论著

文章编号:1000-5404(2008)05-0396-03

从时间序列基因芯片数据中挖掘跨事务关联规则

彭 斌[1,2],李辉智[3],易 东[1] ([1]第三军医大学军事预防医学院卫生统计学教研室,重庆 400038;[2]重庆医科大学卫生统计学教研室,重庆 400016;[3]西南政法大学刑侦学院,重庆 400031)

摘 要:目的 探讨从时间序列基因芯片数据中发掘基因之间的作用关系。方法 采用跨事务关联规则挖掘技术及Gene Ontology标注库对时间序列基因芯片数据进行分析处理。结果 根据6个不同发育时间段的基因芯片表达数据,采用2倍法从10 080个探针中筛选出119个差异表达基因参与规则挖掘,产生的规则达1 300多条,筛选保留J-Measure值最大的10条关联规则并构建基因关联网络图。结论 跨事务关联规则能够描述时序基因之间的作用关系,根据关联规则可以对基因的表达状态进行预测。

关键词:跨事务关联规则;时间序列基因芯片数据;数据挖掘

中图法分类号:R195.1;R318.04 文献标识码:A

Mining inter-transaction association rules from time series microarray data

PENG Bin[1,2], LI Hui-zhi[3], YI Dong[1] ([1]Department of Health Statistics, College of Preventive Medicine, Third Military Medical University, Chongqing 400038; [2]Department of Health Statistics, Chongqing Medical University, Chongqing 400016; [3]Criminal Investigation Institute, Southwest University of Political Sciences and Law, Chongqing 400031, China)

Abstract: Objective To deduce the interactions between genes from time series microarray data. Methods We used inter-transaction association rules mining technique and GO (Gene Ontology) annotation to analyze the microarray data. Results Using 2-fold-change method, 119 differential expression genes were identified from total 10 080 genes or ESTs, whose expression levels varied significantly on 6 periods of fetus cerebellar development. As a result, about 1 300 inter-transaction association rules were extracted and 10 top rules were kept for their maximum J-measure values. A genes association network graph was deduced based on the 10 top rules. Conclusion Inter-transaction association rules are able to deduce the interactions between genes from time series microarray data and the gene expression status can be predicted based on the association rules.

Key words: inter-transaction association rules; time series microarray data; data mining

基因芯片是目前研究基因与基因之间作用关系的一种重要实验技术。它产生的海量基因表达数据蕴含着基因在特定时间点的活动状态及相互作用关系的信息。对芯片实验数据信息的正确提取是当前广泛应用基因芯片技术所面临的巨大挑战,也是生物信息学研究的热点问题之一。

关联规则是数据挖掘领域的一项重要研究内容,已被用于基因芯片数据的分析,它能够推测基因之间的表达关系[1-3],类似规则如"基因A表达上调,基因B表达下调→基因C表达上调"反映了3条基因间的关系。然而,对于具有时间顺序的基因芯片数据,传统的关联规则挖掘方法则无法发现类似"基因A在某时刻表达上调→基因B在下一时刻表达上调,基因C在下一时刻表达下调"的规则。这种跨时间点的关联规则是对传统关联规则的扩展[4],称为跨事务关联规则。本研究探讨从时间序列的基因芯片数据中提取描述基因之间关系的跨事务关联规则,动态地描述基因之间的关系,从而揭示基因之间的相互作用行为。

1 资料与方法

1.1 数据来源

基因芯片实验数据由香港城市大学曾志雄博士研究组提供。每张芯片检测了10 080个基因(片段),共得到12张基因芯片数据。

1.2 数据预处理

取每个阶段2张芯片对应基因片段的平均表达值的对数值(基底为2)作为每个阶段的处理数据,得到6张与发育阶段

基金项目:国家自然科学基金(60371034)
Supported by the National Natural Science Foundation of China (60371034)
作者简介:彭 斌,男,重庆市人,博士研究生,讲师,主要从事生物信息数据挖掘方面的研究。电话:(023)68485259,E-mail:cellepy@163.com
通信作者:易 东,电话:(023)68752344,E-mail:yd_house@hotmail.com
收稿日期:2007-09-05;修回日期:2007-11-19

