基于贝叶斯网络的基因变异间的因果关系发现与验证

蔡瑞初[1] 甄启祺[1] 陈 薇[1] 郝志峰[1,2]

[1](广东工业大学计算机学院 广东 广州 510006)

[2](佛山科学技术学院数学与大数据学院 广东 佛山 528000)

摘 要 基因变异间的相关性是全基因组关联分析等领域中的难点。当前基因变异间关系的研究主要基于基因在染色体上的相对位置展开,从另一个角度研究发现位于同一信号通路中的基因变异之间具有较强的因果性。基于单核苷酸多态(Single Nucleotide Polymorphisms,SNPs)数据对基因状态进行编码,得到离散型的基因变异状态数据;基于大量的基因变异数据构建基因间的因果贝叶斯网模型;用真实信号通路对基因变异数据上发现的因果贝叶斯网络进行验证。在 WTCCC(Wellcome Trust Case Control Consortium)数据集上的实验结果表明,相互调控的基因之间的变异具有较强的因果关系。同时,实验也发现了一批具有较强因果关系的基因变异,可能对相关研究具有一定的启发意义。

关键词 因果关系 基因变异 贝叶斯网 单核苷酸多态性 信号通路

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2020.07.004

DISCOVERY AND VALIDATION OF CAUSALITIES AMONG GENE MUTATIONS BASED ON BAYESIAN NETWORK

Cai Ruichu[1] Zhen Qiqi[1] Chen Wei[1] Hao Zhifeng[1,2]

[1](*School of Computer Science, Guangdong University of Technology, Guangzhou 510006, Guangdong, China*)

[2](*School of Mathematics and Big Data, Foshan University, Foshan 528000, Guangdong, China*)

Abstract The correlation among the gene mutations is still a challenge problem in Genome-wide association analysis (GWAS). In the existing researches, the scientists mainly focus on the relative locations of genes on chromosomes. This article studies from another angle that there is a strong causalities among the gene mutations in the same signal pathway. We coded the gene status based on Single Nucleotide Polymorphisms(SNPs) data to obtain discrete gene mutation status data; a causal Bayesian network model of genes was constructed based on a large amount of gene mutation data; the causal Bayesian network found in the gene mutation data was verified by the real signal pathway. The experimental results on WTCCC(Wellcome Trust Case Control Consortium) show that there is a significant causalities among the mutations of mutually regulated genes. It also finds a group of gene mutations with strong causalities, which may have certain enlightening significance to the related research.

Keywords Causality Gene mutation Bayesian network Single nucleotide polymorphisms Signal pathway

0 引 言

全基因组关联研究(Genome-Wide Association Study,GWAS)当前主要侧重于单核苷酸多态性(Single Nucleotide Polymorphisms,SNPs)与人类疾病的关联研究[1]。研究表明,人类常见疾病的复杂性状涉及多个基因变异之间的相互影响,而基因变异间存在连锁不平衡(Linkage Disequilibrium,LD)现象[2]。由于基因按一定次序在染色体上呈线性排列,关于基因变异间的相

收稿日期:2019-06-04。国家自然科学基金项目(61876043)。**蔡瑞初**,教授,主研领域:因果关系,机器学习。**甄启祺**,硕士。**陈薇**,博士。**郝志峰**,教授。

互作用大多是从基因的相对物理位置来探讨,然而在基因测序数据中,一些位于不同染色体上的基因,其测序数据都反映出了非偶然性的连锁不平衡现象[3],单纯研究相邻基因之间变异的相互作用会很受局限,故多基因变异相互影响是 GWAS 中不可回避的关键问题。

在多基因变异间的相互作用研究中,Yang 等[4]提出了多变量全基因组关联测试的方法,对多变量定量表现型数据进行分析。这是一种线性混合模型,实验表明,该方法较好地把第一类错误控制在合理范围内,并且比不同族群结构和相关性的边缘检验更有效。然而该方法的适用条件为变量之间满足线性关系,该假设较为苛刻。

事实上,由于多个基因变异之间的相互作用效果并不是简单地线性加权,传统的线性模型远不能满足这种多基因模型[5]。Monneret 等[6]着重于边缘因果估计方法,提出了基于高斯结构方程的高斯因果模型,然而该模型是一种线性模型,与真实情况中基因相互影响的非线性并不符合。

