文章编号: 1000-5862(2012)03-0221-09

基因表达模型的研究进展: 概率分布

周天寿

(中山大学数学与计算科学学院, 广东 广州 510275)

摘要: 定量化基因表达(包括数学建模及定性与定量分析)是理解细胞内部过程的重要一步, 也是当今系统生物学的核心研究内容. 基因表达模型已从最初的单状态简单模型发展到考虑细化生物过程、众多生物因素的多状态复杂模型. 基于生物学的中心法则, 综述了有关基因表达模型的最新研究进展, 聚焦于数学模型的完善、*m*RNA 与蛋白质数目的概率分布等研究方面. 通过综述, 试图总结出有关基因表达的某些一般性规律, 并提出今后需要进一步研究的问题与发展方向.

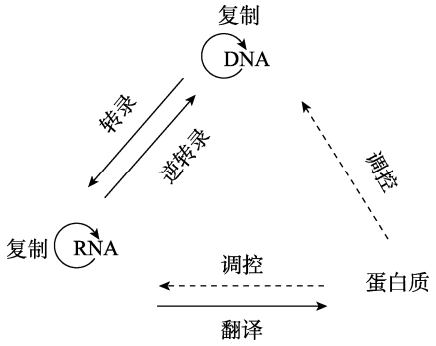
关键词: 基因表达; 基因状态; 生化主方程; 概率密度; 母函数

中图分类号: Q 141　　文献标志码: A

0 引言

基因表达过程是一个复杂的生化过程, 涉及转录、翻译、调控、活性与非活性状态之间的转移、染色质重塑、聚合酶的补充、蛋白质的修饰、DNA 环、甲基化等[1-12]. 本质上, 基因表达是动态和噪声的. 这种分子噪声对细胞功能是重要的, 有益于细胞之间差异性的形成. 在单细胞或单分子实验方法允许对各个活性细胞内基因表达的实时涨落观察的同时, 人们也有相当的兴趣从理论的角度理解基因表达过程中的各种不同机制是如何影响细胞群体水平上细胞之间各 *m*RNA 与蛋白质的差异性. 事实上, 利用基因表达的随机模型定量化分子噪声的不同源是朝着理解基本的细胞过程以及理解群体水平上细胞之间差异性的重要一步.

尽管基因表达过程非常复杂, 但基本的原理还是生物学上的中心法则. 中心法则主要有 2 个版本: 一是克里克(Crick)于 1958 年提出的遗传信息传递的规律, 包括由 DNA 到 DNA 的复制、由 DNA 到 RNA 的转录和由 RNA 到蛋白质的翻译等过程. 20 世纪 70 年代逆转录酶的发现, 表明还有由 RNA 逆转录形成 DNA 的机制, 是对中心法则的补充和丰富; 二是指遗传信息从 DNA 传递给 RNA, 再从 RNA 传递给蛋白质, 即完成遗传信息的转录和翻译过程, 也可以从 DNA 传递给 DNA, 即完成 DNA 的复制过程; DNA 与 RNA 之间遗传信息的传递是双向的(从 RNA 到 DNA 的传递称为反向转录), 而遗传信息只是单向地从核酸流向蛋白质. 所有细胞结构的生物都遵循这种法则. 在某些病毒中的 RNA 自我复制(如烟草花叶病毒等)和能以 RNA 为模板逆转录成 DNA 的过程(某些致癌病毒)是对中心法则的补充, 见图 1.



DNA→DNA (复制), DNA→RNA (转录), RNA–Protein (翻译); RNA→RNA (复制), RNA→DNA (反向转录).

图 1　描述生物学中心法则的示意图

随着基因表达调控系统的深入研究, 已揭示出基于中心法则所描述的调控方式的各种复杂分子机制. 真核细胞的 DNA 首先是由 RNA 聚合酶在转录因子的辅助下转录成信使 *m*RNA, 然后, 通过 *m*RNA 的 5′末端进行 capping(封堵)修饰, 剪切加工去除内含子, 在 3′末端添加 polyA(多聚腺苷酸)尾巴

收稿日期: 2012-03-05

基金项目: 国家自然科学重点基金(60736028)和面上(30973980)资助项目.

作者简介: 周天寿(1962-), 男, 江西新建人, 教授, 博士生导师, 主要从事系统生物学研究.

之后, 使 mRNA 成为成熟的 mRNA, 由相关的输送蛋白将其送到细胞核外. 在细胞质中, mRNA 会同下游的翻译起始因子相结合, 当翻译过程被激活之后, 核糖体将顺着 mRNA 移动, 同时不同的 tRNA 会携带相应的氨基酸结合到核糖体中与 mRNA 上的密码子相互匹配, 而 tRNA 所携带的氨基酸则会有顺序地合成到肽链. 最后, 合成的多肽链会在细胞质中折叠成为一定的构像, 并被其它蛋白质所修饰, 最终成为有功能的成熟蛋白质.

相应地, 基因表达水平的定量化需要发展和建立与实验观察相符合的数学模型. 到目前为止, 有关基因表达水平研究的数学模型可分为下列 5 类:

(i)单状态模型[13-15]: 即假定基因总是处于活性状态, 考虑 DNA 到 mRNA 的转录、mRNA 到蛋白质的翻译. 有时, 仅考虑 DNA 到 mRNA 的转录过程或 DNA 直接到蛋白质过程(而忽视 mRNA 的转录过程, 此时需要假定转录过程比翻译过程快很多);

(ii)双状态模型[16-25]: 即假定基因有 1 个开("ON")状态和 1 个关("OFF")状态且 ON 状态与 OFF 状态之间存在切换, 考虑 DNA 到 mRNA 的转录、mRNA 到蛋白质的翻译. 有时, 仅考虑 DNA 到 mRNA 的转录过程或 DNA 直接到蛋白质过程(而忽视 mRNA 的转录过程, 此时需要假定转录过程比翻译过程快很多);

(iii)多状态模型[26-28]: 即假定基因有若干个 ON 状态和若干个 OFF 状态且 ON 状态与 OFF 状态之间存在多种可能的切换图案, 考虑 DNA 到 mRNA 的转录、mRNA 到蛋白质的翻译. 有时, 仅考虑 DNA 到 mRNA 的转录过程或 DNA 直接到蛋白质过程(而忽视 mRNA 的转录过程, 此时需要假定转录过程比翻译过程快很多).

(iv)自促进模型[13-14, 29-31]: 在前 3 类模型的基础上, 再考虑基因的产物—— 蛋白质作为转录因子促进基因自身的表达, 这种模型叫做基因自促进模型;

(v)自压制模型: 在前 3 类模型的基础上, 再考虑基因的产物—— 蛋白质作为转录因子压制基因自身的表达, 这种模型叫做基因自压制模型.

