PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2008, 35(11): 1239~1246
www.pibb.ac.cn

基因逻辑网络研究进展

王青芸[1)] 易 慧[1)] 刘来福[1)] 孟大志[2)*]
([1)]北京师范大学数学科学学院，北京 100875; [2)]北京工业大学应用数理学院，北京 100022)

摘要 海量生物数据的涌现，使得通过数据分析和理论方法探索生物机理成为理论生物学研究的重要途径．特别是对于基因的复杂的功能系统，建立基因网络这种理论方法的意义更为突出．Bowers 在蛋白质相互作用的分析中引入了高阶逻辑关系，从而建立了系统发生谱数据的逻辑分析(LAPP)的系统方法．LAPP 和通常建立模型的方法不同，它给出了一个从复杂网络的元素(或部件)的表达数据出发，通过逻辑分析，找到元素之间逻辑关联性的建模方法．这种方法能够从蛋白质表达谱数据出发，利用信息熵的算法发现两种蛋白质对一种蛋白质的联合作用，对于发现蛋白质之间新的作用机理有重要意义．由于涉及功能的基因组通常是一个大的群体构成的系统，因此 LAPP 方法也是一个生成复杂的基因逻辑网络的方法．基因逻辑网络的建立，方便实现通过逻辑调控进行基因调控的目的．这种方法可以应用在很多方面，如物种进化、肿瘤诊疗等等．系统阐述并分析了 LAPP 方法，并指出其在方法和应用方面的新进展以及评述．

关键词 系统发生谱的逻辑分析(LAPP)，逻辑相关性，复杂网络，基因网络
学科分类号 TP18

随着基因芯片等高通量数据采集技术的发展，大量生物数据涌现，为系统地从基因水平研究生物学提供了丰富的素材，也发展出许多有效的数据分析方法，例如，在肿瘤研究中对蛋白质的系统发生谱数据分析的各种方法[1~3]．但是生物学的研究仍然明显地面对这样的局面：一方面海量的实验数据在日益增长，另一方面却只有少量的蛋白质(或基因)功能机理为人类所知，导致为功能基因组建立机理模型变得困难．这意味着传统的依赖生物实验的方法已经难以高效率地揭示复杂的生物现象和规律．2004 年 Bowers 等[4]系统地建立了对蛋白质系统发生谱数据的逻辑分析方法(logic analysis of phylogenetic profiles，LAPP)，正在改变这样的局面．由于对一个复杂系统来说，清楚地了解所有元素(或部件)之间的关系是十分困难的(例如基因之间作用机理)，直接进行机理性建模并不是普遍可行的．在 Bowers 的文章中给出的 LAPP 与通常的机理建模方法不同，它是一种从元素的表达数据出发，通过一系列有效的逻辑分析，发现元素之间逻辑关联性的方法．这就使得通过数据分析建立逻辑性的机理模型成为可能．

在生物学中，经常利用大规模网络来研究复杂的生物学现象．随着研究的不断深入，研究者已经不满足于揭示元素(蛋白质、基因)之间简单的相关性，还希望知道两个(或多个)元素之间具有“怎样”的相关性，或者说元素之间“怎样”通过相互影响而联系在一起．LAPP 方法可以揭示这些元素之间更深入的关系，即逻辑关系，而 Bowers 给出的方法能够批量地发现这类逻辑关系．这种方法可以帮助生物医学研究者了解一些未知蛋白或基因之间的可能的逻辑功能[5]，从而通过实验发现它们之间新的生物学机理．更有意义的是，用 LAPP 方法揭示复杂网络中元素(部件)之间的因果关系或制约关系之后，启发人们对网络的调控发展出更多更有效的方法．在研究和揭示生物学现象的复杂性和规律性中，这种系统的方法显然有相当的价值．

* 通讯联系人.
Tel: 010-82255569, E-mail: dzhmeng07@yahoo.com.cn
收稿日期：2008-03-09，接受日期：2008-05-06

由于 LAPP 方法具有上述意义，自 Bowers 开创性工作后的几年来，这个方法在理论、算法和应用上都有了新的发展．发现的逻辑关系很自然地从一阶、二阶发展到三阶和更高阶的逻辑[6]．并且一些作者还提出新型逻辑：一个原因多个结果的逻辑类型和与取值有关的逻辑类型[7]．除了 Bowers 使用的 U 值算法可以发现元素之间的逻辑关系之外，还有贝叶斯方法等概率算法[6, 8]和数据挖掘算法，用于发现数据中的逻辑关系．在应用方面，逻辑分析不仅应用在系统发生谱数据上(离散数据)[4]去分析蛋白质通路，揭示未知蛋白的功能[4, 6, 9~11]，还应用在肿瘤的基因表达数据上[5]分析致病机理，给临床治疗提出建议，甚至结合其他形式的数据一起分析[6]各种类蛋白质之间的关系．同时，LAPP 方法也是生成复杂逻辑网络的一个重要方法[12]，可以用于其他领域的复杂系统的研究[13~15]．当然，LAPP 方法也有其缺陷和困难之处[16~18]，在应用方面也远没有推广．因此，LAPP 方法无论在理论、算法和应用上都具有广阔的发展空间．

