中国科学: 数学 2025 年 第 55 卷 第 ? 期: 1 ∼ 18
SCIENTIA SINICA Mathematica
综 述

单细胞随机基因表达动力学的数学模型与解析理论

贾晨

北京计算科学研究中心应用与计算数学研究部, 北京 100193
E-mail: chenjia@csrc.ac.cn

收稿日期: 2024-08-30; 接受日期: 2025-03-10; 网络出版日期: 2025-06-16
国家自然科学基金 (批准号: 12271020 和 12131005) 资助项目

摘要 基因表达是细胞内最重要的动力学过程. 基因表达与基因调控对细胞分化、细胞增殖、细胞凋亡、信号传导、压力响应, 以及疾病的发生发展有着决定性的影响. 近二十年来, 基因表达随机动力学的理论与实验都取得了突破性进展. 本文介绍单细胞随机基因表达的经典数学模型与数学理论, 尤其是基因产物数量分布的解析理论. 此外, 本文对该领域的重要文献进行了全面性梳理.

关键词 基因调控 基因表达爆发性 Markov 过程 杂交微分方程 平稳分布 瞬时分布 功率谱 特殊函数

MSC (2020) 主题分类 60J27, 60J28, 34A38, 60H10, 92C40, 92B05

1 引言

基因表达是以分子生物学的中心法则为基准、用基因中的遗传信息来合成 mRNA 与蛋白的过程, 是生物体的多功能性与适应性的基础. 基因表达过程极为复杂且精细, 不仅在单个细胞内涉及多层次的调控网络, 而且不同细胞间也存在着错综复杂的相互作用与调控关系. 基因表达与基因调控对细胞分化、细胞增殖、细胞凋亡、信号传导、压力响应, 以及疾病的发生发展有着决定性的影响. 近二十年来, 随着单细胞实验技术的突破性进展, 对于基因表达的研究普遍从定性走向了定量阶段, 人们可以在单细胞水平上测量基因表达量, 并积累了海量的实验数据. 这些实验技术包括流式细胞术、免疫荧光法、活细胞成像、单细胞 RNA 测序、单分子 RNA 荧光原位杂交、微流控装置、母机装置等, 其中有些实验甚至可以达到单分子分辨率 (参见文献 [12, 27, 116]).

在多样化的生物体系中, 基因的表达模式与调控机制均展现出显著的异质性与复杂性, 而这进一步导致基因表达水平呈现出广泛的差异. 研究不同细胞类型的基因表达差异, 能够筛选出在不同生理状况或疾病进程中基因表达发生显著变化的基因. 这些差异表达基因通常与细胞命运决定以及复杂疾病的发生发展密切相关. 通过对差异表达基因的精确筛选, 科研人员得以实现对疾病的准确诊断与预

英文引用格式: Jia C. Single-cell stochastic gene expression dynamics: Mathematical models and analytical theory (in Chinese). Sci Sin Math, 2025, 55: 1–18, doi: 10.1360/SSM-2024-0263

后评估, 并为药物研发提供关键线索. 此外, 研究基因表达噪声有助于理解基本生物过程的随机性以及细胞的个体差异; 研究基因表达量是否呈现双峰或多峰分布, 有助于理解细胞的表型异质性与风险对冲; 研究基因产物间的相关性可以揭示基因转录调控机制, 进而重构基因调控网络; 研究基因表达量从非稳态演化到稳态的弛豫速度, 有助于揭示细胞对刺激的响应机制以及对信号的传导效率等.

大量的单细胞实验表明, 基因表达动力学具有显著的随机性. 基因表达的随机性有两个基本来源: 外噪声与内噪声. 外噪声是由于外部环境 (如细胞体积、RNA 聚合酶浓度、转录因子浓度等) 的随机性而导致的分子数涨落以及由于当前实验技术的局限性而导致的实验误差, 内噪声则是由于细胞内生化分子间的随机相互碰撞与随机生化反应而导致的分子数涨落 (参见文献 [10, 98]). 当参与生化反应的分子数很多时, 根据大数定律, 其随机性可以被忽略. 然而在单细胞内, 参与生化反应的分子数往往很少, 其随机性不可被忽略. 例如, 在细菌细胞中, 基因通常只有几条, 该基因所对应的 mRNA 通常只有不到 50 个分子, 而该基因所对应的蛋白通常只有不到 500 个分子 (参见文献 [79]). 因此随机性对于基因表达动力学的研究是极端重要的.