一般地, 当建立基因表达模型时, 需要考虑染色质模板的结构, 不同的结构导致不同的模型, 如上述 5 类基因模型中的每一种模型均考虑了染色质模板的特定结构. 除了这 5 类模型外, 还有其它基因模型. 本文只综述了这 5 类基因模型中有关概率分布的研究结果.

对于某个给定的基因表达模型, 关注的问题: mRNA 或蛋白质数目的概率分布(包括分析的分布、随机模拟算法); 由有关生化反应所产生的生物噪声对基因表达水平的定性与定量影响; 反馈(假如存在的话, 如上面的第 4、第 5 类模型)对分子噪声的影响与作用; 基因表达的单峰性或双峰性是如何产生的, 等.

本文综述了上述某些方面的研究进展, 从最简单的基因表达模型开始, 介绍 mRNA 或蛋白质的若干常用分布以及介绍导出 mRNA 或蛋白质概率密度分析表达的数学方法. 这种综述将为深入研究基因表达、阐明分子噪声的生物学功能奠定理论基础, 并提供方法论.

1 问题的格式: 数学建模

由于基因表达过程能够由一套生化反应来描述, 而刻画后者的动态变化最常用的模型是化学主方程(或叫做生化主方程). 生化主方程是描述一套生化反应中物种分子数目变化的概率密度函数(probability density function: PDF)的微分方程, 包括 2 种形式: 离散型的主方程与连续型的主方程[32].

1.1 离散的主方程

考虑由 N 个物种 $\{X_1, X_2, \cdots, X_N\}$ 所组成的耦合反应网络, 其状态记为 $\boldsymbol{x} = (x_1, x_2, \cdots, x_N)^{\mathrm{T}}$, 这里 x_i 表示物种 X_i 的分子数目. 构成该网络的 M 个反应由倾向函数 $\boldsymbol{a} = \left(a_1(\boldsymbol{x}), a_2(\boldsymbol{x}), \cdots, a_M(\boldsymbol{x})\right)^{\mathrm{T}}$ 和化学计量矩阵 $\boldsymbol{s} = \left[s_{ij}\right]$ 来描述, 这里 s_{ij} 表示物种 X_j 参加第 i 个反应的分子数目的变化. 记 $\boldsymbol{s}^j = \left(s_{1j}, \cdots, s_{Nj}\right)^{\mathrm{T}}$, 用 $P(\boldsymbol{x}, t)$ 表示物种 X_i 在时刻 t 有 x_i 个分子的概率, 并假定系统状态的变化是一个马氏过程(即当前的状态不依赖于历史), 那么相应的化学主方程[32-33]为

$$\frac{\partial P(\boldsymbol{x};t)}{\partial t} = \sum_{j=1}^{M}\left[a_j\left(\boldsymbol{x}-\boldsymbol{s}^j\right)P\left(\boldsymbol{x}-\boldsymbol{s}^j;t\right) - a_j(\boldsymbol{x})P(\boldsymbol{x};t)\right], \quad (1)$$

这是离散形式的主方程. 假如相应的 x_i 代表第 i 个物种分子的浓度, 那么方程(1)称为连续型的主方程. 注意到静态密度(记为 $P(\boldsymbol{x})$)满足

$$\sum_{j=1}^{M}\left[a_j\left(\boldsymbol{x}-\boldsymbol{s}^j\right)P\left(\boldsymbol{x}-\boldsymbol{s}^j\right) - a_j(\boldsymbol{x})P(\boldsymbol{x})\right] = 0.$$

假如所有物种分子的最大数目已知, 那么方程

(1)能够看出一个庞大的常微分方程(ODE)系统. 例如, 在一个物种情形, 假如它的最大数目为 10, 那么相应的主方程是一个 11 维的 ODE 系统(即 $\boldsymbol{x}$ 可能取值为 0 到 10). 可想而知, 在多个物种情形, 相应的 ODE 系统是多么的庞大.

给定一套生化反应, 写出其离散主方程的关键是确定反应倾向函数和化学计量矩阵. 这里给出一种很一般的写主方程的方法. 考虑一个封闭的体积 Ω (体积的大小亦记为 Ω), 它包含均匀混合的化学物种 X_j $(j=1,2,\cdots,N)$. 让 n_j 表示物种 X_j 的分子数. 一个典型的生化反应式是由一套化学计量系数 s_j, r_j 所决定, 其一般形式为

$$\sum_{j=1}^{m} s_j X_j \underset{k_-}{\overset{k_+}{\rightleftharpoons}} \sum_{j=1}^{m} r_j X_j .$$

这里 $s_j \geqslant 0$, $r_j \geqslant 0$ 是整数. 在化学动力学中, 常习惯以密度或浓度的形式表示物种 $c_j = n_j / \Omega$. 此时, 对于正向反应, 其转移率服从所谓的 Van Hoff 规则 $k_+ \sum_{j=1}^{m} c_j^{s_j}$, 它表示单位体积单位时间内的分子碰撞数目. 这样, 由反应式 $\{n_j\} \to \{n_j + r_j - s_j\}$, 知道第 j 个物种的正向比率方程为

$$\frac{\mathrm{d}c_j}{\mathrm{d}t} = k_+ (r_j - s_j) \prod_{j=1}^{m} (c_j)^{s_j} .$$

类似地, 可写出其反向反应的比率方程. 现在, 容易给出对应于上述反应式的主方程. 让 $P(n;t)$ 表示联合概率密度. 注意到在 $k_+ \sum_{j=1}^{m} c_j^{s_j}$ 中, 对于涉及物种 X_j 的 s_j 个分子的碰撞, 其碰撞的概率正比于因子

$$n_j (n_j - 1)(n_j - 2)\cdots(n_j - s_j + 1) = \frac{n_j!}{(n_j - s_j)!} \equiv \left((n_j)\right)^{s_j},$$

此时, 相应于整个反应式的主方程为

$$\frac{\partial P(n;t)}{\partial t} = k_+ \Omega \left(\prod_{i=1}^{m} E_i^{s_i - r_i} - I \right) \prod_{j=1}^{m} \left\{ \frac{\left((n_j)\right)^{s_j}}{\Omega^{s_j}} \right\} P + k_- \Omega \left(\prod_{i=1}^{m} E_i^{r_i - s_i} - I \right) \prod_{j=1}^{m} \left\{ \frac{\left((n_j)\right)^{r_j}}{\Omega^{r_j}} \right\} P. \tag{2}$$

这里 I 代表恒同算子, E^+ 和 E^- 是平移算子, 其操作规则是 $E^{+k} f(n) = f(n+k)$, $E^{-k} f(n) = f(n-k)$. 方程(2)对于依据生化反应式写出相应的主方程极为有用的, 它是建立随机模型的基础. 对于由多个反应式所组成的系统, 首先根据公式(2)分别写出每个反应对应式的主方程, 然后把这些主方程相加便给出整个系统的主方程.