本文在以下各节中系统地阐述 LAPP 建立复杂逻辑网络的方法以及在各方面的发展，同时也指出在该领域研究中的一些有前景的问题．

1 高阶逻辑和网络

1.1 系统、结构与逻辑网络

复杂系统是当今科学与工程研究中最常见的对象，例如生物学中的各种系统(血液、内分泌、解剖、功能基因等)．在系统中，元素之间和子集之间具有一定关联性．这些关系及其全体的组合成为系统的结构．用连线、箭头或者加权的边表示，其整体构成网络图(无向、有向、加权等等)，因此，复杂系统对应着复杂网络．然而，系统的元素或部件之间的关系是一个泛化的概念，普遍的理解是指统计相关计算所界定的关联关系．但是事实上，这种关联关系中存在着更深层的含义，也就是存在可以细分的逻辑相关性的关联．例如在生物系统的因素之间的激励、制约或限制等关系就具有类似因果含义的逻辑关联．我们可以用符号“¬”表示相关，用符号“$\Rightarrow$”表示逻辑关系，例如“$A\Rightarrow B$”表示 A 影响 B 或 A 的发生导致 B 的发生．元素之间不同关联性的这种表示，形成了一般关联网络(无向网络)或逻辑网络(有向网络)这样两类典型的网络结构．本文关注逻辑网络，特别是含高阶逻辑的基因逻辑网络，并且特别关注这种逻辑关系是怎样通过对基因的数据分析发现的．当然，基于统计的计算方法并非发现确定性的逻辑，而是统计意义下的逻辑，因此我们称之为不确定性逻辑．以下讨论的逻辑网络，就是这种不确定性逻辑整合形成的网络结构．

1.2 逻辑的存在性和不确定性系数

对于网络节点的表达谱数据，通过离散化处理可以变成 0、1 数据．如果将这些数据理解成布尔逻辑值的时候，即 0 表示节点不表达(不发生)而 1 表示节点表达(发生)，则两个节点之间形成了布尔逻辑．本文将两个节点之间的逻辑关系称为一阶逻辑，而将三个、四个节点之间的逻辑称为二阶、三阶逻辑，或统称为高阶逻辑．

元素 A 和 B 之间的一阶逻辑有两类：$A\Rightarrow B$ 即 A 表达导致 B 表达(类似有$\neg A\Rightarrow\neg B$)和 $A\Rightarrow\neg B$ 即 A 表达导致 B 不表达(类似有$\neg A\Rightarrow B$)“¬”表示布尔逻辑的补算子，通常也称为同步和异步逻辑关系．元素 A、B、C 之间的二阶逻辑有 8 类，除了用布尔表达式表示之外，还可用文氏图表示，具体见图 1.

这种二阶逻辑在生物学中是普遍存在的，在 Bowers 的文章中给出一些明显的例子．图 1 中逻辑类型 3，如果脱氢奎尼酸脱水酶(GOC0710)或脱氢奎尼酸脱水酶Ⅱ(GOC0710)中有一种出现，则莽草酸脱氢酶在一个基因组中出现[4]．这说明在芳香氨基酸的合成中，这两类脱氢奎尼酸酶蛋白质家族通过替代或等价的酶合成方式能取得显著优势．类型 5，如果一种 DNA 修复时用到的腺苷三磷酸酶 ATPase (COG0419)出现并且另一种 DNA 错配修复酶 ATPase(COG0249，MutS 家族)不出现，则原核 DNA 粘合蛋白(COG1581)在一个基因组中出现[4]．结果表明，在细菌中 COG1581 可能具有原核 DNA 错配修复功能，与 COG0419 的作用互补．

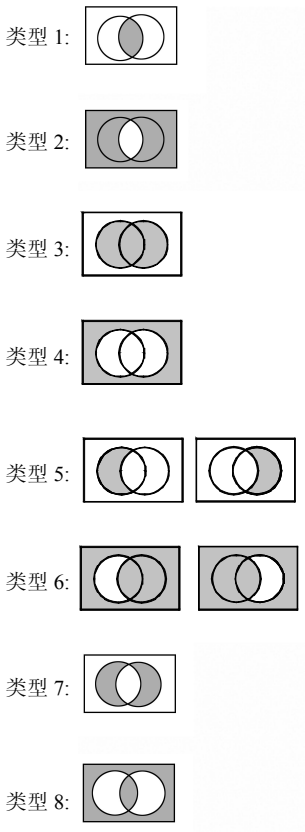
如何通过网络节点的样本数据计算发现节点之间的逻辑关系，是 LAPP 方法的基础．Bowers 使用不确定性系数 U 来界定节点 A、B 之间逻辑关系的存在性：

