

基因启动子甲基化对转录因子结合的抑制作用分析方法*

冯伟兴^{1, 2)**} 王科俊²⁾ 贺波²⁾ 李霞^{1)**}

(¹⁾哈尔滨医科大学生物信息科学与技术学院, 哈尔滨 150081; ²⁾哈尔滨工程大学自动化学院, 哈尔滨 150001)

摘要 基因启动子甲基化对转录因子结合的抑制作用是一种有效的基因转录调控机制。尽管基因启动子甲基化水平已经可以通过实验测量, 但仍未有有效的方法利用这些数据定量分析甲基化对转录因子结合的影响。设计一个通用模型来描述基因启动子甲基化对转录因子结合的抑制作用。在特定细胞环境下, 通过基因表达与转录因子在基因启动子上结合值之间的相关性分析, 实现模型参数求取, 并基于该模型进行甲基化对转录因子结合的抑制作用分析。神经细胞生物实验数据测试证明了该方法的有效性。

关键词 基因启动子, 转录因子, DNA 甲基化, 分析方法
学科分类号 Q332, Q7

DOI: 10.3724/SP.J.1206.2010.00355

众多研究成果表明, 基因启动子甲基化可抑制转录因子在基因启动子区域的结合, 从而导致基因转录水平的下降甚至停止转录, 以达到调控基因转录的目的^[1-2]。神经瘤等很多疾病正是由于基因启动子甲基化变异导致基因转录异常所引发的^[3-4]。因此, 为了准确分析特定细胞中基因转录调控机制, 尤其是很多疾病的致病机理, 基因启动子区域的 DNA 甲基化对转录因子结合的抑制作用是需要考虑的因素。

目前, 通过实验手段, 已经可对细胞中全基因组的 DNA 甲基化水平进行检测。基于检测数据, 很多学者尝试对 DNA 甲基化的调控作用进行分析。Liu 等^[5]通过 *t* 检验、主成分分析、独立成分分析等方法对男女基因的 20 493 个 CpG 位点的甲基化水平进行分析, 发现这些位点甲基化水平具有明显的性别倾向性。Daniela 等^[6]发现, 小鼠 Myogenin 基因启动子上的甲基化水平在不同发育阶段有着显著的变化, 其作用是通过调控转录因子 MEF2 和 SIX 的结合位点来调控不同发育阶段 Myogenin 基因的表达水平。不过, 这类研究均是对基因 DNA 甲基化调控作用进行定性分析。尚没有成熟有效的方法来定量地分析基因启动子甲基化对转录因子结

合的抑制作用。本文设计了一种方法来尝试解决这一问题。该方法首先提出一个通用模型来描述基因启动子甲基化对转录因子结合的抑制作用。然后, 在特定细胞环境下, 考虑启动子甲基化抑制作用的基础上计算转录因子在基因启动子上的结合值, 再通过基因表达与转录因子在基因启动子上结合值之间的相关程度对模型参数进行精确求取, 并最终实现甲基化对转录因子结合抑制作用的定量分析。

通过采用神经细胞实验数据对该方法进行测试, 证实了该方法的有效性。

1 基因启动子甲基化抑制作用分析

基因启动子是一个对于基因转录调控至关重要的区域。各种转录因子在基因启动子上结合并对基因转录进行调控。但如何判定转录因子是否在基因启动子上结合则是一个难题。最直接的方法是统计

* 国家自然科学基金(61071174), 国家博士后科学基金(20100471090)和中央高校基本科研业务费专项资金(HEUCF100411)资助项目。

** 通讯联系人。

Tel: 0451-82569906, E-mail: fengweixing@hrbeu.edu.cn

收稿日期: 2010-07-05, 接受日期: 2010-11-22

分析转录因子易于结合的碱基组合, 再通过基因启动子上进行碱基组合匹配分析来得到转录因子的结合位点. 为此, 根据转录因子所结合碱基的实验测试结果, 形成了转录因子位置频次矩阵(PFM, position frequency matrix)、位置权值矩阵(PWM, position weight matrix)、位置分数矩阵(PSM, position scoring matrix)等统计矩阵. 其中, 位置频次矩阵是基础, 位置权值矩阵和位置分数矩阵则在该基础上进行了不同程度的归一化处理. 基于这些转录因子的碱基组合统计矩阵, 即可以对转录因子是否在基因启动子上结合并调控其转录进行分析. 这类分析主要有两种方式, 一种是静态的分析转录因子结合情况和基因转录程度的相关性^[7], 另一种是动态的分析转录因子结合情况和基因转录程度变化的相关性^[8].