立足于非线性角度,Botta 等[7]对随机森林进行扩展,得到 T 树模型。通过实验与线性模型相比,T 树的预测能力表现出明显优势,并表明了多个 SNPs 之间的非线性效应。在进一步的基因座识别研究中,T 树模型不仅复现了已发现的多数基因座,并且发现了与克罗恩病相关的两个新的符合生物学的易感位点。而在更复杂的非线性关系如网络关系的研究中,T 树的效果有待验证。Lim 等[8]提出了 BTR(BoolTraineR)算法,对布尔模型进行了优化,使用单细胞数据重构了基因调控网络,然而没有进一步考虑基因调控和基因变异之间的关系。

在基因变异和基因调控方面,现有的研究主要关注具体基因与疾病之间的关系。在基因变异上,Pan 等[9]阐明了基因 TNNT2 其新变体 R205Q 的因果作用,Elsaadany 等[10]通过全外显子组测序研究了 WWOX 基因中的新纯合突变,并提出了 WWOX 基因变异与人类癫痫性脑疾病之间的关系。基因调控方面,Wang 等[11]研究了 N6-甲基腺苷对脂肪的生成代谢起到了关键的调控作用,Jin 等[12]研究了 microRNA-192 对膀胱癌细胞的调控,表明 microRNA-192 可能可以通过调控细胞周期,从而抑制膀胱癌细胞。Liu 等[13]在 Sia 等[14]的基础上构建了 CNV-ICC-TRN,结合 KEGG 信号通路进行分析,发现基因变异的影响会发生在通路组件上。Liu 等认为,整合基因变异和调控网络二者的信息可更深层次地研究肝癌致病机理,然而目前基因变异和基因调控这两者关系的相关成果相对较少,是当前亟需进行的研究。

鉴于上述研究状况,针对部分不足,本文对基因变异间的因果关系进行研究,发现在同一信号通路中相互调控的基因,其变异间具有较强的因果性。本文通过改进的 Parents-Children(Improved Parent-Children,IPC)算法,检测出具有直接相互作用的基因对,并通过结构识别的方法判断出因果方向。实验结果表明,相互调控的基因变异之间具有较强因果性。本文还发现了一批具有较强因果性的基因变异可供相关研究参考。

1　相关知识

1.1　SNP 与基因

单核苷酸多态性(Single-Nucleotide Polymorphism,SNP)指在基因组中某特定位点上的单核苷酸所发生的变异,该变异在人类群体中占有一定的规模比例[15]。在多代遗传中,SNPs 一般都会稳定地遗传下去,故在 GWAS 中,一些跟人类疾病相关联基因的变异,可以用 SNPs 作为稳定标记。

基因是一段 DNA 序列,由多个脱氧核糖核苷酸组成,是生物体的遗传分子。人体的性状由基因和生活环境控制调节,其中基因起着决定性作用。不同基因的 DNA 序列长度并不相同,而 SNP 是对单核苷酸状态的描述,由此容易得知,基因与 SNP 之间是属于一对多的关系。

1.2　信号通路

在多细胞生物体内,细胞外的细胞因子、激素等分子信号特异性地与细胞膜结合或直接穿过细胞膜,经过级联转导后多种信号在细胞内相互识别作用,共同调控内部的生化反应。细胞内多种信号的作用途径交织成各种错综复杂的调控网络,进行生化功能的调控,这样的网络称为信号通路(Signal Pathway)[16]。京都基因与基因组百科全书通路(Kyoto Encyclopedia of Genes and Genomes Pathway,KEGG Pathway)是一个手工绘制出这些信号转导网络图的集合。通过 KEGG Pathway,本文可以把处于同一个信号通路的基因聚集起来,验证处于同一信号通路中,相互调控的基因变异间因果关系的存在性。