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

其中 H 表示节点的信息熵．$U(X|Y)$表示当 A 确定的时候对 B 确定性的影响，U 值的大小表示了节点 A、B 之间不确定性的逻辑关系“$A\Rightarrow B$”在统计上存在性．

二阶逻辑关系类似可得，只需计算 $U(c|f(a, b))$ 即可得到 a、b、c 之间某种二阶逻辑关系存在的可

能性，这里函数 f 指二阶逻辑类型中的某一类．要注意的是，一般情况下，只有在不存在一阶逻辑的时候才会去找二阶逻辑，所以常常要同时计算 $U(c|a)$和 $U(c|b)$，只有在 $U(c|a)<< U(c|f(a, b))$ 和 $U(c|a) << U(c|f(a, b))$同时成立时，才会考虑二阶逻辑的存在性．Bowers 的 LAPP 方法的重要性是给出了 3 个因素(或节点)之间的二阶逻辑，并给出计算方法去确定细分的逻辑类型．Bowers 使用的是蛋白质的系统发生谱数据，他通过对 4 873 个蛋白质家族在 67 个物种中表达谱的数据分析以及利用文氏图的表示，确定 2 个因素对 1 个因素发生影响的逻辑关联性的类型共有 8 类[4]，并给出每一类的布尔表达式以及这种二阶逻辑的含义(图 1)，并且利用 U 值计算来确定 8 类逻辑在 3 个因素之间的存在性．



类型 1:	$C=A \wedge B$	C 在基因组中出现，当且仅当 A 和 B 同时出现.
类型 2:	$C= \neg(A \wedge B)$	如果 A 不出现或 B 不出现，则 C 出现.
类型 3:	$C=A \vee B$	如果 A 出现或 B 出现，则 C 出现.
类型 4:	$C= \neg(A \vee B)$	如果 A 不出现并且 B 不出现，则 C 出现.
类型 5:	$C=A \wedge \neg B, C= \neg A \wedge B$	如果 A 出现并且 B 不出现，则 C 出现; 如果 A 不出现并且 B 出现，则 C 出现.
类型 6:	$C=A \vee \neg B, C= \neg A \vee B$	如果 A 不出现或 B 出现，则 C 出现; 如果 A 出现或 B 不出现，则 C 出现.
类型 7:	$C= \neg(A \leftrightarrow B)$	如果 A 和 B 中一个出现，但 A 和 B 不同时出现，则 C 出现.
类型 8:	$C=A \leftrightarrow B$	如果 A 和 B 同时出现，或 A 和 B 同时不出现，则 C 出现.

Fig. 1　Types of two order logic relationships[4]

图 1　二阶逻辑类型图示[4]

图左半部是 8 种逻辑类型的文氏图和布尔表达式，右半部是 8 种逻辑类型的文字说明.

例如在文献[4]中给出的 5 个蛋白质 A，B，C，D，E 的数据向量(分量是 0 或 1，表示表达与不表达)．其中 A 的数据向量里 1 出现的概率是 $p(x=1)$= 0.641 8，0 出现的概率是 $p(x=0)$ = 0.358 2，B 的数据向量里 1 出现的概率是 $p(x=1)$= 0.179 1，0 出现的概率是 $p(x=0)$=0.820 9，A、B 联合发生时，(00), (01), (10), (11)出现的概率分别是 0.197 0、0.164 2、0.626 9、0.014 9，分别计算 A、B 的熵、联合熵和 $U(B|A)$, 得到：$H(A)$=0.652 4, $H(B)$=0.470 0, $H(A, B)$= 0.970 3，则：

$U(B|A)=[H(A)+H(B)-H(A, B)]/H(B)=(0.652\ 4+0.470\ 0-0.970\ 3)/0.470\ 0=0.323\ 6$

同理分别计算出 A，B，C，D，E 之间的所有两两 U 值(表 1 左半部分)和 5 个蛋白质之间三元组的 U 值(即 $U(C|f(A, B))$的值，(见表 1 右半部分)．从表 1 中的 U 值可以看出 5 个蛋白质 A，B，C，D，E 之间的通路最有可能是 A-C-D-E 或者 B-C-D-E[4].

当然，计算逻辑关系的方法不止是上面的 U 值算法，可以应用逻辑分析的数据也不仅仅是系统发生谱数据，这些将在下一节提到.