例 **1**　对于生化反应

$$A \xrightarrow{k_1} X, \quad 2X \xrightarrow{k_2} B,$$

相应的主方程为

$$\frac{\mathrm{d}p_n}{\mathrm{d}t} = k_1 \varphi_A \Omega \left(E^- - I\right) p_n + (k_2/\Omega)\left(E^{+2} - I\right)[n(n-1)p_n],$$

这里 φ_A 表示 A 的浓度, $p_n = P(n;t)$ 代表物种 X 在 t 时刻具有 n 个分子数的概率密度.

例 **2**　考虑酶反应系统

$$E + E^* \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} 2E^*,$$

$$E^* \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} E,$$

它满足保守性条件: $[E] + [E^*] = E_T$ (总浓度, 假定为常数). 用 n 表示 E^* 的分子数目, N_T 表示酶的总数目, 假定为常数. 那么, 相应的主方程

$$\frac{\partial P(n,t)}{\partial t} = -\left[k_1 n(N_T - n) + k_{-1} n(n-1) + k_2 n + k_{-2}(N_T - n)\right] P(n,t) + \left[k_1 (n-1)(N_T - n + 1) + k_{-2}(N_T - n + 1)\right] P(n-1,t) + \left[k_{-1} n(n+1) + k_2 (n+1)\right] P(n+1,t).$$

与概率密度函数密切相关的是母函数(generating function). 对于概率密度 $P(n;t)$, 记其母函数为 $G(z;t)$, 则两者之间的关系是

$$G(z;t) = \sum_{n=0}^{\infty} P(n;t) z^n,$$

反过来, 由母函数可给出概率密度, 即

$$P(n;t) = \frac{1}{n!} \partial_z^n G(z;t)\Big|_{z=0}.$$

对于联合概率密度, 也可类似地引进母函数, 例如, 对于联合概率密度 $P(m,n;t)$, 相应的母函数为

$$G(z', z; t) = \sum_{n=0}^{\infty} P(m,n;t) z'^m z^n.$$

反过来, 有

$$P(m,n;t) = \frac{1}{m!n!} \frac{\partial^{m+1}}{\partial z'^m \partial z^n} G(z', z; t)\Big|_{z'=0, z=0}.$$

1.2　CK 主方程

主方程的另一种形式是 Chapman-Kolmogorov (CK)主方程(需要假定马氏过程, 但连续随机变量可以出现跳跃, 即带跳的随机变量), 其一般形式为

$$\frac{\partial P(\boldsymbol{x},t)}{\partial t}=-\sum_i\frac{\partial}{\partial x_i}\left[A_i(\boldsymbol{x},t)P(\boldsymbol{x},t)\right]+$$
$$\frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i\partial x_j}\left[B_{ij}(\boldsymbol{x},t)P(\boldsymbol{x},t)\right]+$$
$$\int_S\left[W(\boldsymbol{x}|\boldsymbol{z},t)P(\boldsymbol{z},t)-W(\boldsymbol{z}|\boldsymbol{x},t)P(\boldsymbol{x},t)\right]\mathrm{d}\boldsymbol{z},$$

$\boldsymbol{A}(\boldsymbol{x},t)$ 代表漂移向量，$\boldsymbol{B}(\boldsymbol{x},t)$ 代表扩散矩阵，它们的含义为：给定 $\boldsymbol{X}(t)=\boldsymbol{x}$，连续过程的状态增加向量 $\boldsymbol{X}(t+\Delta t)-\boldsymbol{X}(t)$ 以 $\boldsymbol{A}(\boldsymbol{x},t)$ 为平均，以 $\boldsymbol{B}(\boldsymbol{x},t)$ 为方差．$W(\boldsymbol{x}|\boldsymbol{z},t)$ 叫做对应于跳跃($\boldsymbol{z}\to\boldsymbol{x}$)的转移率．下列是几种特殊情形的数学模型：

(i)若 $B_{ij}(\boldsymbol{x},t)=W(\boldsymbol{x}|\boldsymbol{z},t)=W(\boldsymbol{z}|\boldsymbol{x},t)=0$，则 CK 方程变为确定性的刘维尔(Liouville)方程；

(ii)若 $\boldsymbol{A}(\boldsymbol{x},t)=B_{ij}(\boldsymbol{x},t)=0$，则 CK 方程变成描述带跳的且具有不连续道路的主方程；

(iii)若 $W(\boldsymbol{x}|\boldsymbol{z},t)=W(\boldsymbol{z}|\boldsymbol{x},t)=0$，则CK方程变成普通的 Fokker-Planck 方程；

(iv)若 $B_{ij}(\boldsymbol{x},t)=0$，则CK方程变成刘维尔方程．

因此，CK 方程是连续情形下主方程的一般形式．当写这种类型的方程时，关键是理解问题的背景，并确定 $\boldsymbol{A}(\boldsymbol{x},t)$、$\boldsymbol{B}(\boldsymbol{x},t)$ 和 $W(\boldsymbol{x}|\boldsymbol{z},t)$．一般地，求解这种类型的方程是比较困难的，没有有效的方法可循．

为了理解CK方程，这里给出1个例子．考虑下列状态过程：

$$I_1\xrightarrow{\lambda_1}I_2\underset{\lambda_2'}{\overset{\lambda_2}{\rightleftharpoons}}I_3\xrightarrow{\lambda_3}I_1,$$

它构成一个环路，这里 I_i 可理解为状态(如基因的活性或非活性状态)，λ_i ($i=1,2,3$)和 λ_2' 是转移率．让 $P_i(x,t)$ 表示物种如 mRNA、蛋白质、代谢物等(记为 X，假定其降解率为 δ)在 I_i 状态处、t 时刻具有浓度 x 的概率密度，这里 $i=1,2,3$，那么相应的 CK 方程为

$$\frac{\partial P_1(x,t)}{\partial t}+\frac{\partial\left[-\delta xP_1(x,t)\right]}{\partial x}=-\lambda_1P_1(x,t)+\lambda_3P_3(x,t),$$
$$\frac{\partial P_2(x,t)}{\partial t}+\frac{\partial\left[-\delta xP_2(x,t)\right]}{\partial x}=\lambda_1P_1(x,t)+\lambda_2'P_3(x,t)-\lambda_2P_2(x,t),$$
$$\frac{\partial P_3(x,t)}{\partial t}+\frac{\partial\left[-\delta xP_3(x,t)\right]}{\partial x}=\lambda_2P_2(x,t)-(\lambda_3+\lambda_2')\cdot P_3(x,t).$$

感兴趣的问题是：如何给出总的概率密度 $P(x,t)=P_1(x,t)+P_2(x,t)+P_3(x,t)$ 的分析表达，特别是，如何给出其静态概率密度的分析表达?

