PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2007, 34(8): 871~880
www.pibb.ac.cn

一个基于基因表达谱的基因逻辑网络模型的建立与应用 *

阮晓钢 [1)] 王金莲 [1)**] 李 辉 [2)]
([1)]北京工业大学电子信息与控制工程学院，北京 100022；[2)]北京工业大学计算机学院，北京 100022)

摘要 基因之间除了线性关联作用关系外，还存在着非线性的逻辑关系. 用这种逻辑关系构建的生物系统网络模型对研究细胞内的各种生物通路和细胞分子网络非常重要. 首先，根据图着色原理确定了基因的低阶和高阶逻辑关系，然后应用结肠癌基因表达谱数据分析了 51 个癌基因和抑癌基因的逻辑关系，在此基础上构建了结肠癌基因表达的逻辑网络. 通过这个网络模型发现了与 KEGG 数据库中结肠癌通路一致的转化生长因子信号通路，并分析了各生物通路成员之间错综复杂的关系. 实验结果表明，基因逻辑网络模型在一定程度上揭示了结肠癌基因和抑癌基因之间并行、分叉等复杂的相互作用关系，反映了结肠癌发病的复杂分子机制，为分子生物医学家提供了一个参考模型.

关键词 结肠癌，基因表达谱，基因网络，图着色
学科分类号 TP18

结肠癌在我国是一种常见的恶性肿瘤，有着较高的发病率和死亡率. 1990 年 Fearon 和 Vogelstein[1] 提出了结、直肠癌变的多阶段遗传学模型，基本内容是结、直肠癌的发生涉及多层次、多种癌基因的激活以及抑癌基因的失活. Smyrk 和 Lynch[2]在 1997 年对此模型做了新的修改和补充，并且给出了结肠癌恶变的几种途径. 越来越多的研究证明，结肠癌的发病机制不是单个致癌基因激活或者抑癌失活所致，而是多个基因协同作用的结果，癌变过程伴随着基因功能模块和基因表达模式的改变[3]. 由于目前对细胞内正常和病理机制的认识还不够，人们还不能完全解释清楚肿瘤发生的分子机制. 然而结合生物信息学和系统生物学，建立基因 - 基因相互作用网络模型，可为生物分子医学工作者提供一种参考模型，以便于他们提出肿瘤发病机制假设，通过假设再去实验验证，最终发现致病基因的作用机制.

基因表达谱和系统发育谱在肿瘤发病机制分析方面具有一定优势[4,5]. 基于基因表达谱的基因网络模型已经形成了很多方法：无监督学习、贝叶斯、逻辑回归、互信息、布尔网络、Petri 网、微分方程以及概率图模型等. 2002 年 Marcotte 等[6]用去卷积方法建立了细胞动态周期和细胞状态的网络模型. 2004 年 Bowers 等[7]在 Science 上发表了用系统发生谱逻辑分析方法(LAPP)，分析 67 个已完全测序物种的 4 873 个 COG(直向同源蛋白簇)的文章，并发现了 750 000 个新的蛋白质之间的逻辑关系. 2005 年 Zhang 等[8]建立了贝叶斯模型，对 104 个不同家族的 COG 数据进行了四元蛋白之间的逻辑关系分析，发现了 143 057 个新的四元蛋白之间的关系. 2005 年 Bowers 用 LAPP 方法分析了神经胶质瘤基因表达谱和肿瘤预后标志基因之间的相互关系，依此预测疾病状态和分子功能之间的关系.

然而，目前大多数基因网络是基于基因之间线性关联而建立的模型，非线性的基因网络研究刚刚开始，尤其是建立肿瘤基因的逻辑网络模型. 因此，本文结合 LAPP 方法建立了结肠癌基因的逻辑网络模型，首先用图着色原理确定低阶(低阶指 2 个基因的逻辑关系，本文指一阶逻辑)和高阶(高阶指大于 2 个基因的逻辑关系，本文指二阶逻辑)逻辑关

*国家自然科学基金重点资助项目(60234020).
** 通讯联系人. Tel: 13811808114, E-mail: wijinlian1999@gmail.com
收稿日期：2007-01-23，接受日期：2007-03-02

系类型，然后，结合 LAPP 的不确定系数方法，确定出基因之间的低阶和高阶逻辑关系[7]，并且计算出每种逻辑关系在样本中的支持度. 最后根据得到的基因逻辑关系模型，构建了肿瘤基因的逻辑网络模型并对这个网络模型进行分析.

1 材料与方法

1.1 样本和数据

Laiho等[9]收集了 37 个来自芬兰的结肠癌患者的癌组织，其中 29 人被诊断为 2 级肿瘤，4 人为 3 级肿瘤，4 人为 1 级肿瘤；2 人为 Dukes D 期结肠癌，33 人为 Dukes C 期结肠癌，2 人为 Dukes B 期结肠癌；锯齿状结肠癌患者 8 人，非锯齿状结肠癌患者 29 人；近端结肠癌 15 人，远端结肠癌 22 人；19 个女性患者，18 个男性患者. 所有样本的 mRNA 均从新鲜的冷冻肿瘤组织中用 Trizol 法提取出来，用 RNeasy 纯化. RNA 的质量分析由分光光度计和 Agilent2100 生物分析仪完成.