Table 1　The triplet and pairwise uncertainty coefficients, *U*, with ternary relationships, *A* and *B* with *C, D* and *E*[4]

表 1　*A*，*B*，*C*，*D*，*E* 之间一阶逻辑的 *U* 值和部分二阶逻辑 *U* 值[4]

一阶逻辑 $U(x\|y)$					
A		0.2331	0.2983	0.2331	0.0964
B	0.3236		0.0998	0.0838	0.0552
C	0.3797	0.0915		0.7498	0.3938
D	0.3236	0.0838	0.8177		0.6029
E	0.1055	0.0435	0.3385	0.4753	
	A	*B*	*C*	*D*	*E*

二阶逻辑 $U(x\|f(y,z))$		
类型 7	$C=\neg(A\leftrightarrow B)$	0.8550
类型 7	$D=\neg(A\leftrightarrow B)$	0.6647
类型 7	$E=\neg(A\leftrightarrow B)$	0.2933

1.3　逻辑关系的对比验证

通过计算 *U* 值发现的某些元素之间存在逻辑关系，是否具有数据特异性而不是可以随机产生的? 这就需要检验，而检验逻辑关系特异性的方法有很多种，主要都是把实际观测数据和随机数据所进行分析的结果比较，看是否有显著性差异. 研究人员常用来比较的参数有[4, 5]：a. 随机数据的 *U* 值分布和实际数据 *U* 值分布的比较；b. 随机数据和实际数据中高 *U* 值的逻辑数目差异；c. 各种逻辑类型在随机数据和实际数据中所占比例的不同等等. Bowers 的文章中除了用到上面的检验方法，还通过比较观测到的高 *U* 值的二阶逻辑的 3 个蛋白质(基因)和随机数据产生的高 *U* 值的二阶逻辑的 3 个蛋白质(基因)在各种功能型中分布的不同来说明逻辑关系的特异性.

2　高阶逻辑方法的发展

自 2004 年 Bowers 的文章发表以来，关于逻辑关系，特别是分析高阶逻辑的方法已经有了进一步的发展.

2.1　三阶逻辑及生物实例

逻辑关系研究方法的一个非常自然的发展就是把二阶逻辑扩展到了三阶甚至更高阶，三阶逻辑的鉴别也可以根据 **1.2** 节的 *U* 值算法得到. Zhang 等[6]在他的文章里给出了三阶逻辑的分类，三阶逻辑的 *U* 值计算公式：

$$U(d|f(a,b,c))=[H(d)+H(f(a,b,c))-H(c,f(a,b,c))]/H(d) \quad (2)$$

三阶逻辑类型一共有 68 类，它包含了 216 个有效函数，*f* 是其中一个有效函数，是 68 个类型之一. 值得注意的是，同一种类型的逻辑可能有多个不同的有效函数(不同变量)，二阶逻辑的类型是 8 个，但有效函数是 10 个(图 1)，三阶逻辑的各个类型都有多个不同的有效函数.

三阶逻辑比二阶逻辑复杂，在生物现象中，确有一些三阶逻辑关系的例子. 例如一个类型 A 的例子[6]，当蛋白质 COG1121 出现，或 COG3839 不出现，或 COG3842 不出现时，COG0517 才会出现. 在这 4 个蛋白质中，COG1121 是 ABC 型的 Mn/Zn 转运子 ATP 绑定蛋白，COG3839 是 ABC 型的糖转运子 ATP 绑定蛋白，COG3842 是 ABC 型亚精胺 / 丁二胺运输系统中的一员，COG0517 被归类为假设或想像的蛋白质. 这意味着 COG0517 可能在 ABC 型的运输系统中作为 ATP 的绑定蛋白产生作用. 用布尔逻辑表示 COG0517= COG1121 ∨ (¬COG3839) ∨ (¬COG3842). 三阶逻辑关系的存在同样有一些约束条件.

2.2　贝叶斯方法分析

1.2 节中得到的 *U* 值方法并不是唯一鉴别逻辑关系的方法，而且逻辑关系也不是只能用在系统发生谱数据上. Zhang 的文章介绍了一种贝叶斯模型结构方法[6]，它把两种数据类型：蛋白质系统发生谱(Dpp)和 Rosetta Stone 数据(Drs)[19~21]联合应用，得到鉴别蛋白质的逻辑关系 *U* 值的另一种方法.

用 $S^i_{j,k}$ 表示事件蛋白质 *i* 和蛋白质 *j* 和 *k* 相关，从给出的 Dpp 和 Drs 数据得到 $S^i_{j,k}$ 的对数后验概率. 选择适当的逻辑函数 *f*，使 *f* 极小化预测误差，极大化对数似然值. 当每个 $S^i_{j,k}$ 的最佳函数 *f* 找到，那么定义 $P(D_{pp}|S^i_{j,k})$ 的似然：$P(D_{pp}|S^i_{j,k})=U(i|f(j,k))$，其中 $U(x|y)$ 即是所求的不确定性系数，*U* 的阈值取法同 Bowers 的文章. 类似可得到高阶逻辑关系的不确定性系数.

2.3　概率方法和数据挖掘方法

另一种度量系统发生谱的逻辑关系则是概率度

量方法[8]. 这个方法主要是通过比较某种关系的观测频率和随机情况下出现这种关系的概率的差异来判断数据谱与谱之间是否存在着某些关系.