1.3 一般求解方法

对于离散主方程(1)，引入母函数

$$G(z;t)=\sum_{\boldsymbol{n}=(n_1,\cdots,n_N)}P(\boldsymbol{n};t)\boldsymbol{z}^{\boldsymbol{n}},$$

这里 $\boldsymbol{z}^{\boldsymbol{n}}\equiv z_1^{n_1}z_2^{n_2}\cdots z_N^{n_N}$，则得线性微分方程

$$\frac{\partial}{\partial t}G(z;t)=L_z\left[G(z;t)\right],$$

其中 L_z 是某一 Hilbert 空间 H 上的线性算子，且 L_z 由方程(1)的右端函数决定．利用线性算子理论知道：对于线性算子 L_z，存在 1 组特征值 $\{\lambda_j\}$ 和 1 组相应的特征向量 $\{\varphi_j(z)\}$，使得

$$L_z\left[\varphi_j(z)\right]=\lambda_j\varphi_j(z),$$

这里 $\{\varphi_j(z)\}$ 关于 H 的内积是相互正交的(即 1 组正交基)．利用这种正交基，函数 $G(z;t)$ 在 H 上可表示成

$$G(z;t)=\sum_jG_j(t)\varphi_j(z),$$

这里 $\varphi_j(z)$ 亦叫做函数 $G(z;t)$ 的特征函数．同时利用 $\{\varphi_j(z)\}$ 的正交性，有

$$\frac{\mathrm{d}}{\mathrm{d}t}G_j(t)=\lambda_jG_j(t).$$

由于 $\{\lambda_j\}$ 并不依赖于变量 t，因此函数列 $\{G_j(t)\}$ 容易确定且 $G_j(t)=A_j\mathrm{e}^{\lambda_jt}$．一旦 $\{G_j(t)\}$ 被确定，那么母函数 $G(z;t)$ 就完全确定．这就是求解母函数的一般方法．

在这种方法中，关键是

(i)如何由主方程确定算子 L_z；

(ii)如何计算出算子 L_z 的特征值 $\{\lambda_j\}$ 和特征向量 $\{\varphi_j(z)\}$．

值得指出的是：假如采用量子力学符号[34]来表示上述数学操作，那么所有的运算过程将变得更简洁．

主方程的求解问题(包括求静态解与动态解)并没有解决，然而对于某些特殊类型的主方程，还是可以求解的．在下一节中，将简单地介绍上述 5 类基因模型的求解问题．

2 概率密度的分析表达

2.1 单状态基因模型

假定蛋白质以爆发(burst)的方式生成，这里1个 mRNA分子在它消耗前被翻译成几个蛋白质分子，并假定 mRNA的寿命比蛋白质的寿命要短．此时，

蛋白质的生成能够被2个参数特征化: 每个细胞周期内爆发的平均数目, 即 $a=\nu_0/d_1$; 每个爆发中产生蛋白质分子的平均数目, 即 $b=\nu_1/d_0$, 这里 ν_0 是目标基因转录成*m*RNA的比率, ν_1 是目标*m*RNA翻译成蛋白质的比率, 参数 d_0 和 d_1 分别是*m*RNA和蛋白质的降解率. 让 $x(t)$ 代表1个细胞内蛋白质的浓度, 它是一个连续的随机变量, 且当生成爆发发生时, $x(t)$ 随时间可跳跃地变化. 又记 $P(x)$ 为细胞群体中蛋白质的概率密度函数, 那么它可用连续的主方程(即CK主方程)

$$\frac{\partial P(x)}{\partial t}=\frac{\partial}{\partial x}\left[d_1xP(x)\right]+\nu_0\int_0^x w(x,x')P(x')\mathrm{d}x'$$

来刻画, 这里 $x=n/V$ (V 表示细胞的体积), 第 1 项描述蛋白质的减少, 第 2 项描述爆发时蛋白质的生成, 且设

$$w(x,x')=w(x|x')-\delta(x-x'),$$

$w(x|x')$ 是条件概率(给定爆发前的蛋白质浓度为 x'), δ -函数保守总概率密度(可解释为由于蛋白质的爆发生成远离 x 而导致已有蛋白质浓度的密度损失). 假定爆发的大小 $(x-x')$ 独立于目前的蛋白质浓度 (x'), 并服从某个特征密度 $\nu(x-x')$. 在这些假定下, 能够表示

$$w(x,x')=w(x-x')=\nu(x-x')-\delta(x-x').$$

现在, 假定生成爆发的大小服从指数分布

$$\nu(x)=\frac{1}{b}\mathrm{e}^{-x/b},$$

其中 $b=\nu_1/d_0$. 应用拉普拉斯变换, 可求得分析解[13]

$$P(x)=\frac{1}{b^a\Gamma(a)}x^{a-1}\mathrm{e}^{-x/b}.$$

这说明 x 服从 Gamma 分布, 它是最常用的蛋白质分布.

上述基因模型没有考虑转录过程. 假如补充考虑转录过程, 那么相应的主方程为

(注意到相应的生化反应式为

$$\mathrm{DNA}\xrightarrow{\nu_0}\mathrm{DNA}+m,m\xrightarrow{d_0}\varnothing,m\xrightarrow{\nu_1}n\xrightarrow{d_1}\varnothing)$$

$$\frac{\partial P_{m,n}}{\partial t}=\nu_0\left(P_{m-1,n}-P_{m,n}\right)+\nu_1m\left(P_{m,n-1}-P_{m,n}\right)+$$

$$d_0\left[(m+1)P_{m+1,n}-mP_{m,n}\right]+d_1\left[(n+1)P_{m,n+1}-nP_{m,n}\right],\quad(3)$$

其中 m 和 n 分别代表 *m*RNA 和蛋白质数目, $P_{m,n}$ 代表联合概率密度. 若引进母函数

$$G(z',z;t)=\sum_{m,n}z'^mz^nP_{m,n},$$

则方程(3)变成下列偏微分方程

$$\frac{\partial G}{\partial\upsilon}-\gamma\left[b(1+u)-\frac{u}{\upsilon}\right]\frac{\partial G}{\partial u}+\frac{1}{\upsilon}\frac{\partial G}{\partial\tau}=a\frac{u}{\upsilon}G,$$