1.3　贝叶斯网

贝叶斯网是一类概率图模型。该模型借助一个有向无环图,表示一系列的随机变量 $X_1, X_2, \cdots, X_n$ 及其

相互之间的依赖关系,其中每个节点代表一个随机变量,并对应一个概率分布表。图模型直观地反映了各个随机变量之间的依赖关系,而联合概率分布表则具体量化了随机变量之间的概率依赖强度。贝叶斯网如图 1 所示,全部随机变量的联合概率分布可通过这些随机变量的边缘概率分布和条件概率分布相乘而得到,记为:

$$P(X_1,X_2,\cdots,X_n) = \prod_{i=1}^{10} P(X_i \mid Pa(X_i)) \quad (1)$$

式中:$Pa(X_i)$为 X_i的父节点集。

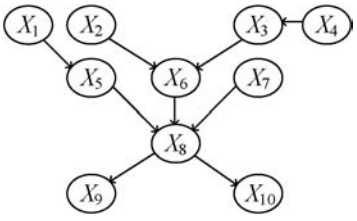


图 1 贝叶斯网模型

2 数据处理

2.1 数据描述

实验所用的 SNPs 数据由英国 WTCCC(Wellcome Trust Case Control Consortium)机构提供,本文所使用的数据是多种复杂疾病整合到一起的数据文件,共有 16 179 个样本,每个样本含有 394 747 个 SNPs 的变异情况。表 1 选取了其中两个 SNPs 的染色体位点和主次等位基因等基本属性的信息,部分 SNPs 变异情况如表 2 所示。

表 1 每个 SNP 的染色体位点及其基本属性

染色体号	SNPs ID	碱基坐标	次等位基因	主等位基因
1	rs3094315	752566	G	A
1	rs4040617	779322	G	A

表 2 部分 SNPs 变异数据

样本	rs2980300_T	rs4040617_G	rs3094315_G
1	0	0	0
2	1	0	0
3	2	2	2

表 2 列出了数据集中 3 个样本的 3 个 SNPs 情况,每一列为一个 SNP 的变异情况,如表 2 中的 rs2980300_T,这是一个命名为 rs2980300 的 SNPs,其后缀"_T"代表该 SNP 位点突变为次等位基因 T;0、1、2 三个取值分别代表样本 1 在该 SNP 位点上没有发生突变,样本 2 有 1 条染色体上的该 SNP 位点突变为此等位基因 T,样本 3 则 2 条染色体全都发生了突变。

2.2 SNPs 数据编码为基因数据

通过 NCBI 可以查询到 rs 命名的 SNPs 是否位于基因区域、属于哪一个基因。根据基因和 SNPs 为一对多的关系,SNPs 数据可以编码成基因粒度,编码方式如表 3 所示,394 747 个 SNPs 中有 3 个位于基因 51150 区域。

表 3 基因 51150 的编码情况表

基因 ID	51150			编码后的基因状态
SNPs ID	rs2887286	rs6603781	rs11260562	
SNPs 的状态	0	1	2	012
	1	1	0	110
	0	1	1	011

3 基于 IPC 算法的基因变异因果关系发现

3.1 IPC 算法框架

基于所得的离散型基因变异数据构建贝叶斯网模型,如果检测到一个基因变量的父子(Parents and Children,PC)节点,则可以推断出调控网络中直接相互作用的基因。通过条件独立性检验构造局部因果结构,可以检测出贝叶斯网中基因变量的 PC 节点[17-19]。受贝叶斯半监督方法(Bayesian Semi-Supervised method,BASSUM)[17]中采用 PC 算法检测变量间因果关系思想的启发,本节提出一种改进的 PC 算法,即 IPC 算法(Improved-PC),检测同一信号通路中基因变异间相互作用的因果关系。

记 $PC(T)$为目标基因 T 的 PC 节点集。由贝叶斯网性质可知,$PC(T)$是导致基因 T 发生变异的直接原因或由基因 T 变异导致的直接结果的集合,即 $PC(T)$中所有基因变量跟 T 都是不独立的。记 v_i为跟基因 T 直接相互作用的基因,由此可得 $v_i \in PC(T)$的一个必要条件:v_i跟 T 不满足相互独立。根据此条件,在检测目标基因 T 的 PC 节点时,遍历 $PC(T)$的候选基因集合 $C(T)$,如果属于 $C(T)$的基因 c_i与 T 不相互独立,则更新 $PC(T)$,使得 $PC(T) = PC(T) \cup c_i$,这里 $c_i \in C(T)$。由于上述条件是 $v_i \in PC(T)$的必要条件,据此检测所得的结果很可能会引入一些非直接相互作用的基因,得到的是 $PC(T)$的一个超集,故检测结束后需剔除掉非直接相互作用的基因。

