完整基因表达模型中的精确概率分布

献给钱敏教授 90 华诞

仇华海¹, 周天寿^{1,2*}

1. 武汉纺织大学数学与计算机学院, 武汉 430073;

2. 中山大学数学学院和广东省计算数学重点实验室, 广州 510275

E-mail: qiuhuahai2006@163.com, mcszhtsh@mail.sysu.edu.cn

收稿日期: 2017-05-18; 接受日期: 2017-10-30; 网络出版日期: 2017-12-01; * 通信作者

国家自然科学基金 (批准号: 91530320)、湖北省科技厅自然科学基金 (批准号: 2017CFB682)、湖北省教育厅科研计划 (批准号: B2016062) 和中山大学广东省计算科学重点实验室开放基金 (批准号: 2016010) 资助项目

摘要 基因表达是细胞内部过程的中心环节, 其建模与分析一直是人们关注的焦点. 转录水平上的两类代表性基因表达模型 (即构成式表达模型和爆发式表达模型) 已经被广泛研究, 特别是 mRNA 的分布已经解析地给出, 但对于同时考虑转录和翻译的完整基因表达模型, 如何解析地导出 mRNA 和蛋白质的分布以及它们的联合分布的问题至今未得到解决. 本文应用作者自己发展起来的二项矩方法解决了此数学问题, 尤其是给出了基因产物分布及有关统计量的解析表示. 此外, 本文还讨论了基因表达噪声的可调性.

关键词 基因表达模型 化学主方程 概率分布 二项矩方法 噪声

MSC (2010) 主题分类 92C42, 97M60

1 引言

众所周知, 基因表达过程遵循“中心法则”, 即储存在 DNA 的信息首先转录成信使 RNA (mRNA), mRNA 然后翻译成具有功能的蛋白质. 由于基因表达本质上是单分子事件^[1], 因此, mRNA 和蛋白质丰度的随机涨落 (即所谓的噪声) 是不可避免的, 这蕴含着基因表达是固有噪声的. 这种分子噪声对许多细胞功能来说是本质性的^[2,3], 且已经被认同为具有相同遗传性质的细胞群体呈现出表型差异性的一个关键因素^[4]. 在单细胞和单分子生物实验方法^[5-9] 允许人们直接观察各个细胞内基因表达水平的实时随机涨落的同时, 人们也有相当的兴趣理论地探索基因表达的不同分子机制是如何影响 mRNA 或蛋白质水平的. 事实上, 利用基因表达的理论模型来定量化不同噪声源对基因表达水平的贡献是理解细胞内部过程和细胞异质性的重要一步.

按照表达方式的不同, 基因表达可区分为构成式表达和爆发式表达^[5,10-13] (参考示意图 1), 前者对应于看家基因, 后者对应于组织特异性基因. 在构成式表达情形 (图 1(a)), 假设基因总是处于活性 (on) 状态, 从 DNA 到 mRNA 的转录率记为 μ , 从 mRNA 到蛋白质的翻译率记为 ρ , mRNA 和蛋白质

英文引用格式: Qiu H H, Zhou T S. Exact distributions in the full models of gene expression (in Chinese). Sci Sin Math, 2017, 47: 1863–1878, doi: 10.1360/N012017-00105

图 1 两种代表性的基因表达模型示意图: (a) 构成式基因表达模型; (b) 爆发式基因表达模型

的降解率分别记为 δ 和 γ . 在爆发式表达情形 (图 1(b)), 假设基因启动子有两个状态^[14-20]: 活性 (on) 状态和非活性 (off) 状态; 基因仅处在 on 状态时才会表达, 处在 off 状态时不会表达; 两个状态之间的转移率分别记为 α 和 β , 转录率记为 μ , 翻译率记为 ρ , mRNA 和蛋白质的降解率分别记为 δ 和 γ . 在以下的分析中, (i) 不考虑转录因子的动态调控但反应比率常数可在生物合理的范围内变化; (ii) 只考虑线性降解; (iii) 假设所有的反应事件都是马氏的.

特征化随机性的最好方式是概率分布. 对于上述两种基因模型, 若仅考虑转录 (或翻译) 或把转录和翻译简化为单步, 则基因产物的分布已经被解析地导出, 例如, 对于构成式基因表达, 若不考虑翻译过程, 则 mRNA 的静态分布为 $P(m) = \frac{1}{m!}(\frac{\mu}{\delta})^m e^{-\mu/\delta}$, 即 mRNA 服从 Poisson 分布; 而对于爆发式基因表达, 若不考虑翻译过程, 则 mRNA 的静态分布为 (参见文献 [6, 21, 22])

$$P(m) = \frac{\Gamma(\frac{\alpha+\beta}{\delta})}{\Gamma(\frac{\alpha}{\delta})} \frac{\Gamma(m + \frac{\alpha}{\delta})}{\Gamma(m + \frac{\alpha+\beta}{\delta})} \frac{(\frac{\mu}{\delta})^m}{m!} {}_1F_1\left(m + \frac{\alpha}{\delta}, m + \frac{\alpha+\beta}{\delta}; -\frac{\mu}{\delta}\right), \quad (1)$$

其中, m 代表 mRNA 的分子数目, $P(m)$ 代表 m 的分布, $\Gamma(\cdot)$ 是普通的伽马函数. 在表达式 (1) 中, 符号 ${}_1F_1(a, b; z)$ 是 Kummer 类型的超几何函数, 其定义为 ${}_1F_1(a, b; z) = \sum_{n=0}^{\infty} \frac{(a)_n}{(b)_n} \frac{z^n}{n!}$, 其中, $(c)_n$ 是 Pochhammer 符号, 其定义为 $(c)_n = c(c+1)\cdots(c+n-1)$. 在有了概率分布的解析表达式之后, 很容易分析 mRNA 的某些统计量, 例如, 它的平均 (或期望) 和噪声强度 (定义为方差与期望的平方之比) 以及分布的峰度和偏度等的特征 (看后面的计算公式). 尽管这样, 但是对于同时考虑转录和翻译过程的基因模型 (图 1(b)), 还没有人解析地导出 mRNA 分布、蛋白质分布以及它们的联合分布, 这是促动本文研究的主要原因.

为清楚起见, 首先写出相应基因模型的化学主方程^[23, 24], 其形式为

(i) 构成式:

$$\begin{aligned} \frac{dP(m, n; t)}{dt} = & \mu(E_m^{-1} - I)[P(m, n; t)] + m\rho(E_n^{-1} - I)[P(m, n; t)] \\ & + \delta(E_m - I)[mP(m, n; t)] + \gamma(E_n - I)[nP(m, n; t)], \end{aligned} \quad (2a)$$

(ii) 爆发式:

$$\begin{aligned} \frac{dP(m, n; t)}{dt} = & \mathbf{A}P(m, n; t) + \Phi(\mathbf{E}_m^{-1} - \mathbf{I})[P(m, n; t)] + m\Psi(\mathbf{E}_n^{-1} - \mathbf{I})[P(m, n; t)] \\ & + D_1(\mathbf{E}_m - \mathbf{I})[mP(m, n; t)] + D_2(\mathbf{E}_n - \mathbf{I})[nP(m, n; t)], \end{aligned} \quad (2b)$$

其中, $\mathbf{P} = (P_1, P_2)^T$ 是一个二维列向量函数 (在本文中, 黑体字母代表向量或矩阵), P_1 和 P_2 分别描述基因在时刻 t 处于 off 和 on 状态时 mRNA 数目 (m) 和蛋白质数目 (n) 的联合概率分布; $\mathbf{E}_m$ 和 $\mathbf{E}_n$ 都是普通的步长算子, 它们的逆分别是 $\mathbf{E}_m^{-1}$ 和 $\mathbf{E}_n^{-1}$, 操作规则如下: 例如, $\mathbf{E}_m \mathbf{E}_n^{-1} [m \mathbf{P}(m, n; t)] = (m+1) \mathbf{P}(m+1, n-1; t)$, 其他情形类似; $\mathbf{I}$ 是单位算子. 在方程 (2b) 中, 矩阵 $\mathbf{A}$ 描述启动子状态之间的转移情况 (将称作启动子矩阵), 它是一个所谓的 $\mathbf{M}$ - 矩阵 (即每一列元素之和为零); 矩阵 Φ 描述基因的转录情况 (因此称作转录矩阵), 它是一个对角矩阵 (由于假设基因在 off 状态不会转录, 因此对角线上的某些元素为零); 矩阵 Ψ 描述 mRNA 的翻译情况, 它也是一个对角矩阵; 矩阵 $\mathbf{D}_1$ 和 $\mathbf{D}_2$ 分别描述 mRNA 和蛋白质的降解, 它们都是对角矩阵. 由于本文仅考虑两状态的启动子, 因此这些矩阵具有下列简单形式:

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} -\alpha & \beta \\ \alpha & -\beta \end{pmatrix}, \quad \Phi = \begin{pmatrix} 0 & 0 \\ 0 & \mu \end{pmatrix}, \quad \Psi = \begin{pmatrix} \rho & 0 \\ 0 & \rho \end{pmatrix}, \\ \mathbf{D}_1 &= \delta \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{D}_2 = \gamma \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \end{aligned} \quad (3)$$

需要指出的是, 由于调控因子 (如转录因子等) 常常调控基因的表达, 因此上述所有涉及的参数或矩阵元素可能依赖于调控子. 当采用这种理解时, 本文仅考虑调控因子的静态调控 (如仅考虑恒定数目的调控子), 此时可认为系统参数取值于有限范围.