结肠癌数据来源于 2006 年 6 月 Laiho 公布了在 GEO 数据库中的 GSE4045[9]系列数据. 所有样本的数据用 Affymetrix 公司的 HG133A 芯片，检测了 22 283 个探针在 37 个样本中 mRNA 的表达丰度水平. 芯片数据经过 MAS 5.0 分析后获得，其中包含每个基因表达量的实数型值和表示表达水平的离散值，和实数型值相比离散值受噪声影响小，不因数据预处理方法不同而对处理结果有影响，离散值适合于只需知道哪些基因在试验中是否被检测到的情况. 离散值由 Presence-Call 记录，它有 3 种不同取值，*A* (Absent)，*P*(Present)，*M*(Marginal)分别代表该基因：“表达水平低于检测阈值”、“表达水平高于检测阈值”、“介于表达和不表达之间”. 设定当探针对 16~20 对时，2 个默认的检测阈值分别为 α_1=0.04，α_2=0.06，则当 $p<\alpha_1$ 时，基因表达，当 $\alpha_1=p<\alpha_2$ 时，基因介于表达或者不表达之间，当 $p=\alpha_2$ 时，基因不表达.其中 p 值为 Wilcoxon's 假设检验中拒绝原假设的概率值(具体计算方法参考文献[10])，p 值越小显著性越好. 由于 M 值在表达谱中很少，所以用 A 值代替. 在本文中 Presence-Call 值用 1 表示基因在样本中表达，Absence用 0 表示该基因在样本中不表达，此值用 Dchip 6.0[10]对原始的芯片激光扫描数据分析得到. 本文采用经 Dchip6.0 处理后的 Presence-Call 默认值作为 0，1 值来计算基因逻辑组合的不确定系数，共得到 22 283 个探针的二进制矢量基因表达谱.

1.2 基于图着色原理的逻辑关系类型

基因之间的逻辑关系不同于基因之间一对一的“对等”关系，它是一种“不对等”的作用关系，一个基因的表达依赖于其他基因的表达，也就是说 2 个基因的逻辑组合共同决定了第 3 个基因的表达行为，通过对基因表达谱的分析发现，基因 C 表达当且仅当基因 A 和 B 同时表达(文中定义的第一种逻辑关系类型)，由此推测，基因 C 的功能可能必须在基因 A 和 B 都表达的时候才得以发挥. 相反，基因 C 表达当且仅当基因 A 或者基因 B 表达(对应第 7 种逻辑关系类型)，可见当生物体可能在 2 个完全不同但功能相同的基因中择其一和第 3 个基因共同完成一个生物功能，即基因 C 和 A 或者基因 B 和 C 的逻辑组合.

对如图 1 所示的韦恩图用图着色原理确定基因的一阶、二阶逻辑关系. 其中 2 个圆分别表示基因 A 表达和 B 表达，阴影表示基因 C 表达. 如果用一种颜色对图 1 中的不同区域进行着色，总共有 2^4 种着色方案，则每一种着色方案对应着一种特定的逻辑关系，基因表达的逻辑关系类型、着色方案、逻辑函数和逻辑关系描述见表 1.

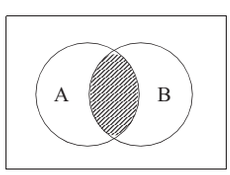
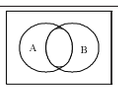
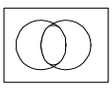
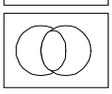
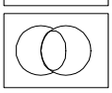
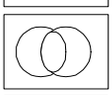
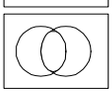
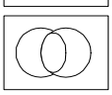
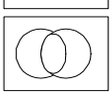
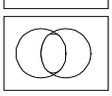
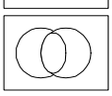
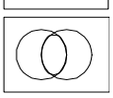
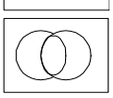
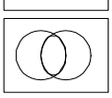
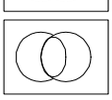
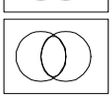
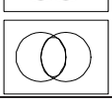


Fig. 1 Graph coloring

Bowers 等[7]将 2 个蛋白质之间的逻辑关系称为蛋白质的低阶逻辑关系，本文称为一阶逻辑关系，表 1 中第 11 到 14 逻辑关系类型是一阶逻辑关系；第 1 到第 10 逻辑关系类型是二阶逻辑关系；第 15 和第 16 逻辑关系类型既不是一阶也不是二阶逻辑关系，C 基因的表达与否与 A、B 基因的表达无关. 本文根据图着色原理确定的 16 种逻辑关系类型和 Bowers 等[7]所确定的 16 种逻辑关系类型完全一致，Bowers 将第 5 和第 9 逻辑关系视为同一种逻辑关系，将第 6 和第 10 种逻辑关系视为同一种逻辑关系，因此 Bowers 将二阶逻辑关系定义为 8 种. 本文将对包括 Bowers 的 8 种逻辑关系在内的 14 种一阶和二阶逻辑关系进行分析.

Table 1 The logic relationships illustrated with Venn diagrams and logic statement

Type	Venn diagram	Logic function	Logic statement
1		C= A∧B	C is present if and only if (iff) A and B are both present
2		C=~(A∧B)	C is present iff A absent or B is absent
3		C＝A∨B	C is present iff A is present or B is present
4		C＝~(A∨B）	C is present iff A is absent and B is absent
5		C＝~A∧B	C is present iff A is absent and B is present
6		C＝~A∨B	C is present iff B is present or A is absent
7		C＝~(A–B)	C is present iff one of either A or B is present
8		C＝(A–B)	C is present iff A and B are both present or A and B are both absent
9		C＝A∧~B	C is present iff A is present and B is absent
10		C＝A∨~B	C is present iff A is present or B is absent
11		C＝A	C is present iff A is present
12		C＝B	C is present iff B is present
13		C＝~A	C is present iff A is present
14		C=~B	C is present iff B is present
15			C is present
16			C is absent

1.3 逻辑关系确定方法

经过如上所述的数据处理得到每个基因的二进制表达值，其中 1 表示某基因在样本中表达，0 表示某基因在样本中不表达. 单独一个基因 A 或者 B 的表达与否都不能预测基因 C 的表达与否，显然只有计算了 A、B 2 个基因组合的逻辑相关的确定性大小，才能确定基因 A、B 对 C 作用的可能性. 另外，基因二阶逻辑关系的确定性必须大于任意 2 个基因一阶逻辑关系的确定性，因为基因 C、A 和 C、B 的确定性大必然导致 ABC 3 个基因之间的二阶逻辑关系确定性大. 本文用不确定系数来计算基因逻辑关系的确定性，一阶逻辑的不确定系数的定义如下：

$$U(c|a)=[H(c)+H(a)-H(a,c)]/H(c) \quad (1)$$

其中，$0 \leqslant U(c|a) \leqslant 1$ 为不确定系数，表示基因 C 和基因 A 相关的可能性大小，当 $U(c|a)$等于 1 表示基因 A 和 C 之间的逻辑关系完全确定，当 $U(c|a)$等于 0 表示基因 A 和 C 之间完全不确定. $H(a)$为基

因 A 的独立熵，$U(c,a)$为基因 A、C 的联合熵：

$$H(a)=\sum p(a)\log p(a) \quad (2)$$

$$H(c,a)=\sum\sum p(c,a)\log p(c,a) \quad (3)$$

其中，$p(a)$, $p(c,a)$分别表示基因 A 在所有样本中表达的概率和基因 A、C 在样本中表达的联合概率.