由于基因启动子甲基化可抑制转录因子在基因启动子区域的结合, 因此, 为了更准确地分析转录因子是否在基因启动子上结合并调控, 除了考虑启动子上碱基组合与转录因子位置矩阵的匹配程度, 还应考虑启动子上 DNA 甲基化对转录因子结合的抑制作用. 由于目前尚没有有效的模型对基因启动子甲基化抑制作用进行模拟和分析, 本文对此进行了尝试.

1.1 转录因子在基因启动子上的匹配值

利用 BIOBASE 公司 TRANSFAC 数据库中提供的转录因子与 DNA 碱基结合的位置权值矩阵(PWM), 可以对转录因子在基因启动子上的碱基匹配情况进行计算. 该矩阵是通过统计大量实验测得的转录因子实际结合位点的碱基组合数据生成的, 例如转录因子 E2F1(M00939)的 PWM 矩阵如表 1 所示.

Table 1 PWM of transcription factor E2F1

No.	A	C	G	T
1	2	6	1	22
2	1	0	2	28
3	1	3	1	26
4	1	12	18	0
5	0	5	26	0
6	0	31	0	0
7	0	0	31	0
8	0	19	12	0
9	1	5	20	5

表 1 显示的 PWM 矩阵中第 i 行第 j 列的数值反映转录因子 E2F1 在结合位点第 i 个位置上与第 j 个碱基(A, C, G, T)的匹配权值. 权值越大, 转录因子在该位置上与该碱基结合的可能性越大. 为了比较同一结合位点上不同转录因子匹配情况的差异, PWM 矩阵没有进行完全的归一化处理, 即矩阵中每行的匹配权值和可以大于 1. 由表 1 可以分析出, E2F1 的结合位点长度为 9 个碱基, 每行的匹配权值和为 31. 根据结合位点匹配权值的最大组合, E2F1 最佳碱基匹配组合为: TTTSGCGSG, 其匹配权值和为 221.

通过转录因子位置权值矩阵和基因启动子碱基序列可以分析转录因子在该启动子上的匹配情况. 本文中, 基因启动子的长度设为基因转录起始位点的前 1 200 个碱基至后 300 个碱基, 共 1 500 个碱基. 这样, 设第 i 个转录因子结合位点长度为 L , 则在第 j 个基因启动子上共有 $1\,500-L+1$ 个假定转录因子结合位点. 在第 k 个假定结合位点上, 可以根据该位点的碱基组合与转录因子 PWM 矩阵的匹配程度计算出转录因子在该位点的匹配权值和 A_{ijk} .

$$A_{ijk} = \sum_{l=1}^L a_{jkl} w_{il}^T \quad (1)$$

式(1)中, a_{jkl} 为描述第 j 个基因启动子第 k 个假定结合位点第 l 个位置上碱基种类的行向量 (1 000 表示 A, 0100 表示 C, 0010 表示 G, 0001 表示 T), w_{il} 为第 i 个转录因子 PWM 矩阵中第 l 行与碱基相匹配的权值行向量. 这样, 对第 j 个基因, 我们可以得到第 i 个转录因子在其启动子上所有 $1\,500-L+1$ 个假定结合位点上的匹配权值和. 那么, 我们将最大的匹配权值和定义为第 i 个转录因子在第 j 个基因启动子上的匹配值 S_{ij} .

$$S_{ij} = \max A_{ijk} \quad (2)$$

计算得到转录因子在基因启动子上的匹配值后, 并不意味着该转录因子一定在其启动子上结合并对该基因转录进行调控. 只有匹配值足够大, 才能认为转录因子在基因启动子上结合.

1.2 转录因子在基因启动子上的匹配阈值分析

为了分析转录因子是否在基因启动子上结合并对该基因转录进行调控, 需进行匹配值的阈值分析, 即当转录因子在基因启动子上的匹配值大于某一阈值时, 才可认为该转录因子在该基因启动子上结合并对该基因转录进行调控.