先对超集中的非直接相互作用的基因分析。图 2 所示的 $PC(T)$超集包含了非直接相互作用的基因 f_i,基因 f_i发生变异是目标基因 T 发生变异的间接原因,

检测时如果基因 f_i 被误加入 $PC(T)$ 中,那么从贝叶斯网络结构上看,节点 f_i 就不可能跟节点 T 直接连接。同时,由于根据必要条件检测所得的集合是真实 $PC(T)$ 的超集,故 f_i 跟 T 必然被 $PC(T)$ 所 d-分隔,从而在给定 $PC(T)$ 的条件下,f_i 跟 T 条件独立,其中 $PC(T)$ 可以通过遍历超集的幂集搜索获得。此分析过程对目标基因 T 变异的间接结果同样适用。

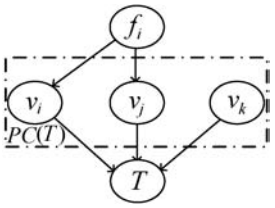


图 2　根据必要条件检测到的 $PC(T)$ 超集

类似地,如果基因 v_i 属于真实 $PC(T)$,则基因 v_i 跟目标基因 T 直接相互作用,故与 f_i 情况相反,基因 v_i 和目标基因 T 不会被任何集合所 d-分隔,即不存在条件集合使 v_i 跟 T 满足条件独立。根据 v_i 和 f_i 在条件独立性上的差异,可以把超集中的非直接相互作用的基因 f_i 和属于真实的 $PC(T)$ 基因 v_i 分辨出来。

根据以上分析,可以得到 IPC 算法步骤如算法 1 所示。

算法 1　IPC 算法

输入:包含所有变量的集合 V、目标变量 T

输出:$PC(T)$

1. 初始化:$PC(T)=\varnothing$,$C(T)=V-\{T\}$;
2. **FOR** $v_i \in C(T)-PC(T)$ **DO**;　　//遍历;
3. $g(v_i)=G^2(v_i,T)$;　　//构造基因 v_i 和 //目标基因 T 的 //G^2 统计量
4. **IF** $g(v_i)>\chi^2_{\text{p-threshold}}(df)$ **THEN** //如果 G^2 统计量落在拒绝域中, //说明基因 v_i 和目标基因 T //不满足相互独立
5. $PC(T)=PC(T)\cup v_i$;
6. **END IF**
7. **END FOR**　　//通过条件独立性检验删除冗余信息
8. **FOR** $f_i \in PC(T)$ **DO**
9. **IF** $\exists S \subset PC(T)-\{f_i\}, f_i \perp T \mid S$ **THEN**
10. $PC(T)=PC(T)-\{f_i\}$; //遍历 $PC(T)$ 减去 f_i 后的幂集,以此 //作为条件进行条件独立性检验
11. **END IF**
12. **END FOR**

IPC 算法检测到基因的 PC 节点后,需要通过因果结构的识别来判定 PC 基因之间的因果方向。由于直接确定两个相互连接的 PC 节点间的方向比较困难,因此,本文通过对三个节点的基本有向无环结构进行识别,进而判断其中的父节点(Parent)和子节点(Child)。三个节点构成的基本结构包括图 3 的 4 种情况。

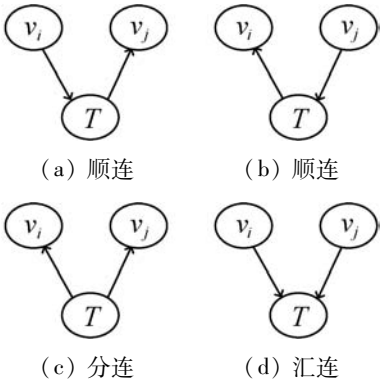


图 3　三个节点构成的基本结构

这 4 种情况可以从独立性和条件独立性上的差异来实现结构的识别。情况(a)-(c)中,节点 v_i 和节点 v_j 并不满足相互独立,而在给定节点 T 这一条件下,节点 v_i 和节点 v_j 满足条件独立。这三种结构在独立性和条件独立性的性质上等价,无法进行有效区分,但情况(d)会呈现出相反的性质。

