计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(1):143−148 [doi: 10.15888/j.cnki.csa.006138]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 R 语言的互信息网络模型在乳腺癌易感基因检测分析中的应用①

王淑栋, 张善强, 贺思程

(中国石油大学 (华东) 计算机与通信工程学院, 青岛 266580)

摘　要: 全基因组关联研究 (Genome-wide association studies, GWAS) 是指在基因水平上进行关联分析来寻找致病基因的方法. 传统的研究方法没有考虑到基因之间的相互作用, 而且在复杂的因素情形下往往效率、准确率较低. 针对上述难题, 本文提出一种基于互信息的结构性关键 SNPs 集合选取方法. 在互信息理论和仿真数据的基础之上, 逆向构建 SNPs 互信息网络, 给定互信息一个阈值范围, 找到对应阈值下相关统计量进行比较分析, 选取出合适的阈值. 根据选取的阈值, 筛选出对网络结构有明显影响效果的“结构性关键 SNPs”. 实验结果表明: 本文采用的参数取值方法能够准确快速地筛选出对网络结构有明显影响效果的关键 SNPs.

关键词: 全基因组关联研究; 互信息; 结构性关键 SNPs; SNP-SNP 相互作用网络

引用格式: 王淑栋,张善强,贺思程.基于 R 语言的互信息网络模型在乳腺癌易感基因检测分析中的应用.计算机系统应用,2018,27(1):143–148. http://www.c-s-a.org.cn/1003-3254/6138.html

Application of Mutual Information Network in Detection and Analysis of Breast Cancer Susceptibility Genes Using R Language

WANG Shu-Dong, ZHANG Shan-Qiang, HE Si-Cheng

(College of Computer and Communication Engineering, China University of Petroleum, Qingdao 266580, China)

Abstract: Genome-wide association studies (GWAS) refer to the method that uses correlation analysis to identify disease associated genes. Traditional research method did not consider the interaction between genes and had low accuracy and efficiency in the case of complex factors. Aimed at these aforementioned problems, this paper presents a key SNPs selecting algorithm based on mutual information. It constructs reversely the SNPs interaction network using simulation data based on the theory of mutual information and compares the difference of the statistics of SNPs interaction networks between case and control groups with the increase of the mutual information threshold. According to the selected threshold, we select the structural key SNPs. The results of experiments show that the method of parameter selection presented in this paper is useful to select the structural key SNPs.

Key words: genome-wide association study; mutual information; structural key SNPs; SNP-SNP interaction network

随着人类基因组测序工作的逐步完成, 大量的数据为全基因组关联分析提供了丰富的素材, 也涌现出许多数据分析方法[1-4]. 人类基因组计划得出人类所有的基因共由 39 000 多个已经编码蛋白的基因序列以及 30 亿碱基组成. 而国际单体型图计划[5]得到了 SNP 的 300 万个位点. 两个计划的实施给生物学领域带来了众多的数据信息, 为全基因组研究中提供了方便. GWAS 因其优势得到了很多的应用. 大量研究成果显示关联研究具有很多的优势[6].

Ghoussaini 等[7]在 2012 年针对乳腺癌相关基因进

① 基金项目: 国家自然科学基金 (61572522)
收稿时间: 2017-03-28; 修改时间: 2017-04-20; 采用时间: 2017-04-26; csa 在线出版时间: 2017-12-22

行研究, 共得到了 3 个致病相关的位点, rs10771399 不仅在乳腺癌的发展中起着关键作用, 在骨转移中也有着同样的重要性. 2013 年, 维尔汉姆等[8]关于躁郁症的数据进行分析, 得出与躁郁症相关的 SNP 位点及致病基因. 2014 年, 广川等[9,10]针对心肌梗塞病设计了病例对照实验, 从实验中得到了有关疾病的致病基因和 SNP, 使心肌梗塞病得到了合理的解释.

GWAS 能够帮助人们更好的解释复杂疾病成因, 但是它也有不足. 一方面, 复杂疾病多种多样, 其中的影响因素也很多, 如何确切地得到与特定的功能相联系的位点是个不小的难题; 另一方面, 对于 GWAS 结果, 它在不同群体中的影响程度并不一样; 目前的大部分研究主要针对简单疾病, 没有涉及到基因间的相互作用.

