用多元逐步回归方法解析转录因子在基因表达调控中的作用

沙 丹 徐海燕 牛 钢[2] 费鹤良 岳荣先 费 俭[1]*

(上海师范大学数理信息学院应用数学系, 上海 200234; [1] 同济大学生命科学与技术学院, 上海 200092; [2] 中国科学院上海生命科学研究院生物化学与细胞生物学研究所, 上海 200031)

摘要 反式作用因子之间的合作关系与顺式调控逻辑对完成基因的正确转录都起到了决定性的作用。本文提出了一种多元逐步回归方法, 对反式作用因子之间的协作关系进行了表达探究, 并将此方法用于鉴定人巨细胞病毒(hCMV)极早期基因增强子/启动子(MIEP)组织的三种转录因子NF-κB (p65)、AP1和CREB之间可能存在的功能合作关系。计算结果表明, 所采用的方法对研究转录因子协作关系是切实可行的。

关键词 转录因子; 合作关系; 多元回归模型; 基因表达

细胞内基因的转录控制是通过利用DNA (调控DNA)编码的顺式调控逻辑, 将存在于当前细胞中的各种转录因子(反式作用因子)按照它们之间的合作关系组装成具有高级结构的转录控制机器, 最终特异地决定目标基因的转录水平[1]。在以往的转录调控研究中, 由于限于方法, 人们往往只能关注于单个转录因子对下游基因表达所起的作用, 无法对转录因子间的相互作用开展简便而有效的研究。为此, 本文提出了一种基于综合性干涉实验分析与数学建模相结合的策略, 对基因表达过程中判断多个转录因子间的协同作用进行了探索性的分析。我们将此方法用于鉴定文献发表的人巨细胞病毒(hCMV)极早期基因增强子/启动子(MIEP)组织的三种转录因子NF-κB (p65)、AP1和CREB之间可能存在的功能合作关系[2]。在文献提及的实验中, 作者将含有特定转录因子结合位点的短DNA片段作为“诱饵(decoy)”, 用于干涉转录因子在基因调控区的集结和组织。在不同“诱饵”干涉条件下, 测量MIEP在体外转录系统中转录出的RNA的量来评价基因转录的功能。将干涉实验的条件进行0,1编码, 采用多维空间中的统计分析方法[3], 构建转录因子与目标基因转录水平之间的关系方程, 揭示转录因子之间的合作关系。计算结果表明, 我们的结论符合以往的研究, 是切实可行的, 同时由于方法简洁, 结果可被检验, 有望在相关的生物学研究中发挥重要作用。

1 方法

1.1 建模的基本思想

目前, 人们普遍承认, 基因调控序列上的顺式元件和细胞中的反式作用因子(转录调控因子)的相互作用, 构成基因转录的分子基础, 但是, 在研究某个特定基因的转录调控时, 各种转录因子对基因表达的贡献常常被独立地研究, 并且单个的转录因子往往被人们自觉地或者不自觉地认为对所调控基因的表达起直接作用。所以, 一种转录因子往往被定义为激活因子或抑制因子。而我们认为, 顺式调控序列组织的转录因子的合作关系才直接决定了基因的表达, 无合作的单个转录因子的独立贡献(如果存在的话)被看成是合作关系的一个特例。常规的思想认为, 同一个基因的调控序列在同等条件下组织一种特定的转录方式, 而我们认为同一个顺式调控序列可以组织起多种不同的转录因子的合作关系, 每种合作关系可以分别独立地、竞争性地对基因的转录做出贡献。因此我们在目前常规的实验中所测定到的基因的表观转录活性是这些存在于同一系统中的由转录因子组成的各种合作关系对基因表达贡献的线性叠加。依据上述假设, 我们可以在基因的转录水平和转录因子合作关系之间建立明确的函数关系, 并且通过解析这个函数关系中的各项参数, 可以推断出隐含着的转录因子间合作关系的种类和在基因表达调控中的作

收稿日期: 2008-06-18 接受日期: 2008-12-01

上海市教育委员会科学项目(No.04DB25)、国家重点基础研究发展规划(973计划)项目(No.2002CB713803)资助

* 通讯作者。Tel: 021-65980334, E-mail: jfei@mail.tongji.edu.cn

用。考虑有 N 个转录因子的基因转录过程，转录因子间可以组织起 2^N-1 种可能的合作关系，由于假设每种合作关系可以独立并同等竞争地对基因转录发挥作用，因此可以获得如下的方程(方程 1):

$$y = b_0 + \sum_{i=1}^{N} b_i x_i + \sum_{1 \le i,j \le N} b_{ij} x_i x_j + \cdots + b_{12\cdots N} x_1 x_2 \cdots x_N$$