用概率的方法度量两个系统发生谱之间的关系：假设有 N 个基因组，蛋白质 X 在 x 个基因组中出现，蛋白质 Y 在 y 个基因组中出现，X 和 Y 共同出现在 N_{xy} 个基因组中. $P(N_{xy} \mid N, x, y)$是随机观测到 N_{xy} 个的共发生的概率，对应一个概率分布. 蛋白质三元组的二阶关系的概率度量类似可以得到.

因为 $P(N_{xy} \mid N, x, y)$计算的是随机出现某些共同发生情况的概率. 如果观测值的 $P(N_{xy} \mid N, x, y)$远远大于随机概率，就可以认为 X、Y 之间存在着某些关系，然后再研究究竟是什么关系(逻辑关系或其他). 这个方法也可以应用到别的数据上. 可以证明，**1.2** 节的 U 值算法是概率方法在 $N \to \infty$ 的一个特殊形式[8]. 但是，这个证明只是理论上的，在实际中哪种方法对于确定逻辑相关性的“可信度”更大并没有明确的结论.

数据挖掘中的关联规则发现同样可以度量系统发生谱的逻辑关系[22, 23]，主要用的是置信度——支持度的基本框架与统计相关性的改进方法.

在两个属性集 A，B 之间若存在关联规则，只要对于给定的两个阈值 S 与 C 使得：

$$\text{Support}(A \Rightarrow B)=P(A \wedge B) > S \quad (3)$$

$$\text{Confidence}(A \Rightarrow B)=P(B \mid A) > C \quad (4)$$

也称两个属性集之间存在强关联规则. 这里阈值 S 与 C 的确定是与问题的目标有关的. 这种方法计算简单，但是要确定两组全都超过阈值的属性集中哪一个规则更“可信”往往无法确定. 另一方面，置信度——支持度方法难于区分正、负相关性. 置信度——支持度方法对确定性很强的逻辑有很好的排序结果，在这方面它可能强于 U 值算法. 但是对于不确定性比较强的数据，阈值 S 与 C 的权值选择就比较困难，这种情况下置信度——支持度方法的区分度会比基于熵的 U 值算法要差一些. 实际应用中要根据具体情况进行选择.

但是无论是基于熵的算法还是概率的算法，事实上各有利弊. 这些算法都可能在关联规则的确定上有不同的不确定性，也就是说是逻辑关系的发现即可能有“不适当的选择”，也可能有“遗漏的选择”，也就是说规则的确定性计算都有不同的误差.

2.4　与取值有关的逻辑关系

逻辑关系的一个有趣的发展是与取值有关的逻辑，最初的例子来源于一个因素多个结果的逻辑关系[7]. 例如，根据控制基因 C 的表达量，另外 2 个基因 A、B 有不同的相互作用关系. 文献[7]给出这样一个例子. 在不考虑基因 C 取值的时候，基因 A 和 B 本无明显的关系.基因 C 的表达量可以取不同的值. 基因 A 和基因 B 随着基因 C 的表达量的不同，呈现出不同的相互关系——正相关和互斥. 在 Zhang 的文章[7]中，基因 A、B、C 之间的这种关系称为三基因交互作用.

目前尚未建立这种逻辑完整系统的理论和计算的方法. Zhang 给出了一个简单的鉴别方法：A, B 和 C 三个基因，假设基因 C 是控制基因，基于基因 C 的表达水平的高低，把样本分成低表达组($n1$ 个样本)和高表达组($n2$ 个样本). 分别计算这两个样本组中基因 A 和基因 B 的对数表达值的相关系数 $r1$, $r2$. 用 t 检验来揭示两组是否有显著差异. 如有显著差异，则基因 A、B、C 之间可能存在与取值有关的逻辑关系. 图 2 的现象在其他的文章[24]中也有类似事例，说明与取值有关的逻辑关系的理论和算法亟待发展和完善.

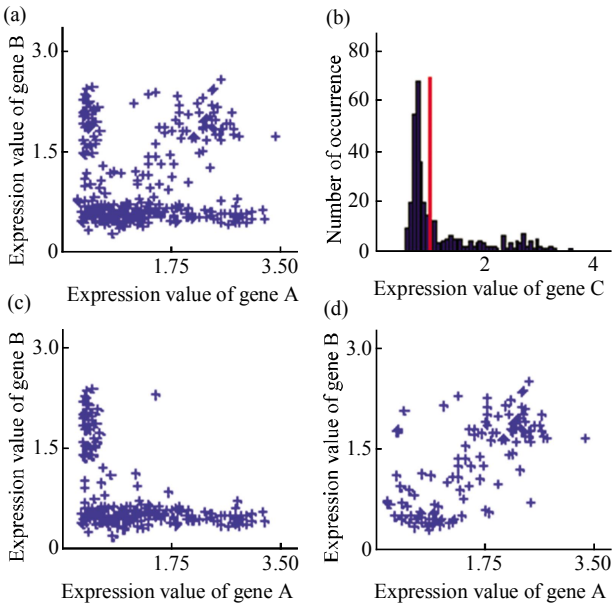


Fig. 2　An example of a three-way gene interactions[7]