而针对基因间的相互作用, 可以通过互信息建立网络进行表达. GWAS 网络方法将 GWAS 数据进行网络建模, 通过比较疾病数据与对照数据得出的网络的不同, 进行后续的相关统计量的分析及解释.

本文试图通过互信息表示 SNP 之间的相互作用关系, 进而建立 SNP 与 SNP 之间的网络. 在此基础上, 进行全基因组关联研究, 找到结构性关键 SNPs.

1 互信息网络建模

随着生物网络的研究深入发展, 研究者对元素之间的相关性的描述越来越准确, 互信息作为两个元素之间的相关信息度量, 具有很多的优势, 其中最具优势的就是它的熵表示, 不仅是对元素出现概率的表示, 更是体现了元素之间的离散程度及相互之间的关系, 对于给定的两个 SNP 表达序列, 他们之间的数据存在着差异, 而利用互信息可以充分表达 SNP 之间的差异性及依赖性, 互信息越大, 说明两个 SNP 之间的关联程度越紧密; 反之, 则说明联系越小, 从而找到跟所有的 SNP 联系较大的节点, 即是关键 SNP. 本文通过互信息建立相互作用网络, 从而分析网络结构的差异性. 设 $X=(x_1,x_2,...,x_n)$, $Y=(y_1,y_2,...,y_m)$是两个 SNP 的基因型数据在个体之间表达形成的向量, $p(x_i,y_j)(i=1,2,...,n;j=1,2,...,m)$是 X 和 Y 的联合概率分布, $H(X,Y)$是他们之间的联合熵, 定义为:

$$H(X,Y)=\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_i,y_j)\log_2\left[\frac{1}{p(x_i,y_j)}\right] \quad (1)$$

对于两个随机变量之间存在的关系, $H(X)$ 表示随机变量 X 蕴含的不确定性, 而条件熵$H(X,Y)$则是已知条件 Y 时随机变量 X 所余下的不确定性, 那样, $H(X)-H(X|Y)$就表示已知条件 Y 后 X 包含的信息量. 进而还可以证明这个值关于 X 和 Y 是对称的, 即 $H(X)-H(X|Y)=H(Y)-H(Y|X)$, 且都等于$H(X)+H(Y)-H(X,Y)$. 由此 X 和 Y 之间的互信息可以计算, 互信息记为$MI(X,Y)$:

$$MI(X,Y)=H(X)+H(Y)-H(X,Y) \quad (2)$$

$MI(X,Y)$越大, 表明 X 和 Y 的相关程度越大; $MI(X,Y)=0$ 表示 X 和 Y 相互独立. 因此, 互信息表达了两个 SNP 之间相互依赖的程度.

因为 SNP 数据是每个 SNP 仿真 1000 组得到的数据, 每三个数据代表一个个体, 首先需要对数据进行处理使得数据能够表示基因型, 我们确定使用 0, 1, 2 三个数来表示每个个体内表达的基因型, 再根据公式 (2) 计算得到所有的 SNP 之间的互信息. 具体计算过程如下:

(1) 我们首先得到每个 SNP 的基因型可能性序列数据, 假设共有 N 个个体, 则每一行包含 $2N$ 个 SNP 碱基可能性数据, 0 代表出现, 1 代表不出现.

例如: 假定两个个体关于 5 个 SNPs 的基因型数据如下:

SNP 1: AA AA
SNP 2: GG GT
SNP 3: CC CT
SNP 4: CT CT
SNP 5: AG GG

输出的正确仿真数据如下所示:

SNP1 rs1 1000 A C 1 0 0 1 0 0
SNP2 rs2 2000 G T 1 0 0 0 1 0
SNP3 rs3 3000 C T 1 0 0 0 1 0
SNP4 rs4 4000 C T 0 1 0 0 1 0
SNP5 rs5 5000 A G 0 1 0 0 0 1

所以, 在 SNP3 上, 两个等位基因上碱基分别为 C 和 T, 所以每个个体与之相对应的碱基组合 CC, CT, TT 出现的可能性序列分别是 100 和 010.

(2) 每个 SNP 的基因型表达数据作为一个向量, x, y 表示来自 SNP 集合 I 中的其中的两个 SNP 向量.

(3) 根据每个 SNP 的基因型表达量的分布, 计算得到每两个 SNP 之间存在的互信息值. 所有 SNP 之间的互信息构成互信息矩阵, 记作$M=\{W_{ij}\}_{m*n}(|I|=n)$, 矩阵中的每行代表一个 SNP, 每一列代表此 SNP 与另一个 SNP 之间的互信息.