二阶逻辑的不确定系数定义如下：

$$U[c|f(a,b)]=\{H(c)+H[f(a,b)]-H[c,f(a,b)]\}/H(c) \quad (4)$$

其中，$U[c|f(a,b)]$表示在基因 A 和 B 逻辑组合下，基因 C 中包含基因 A、B 组合的确定性，函数 $f(a,b)$为表 2 中 14 种逻辑关系类型其中的任意一种逻辑函数. $H(c)$、$H[f(a,b)]$和 $H[c,f(a,b)]$分别为基因 A 的独立熵，基因 A、B 的联合熵和基因 A、B、C 的联合熵.

设一阶逻辑的阈值为 R_1, 二阶逻辑的阈值为 R_2，二阶逻辑关系的确定必须满足以下条件：

$$\begin{cases} U(c|a)<R_1 \\ U(c|b)<R_1 \\ U[c,f(a,b)]\gg U(c|a) \text{ 和 } U(c|b) \\ R_2>R_1 \end{cases} \quad (5)$$

Bowers 分别取阈值 R_1 和 R_2 为 0.3 和 0.6. 本文对阈值 R_1 和 R_2 按照以下步骤确定：

首先，根据计算所得 U 值，画出一阶、二阶逻辑关系的 U 值分布图，分别选择图中的变化最大的 U 值 U_{R_1} 和 U_{R_2} 作为参考阈值.

然后，根据医学先验知识确定阈值 U'_{R_1} 和 U'_{R_2}. 从 PubMed 数据库中选择说明所选基因相互之间作用的文献，以这些挑选出来的相互作用的基因作为先验知识，分别计算这些基因的不确定系数 $U'_{R_1}(c_i|b_i)$，i=1，2，3，…n，和 $U'_{R_2}(c_i|f(a_i,\ b_i))$，$i$=1，2，3，…$n$，根据 $U'_{R_1}(i)$和 $U'_{R_2}(i)$数值的分布范围，如果某个范围 70%以上的数值包含了 U_{R_1} 和 U_{R_2}，则选择这个范围的下界作为 R_1 和 R_2：

$$R_1=\min U(c_i|b_i) \quad (6)$$

$$R_2=\min U(c_i|f(a_i,b_i)) \quad (7)$$

1.4 基因逻辑关系的表达模式和样本支持度

基因逻辑关系的组合由于每个样本个体的不同而导致其逻辑基因表达模式可能相同也可能不同，只有通过统计在多个样本中具有显著性逻辑关系的基因组合，才能找到对肿瘤发病具有贡献的基因网络和模式的改变，进而了解基因组和疾病的关系. 10 种逻辑关系表达模式见表 2. 本文用 0，1 的逻辑组合表示基因之间的逻辑表达模式，如果基因 A 在样本 a 中表达，基因 B 在样本 a 中也表达，基因 C 在样本 a 中也表达，则用 1^1=1 表示 C＝A^B. 每个基因的表达谱包含多个病人 mRNA 表达丰度，每种逻辑关系组合在每个病人中的存在就会不同，因此每种逻辑关系组合就存在对不同样本的支持度，那么确定一种逻辑关系是否和肿瘤相关，不仅需要考虑逻辑关系的置信度，而且还要考虑其在样本中的支持度，支持度指每种逻辑类型在样本中发生的概率.

Table 2 Gene logic expression pattern

Type	Expression pattern				Type	Expression pattern			
	A	1				A	1	0	0
1	B	1			2	B	0	1	0
	C	1				C	1	1	1
	A	1	1	0		A	0		
3	B	0	1	1	4	B	0		
	C	1	1	1		C	1		
	A	1				A	0		
5	B	0			6	B	1		
	C	1				C	1		
	A	1	0	0		A	1	0	1
7	B	1	0	1	8	B	1	0	0
	C	1	1	1		C	1	1	1
	A	1	0			A	1	0	
9	B	0	1		10	B	1	0	
	C	1	1			C	1	1	

按照以上方法获得 2 个和 3 个基因逻辑组合的不确定系数及支持度，以此构建基因的逻辑网络. 本文构建的基因逻辑网络是一个加权有向网，基因表示网络节点，大于给定阈值基因之间的逻辑关系构成网络的边，支持度表示 2 个基因之间逻辑关系的强弱形成边的权重，边的方向表示基因之间的作用方向. 网络的构建借用电子电路图的构成方法，用与、或、非门等符号表示基因之间的逻辑关系.

2 实验结果

2.1 结肠癌基因和抑癌基因选取

从 PubMed[11]数据库记录的与结肠癌 Dukes 分期、结肠癌转移、血管生成、点突变、DNA 错配修复和细胞周期调控功能模块相关的文献中，选取 51 个结肠癌基因和抑癌基因. 所选基因至少有 5 篇以上的文献说明此基因和癌症相关，每个基因在 NCBI 数据库中的 ID，基因名称和探针 ID 号见表 3.