鉴于基因表达与基因调控的复杂性与重要性, 基因表达动力学的数学建模、解析理论以及计算方法研究已成为国际上应用数学与计算系统生物学的前沿课题之一. 基因表达动力学的理论研究不仅为复杂系统的研究提供了可借鉴的理论范式, 而且对理解各种重要的细胞功能起到至关重要的作用. 而随机过程作为研究随机现象的数学工具, 在基因表达动力学的研究中扮演至关重要的作用. 本文介绍单细胞随机基因表达的经典数学模型与数学理论, 并对该领域的重要文献进行梳理.

2 经典随机基因表达模型

生化反应的随机动力学有两种不同的描述方式: 轨道描述与分布描述. 前者刻画了各种生化分子的分子数 (简称为生化分子数) 本身随时间的变化, 而后者刻画了生化分子数的概率分布随时间的变化. 从轨道的角度, 系统的随机动力学由以生化分子数为微观态的 Markov 跳过程 (即连续时间 Markov 链) 所刻画 (参见文献 [2]), 该过程可以通过著名的随机模拟算法 (也称为 Doob-Gellespie 算法) [26] 进行数值模拟. 从概率分布的角度, 系统的随机动力学由生化分子数的概率分布所满足的化学主方程所刻画, 该方程由诺贝尔生理与医学奖得主 Delbrück [19] 于 1940 年所提出, 该方程可以利用有限状态投影算法 [73] 进行数值计算. 从概率论的观点, 化学主方程即是 Markov 跳过程模型的 Kolmogorov 前进方程. 因此, 生化反应随机动力学的轨道描述与分布描述完全统一. 随机模拟算法模拟的是生化系统的随机轨道, 有限状态投影算法模拟的是生化系统的概率分布. 前者适用于一般的反应系统, 而后者仅适用于低维反应系统 [44].

近二十年来, 生化反应的随机动力学理论被广泛应用于基因表达的随机动力学研究, 并取得了一系列丰硕的研究成果. 该领域的核心问题之一是, 基因表达的随机动力学模型是否能够解释实验上观测到的基因表达量的复杂概率分布 (包括平稳分布与瞬时分布) (参见文献 [62, 75, 101]). 设 G 为我们所关注的基因, P 为相应的基因产物 (mRNA 或蛋白). 最简单的随机基因表达模型由如下的反应所描述:

$$G \xrightarrow{\rho} G + P, \quad P \xrightarrow{d} \varnothing, \tag{2.1}$$

其中 ρ 为基因产物的生成速率, d 为基因产物的降解速率, $\varnothing$ 代表蛋白被降解. 当前一个反应发生时, 基因产物分子增加一个; 而当后一个反应发生时, 基因产物分子减少一个. 因此该模型本质上是经典的

生灭过程, 描述了基因产物的生成与降解. 在稳态下, 该模型的化学主方程可以解析求解, 且基因产物分子数服从如下的 Poisson 分布 [79]:

$$p_n^{\rm ss} = \frac{r^n}{n!}{\rm e}^{-r},$$

其中 p_n 的上标 ss 表示稳态 (steady state), $r = \rho/d$ 为基因产物的均值. 我们知道, Poisson 分布最重要的性质之一是其均值等于方差. 然而大量的实验表明, 对于绝大部分基因, 其表达量的方差都显著大于均值 [1], 该现象称为基因表达的过度散布性 (over-dispersion).