转录因子匹配阈值的确定方法有很多. Hertzberg 等^[7]提出一种根据基因表达和转录因子匹

配值相关性分析的阈值确定方法. 设有 n 个基因, $g_1, \dots, g_n$ 是这些基因表达值的对数且遵循正态分布, 其均值为 μ , 均方差为 σ . 设定一个阈值, 根据转录因子在基因启动子上匹配值是否大于该阈值得到一个基因子集 G^i . 设该基因子集 G^i 有 k 个基因, 其序号为 $i_1, \dots, i_k$. 则可如式(3)计算一个 Z 值:

$$Z(TF, G^i) = \left(\frac{1}{k} \sum_{j=1}^k g_{ij} - \mu \right) / (\sigma / \sqrt{k}) \quad (3)$$

由式(3)分析可得, Z 值反映着基因子集 G^i 中基因表达平均值与全部基因表达平均值的差异程度. Z 值的绝对值越大, 差异程度越大, 表明基因子集 G^i 中的基因越可能被该转录因子结合并调控. 而 Z 值的符号则反映着该转录因子的生物功能是促进基因转录还是抑制基因转录. 这样, 通过调整阈值, 我们可得到一系列不同的 Z 值. 其中, 当 Z 值的绝对值最大时, 其对应的阈值即为最大似然阈值, 该 Z 值称为 Z_m . 可以看出, 上述过程是在分析基因表达和转录因子在基因启动子上匹配值的相关性. 相关性越大, 无疑 Z_m 的绝对值也越大.

1.3 基因启动子甲基化对转录因子结合的抑制作用通用模型

在分析转录因子是否在基因启动子上结合并进行调控时, 除了考虑启动子上碱基组合与转录因子位置权值矩阵的匹配程度, 还应考虑启动子上 DNA 甲基化对转录因子结合的抑制作用.

这里设计了一个通用模型来描述基因启动子甲基化对转录因子结合的抑制作用, 该模型采用如下函数形式:

$$E_{ijk} = \frac{e^{-(M_{jk}-C_i)/S_i}}{1 + e^{-(M_{jk}-C_i)/S_i}} \quad (4)$$

其中, M_{jk} 是第 j 个基因启动子上第 k 个假定结合位点的 DNA 甲基化水平, E_{ijk} 是其对第 i 个转录因子结合的抑制作用, 其取值范围为 $[0, 1]$. 当第 k 个假定结合位点的 DNA 甲基化水平 M_{jk} 非常小时, E_{ijk} 将趋近 1, 表示 DNA 甲基化对转录因子的结合无影响; 反之, 当 M_{jk} 足够大时, E_{ijk} 将趋近 0, 表示在 DNA 甲基化影响下, 转录因子无法在该位点结合. 另外, 模型有两个参数, 分别是函数中心值 C 和函数陡度 S . 该模型曲线如图 1 所示. 由于针对不同的转录因子, 模型参数是可以变化的, 因此, 式(4)所描述的模型中, 该两个参数分别是针对第 i 个转录因子的 C_i 和 S_i .

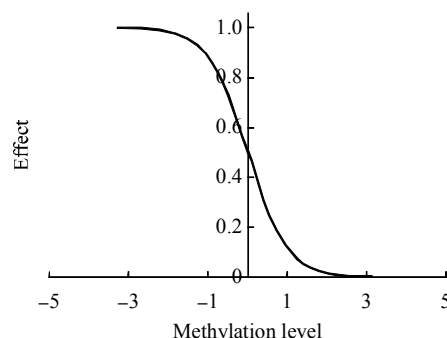


Fig. 1 Function of methylation effect general model
($C=0, S=0.1$)

图 1 中, 横轴是甲基化水平, 纵轴则表示甲基化对转录因子结合的作用. 针对不同的转录因子, 通过取不同模型参数, 通用模型曲线是不同的, 图 1 中模型参数假定选为 $C=0, S=0.1$.