注意到: 对于上述两类基因表达模型, 人们最为感兴趣的是 mRNA 和蛋白质的联合概率分布. 对于模型 (2b) 联合概率应理解为

$$P(m, n; t) = P_1(m, n; t) + P_2(m, n; t). \quad (4)$$

两个未解决的问题是, (i) 如何导出 $P(m, n; t)$ 的边缘分布 $P(m; t)$ 和 $Q(n; t)$, 即如何导出 mRNA 和蛋白质的分布? (ii) 如何导出联合分布 $P(m, n; t)$? 本文应用作者自己建立的二项矩方法^[25] 可彻底 (至少在求解方法方面) 解决了这两个数学问题, 此外还讨论了 mRNA 噪声和蛋白质噪声的可调性.

2 二项矩方法简介

为了求解在静态时的上述方程 (2a) 和 (2b), 这里简要介绍二项矩方法^[25]. 首先, 定义概率分布 $P(m, n; t)$ 的二项矩为

$$b_{m,n}(t) = \sum_{i \geq m, j \geq n} \binom{i}{m} \binom{j}{n} P(i, j; t). \quad (5)$$

为了帮助理解这种定义, 让我们考察概率函数 $P(m, n; t)$ 的生成函数, 其定义为

$$G(x, y; t) = \sum_{m,n} x^m y^n P(m, n; t).$$

由于 $P(x, y; t)$ 是一个幂级数的和函数, 因此, 若假设此级数的收敛半径大于 1, 则和函数 $P(x, y; t)$ 反过来可以在点 $(x, y) = (1, 1)$ 处展开成某一 Taylor 级数, 例如, 设

$$G(x, y; t) = \sum_{m,n} \frac{1}{m!n!} \frac{\partial^{m+n} G(1, 1; t)}{\partial x^m \partial y^n} (x-1)^m (y-1)^n,$$

又设 $P(m, n; t)$ 的二项矩为 $b_{m,n}(t)$, 那么不难证实 (见附录 A)

$$b_{m,n}(t) = \sum_{i \geq m, j \geq n} \binom{i}{m} \binom{j}{n} P(i, j; t) = \frac{\partial^{m+n} G(x, y; t)}{m!n! \partial x^m \partial y^n} \Big|_{x=1, y=1}. \quad (6)$$

简单来说, 二项矩即为概率生成函数在某一点处 Taylor 级数的系数. 其次, 基于方程 (2) 并根据二项矩的定义, 可导出相应的二项矩方程, 即有下列引理:

引理 1 相应于主方程 (2a) 和 (2b) 的二项矩方程分别为

$$\frac{db_{m,n}}{dt} = \mu b_{m-1,n} + (m+1)\rho b_{m+1,n-1} + m\rho b_{m,n-1} - m\delta b_{m,n} - n\gamma b_{m,n} \quad (7a)$$

和

$$\frac{db_{m,n}}{dt} = \mathbf{A}b_{m,n} + \mathbf{\Phi}b_{m-1,n} + (m+1)\mathbf{\Psi}b_{m+1,n-1} + m\mathbf{\Psi}b_{m,n-1} - m\mathbf{D}_1b_{m,n} - n\mathbf{D}_2b_{m,n}, \quad (7b)$$

其中, $\mathbf{b}_{m,n} = (b_{m,n}^{(1)}, b_{m,n}^{(2)})^T$ 是一个二维列向量, 附录 B 给出了导出方程 (7) 的具体细节.

整个二项矩方法分为以下几步:

- (i) 依据感兴趣的生化系统的变量, 定义二项矩;
- (ii) 把有关主方程转化成与之等价的二项矩方程;
- (iii) 一般地, 结果的二项矩方程并不是封闭系统, 为了使之变成一个封闭系统, 需要截取. 由于二项矩当其阶趋于无穷时趋于零或收敛到零, 因此截取是可行的, 而且能够截取;
- (iv) 通过求解截取后的二项矩方程, 可获得近似的二项矩, 再用获得的近似二项矩并根据重构公式来重构一个近似的概率分布函数;
- (v) 在某些情形 (例如, 系统没有反馈, 系统中所有反应式的最高阶均为 1 等), 可以求得静态二项矩方程的解析解 (如本文获得的结果).

此外, 我们指出下列三点: (i) 二项矩方程是相应主方程的一种等价表示, 因此应用起来不需要附加任何条件; (ii) 方程 (7a) 或 (7b) 为导出 mRNA 的分布、蛋白质的分布以及它们的联合分布带来了极大方便, 但对一般的生化反应系统, 精确求解二项矩方程与精确求解相应主方程的难度是一样的 (然而, 二项矩方程可以近似地求解, 看下方程 (8) 的有关内容); (iii) 概率的保守性蕴含着: 对一切时间 t , 零阶二项矩恒等于 1, 例如, 在本文的情形, $b_{0,0}(t) = 1$ 总是成立.

上面引入的二项矩具有许多优势. 例如, 假如通过方程 (7) 已经给出了所有的总的二项矩 $b_{m,n}(t) = b_{m,n}^{(1)}(t) + b_{m,n}^{(2)}(t)$, 那么总的联合概率分布可以由二项矩来重构. 事实上, 有重构公式

$$P(m, n; t) = \sum_{i \geq m, j \geq n} (-1)^{i+j-m-n} \binom{i}{m} \binom{j}{n} b_{i,j}(t). \quad (8)$$

特别地, 可根据二项矩来计算 mRNA 和蛋白质的均值, 以及它们的噪声强度, 例如, 对于 mRNA, 有计算公式

$$\langle m \rangle = b_{1,0}, \quad \eta_m^2 = \frac{2b_{2,0} + b_{1,0} - b_{1,0}^2}{b_{1,0}^2}, \quad (9a)$$

而对于蛋白质, 类似地有

$$\langle n \rangle = b_{0,1}, \quad \eta_n^2 = \frac{2b_{0,2} + b_{0,1} - b_{0,1}^2}{b_{0,1}^2}. \quad (9b)$$

普通矩 (如原点矩和中心矩) 与二项矩存在密切关系, 例如, 蛋白质的第 k 阶中心矩 $\mu_{0,k}(t)$ 与二项矩有如下关系:

$$\mu_{0,k}(t) = [-b_{0,1}(t)]^k + \sum_{i=0}^{k-1} \sum_{j=0}^{k-i} R(k, i, j)(j!)[b_{0,1}(t)]^i b_{0,j}(t), \quad (10)$$

其中, $R(k, i, j) = (-1)^i \binom{k}{i} S(k-i, j)$, 而 $S(n, k) = \sum_{i=0}^k (-1)^{k-i} \binom{k}{i} i^n$ 是第二类 Stirling 数^[26]. 由于非 Gauss 分布的二阶以上的矩不一定为零, 因此, 为了特征化这种类型的分布, 需要计算相应的偏度和峰度, 例如, 对于蛋白质, 偏度和峰度的计算公式分别为

$$\gamma_{0,1}(t) = \frac{\mu_{0,3}(t)}{[\mu_{0,2}(t)]^{3/2}}, \quad \gamma_{0,2}(t) = \frac{\mu_{0,4}(t)}{[\mu_{0,2}(t)]^2} - 3, \quad (11)$$

其中, $\mu_{0,2}(t)$ 、 $\mu_{0,3}(t)$ 和 $\mu_{0,4}(t)$ 由 (10) 给出 (假如有关二项矩是已知). 上述计算公式可以基于二项矩方便地分析 mRNA 和蛋白质的统计指标对参数的依赖性.