这里 $a=\nu_0/d_1$, $b=\nu_1/d_0$, $\gamma=d_0/d_1$, $\tau=d_1t$, $u=z'-1$, $\upsilon=z-1$. 采用特征线求解方程(4), 并利用概率密度与母函数之间的关系可求得蛋白质概率密度的分析表达[17]为

$$P_n(\tau)=\frac{1}{n!}\left(\frac{b}{1+b}\right)^n\left(\frac{1+b\mathrm{e}^{-\tau}}{1+b}\right)^a\cdot$$

$$\frac{\Gamma(a+n)}{\Gamma(a)}{}_2F_1\left(-n,-a;1-a-n;\frac{1+b}{\mathrm{e}^{\tau}+b}\right).\qquad(4)$$

注意这一分析表达仅当 $\gamma\gg1$, $\tau>\gamma^{-1}$, 且 a 和 b 都是有限时才是有效的, 其中 ${}_2F_1(a,b;c;z)$ 是超几何函数. (4)式表明蛋白质数目服从超几何分布. 让 $\tau\to+\infty$, 则得蛋白质的静态概率密度

$$P_n=\frac{\Gamma(a+n)}{\Gamma(n+1)\Gamma(a)}\left(\frac{b}{1+b}\right)^n\left(1-\frac{b}{1+b}\right)^a,$$

这表明蛋白质数目服从负二项分布[35].

2.2　双状态基因模型

单状态基因模型假定基因总是处于活性状态, 但更真实的情况是基因具有两状态: ON 状态和 OFF 状态. 这里, 将导出有关概率密度的分析表达.

假设转录和翻译过程合并成单步, 则简化的两状态基因模型的生化反应式可表示成下列一般格式

$$I\xrightarrow{\lambda}A,\ A\xrightarrow{\gamma}I,$$

$$A\xrightarrow{\mu}A+X,X\xrightarrow{\delta}\varnothing,$$

其中 I 表示基因失活态, A 表示基因激活态, X 代表 *m*RNA 或蛋白质. 当 μ 相对于其它反应比率足够大的时候, 物种(X)的浓度(x)可看作连续变量. 定义 $P_A(x,t)$ 和 $P_I(x,t)$ 分别为 X 处于活性与非活性状态的概率密度, $P=P_A+P_I$ 为 x 的总概率密度 P , 则相应的主方程为

$$\begin{aligned}\frac{\partial P_A}{\partial t}+\frac{\partial J_A}{\partial x}&=\lambda P_I-\gamma P_A,\\ \frac{\partial P_I}{\partial t}+\frac{\partial J_I}{\partial x}&=-\lambda P_I+\gamma P_A,\end{aligned}\qquad(5)$$

其中

$$J_A(x,t)=(\mu-\delta x)P_A(x,t),$$

$$J_I(x,t)=-\delta xP_I(x,t).$$

方程(5)即为基因具有活性和非活性状态时的随

机模型，求解这一方程得静态概率密度[36]

$$P(x)=\left(\frac{\mu}{\delta}\right)^{1-\frac{\lambda}{\delta}-\frac{\gamma}{\delta}}\frac{\Gamma\left((\lambda+\gamma)/\delta\right)}{\Gamma(\lambda/\delta)\Gamma(\gamma/\delta)}x^{\frac{\lambda}{\delta}-1}\left(\frac{\mu}{\delta}-x\right)^{\frac{\gamma}{\delta}-1},$$

这表明 x 服从 Beta 分布，其中 $\Gamma(\cdot)$ 表示 Gamma 函数. 当失活率 γ 远大于激活率 λ，且稍大于 X 的降解速率 δ 时，$P(x)$ 可逼近一个 Gamma 分布，即

$$P(x)=\frac{\gamma/\mu}{\Gamma(\lambda/\delta)}\left((\gamma/\mu)x\right)^{\lambda/\delta-1}\mathrm{e}^{-(\gamma/\mu)x},$$

这时概率密度仅由 λ/δ 和 γ/μ 这 2 个参数决定.

上述模型仅考虑一个过程(转录或翻译)，下面的模型同时考虑了转录与翻译过程. 让 $P_{m,n}^{(0)}$ 表示在 t 时刻且启动子是非活性时有 m 个 mRNA 和 n 个蛋白质，让 $P_{m,n}^{(1)}$ 表示在 t 时刻且 DNA 是活性时有 m 个 mRNA 和 n 个蛋白质. 此时，相应的生化反应式为

非活性(没有转录发生): $A\underset{k_1}{\overset{k_0}{\rightleftharpoons}}I$ (m、n 不变),

$m\xrightarrow{\nu_1}n\xrightarrow{d_1}\varnothing$,

活性: $I\underset{k_0}{\overset{k_1}{\rightleftharpoons}}A\xrightarrow{\nu_0}m\xrightarrow{d_0}\varnothing$, $m\xrightarrow{\nu_1}n\xrightarrow{d_1}\varnothing$,

主方程由 2 个耦合的微分方程构成

$$\begin{aligned}
&\partial P_{m,n}^{(0)}/\partial\tau=\kappa_1P_{m,n}^{(1)}-\kappa_0P_{m,n}^{(0)}+(n+1)P_{m,n+1}^{(0)}-nP_{m,n}^{(0)}+\\
&\gamma\left[(m+1)P_{m+1,n}^{(0)}-mP_{m,n}^{(0)}+bm\left(P_{m,n-1}^{(0)}-P_{m,n}^{(0)}\right)\right],\\
&\partial P_{m,n}^{(1)}/\partial\tau=-\kappa_1P_{m,n}^{(1)}+\kappa_0P_{m,n}^{(0)}+(n+1)P_{m,n+1}^{(1)}-\\
&nP_{m,n}^{(1)}+a\left(P_{m-1,n}^{(1)}-P_{m,n}^{(1)}\right)+\gamma\left[(m+1)P_{m+1,n}^{(1)}-\right.\\
&\left.mP_{m,n}^{(1)}+bm\left(P_{m,n-1}^{(1)}-P_{m,n}^{(1)}\right)\right],
\end{aligned}\tag{6}$$

其中 $\kappa_0=k_0/d_1$，$\kappa_1=k_1/d_1$，$a=\nu_0/d_1$，$b=\nu_1/d_0$，$\gamma=d_0/d_1$. 为了求解此方程，引进 2 个母函数

$$f^{(0)}(z',z)=\sum_{m,n}(z')^m\,z^nP_{m,n}^{(0)},$$

$$f^{(1)}(z',z)=\sum_{m,n}(z')^m\,z^nP_{m,n}^{(1)},$$

则方程(6)变成下列偏微分方程组

$$\begin{aligned}
&\frac{1}{\upsilon}\frac{\partial f^{(0)}}{\partial\tau}=\frac{1}{\upsilon}\left[\kappa_1f^{(1)}-\kappa_0f^{(0)}\right]-\frac{\partial f^{(0)}}{\partial\upsilon}+\\
&\gamma\left[b(1+u)-\frac{u}{\upsilon}\right]\frac{\partial f^{(0)}}{\partial u},\\
&\frac{1}{\upsilon}\frac{\partial f^{(1)}}{\partial\tau}=\frac{1}{\upsilon}\left[-\kappa_1f^{(1)}+\kappa_0f^{(0)}\right]-\frac{\partial f^{(1)}}{\partial\upsilon}+\\
&a\frac{u}{\upsilon}f^{(1)}+\gamma\left[b(1+u)-\frac{u}{\upsilon}\right]\frac{\partial f^{(1)}}{\partial u},
\end{aligned}\tag{7}$$