Table 3 The list of 51 oncogenes and cancer suppressor genes

Probe ID	GeneID	GeneSymbol	Probe ID	GeneID	GeneSymbol
204010_s_0t	3845	k-ras	204748_0t	5743	COX-2
201895_0t	369	Raf	209946_0t	7424	VEGFC
212609_s_0t	10000	AKT	212171_x_0t	7422	VEGFA
209364_0t	572	BAD	203085_s_0t	7040	TGFB
220566_0t	23533	P13K	207334_s_0t	7048	TGFB-RII
212849_0t	8312	Axin	205386_s_0t	4193	MDM2
209189_0t	2353	Fos	202520_s_0t	4292	MLH1
221558_s_0t	51176	LEF	202911_0t	2956	MSH6
208351_s_0t	5594	ERK	210947_s_0t	4437	MSH3
222146_s_0t	6925	TCF4	209421_0t	4436	MSH2
216836_s_0t	2064	C-erbB2	216039_0t	5379	PMS1
209051_s_0t	5900	RalGDS	207004_0t	596	Bcl2
202095_s_0t	332	Survivin	208478_s_0t	581	BAX
208711_s_0t	595	cyclin-D1	209805_0t	5395	PMS2
207839_s_0t	51754	B-Catenin	211300_s_0t	7157	P53
216933_x_0t	324	APC	209644_x_0t	1029	P16
201130_s_0t	999	E-cadherin	206132_0t	4163	MCC
201693_s_0t	1958	EGR1	209588_0t	2048	EPHB2
204489_s_0t	960	CD44	207433_0t	3586	IL-10
210775_x_0t	842	CASP9	207160_0t	3592	IL-12
202763_0t	836	CASP3	205479_s_0t	5328	u-PA
206939_0t	1630	DCC	203076_s_0t	4087	SMAD2
206254_0t	1950	EGF	205396_0t	4088	SMAD3
205828_0t	4314	MMP3	202526_0t	4089	SMAD4
203936_s_0t	4318	MMp9	211551_0t	1956	EGFR(RTK)
			211553_x_0t	317	Apaf

2.2 阈值的确定

根据公式(1)计算的 $U(c|a)$, $U[c|f(a,b)]$和对应的基因逻辑组合个数，如图 2 和图 3 所示，其中图 2 为一阶逻辑基因 $U(c|a)$和对应的基因个数，当 $U(c|a)$= 0.25 时，共有 12 733 个一阶逻辑关系，当 $U(c|a)$= 0.2 时，共有 26 421 个一阶逻辑关系，从医学文献中得到了有相互作用的 22 对基因，7 个相互作用的三元基因. 根据以上阈值确定的方法本文确定的 R_1=0.25，R_2=0.55. 当 U>R_1 时，2 个基因之间存在一阶逻辑关系，当 U>R_2 时，3 个基因之间存在二阶逻辑关系.

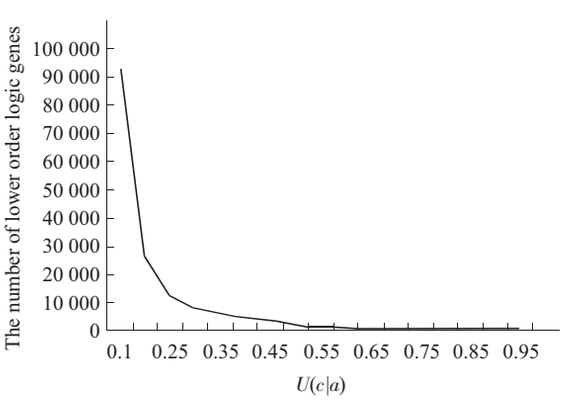


Fig. 2 The distribution of logic association genes and *U* value corresponding to their lower order logic relationship

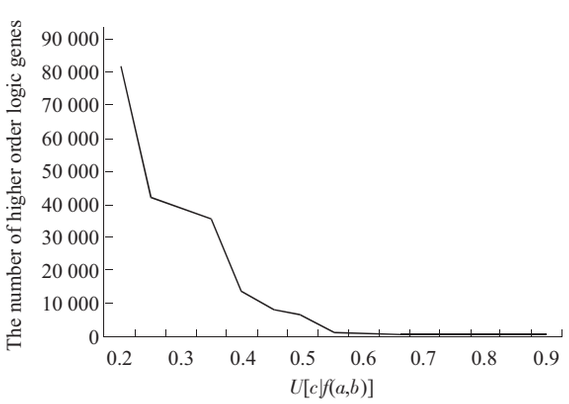


Fig. 3 The distribution of logic association genes and *U* value corresponding to their higher order logic relationship

2.3 显著性检验试验

为了检验所得二阶逻辑关系的统计显著性，假

设具有二阶逻辑关系的基因在扰动试验前后没有差异，本文利用统计模拟产生随机样本，对样本的属性进行扰动试验，以统计二阶逻辑关系基因组合在扰动试验中出现的概率 P，本文给定显著水平 $\alpha=0.05$，主要步骤：a. 产生一个保持原始矩阵 0，1 分布不变的独立同分布的随机矩阵；b. 计算随机矩阵二阶逻辑基因组合的 U^* 值，并统计大于原始矩阵 U 二阶逻辑基因组合个数；c. 重复以上步骤 100 次；d. 计算随机矩阵的逻辑组合的 P 值，P 值为扰动实验中逻辑基因组合的 U^* 值大于或者等于原矩阵中 U 值的概率.

$$P=\frac{\#(|U^*|\geq|U|)}{C_{51}^3\times 100} \tag{7}$$

其中，U 是原始矩阵计算所得的各个 U 值，U^* 是每次产生的随机矩阵的 U 值. 统计所得 $U(c|f(a,b))$ 和观察得到二阶逻辑基因的数目如图 3 所示. 图 3 中逻辑基因个数为 100 次扰动实验中 $U^* > U$ 所有基因逻辑组合的个数，P 值为 100 次扰动中 U^* 大于 U 值的概率，横轴为 U 值，当扰动次数大于 50 次时，逻辑基因的个数趋于稳定值.

图 4 中实线曲线为原始矩阵的逻辑基因组合数，虚线曲线为扰动后逻辑基因的组合数. 可见，原始矩阵中产生的逻辑关系数远大于随机矩阵所产生的逻辑关系数，计算所得 P 值在 0.033 015~0.002 719 之间.