用该通用模型来描述基因启动子甲基化对转录因子结合的抑制作用是基于以下两点生物特性: 首先是结合位点的甲基化水平越高, 其对转录因子结合的抑制作用越强. 其次是结合位点的甲基化水平对转录因子结合的抑制作用并不随甲基化水平的变化而成线性关系, 当甲基化水平较高或较低时, 其对转录因子结合的抑制作用将出现饱和现象, 对甲基化水平的变化也将不敏感, 而在中间值时, 其抑制作用对甲基化水平的变化将较敏感. 图 1 所示的通用模型曲线较好地反映了上述两点生物特性.

为了描述特定细胞环境下基因启动子甲基化对不同转录因子的抑制作用, 该通用模型设有两个可变参数, 即函数中心值 C 和函数陡度 S . 这样, 特定细胞环境下, 通过选取不同的模型参数, 每个转录因子将具有不同的模型来描述其受基因启动子甲基化的影响.

1.4 转录因子在基因启动子上的结合值

考虑基因启动子甲基化对转录因子结合的抑制作用后, 可以按式(5)计算第 i 个转录因子在第 j 个基因启动子上的结合值:

$$S_{ij} = \max(A_{ijk} \times E_{ijk}) \quad (5)$$

其中, E_{ijk} 表示第 j 个基因启动子上第 k 个假定结合位点的 DNA 甲基化水平对第 i 个转录因子结合的抑制作用. E_{ijk} 的值需要在特定细胞环境下进行求取.

1.5 基因启动子甲基化对转录因子结合的抑制作用分析

为了分析特定细胞环境下, 基因启动子甲基化对转录因子结合的抑制作用, 需要对该转录因子受抑制作用模型中的两个参数进行确定. 这可以通过变动模型参数值 C 和 S , 在转录因子结合阈值分析中得到一系列不同的 Z_m 值. 当 Z_m 的绝对值最大时, 此时的模型参数值 C 和 S 即为最大似然模型参数值. 而由最大似然模型参数值 C 和 S 构造的模型即可用于分析基因启动子甲基化对该转录因子结合的抑制作用.

2 方法测试

2.1 数据获取

本文选用的生物实验数据来自 Janine 实验室针对 ATCC 公司生产的 SH-SY5Y 神经细胞所作的生物实验(GSE4600, GSE9568). 实验中, 在 SH-SY5Y 细胞中添加佛波酯(PMA)并在细胞发育 48 h 后, 对细胞进行了生物数据检测. 检测数据包括如下部分:

2.1.1 基因表达数据. 本方法测试所需的 SH-SY5Y 细胞基因表达数据由 Affymetrix 公司生产的人类基因表达检测芯片 HG U133 plus 2.0 生成^[9].

2.1.2 基因启动子甲基化数据.

本方法测试所需的 SH-SY5Y 细胞基因启动子甲基化数据由 NimbleGen 公司生产的人类基因启动子信息检测芯片 1500b_Pomoter_Array 生成^[10]. 利用该数据, 我们采用 Batman 算法^[11]对基因启动子每个位点的甲基化水平进行了推算.

此外, 我们从 Transfac 数据库获得 459 个人类转录因子的 PWM 矩阵. 而人体基因启动子序列信息和人体基因描述信息则从美国 UCSC 网站上获取.

根据人体基因描述信息, 我们从 SH-SY5Y 细胞生物实验数据中共整理出 10 065 个基因的表达数据和启动子甲基化水平数据.

2.2 基因启动子甲基化对转录因子结合的作用

首先, 不考虑基因启动子甲基化对转录因子结合的作用, 通过转录因子在基因启动子的匹配值和基因表达相关性计算 Z_m 值. 其分布直方图如图 2a 所示. 图 2 中, 横轴为 Z_m 值, 纵轴为频数.

与此对应, 考虑基因启动子甲基化对转录因子结合的作用, 通过转录因子在基因启动子的结合值和基因表达相关性计算 Z_m 值. 其分布直方图如图 2b 所示.