3 分析的静态分布: 构成式基因表达模型

这里采用两种方法来导出构成式基因表达模型中的分布. 首先, 介绍二项矩求解方法. 假如考虑静态, 那么方程 (7a) 变成下列迭代形式:

$$(m\delta + n\gamma)b_{m,n} = \mu b_{m-1,n} + (m+1)\rho b_{m+1,n-1} + m\rho b_{m,n-1}, \quad (12)$$

其中, $m, n = 0, 1, 2, \dots$, 若 $m < 0$ 或 $n < 0$, 则定义 $b_{m,n} = 0$. 若在 (12) 中令 $n = 0$, 则容易求得 mRNA 的静态二项矩为

$$b_{m,0} = \frac{1}{m!} \left(\frac{\mu}{\delta} \right)^m, \quad m = 0, 1, 2, \dots, \quad (13)$$

进一步, 根据上面的重构公式 (8), 求得静态 mRNA 的分布为

$$P(m) = \frac{1}{m!} \left(\frac{\mu}{\delta} \right)^m e^{-\mu/\delta}, \quad m = 0, 1, 2, \dots \quad (14)$$

类似地, 可导出蛋白质的分布, 但导出过程较为复杂. 事实上, 若在 (12) 中令 $m = 0$, 则有

$$n\gamma b_{0,n} = \rho b_{1,n-1}, \quad n = 0, 1, 2, \dots, \quad (15)$$

因此, 需要知道 $b_{1,n-1}$. 又若在 (12) 中令 $m = 1$, 则有

$$(\delta + n\gamma)b_{1,n} = \mu b_{0,n} + 2\rho b_{2,n-1} + \rho b_{1,n-1}, \quad n = 0, 1, 2, \dots, \quad (16)$$

这又需要知道 $b_{2,n-1}$. 这里, 省略某些推导细节, 仅给出蛋白质的部分二项矩如下:

$$\begin{aligned} b_{0,1} &= \frac{\tilde{\mu}\tilde{\rho}}{\tilde{\gamma}}, \quad b_{0,2} = \frac{\tilde{\mu}}{2} \left(\frac{\tilde{\rho}}{\tilde{\gamma}} \right)^2 \left(\frac{\tilde{\gamma}}{1+\tilde{\gamma}} + \tilde{\mu} \right), \quad b_{0,3} = \frac{\tilde{\mu}}{6} \left(\frac{\tilde{\rho}}{\tilde{\gamma}} \right)^3 \left[\frac{2\tilde{\gamma}^2}{(1+\tilde{\gamma})(1+2\tilde{\gamma})} + \frac{3\tilde{\mu}\tilde{\gamma}}{1+\tilde{\gamma}} + \tilde{\mu}^2 \right], \\ b_{0,4} &= \frac{\tilde{\mu}}{24} \left(\frac{\tilde{\rho}}{\tilde{\gamma}} \right)^4 \left[\frac{6\tilde{\gamma}^3}{(1+\tilde{\gamma})(1+2\tilde{\gamma})(1+3\tilde{\gamma})} + \frac{\tilde{\mu}\tilde{\gamma}^2(11+14\tilde{\gamma})}{(1+\tilde{\gamma})^2(1+2\tilde{\gamma})} + \frac{6\tilde{\mu}^2\tilde{\gamma}}{1+\tilde{\gamma}} + \tilde{\mu}^3 \right], \end{aligned}$$

其中, $\tilde{\mu} = \mu/\delta$, $\tilde{\gamma} = \gamma/\delta$, $\tilde{\rho} = \rho/\delta$. 类似地, 由迭代式 (12) 还可以给出 mRNA 和蛋白质的联合二项矩, 具体表达式由于复杂性而被省略了.

其次, 介绍另一种求解方程 (2a) 的方法. 考虑静态并引进静态概率函数 $P(m, n)$ 的生成函数 $G(x, y) = \sum_{m,n} x^m y^n P(m, n)$, 则由 (2a) 可导出下列偏微分方程:

$$[\delta(x-1) - \rho x(y-1)] \frac{\partial G}{\partial x} + \gamma(y-1) \frac{\partial G}{\partial y} = \mu(x-1)G, \quad (17)$$

关键是给出 $G(x, y)$ 的解析表达式, 这是因为一旦 $G(x, y)$ 的表达式已知, 那么根据概率分布与生成函数之间的关系, 就容易给出 $P(m, n)$ 的表达式. 为了找到 $G(x, y)$ 的显式表示, 作变换 $x = 1 + u$, $y = 1 + v$, $G(x, y) = e^{H(u, v)}$, 其中函数 $H(u, v)$ 待定, 则方程 (17) 变成

$$[\delta u - \rho(1+u)v] \frac{\partial H}{\partial u} + \gamma v \frac{\partial H}{\partial v} = \mu u, \quad (18)$$

注意到 $G(1, 1) = 1$ 蕴含着 $H(0, 0) = 0$. 现在, 设 $H(u, v) = \sum_{m,n} a_{m,n} u^m v^n$, 则 $a_{0,0} = H(0, 0) = 0$, 且由 (18) 可得下列迭代方程:

$$(m\delta + n\gamma)a_{m,n} = \rho[(m+1)a_{m+1,n-1} + ma_{m,n-1}] + \mu\delta_{m,1}\delta_{n,0}, \quad (19)$$

其中, $\delta_{i,j}$ 的定义为若 $i = j$, 则 $\delta_{i,j} = 1$; 否则, $\delta_{i,j} = 0$. 对 (19), 应用数学归纳法可推出

$$a_{m,n} = 0, \quad m \geq 2, \quad n \geq 0, \quad (20)$$

因此, 关键是给出 $\{a_{0,n}\}_{n \geq 1}$ 和 $\{a_{1,n}\}_{n \geq 0}$ 的表达式. 若在 (19) 中令 $m = 1$, 则有

$$(\delta + n\gamma)a_{1,n} = 2\rho a_{2,n-1} + \rho a_{1,n-1} + \mu\delta_{n,0},$$

由此给出

$$a_{1,n} = \frac{\mu(\rho/\gamma)^n}{\delta(1 + \delta/\gamma)_n}, \quad n \geq 0; \quad (21)$$

又若在 (19) 中令 $m = 0$, 则有 $n\gamma a_{0,n} = \rho a_{1,n-1}$, 由此给出

$$a_{0,n} = \frac{\mu(\rho/\gamma)^n}{\delta n(1 + \delta/\gamma)_{n-1}} = \frac{\mu(\rho/\gamma)^n}{\gamma n(\delta/\gamma)_n}, \quad n \geq 1, \quad (22)$$

由于 (20) 的事实, 因此有 $H(u, v) = \sum_{m,n} a_{m,n} u^m v^n = u \sum_{n \geq 0} a_{1,n} v^n + \sum_{n \geq 1} a_{0,n} v^n$. 利用 (21) 和 (22) 并根据 Kummer 类型的超几何函数的定义, 则有下列解析表示:

$$H(u, v) = \frac{\mu u}{\delta} {}_1F_1\left(1, 1 + \frac{\delta}{\gamma}; \frac{\rho}{\gamma} v\right) + \frac{\mu}{\delta} \frac{\rho}{\gamma} \int_0^v {}_1F_1\left(1, 1 + \frac{\delta}{\gamma}; \frac{\rho}{\gamma} z\right) dz. \quad (23)$$

不难证明这种解析特解是方程 (18) 满足条件 $H(0, 0) = 0$ 的唯一解. 注意到 $G(x, y) = e^{H(u, v)}$, $u = x-1$, $v = y-1$. 这样, 获得生成函数的下列解析表示:

$$G(x, y) = \exp \left\{ \frac{\mu(x-1)}{\delta} {}_1F_1\left(1, 1 + \frac{\delta}{\gamma}; \frac{\rho}{\gamma}(y-1)\right) + \frac{\mu}{\delta} \frac{\rho}{\gamma} \int_0^{y-1} {}_1F_1\left(1, 1 + \frac{\delta}{\gamma}; \frac{\rho}{\gamma} z\right) dz \right\}. \quad (24)$$

下一步, 根据 (24) 来求 mRNA 分布 $P(m)$ 和蛋白质分布 $Q(n)$ (即两个边缘分布). 注意到

$$P(m) = \sum_n P(m, n) = \sum_n \frac{1}{m!n!} \frac{\partial^{m+n} G(x, y)}{\partial x^m \partial y^n} \Big|_{x=0, y=1}$$

$$= \frac{1}{m!} \frac{\partial^m G(x, y)}{\partial x^m} \Big|_{x=0, y=1} = \frac{1}{m!} \left(\frac{\mu}{\delta} \right)^m e^{-\mu/\delta}, \quad (25)$$