假定存在一个集合的 SNP 基因型数据 D, 其中所拥有的 SNP 的集合我们记作 I, $M=\{W_{ij}\}_{m*n}(|I|=n)$可由互信息计算公式 (2) 得到一个互信息矩阵. 定义一个建立在关于 SNP 基因型数据 D 的互信息网络. $G[D]=(V,E;w)$是边赋权图, 其中 V 表示点集合、每个网络中的节点 $i\in V$ 表示一个 SNP, $\forall_{i,j}\in V$, 基因 i 和 j 之间的互信息计算值 w_{ij} 定义为每条边$(i,j)\in E$的权重. 在下面的表述中, 我们将基因 $i\in I$ 以及顶点 $i\in V$ 等同起来看待.

2 基于网络统计量的关键基因选取

利用上述方法得到的 SNP 相关网络中各节点 (SNP) 的网络结构参数来描述特定生物过程中基因的重要性. 首先给出几个重要的能够反映网络结构特点的网络统计量的相关定义[11].

(1) 度 (K): 在网络中, 度指的是与该点相连接的边数目. 节点度可以表示该点的重要程度, 节点度越大, 表示该点在网络中越重要. 而网络的平均度可以通过计算所有的点的度, 后取平均数计算得到.

(2) 平均路径长度 (L): 定义为网络中所有的点之间两两求得的距离的平均数, 网络中的任意两点 i,j 的距离即边的条数, 则两点之间的平均路径长度表示为所有的点之间的平均距离, 记作: $L=\frac{2}{N(N-1)}\sum_{i>j}d_{ij}$, 其中 N 表示网络中的节点数目.

(3) 聚类系数 (C): 网络中节点 i 有 K_i 个边与之连接, 那么与该点可能连接的最大边数为$K_i(K_i-1)/2$, 若这 K_i 个节点之间真实边为 E_i, 则它与总的所有情况下的边比例, 计算得到节点 i 的聚类系数 C_i: $C_i=2E_i/K_i(K_i-1)$. 很显然, $0\le C\le 1$. C=0 代表网络中的点为孤立点; C=1 表示网络中的所有点之间都是互相连接的, 视为全局耦合网络.

(4) 介数 (B): 网络中介数的概念可以分为两类, 一类是点介数, 另一类是边介数. 节点 k 的介数定义为 $B_k=\sum_{i\ne j}C_k(i,j)/C(i,j)$, 其中, $C(i,j)$ 代表 i 与 j 间最短路径总数, $C_k(i,j)$表示中间点为 k 时, i 与 j 间的所有路径总数. 介数反映了节点 k 在 i 和 j 之间的流通量和重要程度. 网络中某个节点的介数越大, 说明该点在网络中信息传播的信息量就越大, 越容易在该点造成网络堵塞. 假设两组连接度很高的网络中间只有少数点连接, 那么这几个少数点介数就会很大, 即很多的信息在流通的过程中经过这几个点, 很容易造成堵塞, 从而造成数据信息丢失. 因此, 最大介数的增大会降低网络同步能力.

(5) 模块度 (Q): 模块度也称作模块化度量值, 是用来衡量网络强度的统计量. 最早是 Newman 提出的, 它用来描述网络社团以及划分的好坏. 假定网络共分为 k 个社团, $E=(e_{ij})$代表一个 $k\times k$ 维的矩阵. 故模块度可以定义为: $Q=\sum_i Q_i=\sum_i(e_{ij}-a_i^2)$, 其中, $a_i=\sum_i e_{ij}$是矩阵中的数值之和 (行或列), e_{ij} 用来表示社区 i 和社区 j 之间的边的数量. 模块度可以区分社区划分的好坏. 若是划分的好, 则社区内部节点相似度较大, 而在社区外边相似度较低. Q 越大, 越接近 1, 代表社区拥有一个很好的划分结构, 使得社区的划分合理化. 通常设定的值是在 0.3 与 0.7 之间.

本文中我们主要选择 5 个参数进行分析比较, 对于给定的参数进行最终的分析, 从而找到影响网络的重要因素, 依据此类统计量进行归纳分析, 得出相应的参数.