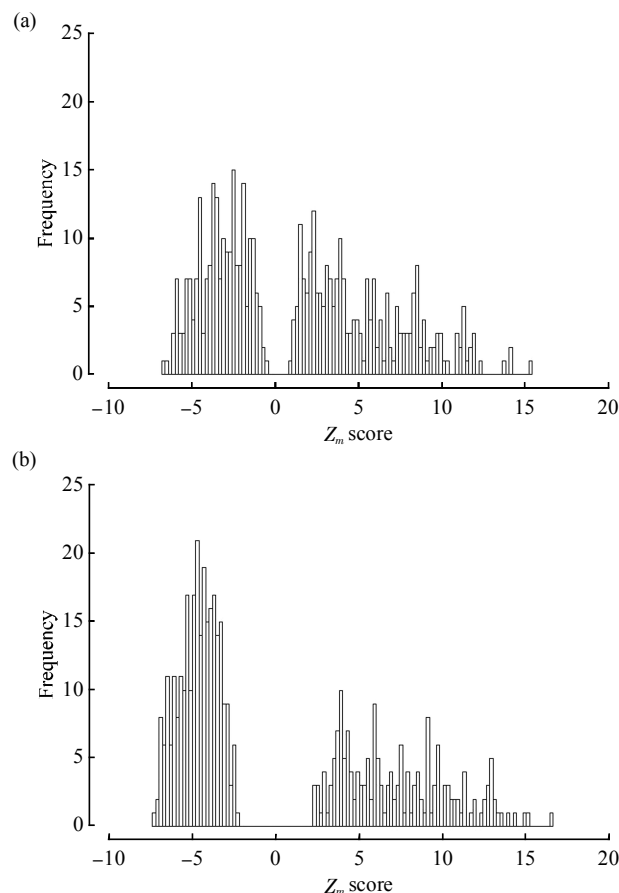


Fig. 2 Distribution of whole genome Z_m scores

(a) Without consideration of methylation effect. (b) With consideration of methylation effect.

由图 2a 和图 2b 可见, 不考虑甲基化作用时, 正 Z_m 集合和负 Z_m 集合的均值分别为 5.33 和 -3.24; 而考虑甲基化作用时, 正 Z_m 集合和负 Z_m 集合的均值分别为 7.23 和 -4.66. 可见, 考虑基因启动子甲基化作用后, Z_m 值的绝对值明显变大, 这表明转录因子结合值和基因表达相关性程度明显提高, 证明本文所提出的基因启动子甲基化对转录因子结合的作用分析方法是可行的.

由于转录因子 E2F1 在 SH-SY5Y 神经细胞中丰度较高^[12], 因此我们针对 E2F1 对本方法进行了详细分析.

首先, 不考虑基因启动子甲基化对转录因子 E2F1 结合的作用, 对 E2F1 在基因启动子的匹配值进行阈值分析. 不同阈值下得到的 Z 值如图 3 所示. 由图 3 可见, Z 值为正值, 这表明 E2F1 对基因转录起促进作用, 这是符合转录因子 E2F1 的生

物功能的. 当匹配阈值取 9.54 时, Z 值取最大, 即 $Z_m=11.17$.

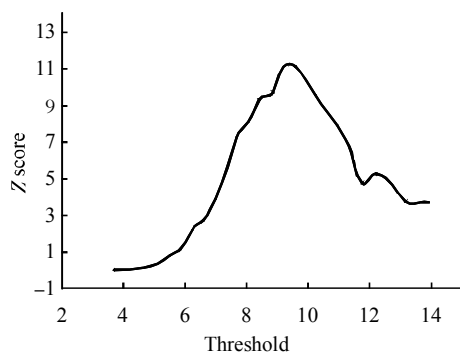


Fig. 3 Analysis of Z score

Without consideration of methylation effect.

图 3 中, 横轴为转录因子匹配阈值, 纵轴为对应的 Z 值.

其次, 考虑基因启动子甲基化对转录因子结合的作用, 通过转录因子在基因启动子的结合值和基因表达的相关性寻找最大的 Z_m 值. 即变动通用模型参数 C 和 S 寻找最大的 Z_m 值. 通过搜寻, 当模型参数 $C=-0.15$, $S=0.05$ 时得到最大的 Z_m 值 ($Z_m=12.531$). 图 4 显示了当 $S=0.05$ 时, 随着 C 值的变化, Z_m 值的变化过程. 图 4 中的横线 ($Z_m=11.17$) 标明不考虑基因启动子甲基化影响时计算得到的 Z_m 值.