为确定方程中的参数，引入最少合作关系的假设，即假设一个合理的转录调控系统应该是一个含有最少合作关系的转录系统。

1.2 数据和实例

我们曾经选择人类巨细胞病毒的极早期基因1/2的启动子(human cytomegalovirus major IE1/2 gene enhancer/promoter，简称 MIEP)作为我们转录控制研究的对象，利用综合干涉的方法研究了 NF-κB (p65)，CREB 以及 AP1 三种转录因子在 MIEP 转录调控中的作用，我们针对上述三种转录因子设计了相应的结合位点双链 DNA 诱饵(decoy)，其中针对 NF-κB 的序列为: 5'-AGTAGGGAATTCCCATAA-3'; 针对AP1的为: 5'-CGCTTGATGAGTCAGCCGGAA-3'; 针对CREB 的为: 5'-AGAGATTGCCTGACGTCAGAGAGCTAG-3'; 一个无关序列作为对照: 5'-TCCAGCACCACGGACAGTTCC-3'。这些诱饵在反应体系中可以特异地阻止目标转录因子结合在调控 DNA 的序列上，从而达到干涉的目的，在获得各种干涉条件下的功能数据后，我们用人工智能的方法分析了上述转录因子可能形成的合作关系。在本文中，我们根据以往发表的数据利用多元逐步回归的方法再次对所获得的数据进行分析，通过对方程 1 中参数的确定，从而获得参与独立调控MIEP转录活性的转录因子合作方式的种类和作用。

由于涉及到三种转录因子，因此它们所有可能存在的合作关系有 7 类(表 1)。

此外，我们还定义 F8，在这种合作关系中不包含上述 3 种转录因子中的任何一种。针对不同的合作关系，我们列出相应的干涉条件(pertubed situations)，见表 2 。

对每一种干涉条件，我们考察了 MIEPE 和 HeLa 细胞核蛋白混合后在体外的相对转录活性(用同位素掺入法测定，以没有干涉情况下的转录活性为 100 计算)。用二进制数给干涉条件进行编码，其中 A 表示实施对 AP1 的干涉，C 表示对 CREB 实施干涉，K 表示对 NF-κB 实施干涉; 以 1 表示干涉存在，0 表示干涉不存在。因此得到编码表和测试数据(表 3) (实验中，每种 DNA 诱饵的加入浓度为 5 pmol，并用对照 DNA 补到 DNA 浓度均为 15 pmol)。

Table 1 The category of functional cooperation among three transcription factors

	Functional cooperation
F1	Among NF-κB (p65), CREB and AP1
F2	Between NF-κB (p65) and AP1
F3	Between NF-κB (p65) and CREB
F4	NF-κB (p65)
F5	Between AP1 and CREB
F6	AP1
F7	CREB

Table 2 The category of perturbation situation

	Perturbation situations (S)
S1	decoy for none specificity
S2	decoy for CREB
S3	decoy for AP1
S4	decoy for CREB and AP1
S5	decoy for NF-κB (p65)
S6	decoy for NF-κB (p65) and CREB
S7	decoy for NF-κB (p65) and AP1
S8	decoy for NF-κB (p65), CREB and AP1

Table 3 The result of perturbation analysis

K	C	A	*in vitro* Transcription			Average	S.D.
0	0	0	100	100	100	100	0
1	0	0	69.84	82.18	76.09	76.0367	6.17017
0	1	0	95.36	142.67	119.07	119.033	23.655
1	1	0	4.86	8.39	6.34	6.53	1.77265
1	0	1	113.06	138.95	126.11	126.04	12.9451
0	1	1	68.49	102.31	85.42	85.4067	16.91
0	0	1	158.31	233	195.61	195.64	37.345
1	1	1	71.69	71.01	71.36	71.3533	0.34005

每次实验我们做了三次重复实验，表中同时给出了三次实验结果的均值和方差，实验结果显示三种 DNA 诱饵均能干涉 DNA 的转录，但是情况不同。在以下的分析中，我们称上述实验结果的平均值为“输出结果”。