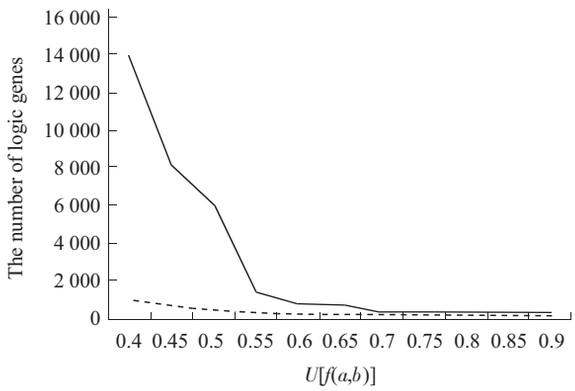


Fig. 4 The cumulative number of gene triplets above certain threshold are more frequent than the number in the random datasets

2.4 一阶和二阶逻辑基因关系

根据上面的分析，确定出一阶逻辑关系的阈值为 0.25，二阶逻辑的阈值为 0.55，通过对逻辑组合的扰动实验，选择 P 小于 0.05 的逻辑基因组合，统计了这些基因组合对应逻辑关系的柱状图，如图 5 所示. 51 个结肠癌基因共有 2 550 种一阶逻辑关系，$U(c|a)$大于 0.25 且支持度大于 60%的逻辑组合共有 39 个. 所得结果如表 4 所示.

Table 4 The list of lower order logic association genes

Gene C	Gene A	$U(c\|a)$	Support	Gene C	Gene A	$U(c\|a)$	Support
C-erbB2	DCC	0.999 778	0.621 622	C-erbB2	AKT	0.316 988	0.864 865
MSH6	EGR1	0.754 507	0.756 757	VEGFA	AKT	0.316 988	0.945 946
u-PA	MSH6	0.698 342	0.864 865	CASP9	APC	0.305 025	0.864 865
u-PA	EGR1	0.584 525	0.756 757	MLH1	TGFB	0.305 025	0.918 919
VEGFA	CASP3	0.584 525	0.945 946	MSH6	TGFB	0.305 025	0.621 622
EGR1	MSH6	0.563 821	0.837 838	u-PA	LEF	0.292 769	0.648 649
CASP9	Fos	0.509 08	0.675 676	MLH1	Raf	0.282 601	0.810 811
u-PA	BAD	0.455 651	0.864 865	MLH1	EPHB2	0.282 601	0.864 865
C-erbB2	Survivin	0.455 651	0.810 811	MSH6	EPHB2	0.282 601	0.864 865
VEGFA	Survivin	0.455 651	0.810 811	u-PA	Bcl2	0.270 98	0.702 703
VEGFA	BAX	0.455 651	0.891 892	MSH2	COX-2	0.261 982	0.783 784
u-PA	TGFB	0.251 178	0.945 946	SMAD2	COX-2	0.261 982	0.783 784
VEGFA	SMAD3	0.455 651	0.810 811	SMAD3	BAX	0.259 795	0.810 811
MSH6	u-PA	0.412 632	0.945 946	BAX	SMAD3	0.259 795	0.756 757
C-erbB2	CD44	0.411 853	0.810 811	CASP3	EGF	0.258 37	0.810 811
C-erbB2	E-cadherin	0.375 43	0.918 919	EGR1	u-PA	0.258 093	0.864 865
C-erbB2	MCC	0.375 43	0.891 892	CASP3	VEGFA	0.258 093	0.621 622
VEGFA	E-cadherin	0.375 43	0.756 757	C-erbB2	COX-2	0.251 178	0.945 946
EGFR	MCC	0.375 43	0.891 892	MMP9	APC	0.251 178	0.891 892
MLH1	LEF	0.356 784	0.621 622				

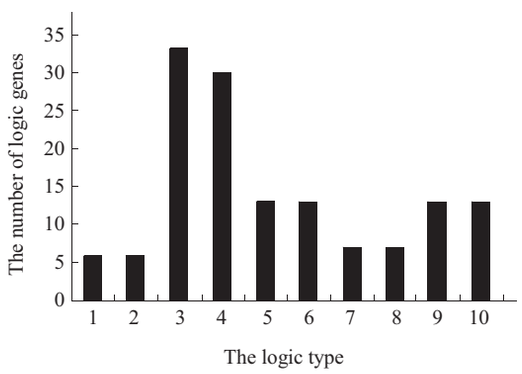


Fig. 5　The histogram showing the number of identified triplets

如图 5 所示，逻辑关系类型 3(基因 C 表达当且仅当 A 或者基因 B 表达)在结肠癌基因中最多，其次是逻辑关系类型 4(基因 C 表达当且仅当基因 A 和 B 不表达). 类型 5 和类型 9 在 Bowers 的文章中属于同一类逻辑类型，即基因 C 表达当且仅当 A 表达而 B 不表达或者基因 C 表达当且仅当基因 A 不表达而 B 表达，类型 6 和类型 10 也属于同一逻辑类型，即基因 C 表达当且仅当基因 A 表达或者 B 不表达；或者基因 C 表达当且仅当基因 A 不表达或者 B 表达. 类型 7 是逻辑异或，基因 C 表达当且仅当基因 A，B 之一表达，类型 8 时逻辑同或关系，基因 C 表达当且仅当基因 A 和 B 都表达. 本文所选的二阶逻辑基因的 U 值、支持度以及具体的逻辑类型见表 5.