其中 $u=z'-1$，$\upsilon=z-1$. 尽管方程(7)是线性方程，但其系数并不是常数，因此一般很难找到它的通解. 然而，感兴趣的是静态概率密度，并求得蛋白质的概率密度[17]

$$P_n=\frac{\Gamma(\alpha+n)\Gamma(\beta+n)(\kappa_0+\kappa_1)}{\Gamma(n+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(\kappa_0+\kappa_1+n)}\left(\frac{b}{1+b}\right)^n\cdot$$

$$\left(1-\frac{b}{1+b}\right)^a{}_2F_1\left(\alpha+n,\kappa_0+\kappa_1-\beta;\kappa_0+\kappa_1+n;\frac{b}{1+b}\right),$$

其中 $P=P_n^{(0)}+P_n^{(1)}$，2 个参数为

$$\alpha,\beta=\frac{1}{2}\left(a+\kappa_0+\kappa_1\pm\sqrt{(a+\kappa_0+\kappa_1)^2-4a\kappa_0}\right).$$

2.3 多状态基因模型

引入下列生化反应

$$\begin{aligned}
&I_k\xrightarrow{\lambda_k}I_{k+1}\,(k=1,2,\cdots,L-1),\\
&I_L\xrightarrow{\lambda_L}A,\quad A\xrightarrow{\gamma}I_1,\\
&A\xrightarrow{\mu_m}A+M,\quad M\xrightarrow{\mu_p}M+P,\\
&M\xrightarrow{\delta_m}\varnothing,\quad P\xrightarrow{\delta_p}\varnothing.
\end{aligned}$$

这里 γ 是基因失活率，λ_L 是基因激活率，λ_k 是从第 k 个“关”状态到第 $(k+1)$ 个“关状态的转移率 ($k=1,2,\cdots,L-1$)，μ_m 是基因处于活性态的转录率，δ_m 是 mRNA 的降解率，μ_p 是 mRNA 的翻译率，δ_p 是蛋白质的降解率. 注意到：假如 $L=1$，那么相应的基因模型将变成熟识的两状态基因模型(或电报模型). 因为 mRNA 并不受蛋白质影响(即没有反馈)，可考虑独立于蛋白质的 mRNA 的主方程.

让 m 表示 mRNA 的数目，$P_0(m,t)$ 和 $P_k(m,t)$ ($k=1,2,\cdots,L$) 分别表示 mRNA 在 t 时刻活性状态和非活性状态有 m 个 mRNA 分子的概率密度. 那么，相应的主方程可表示为

$$\begin{aligned}
&\frac{\mathrm{d}P_0(m,t)}{\mathrm{d}t}=-\gamma P_0(m,t)+\lambda_LP_L(m,t)+\\
&\qquad\left[\delta_m(\boldsymbol{E}-\boldsymbol{I})+\mu_m\left(\boldsymbol{E}^{-1}-\boldsymbol{I}\right)\right]\left[mP_0(m,t)\right],\\
&\frac{\mathrm{d}P_1(m,t)}{\mathrm{d}t}=-\lambda_1P_1(m,t)+\gamma P_0(m,t)+\\
&\qquad\delta_m(\boldsymbol{E}-\boldsymbol{I})\left[mP_1(m,t)\right],\\
&\frac{\mathrm{d}P_k(m,t)}{\mathrm{d}t}=-\lambda_kP_k(m,t)+\lambda_{k-1}P_{k-1}(m,t)+\\
&\qquad\delta_m(\boldsymbol{E}-\boldsymbol{I})\left[mP_k(m,t)\right](k=2,\cdots,L),
\end{aligned}\tag{8}$$

其中 $\boldsymbol{I}$ 是单位算子，$\boldsymbol{E}$ 和 $\boldsymbol{E}^{-1}$ 是平移算子(或步长算子)，即对任意函数 f 和任意整数 n，有

$$\boldsymbol{E}\left[f(n)\right]=f(n+1),\quad\boldsymbol{E}^{-1}\left[f(n)\right]=f(n-1).$$

为了求解方程(8)，对 $P_k(m;t)$ 引入母函数

$$G_i(z,t)=\sum_{m=0}^{\infty}z^m P_i(m,t)\ \ (i=0,1,\cdots,L),$$

这样, 从方程(8), 可获得关于 $G_k(z;t)$ 的微分方程

$$\frac{\partial G_0(z,t)}{\partial t}=-\gamma G_0(z,t)+\lambda_L G_L(z,t)+(1-z)\frac{\partial G_0(z,t)}{\partial z}+\mu(z-1)G_0(z,t),$$

$$\frac{\partial G_1(z,t)}{\partial t}=-\lambda_1 G_1(z,t)+\gamma G_0(z,t)+(1-z)\frac{\partial G_1(z,t)}{\partial z},$$

$$\frac{\partial G_k(z,t)}{\partial t}=-\lambda_k G_k(z,t)+\lambda_{k-1}G_{k-1}(z,t)+(1-z)\frac{\partial G_k(z,t)}{\partial z}\ \ (k=2,\cdots,L),$$

这里所有的参数和时间被 δ_m 所规范化, 即 $\lambda_k/\delta_m\to\lambda_k$ $(k=1,\cdots,L)$, $\gamma/\delta_m\to\gamma$, $\mu_m/\delta_m\to\mu$, $\delta_m t\to t$. 记 $G\equiv\sum_{k=0}^{L}G_k$, 则可求得静态母函数

$$G(z)={}_LF_L\left(a_1,a_2,\cdots,a_L;b_1,b_2,\cdots,b_L;\mu(z-1)\right)C.$$

这里 C 是一个规范化常数, 它由条件 $G(1)=1$ 决定, 蕴含着 $C=1$; ${}_LF_L$ 是广义超几何函数[37]. 在 $z=1$ 点 Taylor 展开函数 $G(z)$, 并利用概率密度于母函数之间的关系, 获得 mRNA 的概率密度的分析表示[28]

$$P(m)=\frac{\prod_{k=1}^{L}\Gamma(a_k+m)\prod_{k=1}^{L}\Gamma(b_k)}{\prod_{k=1}^{L}\Gamma(a_k)\prod_{k=1}^{L}\Gamma(b_k+m)}\frac{\mu^m}{m!}\cdot {}_LF_L(a_1+m,\cdots,a_L+m;b_1+m,\cdots,b_L+m;-\mu).$$

这就是所求的 mRNA 的静态概率密度函数.