相对应的芯片数据。根据广泛采用的 2 倍法确定差异表达基因,即选择对数值为 ±1 作为差异表达基因的截断点,在 2 个截断点之外的基因视为差异表达基因。

由于基因表达结果呈现出 3 种状态:上调、下调、无改变,而跨事务关联规则挖掘算法的输入变量通常是类似"有、无"的 2 值变量,因此,本研究采用 2 个哑变量来表达每条基因的这 3 种状态:设 Zi 是基因 i 的 Ratio 值取对数后的数值。如果 Zi > 1,则 2 个哑变量的取值为 Ui = 1,Di = 0,表示基因表达"上调";相反,如果 Zi < −1,则 2 个哑变量的取值为 Ui = 0,Di = 1,表示基因表达"下调"如果 Zi 的值介于 −1 与 +1 之间,则 2 个哑变量的取值为 0,表示基因表达"无改变",即非差异表达基因。

根据 2 倍法,筛选出至少 4 个时间点呈现差异表达的目标基因共 500 多条,其中只有 119 条具有正式的基因名称。这 119 条基因被提取出来用于跨事务关联规则的挖掘。

1.3　跨事务关联规则挖掘方法介绍

跨事务关联规则是对时序的资料进行分析的数据挖掘方法,可以发现 2 个或多个变量之间在时间上的关联。

设基因集合 $\sum = \{e_1, e_2, \cdots, e_n\}$,也称项集;$T_i$ 为时刻 i 对基因的 1 次检测结果,$T_i = \{e_1(i), e_2(i), \cdots, e_n(i)\}$,$i = 1, 2, \cdots, k$,且 $T_i \subseteq \sum$。一般称 T_i 为第 i 条事务,T_i 是具有时间先后顺序的。全部时间序列基因芯片数据 T_i 的集合构成数据库 D,$D = \{T_1, T_2, \cdots, T_k\}$,称为事务数据库。对此事务数据库进行分析所挖掘出的关联规则属于事务内关联规则,即传统的关联规则。

为了挖掘出类似"{基因 A(t0) ↑} → {基因 B(t1) ↑,基因 C(t1) ↓}"的跨时间点的规则,需要对具有时间顺序的事务数据库 D 进行"滑窗"处理,即设置 1 个时间跨度参数 w,也叫滑窗大小,让滑窗从事务数据库 D 的开始位置依次滑到数据库的结尾位置,每次向后移动 1 个时间点,滑窗每次停留时,位于窗口之内的所有事务就形成 1 条超级事务 T_i^s(图 1)。图 1 中小写英文字母为基因名,带下标的为扩展基因名,即含有时间顺序的基因名。扩展基因名是跨事务关联规则挖掘的对象,其所有可能值的集合为扩展基因集 $\sum'$。如果滑窗大小为 w,则

$$\sum' = \{e_1(0), \cdots, e_1(w-1), e_2(0), \cdots, e_2(w-1), \cdots, e_n(0), \cdots, e_n(w-1)\}$$

全部超级事务的集合形成超级事务数据库 D^s。跨事务关联规则挖掘就是从该超级事务数据库中挖掘规则。

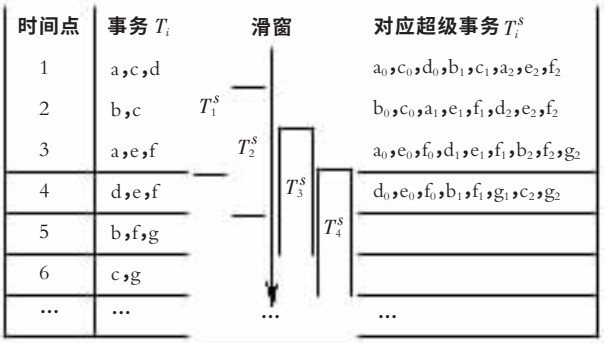


图 1　滑窗大小 w = 3 时形成的超级事务

跨事务关联规则的一般形式是:$X \rightarrow Y$,其中 X 和 Y 是基因的集合,并且满足以下条件:

ⅰ　$X \subset \sum', Y \subset \sum'$

ⅱ　$\exists e_i(0) \in X, 1 \leqslant i \leqslant n$

ⅲ　$\exists e_i(j) \in X, 1 \leqslant i \leqslant n, j \neq 0$

ⅳ　$X \cap Y = \varnothing$

跨事务关联规则与传统关联规则一样是建立在最小支持度和可信度阈值基础之上的。跨事务关联规则的支持度(support)和可信度(confidence)定义为:

$$\text{support} = |T_{XY}|/S$$

$$\text{confidence} = |T_{XY}|/|T_x|$$

其中 S 为事务数据库中事务数,$|T_{XY}|$是包含 XUY 的超级事务数,$|T_x|$是包含 X 的超级事务数。

为了选择最有价值的规则,本研究采用 J-Measure 指标对规则的兴趣度进行度量,根据此值大小对规则进行排序,选择度量值较大的规则。

此外,在选择规则时只考虑 X 为单项的关联规则,如只保留规则 {B} → {C D},而不考虑如下类似的规则 {A,C} → {B}、{B,D} → {A,C}。