我们对由 SNP 数据设定不同的互信息阈值而形成网络, 针对其中大于阈值的边, 做去掉处理, 而针对小于阈值的边进行保留操作, 从网络图可以分析出统计量所对应的参数变化, 得到有益信息量.

根据网络中 SNP 之间互信息计算的值, 选择阈值范围为 0.1 到 0.63. 共设置 63 个阈值, 在每个阈值的条件下, 统计计算相应的网络结果, 从而得到一致性网络, 根据网络的相似性程度选择对实验组和对照组差别较大的统计量进行分析. 我们最终选择了度作为区分依据, 并分析能够区分实验组和对照组的取值范围, 得出最佳的阈值, 对于不同的数据, 得到的互信息值也不同, 所以需要根据数据得到的互信息范围, 由网络统计量得到取值范围, 得到互信息取值的交集, 能够区分对照组和实验组数据, 从而确定最佳的互信息阈值. 这样就能够保证所取的阈值不受样本数量的大小影响, 而是根据样本的不同情况得到相应的阈值. 对于节点 i, 我们定义, $\Delta d=d_i^d-d_i^c$, Δd 代表了这个节点的度差异值, 在该公式中, d_i^d和d_i^c分别代表了这个节点在实验组与对照组网络中节点的度.

我们都知道, 在复杂网络中, 节点度能够代表节点的作用和影响力. 本文从网络结构差异的角度去衡量各个统计量[12], 进而对应到其中的节点, 找到“结构性关键 SNPs”. 这种差异性贡献分为正、负贡献两个方面. 我们用 r 代表度的变化阈值. 正贡献 SNP 代表了该节点在病例组、对照组两个网络中度的贡献$\Delta d\ge r$ 的 SNP; 同理, 负贡献 SNP 代表了该节点在以上两个网络中度的贡献$\Delta d\le -r$ 的 SNP.

本文对基因 BRCA2 仿真数据建立病例组与对照组建立相互作用网络进行数据实验. 对 SNP 互信息设置一个阈值范围, 分析产生的病例组和对照组 SNPs 互

信息网络的统计量: 平均路径长度、聚类系数、平均度、模块度、平均介数随阈值在其变化范围内的增加而变化的情况. 根据计算的网络中 SNP 之间互信息的值, 我们取互信息阈值的范围为 0 至 0.63, 步长 0.01, 分析对应病例组与对照组的 SNP 相互作用网络的上述网络结构参数随变化而变化的情况.

3 数据来源与处理

HapMap 给出了人类基因组单核苷酸多态性 (SNPs) 和拷贝数多态性 (CNPs) 的分布情况. 本文使用 HapMap 提供的三个文件进行实验, 包含了关于 BRCA2 的 88 个 SNPs. 下面是对三个文件的说明.

.hap 文件是已知的单体型数据, 其中行代表 SNP, 列表示单体型. 每一个.hap 文件都需要一个相应的 legend 文件, 所有的等位基因都以 0, 1 作为标记.

.legend 文件是 SNP 标记位点数据, 四列数据分别表示 SNP 的 ID、碱基位置、碱基的 0, 1 表示.

.map 文件包含了小规模的重组率, 共三列分别表示每个 SNP 的物理位置, 距离左标记点的位置和距离右标记点的位置.

在这数据中, 必须去掉全部为 0 或者全部为 1 的数据, 因为这些数据对构建网络结构没有任何帮助. 去掉这些多余的数据, 共得到 45 条 SNP 数据. 把 3 个文件放到一起, 执行 Hapgen2 软件, 代码如下:

./hapgen2 -m BRCA2.map -l BRCA2.legend -h BRCA2.hap -o BRCA2.out -dl 31820136 1 2.5 2 31847382 0 1.5 4.5 -n 5000 5000.