2.4 基因自促进模型

考虑下列生化反应

$$I\underset{\gamma}{\overset{\lambda}{\rightleftharpoons}}A,\quad I+P\xrightarrow{a}A+P,\quad A\xrightarrow{\mu}A+P,\quad P\xrightarrow{\delta}\varnothing,$$

其中 P 代表蛋白质. 让 $P_0(m;t)$ 和 $P_1(m;t)$ 分别表示基因处在 I 和 A 状态于 t 时刻有 m 个蛋白质分子的概率密度, 那么有关的主方程为

$$\frac{\mathrm{d}P_0(m;t)}{\mathrm{d}t}=-\lambda P_0(m;t)+\gamma P_1(m;t)-amP_0(m;t)+\delta\left[(m+1)P_0(m+1;t)-mP_0(m;t)\right],$$

$$\frac{\mathrm{d}P_1(m;t)}{\mathrm{d}t}=\lambda P_0(m;t)-\gamma P_1(m;t)+amP_0(m;t)+\delta\left[(m+1)P_1(m+1;t)-mP_1(m;t)\right]+\mu\left[P_1(m-1;t)-P_1(m;t)\right].$$

为了求解此方程, 可采用泊松表示法[38], 并求得

$$P(m)\sim{}_1F_1\left(\frac{\lambda}{1+a}+m;\frac{\lambda+\gamma}{1+a}+m;-\frac{\mu}{1+a}\right),\tag{9}$$

两者相差 1 个规范化因子, 其中 a 代表反馈强度. 根据(9)式, 可讨论反馈对 mRNA 概率密度以及对分子噪声的影响.

2.5 基因自压制模型

考虑下列生化反应

$$I\underset{\gamma}{\overset{\lambda}{\rightleftharpoons}}A,\quad A+P\xrightarrow{r}I+P,\quad A\xrightarrow{\mu}A+P,\quad P\xrightarrow{\delta}\varnothing.$$

让 $P_0(m;t)$ 和 $P_1(m;t)$ 分别表示基因处在 I 和 A 状态于 t 时刻有 m 个蛋白质分子的概率密度, 那么有关的主方程为

$$\frac{\mathrm{d}P_0(m;t)}{\mathrm{d}t}=-\lambda P_0(m;t)+\gamma P_1(m;t)+rmP_1(m;t)+\delta\left[(m+1)P_0(m+1;t)-mP_0(m;t)\right],$$

$$\frac{\mathrm{d}P_1(m;t)}{\mathrm{d}t}=\lambda P_0(m;t)-\gamma P_1(m;t)-rmP_1(m;t)+\delta\left[(m+1)P_1(m+1;t)-mP_1(m;t)\right]+\mu\left[P_1(m-1;t)-P_1(m;t)\right].$$

为了求解此方程, 可采用泊松表示法, 并求得 mRNA 的静态概率密度

$$P(m)=\frac{\Gamma(\lambda+m)\Gamma(\lambda+\beta)}{\Gamma(\lambda)\Gamma(\lambda+\beta+m)}\left(\frac{\mu}{1+r}\right)^m\cdot\frac{1}{m!}\frac{{}_1F_1\left(\lambda+m;\beta+\lambda+m;-\dfrac{\mu}{(1+r)^2}\right)}{{}_1F_1\left(\lambda;\beta+\lambda;\dfrac{r\mu}{(1+r)^2}\right)},$$

其中 r 代表反馈强度.

3　讨论

本文给出了几种常见随机基因模型中 mRNA 或蛋白质数目的概率密度. 实际的基因模型可能会更复杂, 例如, 可能涉及多个活性状态之间的转移、多

个活性与非活性状态之间的转移、上游基因的产物(如蛋白质)作为转录因子调控下游基因的表达、DNA 环、甲基化、后翻译等. 如何建立更复杂的随机基因模型并导出有关 *m*RNA 与蛋白质数目变化的概率分布是一个值得深入研究的问题, 也是一个十分有趣的问题.

对于某个感兴趣的物种(如 *m*RNA 和蛋白质等), 假如知道或导出了它的概率密度或概率分布, 那么所有相关的随机信息(如平均、方差、Fano 因子等)均可分析地给出. 这样, 就能够进一步研究分子噪声在基因表达过程中的生物学功能与作用. 目前, 关于分子噪声源及分子噪声的生物学功能等已成为系统生物学的一个重要研究方向, 在国际顶尖刊物(如 Nature、Science 等)上经常有相关研究的报道.

任何生物系统或网络本质上可看成是信号系统, 这里物种分子的数目或浓度是动态变化的. 以前的文献采用线性噪声逼近法研究了噪声信号的传播问题[14, 39], 但线性噪声逼近具有诸多局限性, 常常不能够很好地刻画物种分子数目的实际变化. 相对地, 研究信号分子的概率分布传播问题更真实、更有实际意义. 这在数学上可归纳为以下问题:

对于由 $\boldsymbol{Y}=\boldsymbol{f}(\boldsymbol{X})$ 刻画的信号系统, 这里 $\boldsymbol{X}$ 代表输入信号(可以是多维的), $\boldsymbol{Y}$ 代表输出信号(亦可是多维的), 函数(可为向量函数) $\boldsymbol{f}$ 刻画响应情况(注: $\boldsymbol{f}$ 可以代表生化网络). 假定输入变量 $\boldsymbol{X}$ 服从某个分布 $P_I(\boldsymbol{X})$, 那么如何给出 $\boldsymbol{Y}$ 的分布 $P_O(\boldsymbol{Y})$.

目前, 有关信号分子的概率分布传播问题的研究文献几乎还没有, 值得深入研究.

4 参考文献

[1] Blake W J, Kærn M, Collins J J. Noise in eukaryotic gene expression [J]. Nature, 2003, 422: 633-637.

[2] Raser J, O'Shea E. Control of stochasticity in eukaryotic gene expression [J]. Science, 2003, 304: 1811-1814.

[3] Boeger H, Griesenbeck J, Kornberg R D. Nucleosome retention and the stochastic nature of promoter chromatin remodeling for transcription [J]. Cell, 2008, 133: 716-726.

[4] Larson D R. What do expression dynamics tell us about the mechanism of transcription? [J]. Curr Opin Genet Dev, 2011, 21: 591-599.