情况(d)属于汇连结构。该结构跟前面三种结构的区别在于,在节点 T 未知的情况下,节点 v_i 和节点 v_j 满足相互独立,但在给定节点 T 的条件下,节点 v_i 和节点 v_j 不满足条件独立。这跟前面三种结构的性质相反。因此,利用汇连结构在条件独立性上的特性,本文可以通过算法 2 识别出三节点构成的汇连结构,推断其中的父亲节点和孩子节点。

记从节点 v_i 出发,到节点 T 终止的有向边为 $<v_i,T>$,具体的识别过程如算法 2 所示。

算法 2　汇连结构识别算法

输入:目标变量 T、T 的 PC 节点集 $PC(T)$

输出:包含目标变量 T 的有向边集 E

1. 初始化:$E=\varnothing$;
2. **FOR** $v_i, v_j \in PC(T)$ **DO**
3. **IF** $v_i \perp v_j$ **AND NOT** $v_i \perp v_j \mid T$ **THEN**
4. $E=E\cup\{<v_i,T>,<v_j,T>\}$
5. **END IF**
6. **END FOR**

不同于一般的相关性分析,在目标变量 T 和 PC 节点集 $PC(T)$ 中,算法 2 基于条件独立性检验,先有效识别出变量中的汇连结构,进而推断出该结构中的因果方向,保证了检测得到的因果关系的可靠性。

3.2　独立性检验

本文所采用的独立性检验基于卡方检验实现。对一些较为复杂而庞大的数据进行独立性检验时,平方差的计算会增加卡方检验的运算量,所以可以构造 G^2

检验统计量[20]来作为卡方统计量的代替。

可借助列联表来考察两种属性的独立性。如表4所示，O_{mn}表示样本中属性1取值为A_m且属性2取值为B_n的观察计数。其中M为属性1的取值数目，N为属性2的取值数目。

表4 $M\times N$列联表

属性1	属性2			
	B_1	B_2	…	B_N
A_1	O_{11}	O_{12}	…	O_{1N}
A_2	O_{21}	O_{22}	…	O_{2N}
⋮	⋮	⋮		⋮
A_m	O_{M1}	O_{M2}	…	O_{MN}

如果零假设“属性1和属性2相互独立”成立，则两者的联合分布率等于边缘分布率之积。故联合分布率的极大似然估计为：

$$\hat{p}_{mn}=\hat{p}_{m\cdot}\times\hat{p}_{\cdot n} \tag{2}$$

式中：$\hat{p}_{m\cdot}=\frac{O_{m\cdot}}{S}$是属性1取值为$m$的概率的极大似然估计；$\hat{p}_{\cdot n}=\frac{O_{\cdot n}}{S}$是属性2取值为$n$的概率的极大似然估计；$O_{m\cdot}$和$O_{\cdot n}$分别是样本数据中属性1取值为$m$和属性2取值为$n$的样本计数；$S$是本次独立性检验使用的总样本数。

由零假设和式(1)可知，样本数据中属性1取值为m且属性2取值为n的期望样本量为：

$$E_{mn}=\hat{p}_{mn}\times S=\frac{O_{m\cdot}\times O_{\cdot n}}{S} \tag{3}$$

而在备择假设中，两属性的联合分布率为：

$$p_{mn}=\frac{O_{mn}}{S} \tag{4}$$

类似地，在进行条件独立性检验时，如果在给定属性3取值为l的条件下，检验属性1和属性2是否相互独立时，对数据中满足给定属性3的取值l的样本构造属性1和属性2的列联表，由以上推导可得期望样本量为：