为了解释基因表达的过度散布性, 生物信息学家 Anders 和 Huber[1] 及 Robinson 等[88] 于 2010 年首次使用负二项分布拟合单细胞基因表达数据, 并将其应用于基因的差异表达分析. 根据谷歌学术的统计, 文献 [88] 已被引用 40,000 余次, 可见该问题影响力之巨大. 使用负二项分布拟合数据的根本原因在于, 该分布是方差大于均值的最常见的离散概率分布. 有趣的是, 生物数学家 Paulsson 和 Ehrenberg[80] 早在 2000 年就在微观层面上给出了负二项分布的分子机制. 他们认为基因产物分子并不是逐个产生的, 而是会在短时间内积累大量分子, 该现象称为基因表达的爆发性 (bursting). 大量的实验表明, mRNA 与蛋白的产生均可能呈现爆发性, 前者称为转录爆发性, 而后者称为翻译爆发性. 每次基因产物的短时间积累称为一次爆发, 每次爆发所累积的基因产物分子数称为爆发量, 而单位时间内的爆发次数称为爆发频率.

文献 [80] 指出, 如果爆发量服从几何分布, 则相应基因表达模型的平稳分布正好是负二项分布. 具体地, 考虑如下带爆发的基因表达模型:

$$G \xrightarrow{\rho p^k(1-p)} G + kP, \quad k \geqslant 1, \quad P \xrightarrow{d} \varnothing,$$

其中基因产物的产生具有爆发性, 爆发频率为 ρ, 爆发量服从以 p 为参数的几何分布, 即每次爆发产生 k 个基因产物分子的概率为 $p^k(1-p)$. 在稳态下, 该模型的基因产物分子数服从如下的负二项分布[80]:

$$p_n^{\rm ss} = \frac{(r)_n}{n!}p^n(1-p)^r, \tag{2.2}$$

其中 $r = \rho/d$, $(r)_n = r(r+1)\cdots(r+n-1)$ 为 Pochhammer 记号. 该工作为基因表达的负二项分布模型提供了动力学基础. 随着单细胞实验技术的不断发展, 人们可以在单分子水平上观测到基因表达的爆发性. Golding 等[27] 于 2005 年在细菌细胞内首次观测到了转录爆发性, 并验证了 mRNA 的爆发量服从几何分布. 随后, Cai 等[12] 于 2006 年在细菌细胞内首次观测到了翻译爆发性, 并验证了蛋白的爆发量也服从几何分布. 至此, 转录爆发性与翻译爆发性都得到了很好的实验验证.

最近的研究 [27] 表明, 基因并不是永远工作的, 而是会在打开与关闭两种状态之间进行随机切换. 打开状态通常对应于转录机器结合在启动子上的状态, 而关闭状态通常对应于转录机器未结合在启动子上的状态. 当基因处于打开状态时, 基因产物可以快速地积累; 而当基因处于关闭状态时, 基因产物则不积累或较慢地积累. 基于上述生物学事实, Ko[59] 早在 1991 年便提出了基因表达的经典两状态模型 (也称为电报模型, 见图 1(a)), 该模型可以由如下的反应所描述:

$$\begin{aligned} &G \xrightarrow{a} G^*, \quad G^* \xrightarrow{b} G, \\ &G^* \xrightarrow{\rho} G^* + P, \quad P \xrightarrow{d} \varnothing, \end{aligned} \tag{2.3}$$

其中基因可以在关闭状态 G 与打开状态 G^* 之间进行随机切换, 且基因产物 P 只能在基因处于打开状态时生成. 该系统的微观态由二元组 (i,n) 所刻画, 其中 $i = 0,1$ 为基因状态 ($i = 0$ 代表基因的