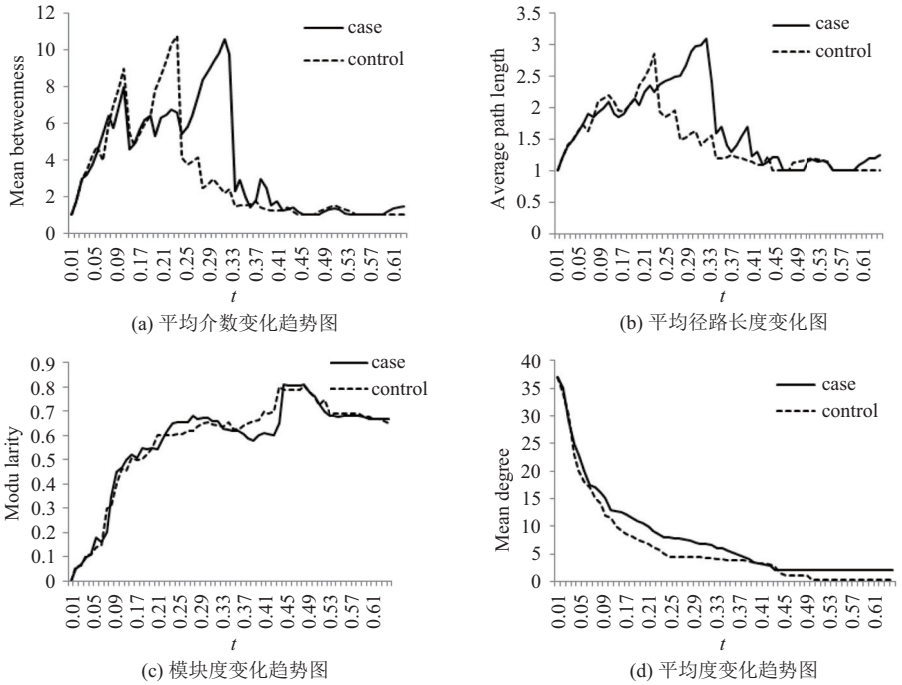
分别仿真了 5000 组实验组和对照组数据. 随机选定 2 个 SNPs 作为致病 SNPs. 它们的信息如下: rs206081 和 rs9534318, 选取杂合子变异率分别是 2.5 和 1.5, 纯合子变异率分别为 2 和 4.5, 上述样本数据都包含 SNP 编号, SNP 位置及 0, 1 表达数据.

本文中, 我们使用.gen 文件, 删除前五列后把数据转换成一个矩阵, 其中每行表示一个向量, 每三个数字代表一个个体, 我们转换成 0, 1, 2 表示.

4 试验方法和结果

4.1 网络统计量的比较

根据得到的互信息矩阵, 大于阈值的向量之间表示相互关系较强, 选定这些 SNP 作为节点建立网络. 分析比较网络的 6 个特性. 每个结构参数都反映着网络的特性, 进而可以显示 SNP 间的互信息的变化, 取 0.01 为步长, 从 0 到 0.63 之间求得每一个阈值下的网络结构特性值, 得到图 1. 图 1 中, 纵坐标表示相应的统计量, 横坐标代表阈值, 虚线表示对照组数据显示效果, 实线表示实验组数据显示效果.



(a) 平均介数变化趋势图 (b) 平均径路长度变化图

(c) 模块度变化趋势图 (d) 平均度变化趋势图

图 1 4 个网络结构的统计量随阈值的增加的变化情况

实验发现 5 个结构特性中, 平均聚类系数 B 交织在一起, 不能区分实验组和对照组.

观察图 1(a), 当 $0<t<0.21$ 时, 网络的平均介数 B 在在两组中的变化趋势走向大体相似. 当 $0.21<t<0.63$ 时, 网络的平均介数 B 逐渐减小. 从图中可以明显的看出, 病例组的平均介数要比对照组的平均介数高. 于是, 我们得到, 随着互信息阈值的增大, 节点的介数也在不断减小, 网络中边越来越稀疏.

观察图 1(b), 当 $0.2<t<0.43$ 时, 实验组与对照组的网络有相对明显的差异. 于是我们可以得到, 在这个变化区间内, 平均路径长度可以很好的区分病例组和对照组, 而当 $t>0.43$ 时, 网络的边越来越少, 平均路径长度趋近于 0.

从模块度 Q 随阈值的变化图 1(c) 看出, 当阈值 $0<t<0.2$ 或 $0.43<t<0.63$ 时, 两组中的模块度 Q 逐步上升, 但变化大致相同, 而当 $0.2<t<0.43$ 时, 实验组模块度与对照组有较大区别.

观察图 1(d), 可以发现, 在很长的一段阈值范围内, 病例组与对照组的网络平均度有很大的区别, 而随着网络的阈值增加, 网络的平均度越来越小, 这与网络的孤立点越来越多也是相对应的.

