2008.
- [7] Basso K, Margolin A A, Stolovitzky G,et al. Reverse Engineering of Regulatory Networks in Human B Cells [J] Nature Genetics, 2005, 37(4) : 382-390
- [8] 刘飞,张绍武,高红艳.基于部分互信息和贝叶斯打分函数的基因调控网络构建算法[J].西北工业大学学报,2017,35(05):876-883.
- [9] Janzing D, Balduzzi D, Grosse-Wentrup M, et al. Quantifying Causal Influences [J] . The Annals of Statistics, 2013, 41(5) : 2324-2358
- [10]Schreiber T. Measuring Information Transfer [J] . Physical Review Letters, 2000, 85(2):461-4



附录

1.基因谱网络建立

%%

%预处理

tep = zeros(9462,90);%正则化矩阵

tep2 = zeros(9462,1);%最大值

tep3 = zeros(1,90);%信息熵

tep4 = zeros(1,90);%权重

tep5 = zeros(9462,90);%第一次归一化矩阵

tep6 = zeros(9462,90);%第二次归一化矩阵

tep7 = zeros(9462,1);%最大值-最小值

tep8 = zeros(9462,1);%最小值

for i = 1:9462

tep2(i) = max(sick(i,:));

end

for i = 1:9462

for j = 1:90

tep(i,j) = sick(i,j)/tep2(i);

end

end

%信息熵

for j = 1:90

for i = 1:9462

tep3(j) = tep3(j) +

tep(i,j)*log(tep(i,j))/log(9462);

end

end



%权重

```
sum = 0;
for j = 1:90
    sum = sum + 1- tep3(j);
end
for j = 1:90
    tep4(j) = (1-tep3(j))/sum;
end
```

%第一次归一化矩阵

```
for j = 1:90
    for i = 1:9462
        tep5(i,j) = tep(i,j)*tep4(j);
    end
end
```

%第二次归一化矩阵

```
for i = 1:9462
    tep7(i) = max(tep5(i,:))-min(tep5(i,:));
end
for i = 1:9462
    tep8(i) = min(tep5(i,:));
end

for j = 1:90
    for i = 1:9462
        tep6(i,j) = (tep5(i,j)- tep8(i))/tep7(i);
    end
end
```

%%

%相关系数计算

%tep6

```
count = zeros(10,1);
```

%计算 $H(g(i))$

```
JiYin = tep6;
f = 0:0.1:1;
low = f(1:end-1);
low(1)=-0.01;
up = f(2:end);
%
```



```
H = [];  
for i = 1:9462  
    J = JiYin(i,:);  
    N=[];  
    for j = 1:size(low,2)  
        v = find(J>low(j)&J<=up(j));  
        n = size(v,2);  
        if n~= 0  
            N = [N n];  
        end  
    end  
    p = N/size(J,2);  
    h = 0;  
    for mm = 1:size(p,2)  
        h = h - p(mm)*log2(p(mm));  
    end  
    H = [H h];  
end  
H = H';  
  
%计算 H()  
HH = [];  
MI = zeros(9462,9462);  
  
for i = 1:9462  
    %    Hh=[];  
    for j = 1:9462  
MI(i,j)=cov(JiYin(i,:))*cov(JiYin(j,:))/det(cov(JiYin(i,:),JiY  
in(j,:)));  
        %    Jh=[JiYin(i,:) JiYin(j,:)];  
        %    Nh = [];  
        %    for k = 1:size(low,2)  
        %        vh = find(Jh>low(k)&Jh<=up(k));  
        %        nh=size(vh,2);  
        %        if nh~=0  
        %            Nh = [Nh nh];  
        %        end  
        %    end  
        %    ph = Nh/size(Jh,2);  
        %    hh = 0;  
        %    for mm = 1:size(ph,2)  
        %        hh = hh - ph(mm)*log2(ph(mm));  
        %    end
```



```
%      Hh = [Hh,hh];  
end  
%  HH=[HH;Hh];  
end  
  
%计算 MI () 和 r ()  
  
r = zeros(9462,9462);  
for i = 1:9462  
    for j = 1:9462  
        % if j~=i  
        %      MI(i,j)=H(i)+H(j)-HH(i,j);  
        r(i,j)=MI(i,j)/max([H(i) H(j)]);  
        % end  
    end  
end  
end
```

2. 度分析方法

```
r2_sick = zeros(9462,9462);  
du_sick = zeros(9462,1);  
for i = 1:9462  
    for j = 1:9462  
        if(r_sick(i,j)>0.4)  
            r2_sick(i,j) = 1;  
        end  
    end  
end  
for i = 1:9462  
    for j = 1:9462  
        du_sick(i) = du_sick(i)+r2_sick(i,j);  
    end  
end  
end
```