跨事务关联规则挖掘所设置的支持度最小阈值为 0.75,可信度最小阈值为 0.9,滑窗大小(w)为 2。

2　结果

2.1　跨事务关联规则

本研究总共挖掘出满足条件的跨事务关联规则 1 300 多条,对所有规则按照 J-Measure 值大小进行排序,寻找兴趣度大的规则。表 1 列出前 10 条 J-Measure 值最大的规则。这些跨事务关联规则相对于经典关联规则的最大区别在于它描述了相邻 2 个时间点(w = 2)基因的表达关系。比如,第 1 条关联规则"*CRYAB*(0) ↓ → *RELN*(1) ↑ & *ZIC2*(1) ↑"明确地提示:如果基因 *CRYAB* 在 0 时刻表达下调,则基因 *RELN* 和 *ZIC2* 将在 1 时刻表达上调。此时,如果知道基因 *CRYAB* 在某个时刻的表达下调,就可以推测下一个时刻基因 *RELN* 和 *ZIC2* 将表达上调。另一方面,为了说明跨事务关联规则提取的基因表达模式的合理性,我们将第 1 条规则涉及到的 3 条基因 *CRYAB*、*RELN* 和 *ZIC2* 的实际表达水平抽取出来,将它们的实际表达水平与挖掘出的规则进行比对(图 2)。显然,表 1 中的每 1 条跨事务关联规则分别描述了不同基因之间的关系。

表 1　J-Measure 值较大的前 10 条跨事务关联规则

序号	跨事务关联规则	可信度	支持度	J-Measure
1	*CRYAB*(0) ↓ → *RELN*(1) ↑ & *ZIC2*(1) ↑	1.0	0.8	0.257 5
2	*ZIC2*(0) ↑ → *LPL*(1) ↓ & *DACH*(1) ↑	1.0	0.8	0.257 5
3	*CRYAB*(0) ↓ → *GAS7*(1) ↓ & *NEUROD1*(1) ↑	1.0	0.8	0.257 5
4	*ENC1*(0) ↓ → *LPL*(1) ↓ & *DACH*(1) ↑	1.0	0.8	0.257 5
5	*NEUROD1*(0) ↑ → *LPL*(1) ↓ & *CKMT2*(1) ↓	1.0	0.8	0.257 5
6	*APP*(0) ↓ → *NEUROD1*(1) ↑ & *ZIC2*(1) ↑	1.0	0.8	0.257 5
7	*RELN*(0) ↑ → *CKMT2*(1) ↓ & *DACH*(1) ↑	1.0	0.8	0.257 5
8	*GAS7*(0) ↓ → *DACH*(1) ↑ & *LIMK2*(1) ↓	1.0	0.8	0.257 5
9	*DDAH1*(0) ↓ → *RELN*(1) ↑ & *ZIC2*(1) ↑	1.0	0.8	0.257 5
10	*ENC1*(0) ↓ → *DACH*(1) ↑ & *LIMK2*(1) ↓	1.0	0.8	0.257 5

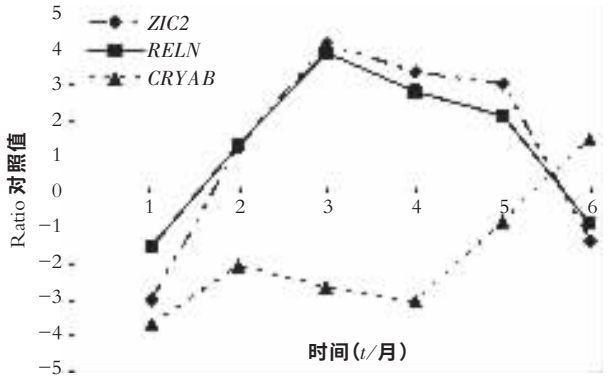


图2 3条基因的表达水平

2.2 基因关联网络图

从表1的跨事务关联规则可以发现,相同的基因可能出现在某条规则的左边,也可能出现在另一条规则的右边,比如基因 *ZIC2* 在第2条规则中出现在左边,在第1、6、9条规则中则出现在右边。因此,当需要分析基因 *ZIC2* 与其他基因的相互表达关系时,就必须同时考查涉及到基因 *ZIC2* 的规则,并将其纳入关联网络中进行分析。我们定义了将关联规则转换为关联网络图的方式:从规则左边基因用方向箭头指向右边基因,从而形成关联网络图。根据表1的规则及上述转换方式,得到关联网络图(图3)。图中实线箭头表示所连接的2个基因表达方向相同,虚线箭头则表示所连接的2个基因表达方向相反。关联网络图清晰地描述了基因之间的表达关系。