即 mRNA 服从 Poisson 分布, 这是已知结果. 类似地, 有

$$Q(n) = \frac{1}{n!} \frac{\partial^n G(x, y)}{\partial y^n} \Big|_{x=1, y=0} = \frac{1}{n!} \frac{\partial^n}{\partial y^n} \left\{ \exp \left[\frac{\mu}{\delta} \frac{\rho}{\gamma} \int_0^{y-1} {}_1F_1 \left(1, 1 + \frac{\delta}{\gamma}; \frac{\rho}{\gamma} z \right) dz \right] \right\} \Big|_{y=0}. \quad (26)$$

尽管这样, 但是蛋白质分布并没有漂亮的解析表达式. 然而, 可以利用 (24) 来给出蛋白质的最初几个二项矩的解析表示, 例如,

$$\begin{aligned} b_{0,1} &= \frac{\partial G(x, y)}{\partial y} \Big|_{x=1, y=1} = \frac{\mu\rho}{\delta\gamma}, \quad b_{0,2} = \frac{1}{2!} \frac{\partial^2 G(x, y)}{\partial y^2} \Big|_{x=1, y=1} = \frac{\mu}{2\delta} \left(\frac{\rho}{\gamma} \right)^2 \left(\frac{\gamma}{\delta + \gamma} + \frac{\mu}{\delta} \right), \\ b_{0,3} &= \frac{1}{3!} \frac{\partial^3 G(x, y)}{\partial y^3} \Big|_{x=1, y=1} = \frac{\mu}{6\delta} \left(\frac{\rho}{\gamma} \right)^3 \left[\frac{2\gamma^2}{(\delta + \gamma)(\delta + 2\gamma)} + \frac{3\gamma}{\delta + \gamma} \frac{\mu}{\delta} + \left(\frac{\mu}{\delta} \right)^2 \right], \\ b_{0,4} &= \frac{1}{4!} \frac{\partial^4 G(x, y)}{\partial y^4} \Big|_{x=1, y=1}, \\ &= \frac{\mu}{24\delta} \left(\frac{\rho}{\gamma} \right)^4 \left[\frac{6\gamma^3}{(\delta + \gamma)(\delta + 2\gamma)(\delta + 3\gamma)} + \frac{\gamma^2(11\delta + 14\gamma)}{(\delta + \gamma)^2(\delta + 2\gamma)} \frac{\mu}{\delta} + \frac{6\gamma}{\delta + \gamma} \left(\frac{\mu}{\delta} \right)^2 + \left(\frac{\mu}{\delta} \right)^3 \right]. \end{aligned}$$

这些结果与前面直接使用二项矩方法导出的结果完全相同.

4 分析的静态分布: 爆发式基因表达模型

在静态处, 上面的二项矩方程 (7b) 变为下列迭代格式:

$$(m\mathbf{D}_1 + n\mathbf{D}_2 - \mathbf{A})\mathbf{b}_{m,n} = \Phi\mathbf{b}_{m-1,n} + (m+1)\Psi\mathbf{b}_{m+1,n-1} + m\Psi\mathbf{b}_{m,n-1}, \quad (27)$$

其中, $m, n = 0, 1, 2, \dots$, 并定义 $\mathbf{b}_{m,-1} = \mathbf{0}$ 和 $\mathbf{b}_{-1,n} = \mathbf{0}$. 展开 (27) 后得

$$(m\delta + n\gamma + \alpha)b_{m,n}^{(1)} - \beta b_{m,n}^{(2)} = (m+1)\rho b_{m+1,n-1}^{(1)} + m\rho b_{m,n-1}^{(1)}, \quad (28a)$$

$$-\alpha b_{m,n}^{(1)} + (m\delta + n\gamma + \beta)b_{m,n}^{(2)} = \mu b_{m-1,n}^{(2)} + (m+1)\rho b_{m+1,n-1}^{(2)} + m\rho b_{m,n-1}^{(2)}, \quad (28b)$$

(28a) 和 (28b) 两边对应相加可得

$$b_{m,n} = \frac{\tilde{\rho}(m+1)}{m+a_n} b_{m+1,n-1} + \frac{\tilde{\rho}m}{m+a_n} b_{m,n-1} + \frac{\tilde{\mu}}{m+a_n} b_{m-1,n}^{(2)}, \quad (29a)$$

因此, 要给出二项矩的表达式, 关键是给出 $b_{m,n}^{(2)}$. 从 (28a) 和 (28b) 可分别得

$$b_{m,n}^{(1)} = \frac{m\delta + n\gamma + \beta}{\alpha} b_{m,n}^{(2)} - \frac{\mu}{\alpha} b_{m-1,n}^{(2)} - \frac{(m+1)\rho}{\alpha} b_{m+1,n-1}^{(2)} - \frac{m\rho}{\alpha} b_{m,n-1}^{(2)}, \quad (29b)$$

$$\begin{aligned} b_{m,n}^{(2)} &= \frac{\tilde{\mu}(m+g_n)}{(m+a_n)(m+c_n)} b_{m-1,n}^{(2)} + \frac{2\tilde{\rho}(m+1)(m+d_n)}{(m+a_n)(m+c_n)} b_{m+1,n-1}^{(2)} + \frac{2\tilde{\rho}(m+e_n)(m+f_n)}{(m+a_n)(m+c_n)} b_{m,n-1}^{(2)} \\ &\quad - \frac{\tilde{\rho}(m+1)}{(m+a_n)(m+c_n)} \left[(m+2)b_{m+2,n-2}^{(2)} + (m+1)b_{m+1,n-2}^{(2)} \right] \\ &\quad - \frac{\tilde{\rho}m}{(m+a_n)(m+c_n)} \left[\tilde{\mu}b_{m-1,n-1}^{(2)} + \tilde{\rho}(m+1)b_{m+1,n-2}^{(2)} + \tilde{\rho}mb_{m,n-2}^{(2)} \right], \end{aligned} \quad (29c)$$

其中, $a_n = n\tilde{\gamma}$, $g_n = n\tilde{\gamma} + \tilde{\alpha}$, $c_n = n\tilde{\gamma} + \tilde{\alpha} + \tilde{\beta}$, $d_n = n\tilde{\gamma} + \frac{1+\tilde{\alpha}+\tilde{\beta}-\tilde{\gamma}}{2}$, $e_n + f_n = n\tilde{\gamma} + \frac{\tilde{\alpha}+\tilde{\beta}-\tilde{\gamma}-\tilde{\mu}}{2}$, $e_n f_n = -\frac{\tilde{\mu}}{2}$, $\tilde{\alpha} = \alpha/\delta$, $\tilde{\beta} = \beta/\delta$, $\tilde{\mu} = \mu/\delta$, $\tilde{\gamma} = \gamma/\delta$, $\tilde{\rho} = \rho/\delta$.

将看到: 当通过 (29c) 给出 $b_{m,n}^{(2)}$ 的表达式时, 需要知道 $b_{0,0}^{(2)}$ 的表达式. 为此, 有下列引理:

引理 2 在任意时刻, 零阶二项矩都满足下列等式:

$$b_{0,0}^{(1)} = \frac{\beta}{\alpha + \beta}, \quad b_{0,0}^{(2)} = \frac{\alpha}{\alpha + \beta}, \quad (30)$$

事实上, 由方程 (27) 和概率的保守性条件分别可得

$$\mathbf{A}b_{0,0} = \mathbf{0}, \quad \mathbf{u}b_{0,0} = 1, \quad (31)$$

其中, $\mathbf{u} = (1, 1)$ 是一个二维行向量. 求解这一代数方程组即得 (30).