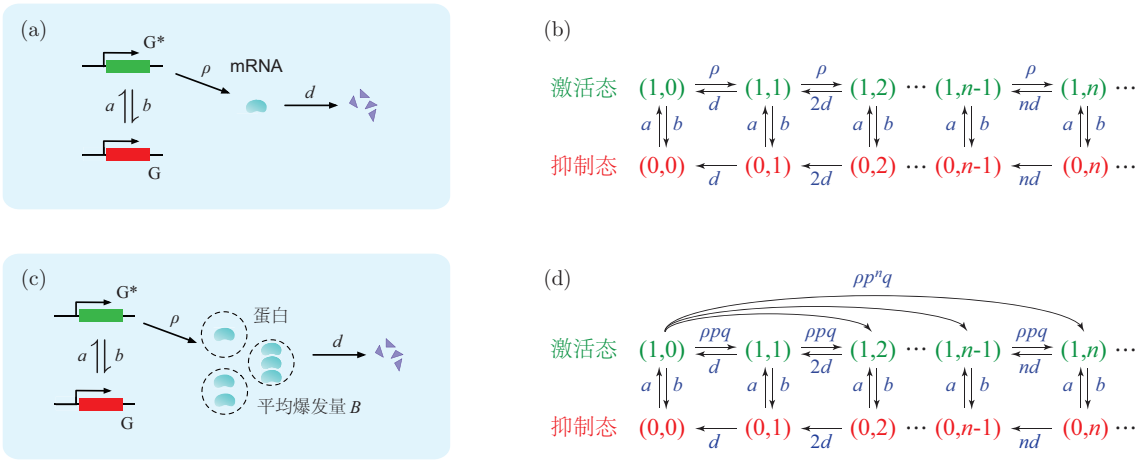

图 1 (网络版彩图) 基因表达的两状态模型 (参见文献 [113]). **(a)** 两状态电报模型. 当基因处于激活态时, 蛋白分子每次只产生一个. **(b)** 两状态电报模型对应的 Markov 跳过程的状态转移图. **(c)** 两状态爆发性模型. 当基因处于激活态时, 蛋白分子以爆发的方式产生. **(d)** 两状态爆发性模型对应的 Markov 跳过程的状态转移图

关闭状态, $i=1$ 代表基因的打开状态), n 为基因产物分子数. 该系统的随机动力学由图 1(b) 所示的 Markov 跳过程所刻画, Markov 跳过程的状态空间由系统的所有微观态所组成. 设 $p_{i,n}$ 为单细胞内基因处于状态 i 且基因产物分子数为 n 的概率, 设 $p_n = p_{0,n} + p_{1,n}$ 为基因产物分子数为 n 的概率. 于是基因表达动力学可以由如下的化学主方程 (即 Markov 跳过程模型的 Kolmogorov 前进方程) 所刻画:

$$\begin{aligned} \dot{p}_{0,n} &= d[(n+1)p_{0,n+1} - np_{0,n}] + [bp_{1,n} - ap_{0,n}], \\ \dot{p}_{1,n} &= \rho[p_{1,n-1} - p_{1,n}] + d[(n+1)p_{1,n+1} - np_{1,n}] + [ap_{0,n} - bp_{1,n}], \end{aligned} \tag{2.4}$$

其中含 ρ 的项表示基因产物生成, 含 d 的项表示基因产物降解, 含 a 和 b 的项表示基因状态切换. 该化学主方程的稳态解析解于 1995 年被 Peccoud 和 Ycart [81] 所得到. 具体地, 基因产物数量的平稳分布为

$$p_n^{\mathrm{ss}} = \frac{r^n}{n!}\frac{(\alpha)_n}{(\beta)_n}{}_1F_1(\alpha+n;\beta+n;-r), \tag{2.5}$$

其中 $\alpha = a/d$, $\beta = (a+b)/d$, $r = \rho/d$, ${}_1F_1(\alpha;\beta;z)$ 为合流超几何函数. Jiao 等 [54] 详细研究了该平稳分布的三种不同形态 (递减分布、钟形分布、双峰分布), 并给出了三种分布形态相图的几何刻画.