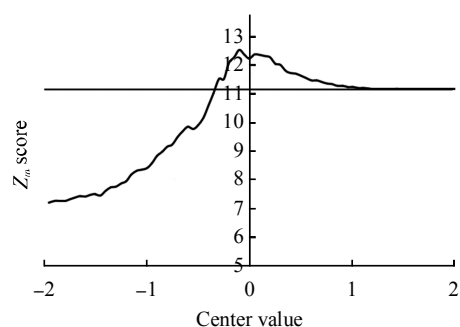


Fig. 4 Analysis of Z_m score

With consideration of methylation effect, $S=0.05$.

图 4 中, 横轴为模型参数 C 值, 纵轴为对应的 Z_m 值.

由图 4 可见, 当模型参数 C 值比较小时, Z_m 值非常小, 表明模型参数取值不合理. 随着 C 值的增大, Z_m 值迅速增大. 当 C 值大于 -0.35 时, Z_m 值开始大于 11.17, 表明其分析结果已优于不考虑

基因启动子甲基化对转录因子 E2F1 作用时得到的结果. Z_m 值在 -0.15 取最大值 ($Z_m=12.531$). 随后, Z_m 值随 C 值增大而下降, 并趋近于 11.17. 通过上述分析, 证明考虑甲基化作用来分析转录因子 E2F1 在基因启动子上的结合是合理的.

另外, 在应用通用模型搜寻最大 Z_m 值过程中, 也同时确定了模型参数 C 和 S 值, 即确定了描述启动子甲基化对转录因子 E2F1 的作用模型.

这里还采用了与式(4)的含义相反的函数来描述基因启动子甲基化对转录因子结合的作用:

$$E_{ijk} = \frac{1}{1 + e^{-((M_{ijk}-C_i)/S_i)}} \quad (6)$$

其函数曲线如图 5 所示. 图 5 中, 横轴是甲基化水平, 纵轴是甲基化对转录因子结合的抑制作用. 与前述不同的是, 这里假设甲基化水平越高, 其对转录因子结合的抑制作用越低.

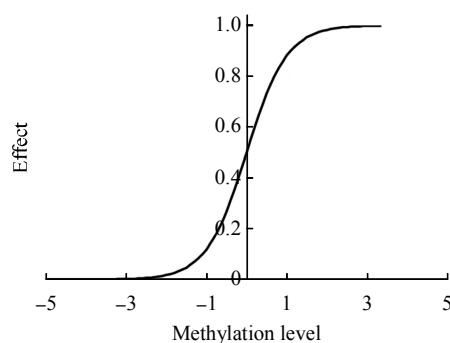


Fig. 5 Methylation effect reverse function

$C=0, S=0.1$.

采用该函数建立的模型来分析甲基化对转录因子 E2F1 在基因启动子上结合的抑制作用, 其结果如图 6 所示.

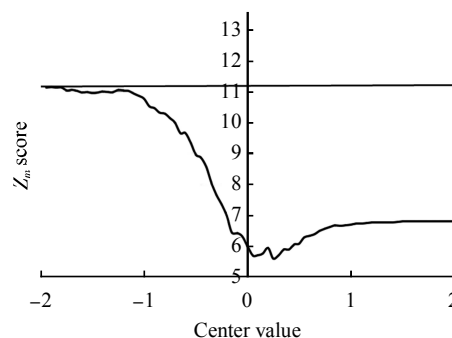


Fig. 6 Analysis of Z_m score

With reverse function model.

由图 6 可见, 不论函数参数 C 取何值, Z_m 值均远小于 11.17. 这表明采用式(6)来描述基因启动子甲基化影响是不合理的. 同时, 也进一步证明了基于本文所设计模型分析甲基化对转录因子结合抑制作用的合理性.

采用上述分析方法, 在 SH-SY5Y 细胞中对 459 个人类转录因子进行了分析, 共发现 91 个转录因子存在该现象, 即采用式(4)所示模型函数分

析得到的 Z_m 值明显高于不考虑甲基化作用得到的 Z_m 值, 而采用式(6)所示模型函数分析得到的 Z_m 值明显低于不考虑甲基化作用得到的 Z_m 值. 其中, 前 10 个转录因子的分析结果见表 2. 表 2 中, Z_{no} 表示不考虑甲基化作用得到的 Z_m 值, Z_{me} 表示考虑甲基化作用得到的绝对值最大的 Z_m 值, C , S 则为针对每一个转录因子所建立的启动子甲基化作用模型参数.