$$E^l_{mn}=\frac{O^l_{m\cdot}\times O^l_{\cdot n}}{S^l_{\cdot\cdot}} \tag{5}$$

式中：各项的右上标l代表给定条件属性的取值为l。

在$M\times N$列联表中，任意单元格中的观察样本数都对应着唯一一个期望样本数，为了便于记录，这里把$M\times N$个单元格的观察样本数记为$O_1,O_2,\cdots,O_{M\times N}$，期望样本数则对应为$E_1,E_2,\cdots,E_{M\times N}$。于是可以得到$G^2$检验统计量作为卡方检验的近似估计，表示为：

$$G^2=2\sum_i^{M\times N}O_i\ln\frac{O_i}{E_i} \tag{6}$$

另一方面是自由度的计算。对每个列联表的G^2检验统计量其所对应的自由度df(degrees of freedom)为：

$$df=(M-1)\times(N-1) \tag{7}$$

在给定条件属性时，自由度为：

$$df=(M-1)\times(N-1)\times S(Con) \tag{8}$$

式中：$S(Con)$表示条件属性值的数目。

4 实 验

4.1 实验方案

要验证同一信号通路中相互调控的基因，其变异间具有因果关系这一假设，需要通过IPC算法检测出具有因果关系的基因对，并将该实验结果分别跟真实信号通路中相互调控的基因对进行匹配。本次实验的卡方独立性检验显著性水平设为0.01。

4.2 实验结果分析

本文使用了KEGG Pathway的真实信号通路数据来评估实验结果。为了评估实验结果的匹配程度和IPC算法的性能，本文使用召回率R(Recall)、准确率P(Precision)和$F1$值(F1-sorce)来评价实验效果。三者的定义如下：

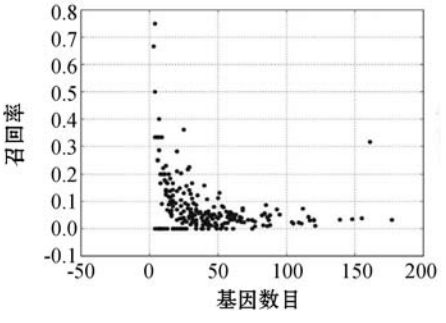
$$R=\frac{TP}{TP+FN} \tag{9}$$

$$P=\frac{TP}{TP+FP} \tag{10}$$

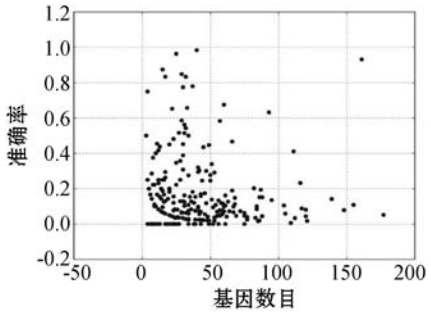
$$F1=2\times\frac{P\times R}{P+R} \tag{11}$$

式中：TP代表IPC算法检测出来的基因变异对和真实信号通路网络结构中相互调控的基因对匹配的数目；FN是真实信号通路网络结构中含有而IPC算法没有检测出的基因变异对的数目；FP为IPC算法检测出而真实信号通路网络结构中没有出现的基因变异对的数目。

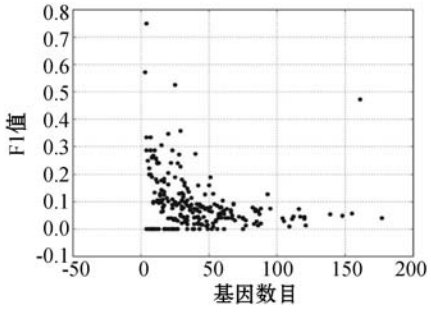
根据上述指标，最后得到的实验结果与真实信号通路的对比效果如图4所示。



(a) 召回率



(b) 准确率



(c) F1 值

图 4　实验检测结果跟真实信号通路网络结构的匹配评分

可以看出,算法的部分检测结果与真实信号通路中相互调控的基因对相匹配,准确率比较理想,其中某个包含 161 个基因的信号通路中,其准确率达到 90% 以上。该实验结果表明,部分相互调控的基因之间的变异具有一定的因果关系。