在上述模型中, 基因产物的生成由单一反应所刻画. 然而基于分子生物学的中心法则, 真实的基因表达过程由 DNA 产生 mRNA 的转录步骤以及 mRNA 产生蛋白的翻译步骤所组成. 为了使理论更贴近生物学实际, Shahrezaei 和 Swain [90] 于 2008 年提出了基因表达的经典两阶段模型 (见图 2(a)) 与三阶段模型 (见图 2(b)). 两阶段模型描述了转录与翻译这两个步骤, 而三阶段模型描述了基因状态切换、转录与翻译这三个步骤. 具体地, 两阶段模型由如下的反应所刻画:

$$\begin{aligned} &G \xrightarrow{\rho} G + M, \quad M \xrightarrow{u} M + P, \\ &M \xrightarrow{v} \varnothing, \quad P \xrightarrow{d} \varnothing, \end{aligned}$$

其中 G 为基因, M 为相应的 mRNA, P 为相应的蛋白. 前两个反应代表转录与翻译, 而后两个反应代表基因产物的降解. 在该模型中, mRNA 数量的平稳分布为 Poisson 分布, 这是由于该模型的 mRNA

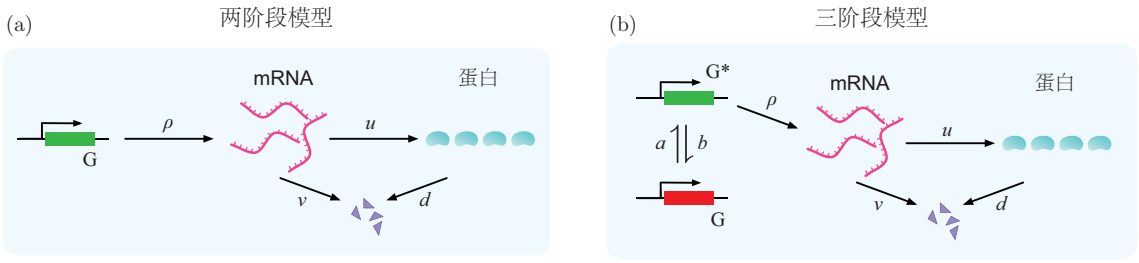

图 2 (网络版彩图) 基因表达的两阶段模型与三阶段模型. (a) 两阶段模型描述了转录与翻译这两个步骤; (b) 三阶段模型描述了基因状态切换、转录与翻译这三个步骤

动力学等效于 (2.1) 中的反应系统. 此外, Bokes 等 [9] 给出了 mRNA 与蛋白数量平稳联合分布的解析表达式.

三阶段模型由如下的反应所刻画:

$$\begin{aligned} &G \xrightarrow{a} G^*, \quad G^* \xrightarrow{b} G, \\ &G^* \xrightarrow{\rho} G^* + M, \quad M \xrightarrow{u} M + P, \\ &M \xrightarrow{v} \varnothing, \quad P \xrightarrow{d} \varnothing, \end{aligned}$$

其中前两个反应代表基因状态切换, 中间两个反应代表转录与翻译, 而后两个反应代表基因产物的降解. 在该模型中, mRNA 数量的平稳分布由合流超几何函数给出, 这是由于该模型的 mRNA 动力学等效于 (2.3) 中的反应系统, 而蛋白数量分布的解析解目前仍未解决. 在细菌与酵母中, mRNA 的降解速率往往远大于蛋白的降解速率. 例如, 在大肠杆菌细胞中, mRNA 的半衰期通常只有不到 10 分钟 (参见文献 [6]), 而蛋白的半衰期则长达 20 小时 (参见文献 [72]), 两者相差了 100 倍以上, 此时基因表达系统具有明显的时间尺度分离特征. 在该时间尺度分离假设下, 蛋白的产生具有爆发性, 于是三阶段模型的蛋白动力学可以简化为如下的反应系统 (见图 1(c)):

$$\begin{aligned} &G \xrightarrow{a} G^*, \quad G^* \xrightarrow{b} G, \\ &G^* \xrightarrow{\rho p^k(1-p)} G^* + kP, \quad k \geqslant 1, \quad P \xrightarrow{d} \varnothing, \end{aligned}$$