为了给出 $b_{m,n}^{(2)}$ 的表达式, 我们建立下列引理:

引理 3 对于一般的迭代系统 $z_n = x_n z_{n-1} + y_n$, 其中, $n = 1, 2, \dots, z_0$ 以及所有的 x_n 和 y_n 都是已知的, 则

$$z_n = \left(\prod_{j=1}^n x_j \right) z_0 + \sum_{k=1}^n \left(\prod_{i=k+1}^n x_i \right) y_k = \left(\prod_{j=1}^n x_j \right) \left(z_0 + \sum_{k=1}^n y_k \prod_{i=1}^k \frac{1}{x_i} \right). \quad (32)$$

对 (29c) 应用引理 3, 可得

$$\begin{aligned} b_{m,n}^{(2)} = & \frac{\tilde{\mu}^m (1 + g_n)_m}{(1 + a_n)_m (1 + c_n)_m} \left\{ b_{0,n}^{(2)} + \frac{2\tilde{\rho}}{a_n c_n} \sum_{k=1}^m \frac{(k+1)(k+d_n)(a_n)_k (c_n)_k}{\tilde{\mu}^k (1 + g_n)_k} b_{k+1,n-1}^{(2)} \right. \\ & + \frac{2\tilde{\rho}}{a_n c_n} \sum_{k=1}^m \frac{(k+e_n)(k+f_n)(a_n)_k (c_n)_k}{\tilde{\mu}^k (1 + g_n)_k} b_{k,n-1}^{(2)} \\ & - \sum_{k=1}^m \frac{\tilde{\rho}(k+1)(a_n)_k (c_n)_k}{a_n c_n \tilde{\mu}^k (1 + g_n)_k} [(k+2)b_{k+2,n-2}^{(2)} + (k+1)b_{k+1,n-2}^{(2)}] \\ & \left. - \sum_{k=1}^m \frac{k\tilde{\rho}(a_n)_k (c_n)_k}{a_n c_n \tilde{\mu}^k (1 + g_n)_k} [\tilde{\mu} b_{k-1,n-1}^{(2)} + \tilde{\rho}(k+1)b_{k+1,n-2}^{(2)} + \tilde{\rho} k b_{k,n-2}^{(2)}] \right\}. \quad (33) \end{aligned}$$

这是本文的主要结果, 因为有了此结果, 可计算出所有的二项矩 (看下面的内容), 从而计算出 mRNA 和蛋白质的分布以及它们的联合分布. 特别地, 由 (33) 可得

$$b_{m,0}^{(2)} = \frac{\tilde{\mu}^m (1 + g_0)_m}{(1 + a_0)_m (1 + c_0)_m} b_{0,0}^{(2)} = \frac{\tilde{\mu}^m (\tilde{\alpha} + 1)_m}{m! (\tilde{\alpha} + \tilde{\beta} + 1)_m} b_{0,0}^{(2)}, \quad m = 1, 2, \dots, \quad (33a)$$

又从 (29a) 可得 $(m + a_0)b_{m,0} = \tilde{\mu} b_{m-1,0}^{(2)}$. 再利用 (33a) 得

$$b_{m,0} = \frac{\tilde{\mu}}{m + a_0} b_{m-1,0}^{(2)} = \frac{\tilde{\mu}}{m + a_0} \frac{\tilde{\mu}^{m-1} (1 + g_0)_{m-1}}{(1 + a_0)_{m-1} (1 + c_0)_{m-1}} b_{0,0}^{(2)}.$$

这样, 若注意到 (30), 则获得 mRNA 的各阶二项矩为

$$b_{m,0} = \frac{\tilde{\mu}^m}{m!} \frac{(\tilde{\alpha})_m}{(\tilde{\alpha} + \tilde{\beta})_m}, \quad (34)$$

再根据上面的重构公式 (8), 由此可得 mRNA 的静态概率分布为

$$P(m) = \frac{\Gamma(\tilde{\alpha} + \tilde{\beta})}{\Gamma(\tilde{\alpha})} \frac{\Gamma(m + \tilde{\alpha})}{\Gamma(m + \tilde{\alpha} + \tilde{\beta})} \frac{\tilde{\mu}^m}{m!} {}_1F_1(m + \tilde{\alpha}, m + \tilde{\alpha} + \tilde{\beta}; -\tilde{\mu}), \quad (35)$$

若恢复原来的参数, 则结果与上面的表达式 (1) 完全相同.

下一步来导出蛋白质的二项矩的解析表示, 但导出过程更为复杂. 从 (29a) 知 $b_{0,1} = \frac{\tilde{\rho}}{a_1} b_{1,0}$. 这样, 利用 (30) 得蛋白质的一阶二项矩为

$$b_{0,1} = \frac{\tilde{\rho}\tilde{\mu}}{a_1} \frac{\tilde{\alpha}}{\tilde{\alpha} + \tilde{\beta}} = \frac{\rho\mu\alpha}{\gamma\delta(\alpha + \beta)}, \quad (36)$$

又从 (29a) 知, $b_{0,2} = \frac{\tilde{\rho}}{a_2} b_{1,1}$, 因此需要知道 $b_{1,1}$. 再从 (29a) 知, $b_{1,1} = \frac{2\tilde{\rho}}{1+a_1} b_{2,0} + \frac{\tilde{\rho}}{1+a_1} b_{1,0} + \frac{\tilde{\mu}}{1+a_1} b_{0,1}^{(2)}$, 因此又需要知道 $b_{0,1}^{(2)}$. 从 (29c) 知,

$$b_{0,n}^{(2)} = \frac{2\tilde{\rho}d_n}{a_n c_n} b_{1,n-1}^{(2)} + \frac{2\tilde{\rho}e_n f_n}{a_n c_n} b_{0,n-1}^{(2)} - \frac{\tilde{\rho}}{a_n c_n} (2b_{2,n-2}^{(2)} + b_{1,n-2}^{(2)}), \quad n = 1, 2, \dots, \quad (37)$$

特别地, $b_{0,1}^{(2)} = \frac{2\tilde{\rho}d_1}{a_1 c_1} b_{1,0}^{(2)} + \frac{2\tilde{\rho}e_1 f_1}{a_1 c_1} b_{0,0}^{(2)}$. 若把这种表示代入 $b_{1,1}$ 并利用 (34), 则首先获得 mRNA 和蛋白质的交叉二项矩 (代表它们之间的一阶关联)

$$b_{1,1} = \frac{\tilde{\mu}^2 \tilde{\rho} \tilde{\alpha}}{(1 + \tilde{\gamma})(\tilde{\alpha} + \tilde{\beta})(1 + \tilde{\alpha} + \tilde{\beta})} \left[1 + \tilde{\alpha} + \frac{1 + \tilde{\alpha} + \tilde{\beta}}{\tilde{\mu}} + \frac{\tilde{\alpha}\tilde{\beta} + (1 + \tilde{\alpha})(\tilde{\gamma} + \tilde{\alpha})}{\tilde{\gamma}(\tilde{\gamma} + \tilde{\alpha} + \tilde{\beta})} \right], \quad (38a)$$

然后获得蛋白质的二阶二项矩

$$b_{0,2} = \frac{\tilde{\mu}^2 \tilde{\rho}^2 \tilde{\alpha}}{2\tilde{\gamma}(1 + \tilde{\gamma})(\tilde{\alpha} + \tilde{\beta})(1 + \tilde{\alpha} + \tilde{\beta})} \left[1 + \tilde{\alpha} + \frac{1 + \tilde{\alpha} + \tilde{\beta}}{\tilde{\mu}} + \frac{\tilde{\alpha}\tilde{\beta} + (1 + \tilde{\alpha})(\tilde{\gamma} + \tilde{\alpha})}{\tilde{\gamma}(\tilde{\gamma} + \tilde{\alpha} + \tilde{\beta})} \right], \quad (38b)$$

再来给出蛋白质的三阶二项矩. 从 (29a) 可知 $b_{0,3} = \frac{\tilde{\rho}}{a_3} b_{1,2}$, 因此需要知道 $b_{1,2}$. 但从 (29a) 可得 $b_{1,2} = \frac{2\tilde{\rho}}{1+a_2} b_{2,1} + \frac{\tilde{\rho}}{1+a_2} b_{1,1} + \frac{\tilde{\mu}}{1+a_2} b_{0,2}^{(2)}$ 和 $b_{2,1} = \frac{3\tilde{\rho}}{2+a_1} b_{3,0} + \frac{2\tilde{\rho}}{2+a_1} b_{2,0} + \frac{\tilde{\mu}}{2+a_1} b_{1,1}^{(2)}$. 这样, 把上述各表达式代入 $b_{0,3} = \frac{\tilde{\rho}}{a_3} b_{1,2}$ 后得蛋白质的三阶二项矩

$$b_{0,3} = \frac{\tilde{\rho}}{a_3(1 + a_2)(2 + a_1)} [6\tilde{\rho}^2 b_{3,0} + 4\tilde{\rho}^2 b_{2,0} + \tilde{\rho}(2 + a_1) b_{1,1} + 2\tilde{\mu}\tilde{\rho} b_{1,1}^{(2)} + \tilde{\mu}(2 + a_1) b_{0,2}^{(2)}], \quad (39)$$

因此需要知道 $b_{1,1}^{(2)}$ 和 $b_{0,2}^{(2)}$. 然而, 从 (29c) 可知,

$$b_{1,1}^{(2)} = \frac{\tilde{\mu}(1 + g_1)}{(1 + a_1)(1 + c_1)} b_{0,1}^{(2)} + \frac{4\tilde{\rho}(1 + d_1)}{(1 + a_1)(1 + c_1)} b_{2,0}^{(2)} + \frac{2\tilde{\rho}(1 + e_1)(1 + f_1)}{(1 + a_1)(1 + c_1)} b_{1,0}^{(2)} - \frac{\tilde{\rho}\tilde{\mu}}{(1 + a_1)(1 + c_1)} b_{0,0}^{(2)}, \quad (39a)$$

以及从 (33) 得

$$b_{0,2}^{(2)} = \frac{2\tilde{\rho}d_2}{a_2 c_2} b_{1,1}^{(2)} + \frac{2\tilde{\rho}e_2 f_2}{a_2 c_2} b_{0,1}^{(2)} - \frac{\tilde{\rho}}{a_2 c_2} (2b_{2,0}^{(2)} + b_{1,0}^{(2)}), \quad (39b)$$

因此, $b_{1,1}^{(2)}$ 和 $b_{0,2}^{(2)}$ 都是已知的. 由于最后的 $b_{0,3}$ 表达的复杂性, 因此这里省略. 完全类似地可导出其他蛋白质二项矩的表达式. 由于导出过程以及最后表达式的复杂性, 这里也省略.