1.3 数学计算

干涉条件的编码为进行数学分析提供了基础，由于我们采用了大大过量的干涉DNA，因此我们假定被干涉的因子被完全排除在转录复合物的组成之外。设 x_1、x_2 和 x_3 分别代表三个转录基因 NF-κB (p65)、CREB 和 AP1。实验的输出结果(表达强度)为 y (因变量)。方程 1 即为如下形式:

$$y = b_0 + \sum_{i=1}^{3} b_i x_i + \sum_{1 \le i,j \le 3} b_{ij} x_i x_j + b_{123} x_1 x_2 x_3$$

这里, x_1 的取值为 0 和 1。当取值为 0 时, 表示该因子被干涉而不存在; 反之, 表示该因子没被干涉而存在(或可被分别称为“缺失”和“在场”)。因此, 若在所得的方程中, 相应的系数不为零, 则表明该项因子对基因表达有贡献; 否则, 则表示不具贡献。变量的乘积项表示了因子之间的合作关系的存在。比如, x_1x_2 表示转录因子 NF-κB (p65)和 CREB 的合作。如果 x_1x_2 前的系数 b_{12} 不为零, 则表明该种合作对下游的表达是有作用的。b_0 表示关系 F8。

为了得到上述的函数关系式, 并确定有哪些合作作用对下游基因的表达有作用, 我们分别采用了多元数量化理论中的全回归和逐步回归分析方法, 这是基于下述理由:

1. 目前我们所掌握的实验数据是完备的, 它是对数据空间的一个完整描述, 据此得到的回归方程将具有较好的代表性;

2. 采用逐步回归能最大程度地筛选出显著的因子, 剔除不显著的因子, 从而使所得的方程具有较高的稳定性;

3. 回归分析中的统计检验, 特别是对偏回归系数的检验能有效地判断该因子对总体的贡献。

对于 RNA 的实验结果得到回归方程:

$$y_{ivt} = 71.35333 + 14.05333x_1 + 54.68667x_2 - 64.82333x_3 + 55.54667x_1x_2 + 98.45000x_1x_3 + 14.82000x_2x_3 - 144.08667x_1x_2x_3$$

方程的复相关系数为 0.96239, 可决系数为 0.9262。这些数据表明回归方程是有效的。当显著性水平 α=0.01 时, 方程的总体是显著的(P 值<0.0001)。表 4 显示了对各回归系数的 t 检验结果。

表 4 中显示, 因子 x_2 和 x_1x_3 均为通过 t 检验, 其余因子都对方程做出了显著的贡献。注意到在全回归方程中, 我们所关注的各合作关系并不都显著, 表明, 一些假设的关系并不真实存在。为此, 我们需要将不显著的因子从两方程中剔除, 来获得可靠的转录因子的合作关系。我们进一步采用选回归的方法来达到这一目的。

选回归计算得到的方程为:

$$y_{ivt} = 75.25905 + 57.02286x_2 - 62.48714x_3 + 63.35810x_1x_2 + 106.26143x_1x_3 - 139.41429x_1x_2x_3$$

方程的复相关系数为 0.95681, 可决系数为 0.9155。这些都表示方程已然包含了影响 RNA 合成的因子。此外, 当显著性水平时, 方程总体和各回归系数都通过了假设检验(方程总体的 P 值<0.0001)。表 5 是对各回归系数的检验结果:

Table 4　The parameter estimated from the first round of regression analysis

		Parameter estimation			
Variable	DF	Parameter estimated	Standard error	*t* Value	Pr > |t|
Intercept	1	71.35333	10.79917	6.61	<0.0001
×1	1	14.05333	15.27233	0.92	0.3731
×2	1	54.68667	15.27233	3.58	0.0030
×3	1	−64.82333	15.27233	−4.24	0.0008
×12	1	55.54667	21.59834	2.57	0.0222
×13	1	98.45000	21.59834	4.56	0.0004
×23	1	14.82000	21.59834	0.69	0.5038
×123	1	−144.08667	34.14997	−4.22	0.0009

Table 5　The parameter estimated from second round of regression analysis

		Parameter estimation			
Variable	DF	Parameter estimated	Standard error	*t* Value	Pr > |t|
Intercept	1	75.25905	7.07326	10.64	<0.0001
×2	1	57.02286	10.00309	5.70	<0.0001
×3	1	−62.48714	10.00309	−6.25	<0.0001
×12	1	63.35810	14.14651	4.48	0.0004
×13	1	106.26143	14.14651	7.51	<0.0001
×123	1	−139.41429	25.17390	−5.54	<0.0001