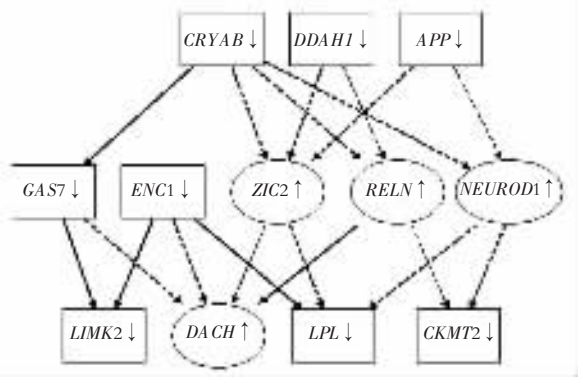


图3 基因关联网络图

3 讨论

关联规则挖掘是数据挖掘和知识发现领域中的一种重要技术手段。它是建立在一定支持度与可信度下的条件规则,在规则左边出现的基因很可能扮演调控的角色,而右边的基因则是被调控的角色[4-6]。跨事务关联规则是传统关联规则的扩展。

跨事务关联规则加入了时间的因素,因而它具有一定的预测作用。例如,从时序芯片数据中挖掘出规则"*CRYAB*(0)↓→*RELN*(1)↑,*ZIC2*(1)↑",它描述了这3条基因之间在时间上的相互表达关系。如果某时刻基因 CRYAB 表达下调,则有很大的把握度可以推测基因 *RELN* 和 *ZIC2* 将在下一时刻表达上调。因此,在实际中可以根据跨事务关联规则来预测它所涉及到的基因之间在时间上的表达关系。

跨事务关联规则所反映的基因之间的关系是在数据挖掘的层面上体现出来的,是建立在已有数据基础之上的,而这些数据是基因活动行为的外在表现,因此规则只描述了基因之间相互作用关系的一个侧面。比如,从图2可以看出,表1中的第1条关联规则从实际表达数据上来说是合理的。为了寻找该规则的生物学基础,将其涉及的 *CRYAB*、*ZIC2*、*RELN* 3条基因提交GO(Gene Ontology)数据库,查询结果显示:*CRYAB* 基因主要参与发育过程,涉及器官、肌肉、眼、骨骼的发育等;*ZIC2* 基因主要参与发育过程,涉及系统发育、器官发育、胚胎发育,神经系统及中枢神经系统发育、神经管发育等;*RELN* 基因主要参与发育过程,涉及系统发育、器官发育、神经系统及中枢神经系统发育,神经元发育、分化、迁移等。显然,3条基因均参与了发育过程,从这个意义上来看,它们之间存在表达关系是合理的;此外,可以看出 *RELN* 与 *ZIC2* 均参与了系统、器官及中枢神经系统的发育,它们表达一致很可能是由于参与了共同的生物学过程。所以该关联规则所表达的信息背后具有一定的生物学基础。当然,在目前的知识背景下对所有的规则进行合理的解释还存在一定的难度。

总之,跨事务关联规则可以从时间序列基因芯片数据中发掘基因之间的相互关系,而且规则可以用于基因表达的预测,更重要的是规则所提示的信息可以为进一步研究提供假说依据。

参考文献:

[1] Becquet C, Blachon S, Jeudy B, *et al*. Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data[J]. Genome Biology, 2002, 3(12): RESEARCH0067.

[2] Georgii E, Richter L, Ruckert U, *et al*. Analyzing microarray data using quantitative association rules [J]. Bioinformatics, 2005, 21(Suppl 2): ii123 - ii129.

[3] Tung A K H, Lu H J, Han J, *et al*. Efficient Mining of Intertransaction Association Rules[J]. IEEE Trans Knowl Data Eng, 2003, 15(1): 43 - 56.

[4] 秦亮曦, 刘新峰, 史忠植. 基于片段模式的多时间序列关联分析[J]. 计算机科学, 2006, 33(1): 232 - 235.

[5] Carmona-Saez P, Chagoyen M, Rodriguez A, *et al*. Integrated analysis of gene expression by Association Rules Discovery [J]. BMC Bioinformatics, 2006, 7: 54.

[6] Creighton C, Hanash S. Mining gene expression databases for association rules [J]. Bioinformatics, 2003, 19(1): 79 - 86.

(编辑 薛国文)