[5] Chubb J R, Trcek T, Singer R H. Transcriptional pulsing of a developmental gene [J]. Curr Biol, 2006, 16: 1018-1025.

[6] Pedraza J M, Paulsson J. Effects of molecular memory and bursting on fluctuations in gene expression [J]. Science, 2008, 319: 339-343.

[7] Sánchez A, Kondev J. Transcriptional control of noise in gene expression [J]. Proc Natl Acad Sci U S A, 2008, 105: 5081-5086.

[8] Sanchez A, Garcia H G, Kondev J. Effect of promoter architecture on the cell-to-cell variability in gene expression [J]. PLoS Comput Biol, 2011, 7: e1001100.

[9] Mao C, Brown C R, Boeger H. Quantitative analysis of the transcription control mechanism [J]. Mol Syst Biol, 2010, 6: 431-438.

[10] Mariani L, Schulz E G, Höfer T. Short-term memory in gene induction reveals the regulatory principle behind stochastic IL-4 expression [J]. Mol Syst Biol, 2010, 6: 359-365.

[11] Miller-Jensen K, Dey S S, Arkin A P. Varying virulence: epigenetic control of expression noise and disease processes [J]. Trends Biotechnol, 2011, 29: 517-525.

[12] 周天寿, 胡长春. 3 类基因振子和它们的基本动力学 [J]. 江西师范大学学报: 自然科学版, 2008, 32(1): 1-5.

[13] Friedman N, Cai L, Xie X S. Linking stochastic dynamics to population-distribution: an analytical framework of gene expression [J]. Phys Rev Lett, 2006, 97: 168302.

[14] Paulsson J. Summing up the noise in gene networks [J]. Nature, 2004, 427: 415-418.

[15] Thattai M, Van Oudenaarden A. Intrinsic noise in gene regulatory networks [J]. Proc Natl Acad Sci U S A, 2001, 98: 8614-8619.

[16] Peccoud J, Ycart B. Markovian modelling of gene product synthesis [J]. Theor Popul Biol, 1995, 48: 222-234.

[17] Shahrezaei V, Swain P S. Analytical distributions for stochastic gene expression [J]. Proc Natl Acad Sci U S A, 2008, 105: 17256-17261.

[18] Paulsson J. Models of stochastic gene expression [J]. Phys Life Rev, 2005, 2: 157-175.

[19] Kepler T B, Elston T C. Stochasticity in transcriptional regulation: origins, consequences and mathematical representations [J]. Biophys J, 2001, 81: 3116-3136.

[20] Paulsson J, Ehrenberg M. Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks [J]. Phys Rev Lett, 2000, 84: 5447-5450.

[21] Karmakar R, Bose I. Graded and binary responses in stochastic gene expression [J]. Phys Biol, 2004, 1: 197-204.

[22] Iyer-Biswas S, Hayot F, Jayaprakash C. Stochasticity of gene products from transcriptional pulsing [J]. Phys Rev E, 2009, 79: 031911.

[23] Mugler A, Walczak A M, Wiggins C H. Spectral solutions to stochastic models of gene expression with bursts and regulation [J]. Phys Rev E, 2009, 80: 041921.

[24] Mackey M C, Tyran-Kamińska M, Yvinec R. Molecular distributions in gene regulatory dynamics [J]. J Theor Biol, 2011, 274: 84-96.

[25] Assaf M, Roberts E, Luthey-Schulten Z. Determining the stability of genetic switches: explicitly accounting for *m*RNA noise [J]. Phys Rev Lett, 2011, 106: 248102.

[26] Suter D M, Molina N, Naef F. Mammalian genes are transcribed with widely different bursting kinetics [J]. Science, 2011, 332: 472-474.

[27] Harper C V, Finkenstädt B, White M R. Dynamic analysis of stochastic transcription cycles [J]. PLoS Biol, 2011, 9: e1000607.

[28] Zhang Jiajun, Zhou Tianshou. Analytical distribution and tunability of noise in a model of promoter progress [J]. Biophys J, 2012, 102: 1247-1257.

[29] Zhang Jiajun, Yuan Zhanjiang, Zhou Tianshou. Physical limits of feedback noise-suppression in biological networks [J]. Phys Biol, 2009, 6: 046009.

[30] 苑占江, 张家军, 周天寿. 基因自调控环路的功能 [J]. 生物物理学报, 2010, 26(6): 457-471.

[31] Zhang Jiajun, Yuan Zhanjiang, Li Hanxiong, et al. Architecture-dependent robustness and bistability in a class of genetic circuits [J]. Biophys J, 2010, 99: 1034-1042.

[32] Ullah M, Wolkenhauer O. Family tree of Markov models in systems biology [J]. IET Syst Biol, 2007, 1(4): 247-254.

[33] Van Kampen N G. Stochastic process in physics and chemistry [M]. Amsterdam: North-Holland, 1992.

[34] Sakurai J. Modern quantum mechanics [M]. India: Pearson Education, 1985.

[35] McAdams H H, Arkin A. Stochastic mechanisms in gene expression [J]. Proc Natl Acad Sci U S A, 1997, 94(3): 814-819.

[36] Raj A, Peskin C S, Tranchina D, et al. Stochastic *m*RNA synthesis in mammalian cells [J]. PLoS Biol, 2006, 4(10): e309.

[37] Slater L J. Confluent hypergeometric functions [M]. Cambridge: Cambridge University Press, 1960.

[38] Iyer-Biswas S, Jayaprakash C. Mixed poisson distribution in exact solution of stochastic auto-regulation models [J]. arXiv, 1110.2804v1 [q-bio.QM] 12 Oct. 2011.

[39] Tanase-Nicola S, Warren P B, Ten Wolde P R. Signal detection, modularity, and the correlation between extrinsic and intrinsic noise in biochemical networks [J]. Phys Rev Lett, 2006, 97: 068102.

Review on Gene Expression Models: Probability Distribution

ZHOU Tian-shou

(School of Mathematics and Computational Sciences, Sun Yat-Sen University, Guangzhou Guangdong 510275, China)

Abstract: Quantifying gene expression (including mathematical modeling and qualitative and quantitative analysis) is not only an important step toward to understanding intracellular processes but also the core of the current systems biology. Gene expression models have been developed to complicated multi-state models considering detailed biological processes and a number of biological factors from the initial simple single-state models. Based on central dogma in biology, The proceeding in the study of gene expression models, focusing on improvement of mathematical models, probability distribution of *m*RNAs and proteins, etc. are simply reviewed. Consequently, some general laws related to gene expression are summarized. In addition, some issues to deserve further studies are discussed and potential directions are pointed out.

Key words: gene expression; gene state; biochemical master equation; probability density; generating function

(责任编辑: 王金莲)