其中 ρ 为蛋白的爆发频率, 爆发量服从以 $p = u/(u+v)$ 为参数的几何分布. 类似地, 该系统的微观态由二元组 (i, n) 所刻画, 其中 $i = 0, 1$ 为基因状态, n 为基因产物分子数. 该系统的随机动力学由图 1(d) 所示的 Markov 跳过程所刻画, Markov 跳过程的状态空间由系统的所有微观态所组成. 文献 [90] 表明该模型蛋白数量的平稳分布具有以下形式:

$$p_n^{\mathrm{ss}} = \frac{B^n}{n!} \frac{(\alpha_1)_n (\alpha_2)_n}{(\beta)_n} {}_2F_1(\alpha_1 + n, \alpha_2 + n; \beta + n; -B),$$

其中 ${}_2F_1(\alpha_1, \alpha_2; \beta; z)$ 为 Gauss 超几何函数, 且

$$\alpha_1 + \alpha_2 = \frac{a + b + s}{d}, \quad \alpha_1 \alpha_2 = \frac{as}{d^2}, \quad \beta = \frac{a+b}{d}, \quad B = \frac{p}{1-p}.$$

上面所提到的结果均只涉及基因产物的平稳分布. 系统未达到平稳时基因产物的含时间的分布通常称为瞬时分布. 事实上, 关于基因产物的瞬时分布也有大量的研究. 在上述时间尺度分离假设下, 两阶段模型 mRNA 与蛋白数量瞬时分布的解析表达式首次由 Shahrezaei 和 Swain [90] 所得到, 三阶段模

型 mRNA 数量瞬时分布的解析表达式首次由 Iyer-Biswas 等[35] 所得到, 而三阶段模型蛋白数量瞬时分布的解析表达式首次由 Cao 和 Grima[13] 所得到.

基因表达爆发性的背后也存在着深刻的数学理论. 翻译爆发性产生的根本原因是生存时间很短的 mRNA 快速合成大量蛋白, 而转录爆发性产生的根本原因是基因长时间处于关闭状态, 但当基因处于打开状态时快速合成大量 mRNA[79]. 文献 [38] 基于以上假设, 利用 Markov 跳过程的多尺度简化技术 (参见文献 [8,36,37,115]), 证明了转录爆发性、翻译爆发性, 以及爆发量服从几何分布这些事实均可以从经典的三阶段模型中自然地推导出. 这为基因表达爆发性的产生提供了严格的数学基础.

3 多状态随机基因表达模型

在基因表达的经典两状态模型 (见图 2) 中, 基因的打开与关闭时间均服从指数分布. 然而最近的实验[33,97] 表明, 在哺乳动物细胞中, 很多基因的关闭时间并不服从指数分布, 这暗示着这些基因可能存在两个或两个以上的关闭状态. 为了解释上述现象, Suter 等[97] 提出了三状态环状转移模型 (也称为 refractory 模型), Jiao 和 Zhu[55] 提出了三状态线性转移模型 (也称为交互路径模型). 文献 [39] 指出三状态模型在特定的时间尺度分离假设下的平稳 mRNA 数量服从零膨胀负二项分布, 该结果为单细胞 RNA 测序分析中广泛使用的零膨胀负二项模型[71,87] 提供了动力学基础. 更一般地, Zhou 和 Zhang[121] 考虑了复杂的基因切换机制, 认为基因可以在任意多个状态之间进行随机切换. 如果多个状态中只有一个是打开状态, 而其余的都是关闭状态, 则基因产物数量的平稳分布包含广义超几何函数.