第三, 导出 mRNA 和蛋白质的联合二项矩. 前面已经导出了 $b_{1,1}$ 的表达式, 见 (38a). 下面导出其他交叉二项矩的表达式. 从 (33) 可得

$$b_{m,1}^{(2)} = \frac{\tilde{\rho}\tilde{\mu}^m(1 + g_1)_m}{(a_1)_{m+1}(c_1)_{m+1}} \times \left[\frac{a_1 c_1}{\tilde{\rho}} b_{0,1}^{(2)} + 2 \sum_{k=1}^m \frac{(k+1)(k+d_1)(a_1)_k(c_1)_k}{\tilde{\mu}^k(1 + g_1)_k} b_{k+1,0}^{(2)} \right]$$

$$+ 2 \sum_{k=1}^m \frac{(k+e_1)(k+f_1)(a_1)_k(c_1)_k}{\tilde{\mu}^k(1+g_1)_k} b_{k,0}^{(2)} - \sum_{k=1}^m \frac{k(a_1)_k(c_1)_k}{\tilde{\mu}^k(1+g_1)_k} b_{k-1,0}^{(2)} \Big], \quad (40)$$

其中,

$$b_{0,1}^{(2)} = \frac{2\tilde{\rho}d_1}{a_1c_1} b_{1,0}^{(2)} + \frac{2\tilde{\rho}e_1f_1}{a_1c_1} b_{0,0}^{(2)} \quad (40a)$$

是已知的. 此时, 从 (29a) 得

$$b_{m,1} = \frac{\tilde{\rho}(m+1)}{m+a_1} b_{m+1,0} + \frac{\tilde{\rho}m}{m+a_1} b_{m,0} + \frac{\tilde{\mu}}{m+a_1} b_{m-1,1}^{(2)}, \quad (40b)$$

其中, $b_{m-1,1}^{(2)}$ 已由 (40) 给出, 而所有的 $b_{k,0}$ 已由 (34) 给出, 因此 $b_{m,1}$ 也是已知的. 又从 (33) 可得

$$\begin{aligned} b_{m,2}^{(2)} = & \frac{\tilde{\rho}\tilde{\mu}^m(1+g_2)_m}{(a_2)_{m+1}(c_2)_{m+1}} \left\{ \frac{a_2c_2}{\tilde{\rho}} b_{0,2}^{(2)} + 2 \sum_{k=1}^m \frac{(k+1)(k+d_2)(a_2)_k(c_2)_k}{\tilde{\mu}^k(1+g_2)_k} b_{k+1,1}^{(2)} \right. \\ & + 2 \sum_{k=1}^m \frac{(k+e_2)(k+f_2)(a_2)_k(c_2)_k}{\tilde{\mu}^k(1+g_2)_k} b_{k,1}^{(2)} - \sum_{k=1}^m \frac{(k+1)(a_2)_k(c_2)_k}{\tilde{\mu}^k(1+g_2)_k} [(k+2)b_{k+2,0}^{(2)} + (k+1)b_{k+1,0}^{(2)}] \\ & \left. - \sum_{k=1}^m \frac{k(a_2)_k(c_2)_k}{\tilde{\mu}^k(1+g_2)_k} [\tilde{\mu}b_{k-1,1}^{(2)} + \tilde{\rho}(k+1)b_{k+1,0}^{(2)} + \tilde{\rho}kb_{k,0}^{(2)}] \right\}, \quad (41) \end{aligned}$$

其中, $b_{0,2}^{(2)}$ 已由 (39b) 给出. 此时, 利用 (29a) 得

$$b_{m,2} = \frac{\tilde{\rho}(m+1)}{m+a_2} b_{m+1,1} + \frac{\tilde{\rho}m}{m+a_2} b_{m,1} + \frac{\tilde{\mu}}{m+a_2} b_{m-1,2}^{(2)}, \quad (42)$$

其中, 所有的 $b_{k,1}$ 已由 (40b) 给出, 而 $b_{m-1,2}^{(2)}$ 已由 (41) 给出, 因此 $b_{m,2}$ 也是已知的.

类似地, 对于固定的 m , 可给出 $n > 2$ 时所有 $b_{m,n}$ 的表达式. 由于导出过程及有关表达的复杂性, 这里就省略了. 假如所有的 $b_{m,n}^{(2)}$ 均以上述间接迭代的方式给出, 那么 mRNA 和蛋白质的联合二项矩 $b_{m,n}$ 由 (29a) 给出. 进一步, mRNA 和蛋白质的联合分布 $P(m, n)$ 由重构公式 (8) 给出.

5 基因表达噪声的可调性

基因表达是固有噪声的, 且常常受到调控子的调控. 生物学家常常用噪声强度或 Fano 因子^[27] (定义为方差除以期望) 来量化这种噪声. 尽管这种简化有时是方便的且是有效的, 但明显是一种近似刻画. 以下, 主要仅针对爆发式表达情形讨论基因表达噪声的可调性.

为了方便分析 mRNA 或蛋白质噪声的性质, 根据上面的公式 (9), 需要知道一阶和二阶二项矩的具体表达式. 由于上面已经给出 mRNA 和蛋白质的分布以及它们的联合分布, 因此给出 mRNA 或蛋白质噪声的解析计算公式并不困难. 以下, 仅列出有关结果.

mRNA 的静态平均 (或期望) 为

$$\langle m \rangle = b_{1,0} = \frac{\mu\alpha}{\delta(\alpha+\beta)}, \quad (43)$$

而 mRNA 噪声的计算公式为

$$\eta_m^2 = \frac{1}{\langle m \rangle} + \frac{\delta\beta}{\alpha(\delta+\alpha+\beta)}. \quad (44)$$

蛋白质的静态平均为

$$\langle n \rangle = b_{0,1} = \frac{\rho\mu\alpha}{\gamma\delta(\alpha+\beta)}, \quad (45)$$

蛋白质的二阶二项矩为

$$b_{0,2} = \frac{\rho^2\mu^2\alpha}{2\gamma\delta(\alpha+\beta)(\delta+\gamma)(\delta+\alpha+\beta)} \left[\frac{\delta+\alpha}{\delta} + \frac{\delta+\alpha+\beta}{\mu} + \frac{\alpha\beta+(\delta+\alpha)(\gamma+\alpha)}{\gamma(\gamma+\alpha+\beta)} \right], \quad (46)$$

这样, 根据蛋白质噪声的计算公式 (9b), 获得蛋白质的噪声强度为

$$\eta_n^2 = \frac{1}{\langle n \rangle} + \frac{\gamma\delta(\alpha+\beta)}{\alpha(\delta+\gamma)(\delta+\alpha+\beta)} \left[\frac{\delta+\alpha+\beta}{\mu} + \frac{\beta}{\alpha+\beta} \left(1 + \frac{\delta}{\gamma+\alpha+\beta} \right) \right], \quad (47)$$

数值计算证实了理论预测, 见图 2.