为了进一步验证这种匹配的非偶然性和 IPC 算法的有效性,本文进一步进行了随机对照实验作为对比实验。实验保持真实信号通路的拓扑结构不变,将同一信号通路中的所有节点所对应的基因标签随机混淆,信号通路中原有的调控信息破坏掉,并以随机信息替换,得到了随机对照的网络结构。结果发现其中并没有基因对可以与 IPC 算法结果相匹配,部分实验结果如表 5 所示。这一结果验证了这种匹配的非偶然性,基因变异间的因果关系跟信号通路中的基因调控关系存在某种关联。

表 5　本文 IPC 算法实验、LD 算法对比实验和随机对照实验的指标对比

信号通路 ID	准确率		召回率		F1 值	
	本文算法	随机对照实验	本文算法	随机对照实验	本文算法	随机对照实验
hsa00830	0.983 6	0	0.158 3	0	0.272 7	0
hsa00190	0.961 5	0	0.347 2	0	0.510 2	0
hsa04740	0.932 8	0	0.309 5	0	0.464 8	0
hsa00051	0.875 0	0	0.184 2	0	0.304 3	0
hsa05217	0.848 5	0	0.225 8	0	0.356 7	0
hsa00565	0.840 9	0	0.100 0	0	0.178 7	0

续表 5

信号通路 ID	准确率		召回率		F1 值	
	本文算法	随机对照实验	本文算法	随机对照实验	本文算法	随机对照实验
hsa00670	0.833 3	0	0.103 4	0	0.184 0	0
hsa00140	0.775 5	0	0.059 8	0	0.111 1	0
hsa00600	0.774 1	0	0.072 1	0	0.131 9	0
hsa04978	0.750 0	0	0.750 0	0	0.750 0	0
hsa00564	0.683 5	0	0.058 5	0	0.107 8	0
hsa05032	0.676 5	0	0.050 2	0	0.093 5	0
hsa00910	0.666 7	0	0.666 7	0	0.666 7	0
hsa00410	0.636 4	0	0.138 6	0	0.227 6	0

在图 4 中一个具有 161 个基因的信号通路,其召回率、准确率和 F1 值都突然明显升高,其中原因可能跟信号通路的类型有关。进一步分析,基因数目超过 10 个的信号通路共有 181 个,总体情况如表 6 所示。其中,细胞过程和环境信息处理这两类信号通路的平均基因数较多,其平均准确率也偏低;遗传信息处理类型的信号通路只有 4 个,代表性较弱;人类疾病和生物体系统这两类的平均基因数都在 48 左右,然而平均准确率分别是 0.11 和 0.18,相差较大,可能还有其他因素本实验中没有考虑到;实验效果比较理想的是新陈代谢类型的信号通路,相较于人类疾病和生物体系统类型,虽然其平均基因数较少,但平均准确率增幅很大,达到 0.38,且样本量有 40,具有一定的代表性。这也说明了对于某些信号通路,通过本文算法可以有效地检测出其基因变异间所蕴含的因果关系。

表 6　基因数超过 10 个的信号通路的情况

信号通路类型	信号通路数目	信号通路的平均基因数	平均准确率
细胞过程	11	64.55	0.13
环境信息处理	27	73.93	0.18
遗传信息处理	4	23.25	0.16
人类疾病	47	48.43	0.11
新陈代谢	40	30.23	0.38
生物体系统	52	47.33	0.18

表 7 给出了部分跟真实信号通路网络结构相匹配的 IPC 算法检测结果,调控类型列中 ECrel 代表酶与酶的催化逐次反应,PPrel 代表该基因对的相互作用表现为蛋白质与蛋白质的相互作用。

表 7 检测基因变异的因果关系所验证到的基因的调控情况

信号通路 ID	基因 1 ID	基因 1 所属染色体	基因 2 ID	基因 2 所属染色体	基因 1 和基因 2 的调控类型
hsa00140	1588	15	54659	2	ECrel
hsa00670	1719	5	2618	21	ECrel
hsa00562	5333	3	51763	17	ECrel
hsa03460	2177	3	22909	15	PPrel
hsa00565	30814	1	56994	12	ECrel
hsa03460	2177	3	675	13	PPrel
hsa04973	207	14	6518	1	PPrel
hsa00512	55808	17	63917	7	ECrel

5 结 语

本文研究了相互调控的基因之间的变异的因果关系。把 WTCCC 研究机构的 SNPs 数据编码为基因粒度的数据,并使用因果关系发现 IPC 算法检测出了数据中的因果关系。将检测结果分别跟 KEGG Pathway 的真实信号通路网络结构数据和随机对照实验网络结构对比,发现因果结构在部分真实信号通路网络中获得了验证,而跟基于连锁不平衡的对比实验则体现了 IPC 算法具备更良好的性能。

参 考 文 献

[1] Genome-wide association study[OL]. https://en.wikipedia.org/wiki/Genome-wide_association_study.