Table 5　The list of higher order logic association genes

Type	Gene C	Gene A	Gene B	$U(c\|a)$	$U(c\|b)$	$U(C\|f(a,b)$	Support
3	CASP3	k-ras	Survivin	0.239 09	0.249 414	0.563 821	0.918 919
3	CASP3	EPHB2	SMAD3	0.064 709	0.249 414	0.563 821	0.918 919
3	CASP3	Survivin	SMAD3	0.249 414	0.249 414	0.563 821	0.918 919
3	CASP3	RalGDS	SMAD3	0.010 277	0.249 414	0.563 821	0.918 919
6	u-PA	Axin	EPHB2	0.200 732	0.233 031	0.584 525	0.918 919
6	MMp9	Axin	EPHB2	0.200 732	0.233 031	0.584 525	0.918 919
6	MMp9	Axin	EGF	0.200 732	0.186 219	0.510 613	0.891 892
6	MMp9	MDM2	EPHB2	0.233 031	0.233 031	0.510 613	0.891 892
3	MMp9	COX-2	EPHB2	0.216 282	0.233 031	0.510 613	0.891 892
10	VEGFA	EGF	P16	0.186 219	0.159 813	0.510 613	0.891 892
10	MMp9	EGF	P16	0.186 219	0.159 813	0.510 613	0.891 892
3	MMp9	EGF	COX-2	0.186 219	0.216 282	0.510 613	0.891 892
6	u-PA	MDM2	EPHB2	0.233 031	0.233 031	0.510 613	0.891 892
6	MMp9	ERK	VEGFC	0.159 813	0.172 616	0.510 613	0.891 892
7	SMAD4	TGFB-RII	MSH3	0.238 194	0.174 724	0.567 435	0.864 865
5	SMAD4	TGFB-RII	MSH4	0.238 194	0.174 724	0.567 435	0.864 865
8	MSH2	cyclin-D1	SMAD2	0.174 724	0.238 194	0.567 435	0.864 865
10	SMAD4	k-ras	P16	0.208 521	0.192 918	0.509 08	0.837 838
8	MSH2	Survivin	E-cadherin	0.106 488	0.067 519	0.509 08	0.837 838
7	MLH1	BAD	MSH3	0.106 488	0.174 724	0.509 08	0.837 838
3	CASP9	k-ras	VEGFC	0.208 521	0.208 521	0.509 08	0.837 838
8	cyclin-D1	Fos	SMAD3	0.208 906	0.041 157	0.540 825	0.810 811
3	CASP3	k-ras	PMS1	0.239 09	0.204 024	0.540 825	0.810 811

51个结肠癌基因的二阶逻辑关系类型总共有 565 696 种，但满足约束条件 $U(c|a)<0.25$，$U(c|b)<0.25$ 且 $U(c|f(a,b))>0.55$ 的基因有 30 个基因，这 30 个基因之间存在 23 种二阶逻辑关系，每种逻辑关系的支持度大于 60%. 其中逻辑类型 3 最多. 具体基因和数值见表 5.

2.5　结肠癌基因表达逻辑网络

通过上面的分析确定了结肠癌基因之间的一阶和二阶逻辑关系，依此构建基因表达的逻辑. 如前所述，这是一个加权有向网，权重表示逻辑关系在样本中的支持度，方向表示一个基因对第 3 个基因的作用关系，如图 6 所示. 图 6 中一阶逻辑关系用

虚线表示，二阶逻辑用实线表示. 从图 6 中可见，存在二阶逻辑关系的基因之间并不存在一阶逻辑关系，因此二阶逻辑关系分析可揭示更为复杂的基因之间的关系. 虚线框中的基因表示只有一阶逻辑关系的基因，实框中的基因表示二阶逻辑基因，椭圆框表示一阶二阶逻辑关系输出较多的基因. 蓝色表示微卫星不稳定基因，红色表示癌基因，绿色表示抑癌基因.