当 $t>0.62$ 时, 病例、对照组中都只有一个包含四个节点的全耦合子网, 聚类系数 C、平均路径长度 L 两者相等, 且都为 1. 当 $t>0.63$ 时, 平均路径长度 L、聚类系数 C 是缺失的, 平均介数 B 以及其他三个统计量值均为 0.

总之, 平均聚类系数 C 不能区分两组数据, 平均路径长度 L 和平均介数 B 能够区分但是阈值具有一定局限性. 平均度可以在很大的范围内把实验组和对照组分别出来, 我们选择平均度作为区分的依据.

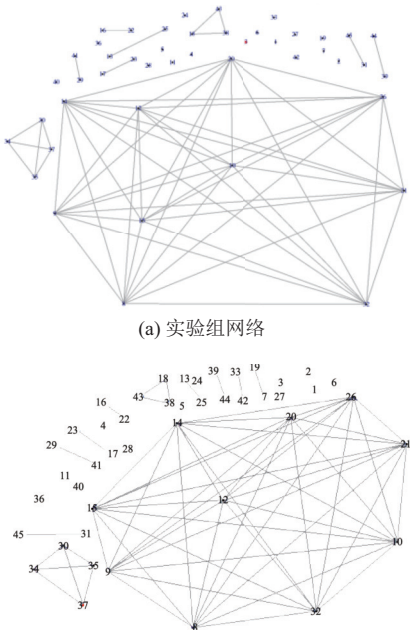
从图 1 中我们得到每个统计量能够区分两组的阈值范围, 如表 1.

表 1 各统计量能够区分实验组和对照组的阈值范围

网络统计量	阈值范围
平均度(K)	0.08–0.35
模块度(Q)	0.18–0.33
平均路径长度(L)	0.19–0.31
平均介数(B)	0.24–0.33

从表 1 可以看出, 每一个统计量都有不同的阈值范围, 平均度 K 的范围较大, $0.08<K<0.35$; 其他的统计量阈值范围相差不大, 基本在 0.2 到 0.3 之间. 结合图 1, 选择 0.28 为阈值构建网络.

依据图 2, 实验组和对照组的图像是有很大差异的. 在对照组, 节点之间联系较弱且存在更多的孤立点. 但是在实验组中, 很多的孤立点不再是独立的, 并且拥有了更多的联系. 对照组中存在 36 个连接点和 9 个孤立点, 而实验组中存在 39 个连接点喝 6 个孤立点. 这表明我们选取的阈值 0.28 是合适的. 经过多次仿真数据试验, 对于结合数据互信息得到阈值范围, 而后确定互信息阈值的方法都是有效的.



(a) 实验组网络

(b) 对照组网络

图 2 阈值为 0.28 的条件下, 实验组和对照组互信息网络

4.2 获得“结构性关键 SNP”

结构决定功能, 而结构的差异决定了功能的差异, 本文将这种差异细化到每个节点上, 而平均度可以很好的区分病例组和病例组, 所以我们选择每个 SNP 位点的平均度来刻画 SNP 在病例组和对照组的差异, 计算每个网络的每个节点的节点度差异, 当节点的度在病例、对照组中的变化差异比较大时, 说明这两个组的网络结构差异较大. 从两组网络的数据分析来说, 节点度的增量有正有负, 所以, 节点在病例组中的度也有增减之分, 即存在正、负贡献 SNPs. 度变化量增加最大的是节点 39, 增加值的大小是 5, 同理, 减少量最大的是 16, 41, 减少值的大小是 2.

当阈值为 0.28 时, 对照组网络中的平均度大致等于 2, 从而可以得到, 当病例、对照组网络中节点度的变化值大于等于 3 时, 其对网络结构影响较大. 故可设 $\Delta d=3$, 由此, 我们可以获得对网络结构有显著影响 4 个 SNPs, 如表 2, 其中 rs206081, rs9534318 为预设致病 SNPs.

表 2 给定参数为 3 的条件下, 部分结构性关键 SNPs 的信息及度的变化量

SNPs	Δd
rs206081	3
rs4942448	4
rs9534262	3
rs9534318	5

4.3 参数评估

在查找“结构性关键 SNPs”时, 我们需要从网络平均度出发, 对选取网络中的关键 SNPs 设置合适的差值参数. 如果选取的差值参数比较小, 对 SNPs 选取限制比较宽泛, 一些不相关的 SNPs 也会选取到 SNPs 集合内, 从而导致假阳性. 反之, 如果选取过于严苛, 反而会遗漏一些比较重要的节点, 导致假阴性.