2　结果与讨论

在本文中, 我们探索了转录因子以及转录因子间相互作用方式对基因表达的数学关系, 并具体地用这一数学模型分析了针对 MIEP 功能干涉的实验数据, 用多元逐步回归方法对方程中的参数进行了确定, 由于我们事实上并无依据来计算或者测量各种合作关系在被干涉后在基因表达过程中权重的变化, 而我们的目的也并非要求解各种合作关系在基因表达中的绝对贡献, 我们只是要了解有多少合作关系, 以及这种合作关系在基因表达中的相对地位和贡献, 因此我们在建模过程中完全忽略了干涉对合作关系在基因表达中的权重变化, 我们的结果表明, 所考察的 3 种转录因子共形成的所有 7 种合作关系中, 只有 5 种参与了 MIEP 功能的发挥, 它们分别为 F1、F2、F3、F6、F7, 这个结果和我们先前采用的算法在结论上基本相符, 其中, F1、F2、F3 和 F6 均出现在两次计算结果中, 并且符号一致。但是两次计算结果也有一定的区别, 其中 F4 没有出现在本次的计算结果中, 取而代之的是 F7。这可能来源于两次计算的假设有一定的差异, 计算过程也不尽相同, 在前次的建模中, 我们引入了干涉系数, 以求将干涉后对各合作

关系对基因表达的影响(权重改变)数量化, 进而求出每种合作关系对基因表达的最大贡献值, 但是其中的干涉系数只能采用非常人为的一个估计值, 因此给方程的求解带来一定的影响。

方程中针对F1和F6的参数为负值, 表明这两种关系对应的基因转录是低效率的(低于表观的系统输出数值), 如果它们在转录比例中占优势将导致总体转录活性的下降, 其中由AP1、NF-κB和CREB共同组成的转录合作关系是效率最低的。其他3种参数为正的关系则表示对应的基因表达呈高效率(高于表观的系统输出数值), 其中NF-κB和AP1组成的转录合作关系是效率最高的。从我们的分析结果可以知道, 一个转录因子其作用在不同场合下是不同的, AP1和NF-κB的合作表现出最强的基因转录活性, 但是在和NF-κB以及CREB共同组成三元复合物时, 则形成了效率最低的转录合作关系。同样, CREB和NF-κB都具有这种特性。因此, 与人们通常使用的比较分析策略相比, 本文所提出的基于合作关系的分析策略, 以及多元逐步回归方法可望使研究者对转录控制系统的理解更加接近全局和本质。本文的贡献在于对基因转录调控的方式提出了新的看法, 即一个特定的基因转录是可以由多种不同的转录因子的组合来独立的和竞争性地完成的, 我们所检测到的是这些过程的总和。同时, 我们也为鉴定这些合作关系提供了一种简便有效的系统生物学与生物信息学的方法。

参考文献(References)

[1] 特怀曼。陈 淳, 徐 沁, 等译。*高级分子生物学要义*, 科学出版社, 2000

[2] Niu G, Huang L, Wang Q, *et al*. A novel strategy to identify the regulatory DNA-organized cooperations among transcription factors, *FEBS Lett*, 2006, 580(2): 415-24

[3] 周纪芗。*回归分析*, 华东师范大学出版社, 1993

Identification of the Cooperation Patterns among Transcription Factors in Gene Expression Regulation by the Multiple Stepwise Regression Analysis

Dan Sha, Hai-Yan Xu, Gang Niu[2], He-Liang Fei, Rong-Xian Yue, Jian Fei[1]*

(*Mathematics and Science College, Shanghai Normal University, Shanghai 200234, China;*
[1]*College of Life Science and Technology, Tongji University, Shanghai 200092, China;*
[2]*Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*)

Abstract The cooperative relations among transcription factors and the *cis*-regulatory logic are both fundamental in DNA transcription activation. To predict the functional cooperative relationship patterns among *trans*-cooperative factors existing in a gene expression regulation, we propose a new method, which is based on the multiple stepwise regression analysis, to identify cooperation among NF-κB (p65), AP1, and CREB in transcription activation of human cytomegalovirus major IE1 promoter/enhancer (MIEP). The computational results show that the method is effective and can be applied on understanding the transcription control systems.

Key words transcription factor; cooperation; multiple stepwise regression analysis; gene expression

Received: June 18, 2008 Accepted: December 1, 2008

This work was supported by the grants from Shanghai Municipal Education Commission (No.04DB25) and the National Basic Research Program of China (973 Program) (No.2002CB713803)

*Corresponding author. Tel: 86-21-65980334, E-mail: jfei@mail.tongji.edu.cn