图 2　一个三基因交互作用的例子[7]

三个基因是 CFTR(基因 C)，MYB(基因 A)和 USH1C(基因 B). 这一组的相关值是 11.6. (a) MYB 和 USH1C 的对数表达值，没有明显相关(r=0.09). (b) CFTR 的对数表达值分布出现双峰. 把样本按照 CFTR 的表达量分为两类，MYB 和 USH1C 出现了关系. (c) 当 CFTR 的表达量低时(ln(CFTR) < 0.98)，MYB 和 USH1C 的对数表达值负相关(r=–0.434)，出现 L 型关系. (d) 当 CFTR 的表达量高时(ln(CFTR) > 0.98), MYB 和 USH1C 的对数表达值正相关(r=0.7) .

任何计算方法都有其局限性．Shoemaker 等[16]提到系统发生谱法的一些缺陷：大计算量，依赖高信息谱和遗传距离远的物种的同源检测，这是LAPP 算法应用中的一些问题．此外，还有一些人也提到了 LAPP 算法在应用上要注意的问题，即原核生物和真核生物的蛋白质不能放在一起比较[17]．Uetz 等[18]则提到通过计算方法找到的蛋白质之间的相互作用可能是假阳性的，因为一个基因可能编码多个蛋白质，并不知道是哪个蛋白质之间在发生作用．因此，Lu 等[25]给出了一个半监督学习的算法来验证我们找到的关系的可靠性．

3 含高阶逻辑网络的应用

利用 **1.2**、**2.2** 与 **2.3** 节鉴别逻辑的方法，可以得到不同元素之间的显著的逻辑关系．这些逻辑关系进一步可以生成网络——含有高阶逻辑的网络．这种网络作为揭示复杂系统结构的模型，在各方面都有着广泛的应用．

Bowers 的方法应用在物种上，它可以揭示一些未知的蛋白质通路、网络或者未知蛋白质的功能[4]．同样的数据(67 个物种的 4 873 个 COG 系统发生谱)，Robbins 等[26]则分析其中的一些特殊谱(有重复的谱)，来说明不确定性预测(U 值)对蛋白质网络生成的意义．发现重复谱更具有相似性，距离更近，保守性可能越强，重复谱更有可能发现相互关系．重复谱之间更容易产生二阶关系或者网络，它们生成的二阶关系或网络也可能更保守．

逻辑关系应用在肿瘤样本的基因表达数据上则可能发现一些致病机制，这对临床治疗有很大的意义．Bowers 分析了神经胶质瘤 85 个样本的基因芯片，得到一些和致病紧密相关的基因[5]．Bowers 给出的一个例子是，GLUT10 蛋白(A)和 GFMG、NCF2、LAMA5(B)的逻辑关系是 AND；GLUT10 蛋白(A)和 TRHDE、CDH12、PTPRT、CDK5R2、SPTBN2(B)的逻辑关系是 AND!．这两种逻辑关系产生的结果都是神经胶质瘤表型之一的 HC2B．那么，如果可以控制 GLUT10 蛋白的表达，对治疗或者控制神经胶质瘤可能起到一定的作用(图 3)．

逻辑关系的方法还有其他方面的应用：它可以鉴别蛋白质之间的功能联系，这些联系可以用来推断无特征蛋白可能具有的功能[9] (在很多情况下，功能相关的蛋白质(基因)保守性较强[18, 27])；分析数据之间的(高阶)逻辑关系的方法可以帮助发现类蛋白质之间相互作用[10]；逻辑关系的方法也是研究相似性的新方法[11]；通过研究高通量数据的这种新的方法[28]，帮助生成基因网络[11]，等等．

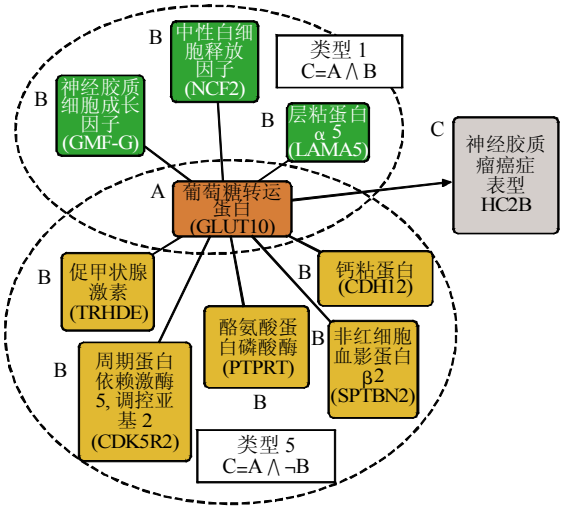


Fig. 3 Proteins logically related to the presence or absence of the glucose transport protein GLUT10[5]

图 3 与 GLUT10 蛋白通过逻辑关系产生癌症表型 HC2B 的蛋白质[5]