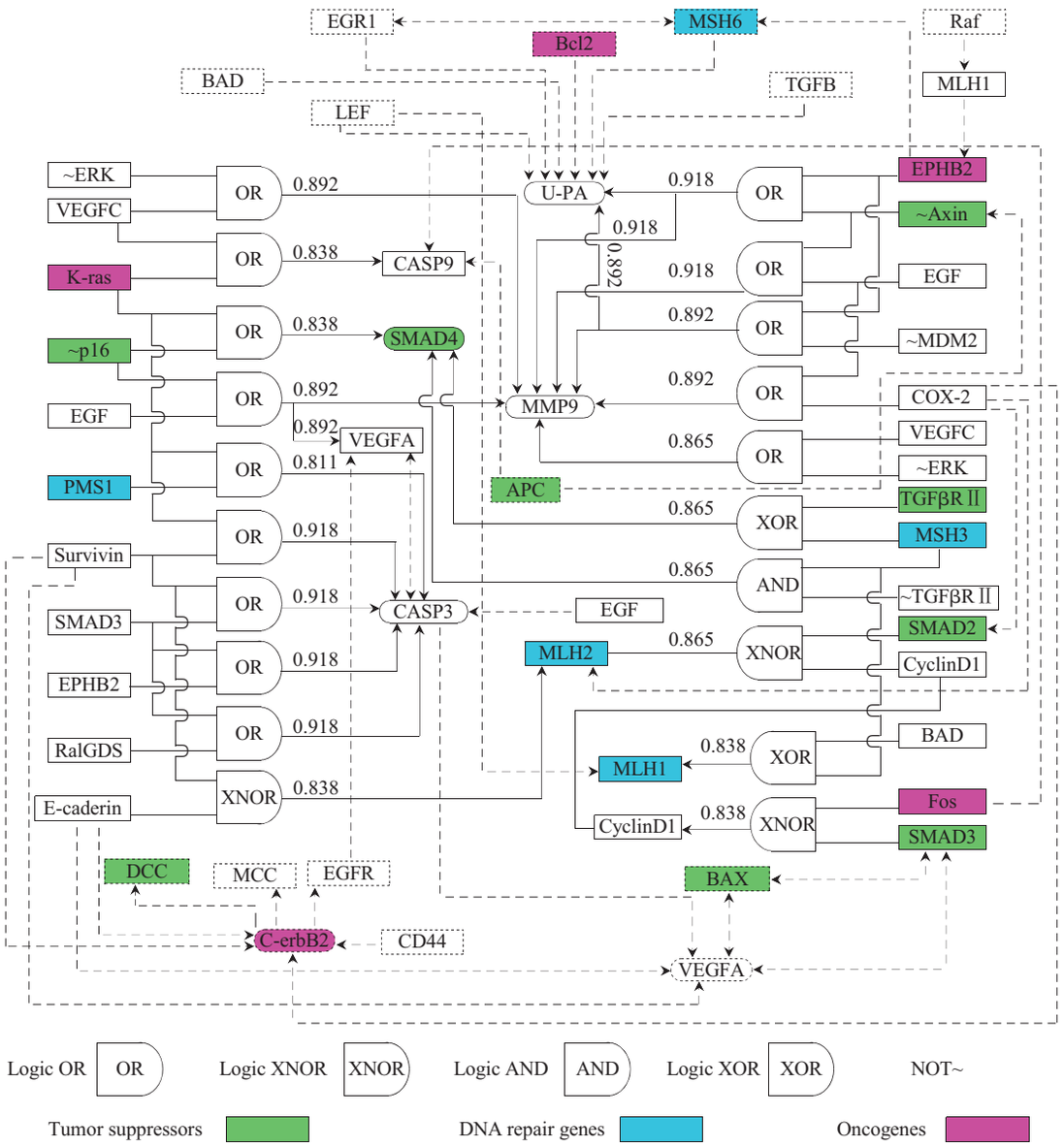


Fig. 6 The gene expression logic network of colon cancer

如图 6 所示，参与 MAPK 通路的 k-ras, Ral 鸟嘌呤核苷酸分裂刺激因子(RalGDS)与凋亡抑制基因其表达产物命名为生存蛋白(Survivin)、血管内皮生长因子 C(VEGFC)、细胞周期调节抑制基因(P16)通过逻辑或与结肠癌细胞抑制凋亡通路的(CASP9)基因、CASP3 以及转化生长因子 β(TGF-Beta)信号通路基因 SMAD4 相互作用. 基质金属蛋白酶 -9 (MMP9)在癌症发生发展过程中主要通过切割一系列的功能性组件(包括生长因子或者细胞外基质，细胞黏附分子等)的底物来发挥其在促进癌细胞生长、迁移、侵袭、转移以及血管生成多方面的作用[12]，MMP9 基因在结肠癌中的表达与 WNT 信号转导通路的轴抑制基因(Axin[13])、MDM2、细胞周期调节抑制基因(P16)、细胞外信号调节激酶

(ERK)，以及受体酪氨酸激酶(EphB2[14])、表皮生长因子(EGF)、环氧化酶诱导型基因(COX-2[CoX-2])、血管内皮生长因子 C(VEGFC)之间存在逻辑相互作用. 微卫星不稳定通路成员基因 APC (结肠腺瘤性息肉病基因) 和 WNT 信号转导通路成员基因 Axin (轴抑制基因[13])、抑制凋亡通路成员基因 CASP9 有逻辑依赖关系. 而细胞抑制凋亡通路成员基因 CASP3(天冬氨酸特异的半胱氨酸蛋白酶)的表达，可能会受到 7 个基因 5 种逻辑关系基因组合表达的影响，CASP3 的激活是调控细胞凋亡的重要步骤[15]，它的激活可能会受到 k-ras，survivin，EPHB2，SMAD3，RalGDS，PMS1(错配修复基因)的影响. 图 6 中，SMAD4、TGFβ-R Ⅱ和微卫星编码基因 MSH3 之间有 2 种逻辑关系类型：类型 8 和类型 5，而且 2 种逻辑类型的支持度一样，类型 5 指的是 SMAD4 表达当且仅当 TGFβ-R Ⅱ不表达且 MSH3 表达，类型 8 指的是 SMAD4 表达当且仅当 TGFβ-R Ⅱ、MSH3 不同时表达，类型 8 包含了类型 5 的情况，根据 KEGG 结肠癌信号通路 05210 分析[16]，这个逻辑关系与我们分析的结果一致，TGFβ 是一种作用很强的生长抑制因子，它通过 TGFβ Ⅰ型受体结合发挥细胞信号传递作用，该通路由转化生长因子 TGFβ 信号引发，通过 SMAD 传导. 细胞周期素 D(Cyclin-D)上调表达当且仅当 Fos 和 SMAD3 同时表达或者不表达，而 Cyclin-D 和 SMAD2 基因同时表达或者不表达导致错配修复基因 MSH2 表达上调，这是一个带反馈的逻辑关系. 一阶逻辑关系如图 5 所示，主要有以尿激酶型纤溶酶原激活物(u-PA)为中心的 u-PA 和 LEF、BAD、EGR1、Bcl2、MSH6、TGFβ 基因之间的逻辑关系，这些基因的表达都可能会使 u-PA 表达上调，除此之外，u-PA 和 EHPB2、Axin、EGF 和 MDM2 之间存在二阶逻辑关系. C-erbB2 基因表达和 MCC、CD44、Cox-2、DCC、Survivin 基因的表达有关系.

3　讨　论

一个细胞内的 2 个蛋白质或者基因，其中一个的存在缺失依赖于另一个的存在缺失，它们之间究竟关系如何？这种简单的存在与缺失的二元关系模式不能充分描述细胞内各种通路之间分支、并行、交叉等形成细胞网络的复杂程度，比如，C 蛋白表达当且仅当蛋白 A 和 B 表达、C 蛋白表达当且仅当 A 或者 B 表达等. 由于细胞内分子调控网络的复杂性，引发人们去思考具有多样性的蛋白质是否受到更高级的逻辑关系的控制，使人们探求用更高阶的复杂的逻辑关系去分析细胞通路和细胞网络. 当然，这种复杂的逻辑关系通过生物实验来发现几乎是不可行的，然而，这种客观存在的蛋白质或者基因之间的逻辑关系对于研究和揭示细胞复杂的基因网络和分子之间的相互作用是至关重要的，尤其是在研究肿瘤发生的分子机制方面. 本文所建立的基因逻辑网络关系模型从一定程度上反映了肿瘤基因之间复杂的相互作用关系，由于肿瘤的发生不是单个基因突变、抑制或者激活导致的，再加上基因表达受时间、空间方位和条件等的影响，使得肿瘤基因之间的关系类型错综复杂，所以简单的线性关系模型是不足以解释肿瘤细胞癌变的复杂机制的. 正如前文所述逻辑网络模型在基因复杂关系模型分析中显示了优势.