具体地, 考虑如下的多状态基因表达系统 (见图 3):

$$G_i \xrightarrow{k_{ij}} G_j, \quad i,j=0,1,\dots,L,$$
$$G_0 \xrightarrow{\rho p^k(1-p)} G_0 + kP, \quad k \geqslant 1, \quad P \xrightarrow{1} \varnothing.$$

这里假定基因可以在 $L+1$ 个不同状态 $G_0, G_1, \dots, G_L$ 之间进行随机切换, 基因具有唯一的打开状态 G_0 以及多个关闭状态 $G_1, \dots, G_L$. 当基因处于打开状态 G_0 时, 基因产物 P 的生成具有爆发性, 爆发频率为 ρ, 爆发量服从以 p 为参数的几何分布. 为了简单起见, 这里假定蛋白降解速率 $d=1$, 该假定实际上是对时间和参数进行了标准化.

设 $K=(k_{ij})_{(L+1)\times(L+1)}$ 为基因状态切换的转移速率矩阵 (即生成元矩阵), 设 H 为去掉 K 的第一行与第一列所得到的 $L\times L$ 矩阵. 如果 K 不可约, 则由 Perron-Frobenius 定理可知, K 一定具有唯一的零特征值. 设 $\alpha_1, \dots, \alpha_L$ 为 $-K$ 的所有非零特征值, 设 $\beta_1, \dots, \beta_L$ 为矩阵 $-H$ 的所有特征值. 文

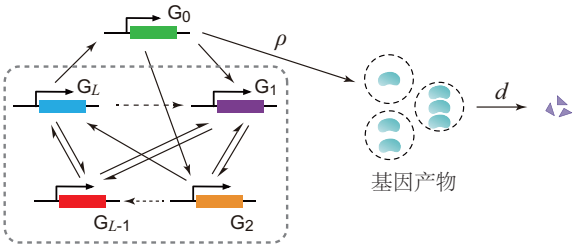

图 3 (网络版彩图) 多状态基因表达模型 (参见文献 [45]). 该模型假定基因可以在多个状态之间进行随机切换, G_0 为激活态, 其余状态 $G_1, \dots, G_L$ 均为抑制态

献 [17] 证明了该模型基因产物数量的平稳分布为

$$p_n^{\rm ss}=\frac{B^n}{n!}\frac{(\gamma_1)_n\cdots(\gamma_{L+1})_n}{(\beta_1)_n\cdots(\beta_L)_n}{}_{L+1}F_L(\gamma_1+n,\ldots,\gamma_{L+1}+n;\alpha_1+n,\ldots,\alpha_L+n;-B),$$

其中 ${}_{L+1}F_L(\gamma_1,\ldots,\gamma_{L+1};\alpha_1,\ldots,\alpha_L;x)$ 为广义超几何函数, $B=\sum_{k=1}^{\infty}kp^k(1-p)=p/(1-p)$ 为平均爆发量, 且常数 $\gamma_1,\ldots,\gamma_{L+1}$ 满足如下方程:

$$\sigma_k(\gamma_1,\ldots,\gamma_{L+1})=\sigma_k(\alpha_1,\ldots,\alpha_L)+\rho\sigma_{k-1}(\beta_1,\ldots,\beta_L),\quad k=1,\ldots,L+1,$$

其中 σ_k 为 k 阶基本对称多项式, 即

$$\sigma_k(\alpha_1,\ldots,\alpha_L)=\sum_{1\leqslant i_1<\cdots<i_k\leqslant L}\alpha_{i_1}\alpha_{i_2}\cdots\alpha_{i_k}.$$

对于一般的多状态模型, Zhang 和 Zhou [117] 指出基因产物数量的平稳分布则可以通过二项矩方法进行重构. Shi 等 [92] 和 Zhang 等 [120] 将该结果推广到了具有任意打开与关闭时间分布的非 Markov 基因表达模型中. Xu 等 [114]、Szavits-Nossan 和 Grima [99] 进一步将该结果推广到了具有确定性降解时间的非 Markov 基因表达模型中. 在基因产物的瞬时分布方面, Chen 和 Jiao [16] 在基因产物没有爆发性时给出了三状态基因表达模型瞬时分布的解析解. Jia 和 Li [45] 给出了任意多状态基因表达模型瞬时分布的解析解, 该结果同时适用于基因产物没有爆发性与有爆发性的情况.