我们选取基因 BRCA2, 得到它在阈值为 0.28 时候的网络, 如图 2 所示. 选择不同的差值参数, 得到一系列不同的结构性关键 SNPs, 如表 3 所示.

表 3 不同参数 r 的取值下关键 SNPs 个数

r	SNPs(个)
1	18
2	11
3	4
4	2
5	1

当互信息阈值设定为 0.28 时, 网络中度的最大变化量是 5. 当 $r\geq 5$ 时, 所得的关键 SNPs 只有节点 39, 对网络影响较大的节点 25 却被忽略. 当 $r\leq 2$ 时, 所得的关键 SNPs 只有 13 个, 这里面也包括了其中的非零点.

5 结论与分析

本文通过国际项目 HapMap3 中以及 Hapgen2 软件生成的 13 号染色体上 BRCA2 基因生成仿真数据. 利用互信息表示 SNPs 间的相互作用. 构建实验组和对照组的网络, 根据阈值及差值参数筛选出关键 SNPs. 最后, 对我们所选择的参数进行了评估, 证明我们所选定的参数能够反映结构的变化, 能够较好地选择出预设的关键 SNPs. 通过数值实验发现: 样本数目会影响互信息的大小, 样本数较小时, 互信息较高, 样本数较大时, 互信息逐渐降低, 本文认为, 样本数偏少, 则特异性个体数目不完备, 样本数过多, 又会造成冗余, 增加了计算复杂度. 目前, 确定合适的上下界仍然是一个具有挑战的问题.

参考文献

1 Pharoah PDP, Tsai YY, Ramus SJ, *et al*. GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. Nature Genetics, 2013, 45(4): 362–370e2. [doi: 10.1038/ng.2564]

2 Xu ZL, Taylor JA. SNPinfo: Integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. Nucleic Acids Research, 2009, 37(S2): W600–W605.

3 Larsson M, Duffy DL, Zhu G, *et al*. GWAS findings for human iris patterns: Associations with variants in genes that influence normal neuronal pattern development. The American Journal of Human Genetics, 2011, 89(2): 334–343. [doi: 10.1016/j.ajhg.2011.07.011]

4 Jia PL, Zheng SY, Long JR. dmGWAS: Dense module searching for genome-wide association studies in protein-protein interaction networks. Bioinformatics, 2011, 27(1): 95–102. [doi: 10.1093/bioinformatics/btq615]

5 Collins FS, Morgan M, Patrinos A. The human genome project: Lessons from large-scale biology. Science, 2003, 300(5617): 286–290. [doi: 10.1126/science.1084564]

6 Yong Y, He L. SHEsis, a powerful software platform for analyses of linkage disequilibrium, haplotype construction, and genetic association at polymorphism loci. Cell Research, 2005, 15(2): 97–98. [doi: 10.1038/sj.cr.7290272]

7 Ghoussaini M, Fletcher O, Michailidou K. Genome-wide association analysis identifies three new breast cancer susceptibility loci. Nature Genetics, 2012, 44(3): 312–318. [doi: 10.1038/ng.1049]

8 Winham SJ, Cuellar-Barboza AB, Oliveros A. Genome-wide association study of bipolar disorder accounting for effect of body mass index identifies a new risk allele in TCF7L2. Molecular Psychiatry, 2014, 19(9): 1010–1016. [doi: 10. 1038/mp.2013.159]

9 Hirokawa M, Morita H, Tajima T. A genome-wide association study identifies PLCL2 and AP3D1-DOT1L-SF3A2 as new susceptibility loci for myocardial infarction in Japanese. European Journal of Human Genetics, 2015, 23(3): 374–380. [doi: 10.1038/ejhg.2014.110]

10 Goh KI, Cusick ME, Valle D. The human disease network. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(21): 8685–8690. [doi: 10.1073/pnas.0701361104]

11 汪小帆, 李翔, 陈关荣. 复杂网络理论及其应用. 北京: 清华大学出版社, 2006: 35–38.

12 贾华仟. 复杂网络分析方法在全基因组关联研究中的应用 [硕士学位论文]. 青岛: 山东科技大学, 2015.