现在来分析表达噪声的可调性. 首先, 考虑 mRNA 的噪声. 计算公式 (44) 可改写为

$$\eta_m^2 = \frac{1}{\langle m \rangle} + \frac{\delta\langle\tau_{\text{off}}\rangle^2}{\delta\langle\tau_{\text{on}}\rangle\langle\tau_{\text{off}}\rangle + \langle\tau_{\text{on}}\rangle + \langle\tau_{\text{off}}\rangle}, \quad (48)$$

其中, $\langle\tau_{\text{off}}\rangle = 1/\alpha$ 和 $\langle\tau_{\text{on}}\rangle = 1/\beta$ 分别代表基因处于 off 和 on 状态的平均驻留时间. 尽管这两个时间都是实验可测量的量, 但是由于调控子的调控, 因此在理论上可假设在某些范围内变化, 导致在平均 mRNA 固定的情形下 mRNA 噪声是可调性的. 计算公式 (48) 蕴含着: 对于上述两状态基因模型, mRNA 的分子数目总是服从超 Poisson 分布^[27].

其次, 考虑蛋白质噪声的可调性. 相应的 Fano 因子 (定义为方差除以平均) 为

$$FF = 1 + \frac{\mu\rho}{(\delta+\gamma)(\delta+\alpha+\beta)} \left[\frac{\delta+\alpha+\beta}{\mu} + \frac{\beta}{\alpha+\beta} \left(1 + \frac{\delta}{\gamma+\alpha+\beta} \right) \right] > 1. \quad (49)$$

这蕴含着蛋白质同 mRNA 一样总是服从超 Poisson 分布, 独立于翻译过程和蛋白质的降解过程. 对于固定的平均蛋白质数目, 假如 mRNA 的降解率 δ 相对于其他参数而言很大, 那么由 (47) 可知蛋白质的噪声强度可近似为

$$\eta_n^2 \approx \frac{1}{\langle n \rangle} + \frac{\gamma(\alpha+\beta)(\mu+\gamma+\alpha+\beta)}{\mu\alpha(\gamma+\alpha+\beta)}. \quad (50)$$

特别地, 假如 γ 相对于其他参数而言很小或假设 $\gamma \approx 0$, 但保持固定的平均蛋白质数目, 则有

$$\eta_n^2 \approx \frac{1}{\langle n \rangle}, \quad (51)$$

图 2 基因表达噪声的可调性: 爆发式表达模型. (a) 翻译率对蛋白质噪声的影响, 其中, $\delta = 100$, $\gamma = 5$; (b) 蛋白质降解率对蛋白质噪声的影响, 其中, $\rho = 50$, $\delta = 100$; (c) mRNA 降解率对蛋白质噪声的影响, 其中, $\rho = 50$, $\gamma = 5$. 图中实线代表理论结果, 空圆代表 Gillespie 随机模拟结果. 其他参数值设为 $\alpha = 3$, $\beta = 1$, $\mu = 200$

这蕴含着, 此时蛋白质分布可退化为简单的 Poisson 分布.

总之, mRNA 和蛋白质一般都服从超 Poisson 分布, 而在某些极端情形, 相应的复杂分布可以退化到简单的 Poisson 分布.

6 简单讨论

由于基因表达式是固有的噪声过程, 而这种随机性的最佳描述方式是概率分布, 因此基因产物的概率分布一直是令人关注的问题. 尽管以前的研究均在某些假设条件下 (例如, 仅考虑转录模型, 或者把转录和翻译简化为单步过程等) 已经导出了某些比较实用的基因产物分布 (如 Gamma 分布^[28] 和 Beta 分布^[10] 等), 且用这些分布较好地解释了某些实验现象, 但是对于同时考虑转录和翻译这种更符合生物实际的基因模型中 mRNA 的分布、蛋白质的分布以及它们的联合分布如何导出等问题并未在数学上得到解决. 本文采用作者自己建立的二项矩方法, 给出了这些分布 (包括有关统计量) 的精确计算公式.

值得指出的是, (i) 尽管本文中的数学模型考虑了启动子状态之间的转移, 但是所获得的结果没有考虑转录因子的动态调控. 由于这种调控常常以随机的、动态的方式既可改变上述模型的某些参数 (如启动子状态之间的转移率) 又可改变启动子的结构, 因此相应的化学主方程一般无法求得解析解 (即使是应用本文介绍的二项矩方法), 而只能借助数值模拟方法, 如 Gillespie 随机模拟方法^[29] 和有限状态映射法^[30]. (ii) 二项矩方法除能够导出某些基因表达调控系统 (如本文所考虑的基因模型) 中的概率分布外, 还可以由此建立求解基因表达系统中联合概率分布的数值算法. 这是可行的, 因为二项矩方程是线性的, 因此通过截取 (由于二项矩当其阶趋于无穷时趋于零^[25]) 可获得封闭系统, 再根据上面给出的重构公式 (8), 可用数值方法获得的二项矩来重构近似的基因产物分布. (iii) 尽管本文仅分析了两类经典的基因表达模型, 但是相应的分析方法很容易推广到启动子结构是一般的情形. 事实上, 生物学实验证实: 甚至在细菌细胞中, 启动子可以包含许多个状态, 例如, 大肠杆菌中噬菌体的 PRM 启动子由于可被两个不同的转录因子 (它们均可绑定到两套三个操作子中的每一个) 所调控, 因此 PRM 启动子状态的数目可以高达 128 个^[31]. 相对地, 真核细胞中的启动子结构可能更复杂 (主要是由于额外地受到组蛋白修饰的表观遗传调控等), 启动子状态的确切数目还是个未知数. (iv) 许多基因表达模型可用排队论来描述^[32-34], 包括建模与分析, 但如何建立排队论意义下基因表达模型的二项矩方程, 以及如何求解有关二项矩方程进而导出有关基因产物分布等问题值得深入研究.

参考文献

- 1 Sanchez A, Golding I. Genetic determinants and cellular constraints in noisy gene expression. *Science*, 2013, 342: 1188–1193
- 2 Boettiger A, Levine M. Synchronous and stochastic patterns of gene activation in the *Drosophila* embryo. *Science*, 2009, 325: 471–473
- 3 Raj A, Rifkin S, Andersen E, et al. Variability in gene expression underlies incomplete penetrance. *Nature*, 2010, 463: 913–918
- 4 Eldar A, Elowitz M. Functional roles for noise in genetic circuits. *Nature*, 2010, 467: 167–173
- 5 Golding I, Paulsson J, Zawilski S, et al. Real-time kinetics of gene activity in individual bacteria. *Cell*, 2005, 123: 1025–1036
- 6 Suter D, Molina N, Gatfield D, et al. Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 2011, 332: 472–474

- 7 Harper C, Finkenzstädt B, Woodcock D, et al. Dynamic analysis of stochastic transcription cycles. *PLoS Biol*, 2011, 9: e1000607
- 8 Spiller D, Wood C, Rand D, et al. Measurement of single-cell dynamics. *Nature*, 2010, 465: 736–745
- 9 Raj A, van Oudenaarden A. Single-molecule approaches to stochastic gene expression. *Annu Rev Biophys*, 2009, 38: 255–70
- 10 Raj A, Peskin C S, Tranchina D, et al. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol*, 2006, 4: e309
- 11 Chubb J R, Treck T, Shenoy S, et al. Transcriptional pulsing of a developmental gene. *Curr Biol*, 2006, 16: 1018–1025
- 12 Yu J, Xiao J, Ren X, et al. Probing gene expression in live cells, one protein molecule at a time. *Science*, 2006, 311: 1600–1603
- 13 Cai L, Friedman N, Xie X. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 2006, 440: 358–362
- 14 Blake W, Kærn M, Cantor C, et al. Noise in eukaryotic gene expression. *Nature*, 2003, 422: 633–637
- 15 Raser J, O'Shea E. Control of stochasticity in eukaryotic gene expression. *Science*, 2003, 304: 1811–1814
- 16 Larson D R. What do expression dynamics tell us about the mechanism of transcription? *Curr Opin Genet Dev*, 2011, 21: 591–599
- 17 周天寿. 基因表达系统的研究进展: 概率分布. *江西师范大学 (自然科学版)*, 2012, 36: 221–229
- 18 Mao C, Brown C, Knox K, et al. Quantitative analysis of the transcription control mechanism. *Mol Syst Biol*, 2013, 6: 431
- 19 Mariani L, Schulz E, Lexberg M, et al. Short-term memory in gene induction reveals the regulatory principle behind stochastic IL-4 expression. *Mol Syst Biol*, 2010, 6: 359
- 20 Miller-Jensen K, Dey S, Schaffer D, et al. Varying virulence: Epigenetic control of expression noise and disease processes. *Trends Biotechnol*, 2011, 29: 517–525
- 21 Zhou T, Zhang J. Analytical results for a multistate gene model. *SIAM J Appl Math*, 2010, 72: 789–818
- 22 Zhang J, Zhou T. Promoter-mediated transcriptional dynamics. *Biophys J*, 2014, 106: 479–488
- 23 van Kampen N. *Stochastic Process in Physics and Chemistry*. Amsterdam: North-Holland, 1992
- 24 周天寿. 概率主方程的研究综述. *江西师范大学学报 (自然科学版)*, 2015, 39: 1–6
- 25 Zhang J, Nie Q, Zhou T. A moment-convergence method for stochastic analysis of biochemical reaction networks. *J Chem Phys*, 2016, 144: 018620
- 26 Balakrishnan N, Johnson N, Kotz S. A note on relationships between moments, central moments and cumulants from multivariate distributions. *Statist Probab Lett*, 1998, 39: 49–54
- 27 He Y, Barkai E. Super- and sub-Poissonian photon statistics for single molecule spectroscopy. *J Chem Phys*, 2005, 122: 184703
- 28 Friedman N, Cai L, Xie X. Linking stochastic dynamics to population distribution: An analytical framework of gene expression. *Phys Rev Lett*, 2006, 97: 168302
- 29 Gillespie D T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J Comput Phys*, 1976, 22: 403–434
- 30 Munsky B, Khammash M. The finite state projection algorithm for the solution of the chemical master equation. *J Chem Phys*, 2006, 124: 044104
- 31 Vilar J, Saiz L. Cplex A: A mathematica package to study macromolecular-assembly control of gene expression. *Bioinformatics*, 2010, 26: 2060–2061
- 32 Kumar N, Singh A, Kulkarni R V. Transcriptional bursting in gene expression: Analytical results for general stochastic models. *PLoS Comput Biol*, 2015, 11: e1004292
- 33 Jia T, Kulkarni R. Intrinsic noise in stochastic models of gene expression with molecular memory and bursting. *Phys Rev Lett*, 2011, 106: 058102
- 34 Liu L, Kashyap B R K, Templeton J G C. On the $GI^X/M/\infty$ system. *J Appl Prob*, 1990, 27: 671–683