4 高阶逻辑网络的理论方法及评述

LAPP 方法的核心是通过对系统中元素的表达数据分析发现元素之间的逻辑关系．目前使用的数据谱是离散的二值 0 和 1，从而发现的是比较简单的布尔逻辑．对于元素(等同地是节点)布尔逻辑是研究因素只有两个状态下的逻辑：发生或没有发生．事实上，还有元素表现出三种状态：上升、保持与下降，可以对应到数据谱是离散的三值 1、0、–1．一般而言，元素的发生谱数据甚至可能是多值以至于连续的，那么如何建立判别多态数据的逻辑关联性的算法，变得很有研究的必要．和目前多种不确定性参数的计算并没有明确的优劣类似，面对上述数据类型的元素，如何确定信息论的方法和概率统计方法之间的优劣，仍然是一个未决的问题．值得注意的是，无论是一阶、二阶、三阶以及 n 阶逻辑有多少个不同的类型，都没有文献给出严格的数学证明，Bowers 的文氏图只是一种直观的表示．

针对目前多个因素影响一个因素的逻辑分析研究，事实上还存在一个因素影响多个因素的逻辑类型，而且更一般地说，多个因素与多个因素之间存在逻辑影响的关系也是存在的．因此，如何将目前的 LAPP 方法推广到“多对多”情况也是一个有趣

的问题．当然一些作者提示，逻辑关系的存在与分类还和因素的取值范围有关，也许这种情况并非罕见，因此进一步的研究不应忽视这个问题．

LAPP 方法是系统发生谱方法的发展[29]，与传统的系统发生谱方法相比，它有更为广泛的应用，如在药物学上的应用[13, 14]，Rice 等[15]还提到，逻辑关系分析的方法可以在古生物学上应用(研究脚印)．系统发生谱方法可以应用在人类基因组学和研究进化过程上(研究染色体重复块之间的关系和研究突变体)[30, 31]，LAPP 方法也可以在这些方面进行推广．LAPP 方法的本质并非只限于生物问题，这个方法应该可以应用在更广泛的领域中，例如经济领域，股票或证券方面也许有不错的应用．

最后，我们特别关注的研究目标是复杂网络．目前人们研究的大量是无向网和理解为一阶逻辑的有向网络．用 LAPP 方法生成的网络是含有高阶逻辑的，这类逻辑网络更全面地表达了复杂系统的结构，这种复杂结构又如何建立各种统计量来描述，如度分布、聚类系数等等，显然是很有趣的．很明显，高阶逻辑网络作为复杂系统的模型比普通的相关网络更加准确，它除了揭示系统的结构，还启发研究者发现更有效的方法来调控系统．在未来的研究中，高阶逻辑网络的研究将对复杂系统的研究发挥越来越重要的作用．

参 考 文 献

1 Segal E, Friedman N, Kaminski N, *et al*. From signatures to models: understanding cancer using microarrays. Nat Genet, 2005, **37** (Suppl): S38～S45

2 Bhattacharjee A, Richards W G, Staunton J, *et al*. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA, 2001, **98**(24): 13790～13795

3 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

4 Bowers P M, Cokus S J, Eisenberg D, *et al*. Use of logic relationships to decipher protein network organization. Science, 2004, **306**(5706): 2246～2249

5 Bowers P M, O′Connor B D, Cokus S J, *et al* . Utilizing logical relationships in genomic data to decipher cellular processes. FEBS J, 2005, **272** (20): 5110～5118

6 Zhang X, Kim S, Wang T, *et al*. Joint learning of logic relationships for studying protein function using phylogenetic profiles and the Rosetta Stone method. IEEE Transactions on Signal Processing, 2006, **54**(6): 2427～2435

7 Zhang J X, Ji Y, Zhang L. Extracting three-way gene interactions from microarray data. Bioinformatics (Oxford, England), 2007, **23** (21): 2903～2909

8 Wu J, Mellor J C, DeLisi C. Deciphering protein network organization using phylogenetic profile groups. genome informatics. International Conference on Genome Informatics, 2005, **16** (1): 142～149

9 Anderson M W. Functional genomics annotation: it′s logical!. Scientist, 2005, **19** (5): 33～33

10 Spirin V, Gelfand M S, Mironov A A, *et al*. A metabolic network in the evolutionary context: Multiscale structure and modularity. Proc Natl Acad Sci USA, 2006, **103** (23): 8774～8779

11 Glazko G, Coleman M, Mushegian A. Similarity searches in genome-wide numerical data sets. Biology Direct, 2006, **1** (1): 13 (doi: 10.1186/1745-6150-1-13)

12 Srinivasan B S, Novak A F, Flannick J A, *et al*. Integrated protein interaction networks for 11 microbes. Lecture Notes in Computer Science, 2006, **3909**: 1～14

13 Aksenov S V, Church B, Dhiman A, *et al*. An integrated approach for inference and mechanistic modeling for advancing drug development. FEBS Letters, 2005, **579** (8): 1878～1883