Table 2 Result of methylation effect on transcription factors

No.	AC	Name	Z_{no}	Z_{me}	C	S
1	M00716	ZF5	11.760	13.164	0	0.01
2	M00939	E2F1	11.173	12.531	-0.35	0.01
3	M00189	AP2	9.991	11.713	0.2	0.1
4	M00196	Sp1	9.908	11.308	0.05	0.01
5	M00466	HIF1	9.620	11.213	-0.15	0.01
6	M00332	Whn	8.891	10.885	-0.1	0.01
7	M00778	AhR	9.384	10.767	0.05	0.01
8	M00801	CREB	8.074	10.561	0.1	0.05
9	M00245	Egr3	8.965	10.307	0.1	0.1
10	M00982	KROX	8.700	10.193	0	0.01

在 SH-SY5Y 神经细胞这一特定环境下, 我们尝试证明存在上述现象的转录因子确实存在. 我们发现在 Katarzyna 采用 SH-SY5Y 细胞研究哺乳动物时控基因调控机制的过程中, 通过实验测出 17 个转录因子参与神经系统的时控基因调控^[13]. 与之相对照, 本文发现的 SH-SY5Y 细胞中明显受甲基化影响的前 10 个转录因子中, 除 Whn 和 KROX 外, 有 8 个转录因子出现在 Katarzyna 的研究结果中. 而 Whn(亦称 Foxn1) 和 KROX 也分别在吴云舟和 Merrill 采用 SH-SY5Y 细胞所作的研究中提及^[14-15].

3 结 论

本文提出一个通用模型来描述基因启动子 DNA 甲基化对转录因子结合的抑制作用, 然后利用转录因子在基因启动子上的结合值与基因表达的相关性分析对模型参数进行求取, 从而实现基因启动子甲基化对转录因子结合的抑制作用分析. 采用上述方法, 在 SH-SY5Y 神经细胞中, 共发现 91 个转录因子, 其在基因启动子上的结合明显受到启动子甲基化的影响.

此外, 本文所提出的基因启动子甲基化对转录

因子结合的作用模型可以为准确判定转录因子在基因组中的结合位点提供了一个研究途径, 并为进一步研究转录因子的调控基因以及基因转录的调控机制提供了一个研究思路.

不过, 我们也应认识到, 细胞中基因调控是一个非常复杂的过程, 本文所研究的基因启动子甲基化对转录因子结合的抑制作用仍然只是影响基因表达的部分因素. 对于基因调控机制的分析, 还有很多工作要做.

参 考 文 献

- [1] Jones P, Takai D. The role of DNA methylation in mammalian epigenetics. *Science*, 2001, **293**(5532): 1068-1070
- [2] Zilberman D, Gehring M, Tran R K, *et al.* Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genetics*, 2006, **39**(11): 61-69
- [3] Robertson K. DNA methylation and human disease. *Nature Reviews Genetics*, 2005, **6**(8): 597-610
- [4] Banelli B, Vinci A, Gelvi I, *et al.* DNA methylation in neuroblastic tumors. *Cancer Letters*, 2005, **228**(1-2): 37-41
- [5] Liu J, Morgan M, Hutchison K, *et al.* A study of the influence of sex on genome wide methylation. *PLoS One*, 2010, **5**(4): 1-8
- [6] Daniela P, Dennis S, Peter W J R, *et al.* Interplay between DNA