附录 A 方程 (6) 的导出

对于函数 $G(x, y; t) = \sum_{m,n} x^m y^n P(m, n; t)$, 有

$$\left. \frac{\partial^{m+n} G(x, y; t)}{m!n! \partial x^m \partial y^n} \right|_{x=1, y=1} = \frac{1}{m!n!} \sum_{i \geq m, j \geq n} \frac{i!}{(i-m)!} \frac{j!}{(j-n)!} P(i, j; t).$$

根据二项矩的定义可知,

$$b_{m,n}(t) = \sum_{i \geq m, j \geq n} \binom{i}{m} \binom{j}{n} P(i, j; t) = \frac{1}{m!n!} \frac{\partial^{m+n} G(x, y; t)}{\partial x^m \partial y^n} \Big|_{x=1, y=1}.$$

附录 B 二项矩方程 (7b) 的导出

对于正文中的概率主方程 (2b), 根据二项矩的定义 (5), 首先有

$$\begin{aligned} \frac{db_{m,n}(t)}{dt} = & \mathbf{A}b_{m,n}(t) + \Phi \sum_{i \geq m, j \geq n} \binom{i}{m} \binom{j}{n} [P(i-1, j; t) - P(i, j; t)] \\ & + \Psi \sum_{i \geq m, j \geq n} i \binom{i}{m} \binom{j}{n} [P(i, j-1; t) - P(i, j; t)] \\ & + D_1 \binom{i}{m} \binom{j}{n} [(i+1)P(i+1, j; t) - iP(i, j; t)] \\ & + D_2 \binom{i}{m} \binom{j}{n} [(j+1)P(i, j+1; t) - jP(i, j; t)]. \end{aligned} \quad (\text{B1})$$

然后, 注意到

$$\begin{aligned} & \sum_{i \geq m, j \geq n} \binom{i}{m} \binom{j}{n} [P(i-1, j; t) - P(i, j; t)] \\ &= \sum_{i \geq m, j \geq n} \left[\binom{i+1}{m} - \binom{i}{m} \right] \binom{j}{n} P(i, j; t) + \sum_{j \geq n} \binom{j}{n} P(m-1, j; t) \\ &= \sum_{i \geq m, j \geq n} \binom{i}{m-1} \binom{j}{n} P(i, j; t) + \sum_{j \geq n} P(m-1, j; t) \\ &= \sum_{i \geq m-1, j \geq n} \binom{i}{m-1} \binom{j}{n} P(i, j; t) \\ &= b_{m-1,n}(t), \\ & \sum_{i \geq m, j \geq n} \binom{i}{m} \binom{j}{n} [(i+1)P(i+1, j; t) - iP(i, j; t)] \\ &= \sum_{i \geq m+1, j \geq n} i \left[\binom{i-1}{m} - \binom{i}{m} \right] \binom{j}{n} P(i, j; t) - m \sum_{j \geq n} P(m, j; t) \\ &= -m \sum_{i \geq m+1, j \geq n} \binom{i}{m-1} \binom{j}{n} P(i, j; t) - m \sum_{j \geq n} P(m, j; t) \end{aligned} \quad (\text{B2})$$

$$\begin{aligned}
&= -m \sum_{i \geq m, j \geq n} \binom{i}{m} \binom{j}{n} P(i, j; t) \\
&= -m b_{m,n}(t).
\end{aligned} \tag{B3}$$

类似地, 可得

$$\sum_{i \geq m, j \geq n} \binom{i}{m} \binom{j}{n} [(j+1)P(i, j+1; t) - jP(i, j; t)] = -n b_{m,n}(t). \tag{B4}$$

为方便, 定义 $I \equiv \sum_{i \geq m, j \geq n} i \binom{i}{m} \binom{j}{n} [P(i, j-1; t) - P(i, j; t)]$, 它可改写为

$$I \equiv \sum_{i \geq m, j \geq n-1} i \binom{i}{m} \binom{j+1}{n} P(i, j; t) - \sum_{i \geq m, j \geq n} i \binom{i}{m} \binom{j}{n} P(i, j; t). \tag{B5}$$

为了化简 (B5), 注意到

$$\binom{j+1}{n} = \binom{j}{n} + \binom{j}{n-1}.$$

因此,

$$\begin{aligned}
I &\equiv \sum_{i \geq m, j \geq n-1} i \binom{i}{m} \left[\binom{j}{n} + \binom{j}{n-1} \right] P(i, j; t) - \sum_{i \geq m, j \geq n} i \binom{i}{m} \binom{j}{n} P(i, j; t) \\
&= \sum_{i \geq m, j \geq n-1} i \binom{i}{m} \binom{j}{n-1} P(i, j; t).
\end{aligned}$$

又注意到

$$\begin{aligned}
i \binom{i}{m} &= (i+1) \binom{i}{m} - \binom{i}{m} = (m+1) \binom{i+1}{m+1} - \binom{i}{m} \\
&= (m+1) \left(\binom{i}{m+1} + \binom{i}{m} \right) - \binom{i}{m} = m \binom{i}{m} + (m+1) \binom{i}{m+1}.
\end{aligned}$$

最后,

$$\begin{aligned}
I &= \sum_{i \geq m, j \geq n-1} i \binom{i}{m} \binom{j}{n-1} P(i, j; t) \\
&= \sum_{i \geq m, j \geq n-1} \left[m \binom{i}{m} + (m+1) \binom{i}{m+1} \right] \binom{j}{n-1} P(i, j; t) \\
&= m \sum_{i \geq m, j \geq n-1} \binom{i}{m} \binom{j}{n-1} P(i, j; t) + (m+1) \sum_{i \geq m+1, j \geq n-1} \binom{i}{m+1} \binom{j}{n-1} P(i, j; t) \\
&= m b_{m,n-1} + (m+1) b_{m+1,n-1}.
\end{aligned}$$

把上述分析结果代入到方程 (B1), 即得方程 (7b).

Exact distributions in the full models of gene expression

QIU HuaHai & ZHOU TianShou

Abstract Gene expression is the core of intracellular processes, and its modeling and analysis has been being a focus. The two-state model of gene expression at the transcription level has been extensively studied, and mRNA distribution has analytically been derived, but questions of, e.g., how mRNA and protein distributions and the joint distribution of mRNA and protein in the full gene model that considers transcription and translation simultaneously, remain unsolved. This paper thoroughly solve these mathematical questions by using a binomial moment approach. In particular, formula for calculating these distributions and their statistical quantities are analytically derived. In addition, tunability of gene expression noise is discussed.

Keywords gene expression model, chemical master equation, probability distribution, binomial moment approach, noise

MSC(2010) 92C42, 97M60

doi: 10.1360/N012017-00105