本文分析的 10 种逻辑关系和 2004 年 Bowers 等发现的蛋白质逻辑关系略有不同，他发现逻辑类型 1，3，5，7 在蛋白质数据中发生的频率较高，2005 年当 Bowers 等把 LAPP 方法用于神经胶质瘤基因表达谱数据中时发现，逻辑类型 5，1，4，6 出现的频率比较高，而本文在综合考虑逻辑关系支持度的情况下，发现类型 3，5，6，7，8 逻辑关系类型出现的频率比较高. 类型 2 和 1 出现的频率最低. 虽然类型 4 仅次于类型 3，但是由于其样本支持度太低，所以本文认为它不足以说明基因之间的二阶逻辑关系. 通过本文构建的癌基因表达逻辑网络模型可以看出，每个基因并不是一成不变地参与一种生物通路和承担一种角色，而是参与了各种不同的生物通路，影响了多个基因的表达，从而形成一个错综复杂的网络. 比如 MMP9 和多个通路成员基因有逻辑依赖关系，k-ras、Survivin、EPHB2、C-erbB2 等也参与了多个生物通路等等，由此可见，这些癌基因的激活和抑癌基因的失活之间存在复杂的相互作用关系，而本文构建的这种基因逻辑网络模型仅仅为分子生物学家提供了一个可供参考的生物学假说，结肠癌基因和抑癌基因之间的关系是否真的如此，还需要具体的实验验证.

癌症基因组研究的最终目的是描述细胞的分子网络完成生物功能和疾病过程的相互作用，癌症本身是一个复杂系统，网络模型是对复杂系统的高度抽象，本文提出的模型中除了一阶二阶逻辑关系外，还可以再考虑三阶、四阶等更高阶的逻辑关系，因为研究这些更高阶逻辑关系对理解癌症的分

子机制和基因之间的复杂相互关系是非常有价值的.

参 考 文 献

1 Fearon E R, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell, 1990, **61** (5): 759～767

2 Smyrk T, Lynch H T. Colorectal cancer: molecular and cellular abnomalities. In: Bertino J R. Encyclopedia of Cancer, 1. San Deigo: Academic Press, 1997. 451～463

3 Hartwell L H, Hopfield J J, Leibler S, *et al*. From molecular to modular cell biology. Nature, 1999, **402** (6761 Suppl): C47～52

4 Van't Veer L J, Dai H, Van de Vijver M J, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. Nature, 2002, **415** (6871): 530～536

5 Freije W A, Castro-Vargas, Fang A, *et al*. Gene expression profiling of gliomas strongly predicts survival. Cancer Research, 2005, **64** (10): 6503～6510

6 Lu P, Nakorchevskiy A, Marcotte E M. Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations. Proc Natl Acad Sci USA, 2003, **100** (18): 10370～10375

7 Bowers P, Cokus S, Eisenberg D, *et al*. Use of logic relationship to decipher protein network organization. Science, 2004, **306**(5705): 2246～ 2249 [DOI: 10.1126/science.1103330] (in Reports)

8 Zhang X, Kim S C, Wang T, *et al*. Joint learning of logic relationships for studying protein function using phylogenetic profiles and the Rosetta Stone method. IEEE Transaction on Signal Processiong. 2005, **54** (6): 2427～2435

9 Laiho P, Kokko A, Vanharanta S, *et al*. Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gds&cmd=search&term=GSE4045, 2006-07-03

10 Li C, Wong W H. DNA-Chip Analyzer (dChip). In: Parmigiani G, Garrett E S, Irizarry R, ZegerIn S L, eds. The Analysis of Gene Expression Data: Methods and Software. New York: Springer, 2003. 120～141

11 Heerdt B G, Houston M A, Augenlicht L H. Growth properties of colonic tumor cells are a function of the intrinsic mitochondrial membrane potential. Cancer Res, 2006, **66** (3): 1591～1596

12 Egeblad M, Werb Z. New functions for the matrix metalloproteinases in cancer progression. Nature Reviews, 2002, **2** (3):161～174

13 Zeng L, Fagotto F, Zhang T, *et al*. The mouse Fused locus encodes Axin, an inhibit or of the Wnt signaling pathway that regulates embryonic axis formation. Cell, 1997, **90** (1): 181～192

14 Liu W, Ahmad S A, Jung Y D, *et al*. Coexpression of ephrin-Bs and their receptors in colon carcinoma. Cancer, 2002, **94** (4): 934～939

15 Budihardjo, Oliver H, Lutter M, *et al*. Biochemical pathways of caspase activation during poptosis. Annu Rev Cell Dev Biol, 1999, **15** (1): 269～290

16 Grady WM. Genomic instability and colon cancer. Cancer Metastasis Rev, http://www.genome.jp/kegg-bin/path_ref_list?pathway=05210. 2004, **23**:11~27

Modeling Colon Cancer Gene Logic Network With mRNA Microarray Data*

RUAN Xiao-Gang[1], WANG Jin-Lian[1)**], LI Hui[2)]

([1)]*College of Electronic Information and Control Engineering, Beijing University of Technology, Beijing* 100022, China ;

[2)]*College of Computer Science and Technology, Bejing University of Technology, Beijing* 100022, China)

Abstract Analysis of cellular pathways and networks in terms of logic relations is important to decipher the networks of molecular interactions that underlie cellular function. A computational approach for identifying lower and higher order gene logic associations was presented on the base of graph coloring theory and applied it to the colon cancer mRNA microarray data. Then the logic relationships of 51 oncogenes and cancer suppressor genes are analyzed and the logic association network of them was constructed. The signal pathway of TGFβ from the network model was found and verified by the colon cancer pathway of KEGG. The model reveals many higher order logic relationships of cancer genes. These relationships illustrate the complexities that arise in cancer cellular networks because of interacting pathways. The results show that this method is feasible and is expected to give a reference to the medical molecular biologist.

Key words colon cancer, mRNA microarray, gene network，graph coloring

*This work was supported by a grant from The National Natural Science Foundation of China (60234020).

**Corresponding author . Tel: 13811808114, E-mail: wijinlian1999@gmail.com

Received: January 23, 2007 Accepted: March 02, 2007