[2] Linkage disequilibrium[OL]. https://en.wikipedia.org/wiki/Linkage_disequilibrium.

[3] Rodley C D M, Bertels F, Jones B, et al. Global identification of yeast chromosome interactions using genome conformation capture[J]. Fungal Genetics & Biology, 2009, 46(11):879 - 886.

[4] Yang J J, Williams L K, Buu A, et al. Identifying pleiotropic genes in genome-wide association studies from related subjects using the linear mixed model and fisher combination function[J]. BMC Bioinformatics, 2017, 18(1): 376.

[5] Moore J H, Williams S M. New strategies for identifying gene-gene interactions in hypertension[J]. Annals of Medicine, 2002, 34(2):88 - 95.

[6] Monneret G, Jaffrézic F, Rau A, et al. Identification of marginal causal relationships in gene networks from observational and interventional expression data[J]. Plos One, 2017,12(3):e0171142.

[7] Botta V, Louppe G, Geurts P, et al. Exploiting SNP correlations within random forest for genome-wide association studies[J]. Plos One, 2014, 9(4):e93379.

[8] Lim C Y, Wang H, Woodhouse S, et al. BTR: training asynchronous boolean models using single-cell expression data[J]. BMC Bioinformatics, 2016, 17(1): 355.

[9] Pan S, Sommese R F, Sallam K I, et al. Establishing disease causality for a novel gene variant in familial dilated cardiomyopathy using a functional in-vitro assay of regulated thin filaments and human cardiac myosin[J]. Bmc Medical Genetics, 2015, 16(1):97.

[10] Elsaadany L, Elsaid M, Ali R, et al. W44X mutation in the WWOX gene causes intractable seizures and developmental delay: a case report[J]. Bmc Medical Genetics, 2016, 17(1):53.

[11] Wang X X, Zhu L N, Chen J Q, et al. MRNA m6A methylation downregulates adipogenesis in porcine adipocytes[J]. Biochemical and Biophysical Research Communications, 2015, 459(2): 201 - 207.

[12] Jin Y C, Lu J S, Wen J L, et al. Regulation of growth of human bladder cancer by miR-192[J]. Tumour Biol, 2015, 36(5):3791 - 3797.

[13] Liu X Q, Lian B F, Lin Y. Gene regulatory network of hepatocellular carcinoma: a review[J]. Chinese Journal of Biotechnology, 2016, 32(10):1322 - 1331.

[14] Sia D, Hoshida Y, Villanueva A, et al. Integrative molecular analysis of intrahepatic cholangiocarcinoma reveals 2 classes that have different outcomes[J]. Gastroenterology, 2013, 144(4):829 - 840.

[15] Single-nucleotide_polymorphism[OL]. https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism.

[16] Signal_transduction[OL]. https://en.wikipedia.org/wiki/Signal_transduction.

[17] Cai R C, Zhang Z J, Hao Z F. BASSUM: a bayesian semi-supervised method for classification feature selection[J;. Pattern Recognition, 2011, 44(4):811 - 820.

[18] Aliferis C F, Tsamardinos I, Statnikov A. HITON: a novel markov blanket algorithm for optimal variable selection[C]//Proceedings of the 2003 American Medical Informatics Association Annual Symposium. Washington, DC, USA, 2003: 21 - 25.

[19] Tsamardinos I, Aliferis C F, Statnikov A. Time and sample efficient discovery of markov blankets and direct causal relations[C]//Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2003: 673 - 678.

[20] Mcdonald J H. Handbook of biological statistics[M]. Baltimore, MD: sparky house publishing, 2009.