14 Ofran Y, Punta M, Schneider R, *et al*. Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. Drug Discovery Today, 2005, **10** (21): 1475～1482

15 Rice J J, Kershenbaum A, Stolovitzky G. Lasting impressions: Motifs in protein-protein maps may provide footprints of evolutionary events. Proc Natl Acad Sci USA, 2005, **102** (9): 3173～3174

16 Shoemaker B A, Panchenko A R. Deciphering protein-protein interactions. Part Ⅱ. computational methods to predict protein and domain interaction partners. PLoS Computational Biology, 2007, **3** (4): e43

17 Jothi R, Przytycka T M, Aravind L. Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. BMC Bioinformatics, 2007, **8** (1): 173(doi: 10.1186/1471-2105-8-173)

18 Uetz P, Finley R L. From protein networks to biological systems. FEBS Letters, 2005, **579** (8): 1821～1827

19 Enright A J, Iliopoulos L, Kyrpides N C, *et al*. Protein interaction maps for complete genomes based on gene fusion events. Nature, 1999, **402** (6757): 86～90

20 Marcotte E M, Pellegrini M, Ng H L, *et al*. Detecting protein function and protein-protein interactions from genome sequences. Science, 1999, **285** (5428): 751～753

21 Bowers P M, Pellegrini M, Thompson M J, *et al*. Prolinks: a database of protein functional linkages derived from coevolution. Genome Biology, 2004, **5** (5): R35

22 Han J W, Kamber M. Data Mining: Concepts and Techniques. San Franciso: Morgan Kaufmann Publishers, Inc., 2001. 23～37

23 Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large database. In: Buneman P, ed. Proc of the 1993 ACM SIGMOD Conference on Management of Data. New York: ACM, 1993. 207～216

24 Shedden K, Taylor J. Differential correlation detects complex

associations between gene expression and clinical outcomes in lung adenocarcinomas. In: Shoemaker J, ed. Methods of Microarray Data Analysis Ⅳ. New York: Springer, 2004.121～131

25 Lu Y, Tian Q, Liu F, *et al*. Interactive semisupervised learning for microarray analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2007, **4** (2): 190～203

26 Robbins K A, Zhao L. Visualization of logic analysis of phylogenetic profiles (LAPP). In Computational Systems Bioinformatics Conference, 2005. Workshops and Poster Abstracts. IEEE 2005. 275～276

27 Sardiello M, Annunziata I, Roma G, *et al*. Sulfatases and sulfatase modifying factors: an exclusive and promiscuous relationship. Human Molecular Genetics, 2005, **14**(21): 3203～3217

28 Weber A P M, Fischer K. Making the connections -The crucial role of metabolite transporters at the interface between chloroplast and cytosol. FEBS letters, 2007, **581** (12): 2215～2222

29 Ranea J A G, Yeats C, Grant A, *et al*. Predicting protein function with hierarchical phylogenetic profiles: The Gene3D phylo-tuner method applied to eukaryotic Genomes. Plos Computational Biology, 2007, **3** (11): 2366～2378

30 Jiang Z, Tang H, Ventura M, *et al*. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. Nature Genetics, 2007, **39** (11): 1361～1368

31 Masso M, Vaisman Ⅱ. Accurate prediction of enzyme mutant activity based on a multibody statistical potential. Bioinformatics, 2007, **23** (23): 3155～3161

Progress in Gene Logic Networks

WANG Qing-Yun[1)], YI Hui[1)], LIU Lai-Fu[1)], MENG Da-Zhi[2)*]

([1)]*School of Mathematical Sciences,Beijing Normal University, Beijing* 100875, China;

[2)]*College of Applied Sciences, Beijing University of Technology, Beijing* 100022, China)

Abstract As tremendous genomic data avalanches, exploring biological mechanism by data analysis and theory methods has become important for theoretical biology research. This method is significant for the study of complex gene functions and gene networks. Bowers used higher order logic relationships to decipher protein network organization, which is a systemic method called logic analysis of phylogenetic profiles (LAPP). LAPP is a data modeling and different from traditional computational methods. This computational approach identifies logic relationships of the elements (or components) in complex networks through the logic analysis of their expression data. The method can be used to infer functional relationships of two associated proteins to one another. It is important for discovering the new function mechanism of the protein. The clusters of orthologous groups (COGs) involved in a gene network usually are large groups and therefore LAPP is also an approach for complex gene logic networks. After the establishment of the gene logic network, it is convenient for the regulation of gene through the network. The method can used in many fields, such as species evolution, oncologic diagnosis and so on. LAPP was systematically described and analyzed and recent developments in methodologies and applications were highlighted. Some opinions of them were also given.

Key words logic analysis of phylogenetic profiles (LAPP), logic relationships, complex networks, gene networks

*Corresponding author.

Tel: 86-10-82255569, E-mail: dzhmeng07@yahoo.com.cn

Received: March 9, 2008 Accepted: May 6, 2008