- methylation and transcription factor availability: implications for developmental activation of the mouse myogenin gene. *Mol Cell Biol*, 2010, **30**(15): 3805–3815
- [7] Hertzberg L, Izraeli S, Domany E, *et al.* STOP: searching for transcription factor motifs using gene expression. *Bioinformatics*, 2007, **23**(14): 1737–1743
- [8] Guohua W, Yadong W, Weixing F, *et al.* Transcription factor and microRNA regulation in androgen-dependent and -independent prostate cancer cells. *BMC Genomics*, 2008, **9**(suppl 2): S22
- [9] Peddada S, Yasui D, LaSalle J, *et al.* Inhibitors of differentiation (ID1, ID2, ID3 and ID4) genes are neuronal targets of MeCP2 that are elevated in Rett syndrome. *Human Molecular Genetics*, 2006, **15**(12): 2003–2014
- [10] Yasui D, Peddada S, Bieda M, *et al.* Integrated epigenomic analyses of neuronal MeCP2 reveal a role for long-range interaction with active genes. *Proc Natl Acad Sci USA*, 2007, **104** (49): 19416–19421
- [11] Thomas A D, Vardhman K R, Daniel J T, *et al.* A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature Biotechnology*, 2008, **26**(7): 779–785
- [12] Kim K, Oh M, Ki H, *et al.* Identification of E2F1 as a positive transcriptional regulator for δ -catenin. *Biochem Biophys Res Communications*, 2008, **369**(2): 414–420
- [13] Katarzyna B, Angela R, Szymon M K, *et al.* Regulation of clock-controlled genes in mammals. *PLoS ONE*, 2009, **4**(3): e4882
- [14] 吴云舟, 郝景波, 李德山, 等. 应用新城疫病毒治疗肿瘤的研究进展, *生物工程学报*, 2010, **26**(8): 1031–1036
Wu Y Z, Hao J B, Li D S, *et al.* *Chin J Biotechnology*, 2010, **26**(8): 1031–1036
- [15] Merrill R A, Ahrens J M, Kaiser M E, *et al.* All-trans retinoic acid-responsive genes identified in the human SH-SY5Y neuroblastoma cell line and their regulated expression in the nervous system of early embryos. *Biological Chemistry*, 2004, **385**(7): 605–614

Method to Analyze Gene Promoter Methylation Inhibition Effect on Binding of Transcription Factors*

FENG Wei-Xing^{1,2)**}, WANG Ke-Jun²⁾, HE Bo²⁾, LI Xia^{1)**}

⁽¹⁾ College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China;

²⁾ Automation College, Harbin Engineering University, Harbin 150001, China)

Abstract DNA methylation is identified as an elaborate epigenetic element to regulate binding of transcription factor to gene promoter region. With latest highthroughput technology, it is convenient to accurately test methylation level in experiment, which opens a door to investigate how methylation affects transcription factor. A general model is presented to sense methylation effect on transcription factor in a specific cell. In the model, an inverse sigmoid function is adopted to depict effect of DNA methylation to binding ability of transcription factors with two parameters as center C and steepness S . For each transcription factor, the parameters of model can be fixed by analysis of relativity between transcription factor binding scores in promoter regions and gene expression levels. Here three relativity values should be computed while different formula is used to calculate transcription factor binding score. Relativity value A is obtained when transcription factor binding scores are calculated without considering methylation effect. Relativity value B is analyzed from transcription factor binding scores considering methylation effect with the proposed model. On the contrary, normal sigmoid function is used to depict effect of DNA methylation and relativity value C is just calculated with transcription factor binding scores considering methylation effect using such model. For a transcription factor, if relativity value B is found obviously larger than relativity value A and relativity value C is always less than relativity value A, the transcription factor can be figured out to be apparently affected by DNA methylation and the model with optimal fixed parameters can be used to depict the methylation effect. In neuroblastoma cell, with the proposed model, 10 transcriptional factors were found to be apparently affected by methylation of promoter regions which proves the effectiveness of the model. Based on the proposed model, TF binding status in genome promoter region can be presumed to further investigate how a gene is regulated by a specific group of TFs organized in a particular pattern, which should be helpful in building of gene regulation network. Moreover, if it is exactly obtained how methylation exerts quite different effect on the same TF between normal and cancer cells, the proposed method also can be used as a way to search mutant TFs which would be responsible for the cancer. But, as so many factors including DNA methylation are involved in gene transcription process and to investigate regulation mechanism of gene transcription is still a hard job.

Key words gene promoter, transcription factor, DNA methylation, analysis method

DOI: 10.3724/SP.J.1206.2010.00355

*This work was supported by grants from The National Natural Science Foundation of China(61071174), China Postdoctoral Science Foundation (20100471090) and Fundamental Research Funds for the Central Universities (HEUCF100411).

**Corresponding author.

Tel: 86-451-82569906, E-mail: fengweixing@hrbeu.edu.cn

Received: July 5, 2010 Accepted: November 22, 2010