4 连续随机基因表达模型

现有的随机基因表达模型可以分为离散与连续模型两种. 在离散模型中, 基因表达量只能取非负整数值, 其状态空间是离散的; 而在连续模型中, 基因表达量可以取任意非负实数值, 其状态空间是连续的. 以上提到的模型均为离散模型, 而最著名的连续模型包括基于质量作用定律的确定性常微分方程模型 [89] 以及 Friedman 等 [21] 所提出的连续爆发模型. 在连续爆发模型中, 基因产物的生成由复合 Poisson 过程所刻画, 因此该模型本质上是复合 Poisson 过程所驱动的带跳随机微分方程. Jia 等 [52] 发现这两种连续模型均可以看作离散模型当系统的尺度参数趋于无穷时的宏观极限, 常微分方程模型对应于爆发频率与尺度参数成正比的情形, 而连续爆发模型对应于爆发量与尺度参数成正比的情形. 至此, 基因表达的各种离散模型与连续模型得以统一到一套完整的理论框架之中.

接下来给出连续模型的两个例子. 我们仍考虑如下的简单带爆发的基因表达系统:

$$G\xrightarrow{\rho w(x)}G+xP,\quad x\geqslant 0,\quad P\xrightarrow{d}\varnothing.$$

在该系统的离散模型中, 爆发量服从均值为 $B=p/(1-p)$ 的几何分布, 基因产物数量的平稳分布为负二项分布, 参见方程 (2.2). 在该系统的连续模型中, 由于基因产物数量是连续变化的, 我们需要假定爆发量服从指数分布, 其概率密度为 $w(x)=\mathrm{e}^{-x/B}/B$, 这是由于指数分布是几何分布的连续版本. 连续爆发模型由如下的复合 Poisson 过程驱动的随机微分方程所描述:

$$\dot{x}=\dot{\xi}(t)-dx,$$

其中 x 为基因产物水平, $\xi(t)$ 为以 ρ 为爆发频率、以 $w(x)$ 为爆发量分布的复合 Poisson 过程. Friedman 等 [21] 指出, 在稳态下, 该连续模型的基因产物数量服从如下的 Gamma 分布:

$$p^{\rm ss}(x)=\frac{1}{B^s\Gamma(s)}x^{s-1}\mathrm{e}^{-x/B}.$$

文献 [50] 给出了基因产物数量瞬时分布的解析解. 综上所述, 可以看到对于相同的基因表达系统, 离散模型的平稳分布 p_n^{ss} 为负二项分布, 而连续模型的平稳分布 $p^{\mathrm{ss}}(x)$ 为 Gamma 分布. 容易验证, 这两个分布可以通过 Chaturvedi 和 Gardiner [15] 以及 Gardiner 和 Chaturvedi [23] 所提出的 Poisson 表示相连, 即

$$p_n^{\mathrm{ss}} = \int_0^\infty \frac{\mathrm{e}^{-x}x^n}{n!}p^{\mathrm{ss}}(x)dx, \quad n = 0, 1, 2, \ldots \tag{4.1}$$

因此, 负二项分布也称为 Gamma-Poisson 分布 [30, 32].

如果考虑基因状态切换, 则基因表达过程可以由如下的两状态电报模型所描述:

$$G \xrightarrow{a} G^*, \quad G^* \xrightarrow{b} G, \quad G^* \xrightarrow{u} G^* + P, \quad P \xrightarrow{d} \varnothing.$$

该系统离散模型的平稳分布由方程 (2.5) 给出, 其中包含合流超几何函数. 由于该系统没有考虑基因产物的爆发性, 所以其连续模型可以由如下带切换的杂交常微分方程所描述:

$$\begin{array}{l} \dot{x} = u - dx \quad (\text{基因处于激活态}), \\ \quad a \uparrow \Big\| \downarrow b \\ \dot{x} = -dx \quad (\text{基因处于抑制态}). \end{array}$$

该杂交常微分方程模型也被称为分段确定性 Markov 过程 (piecewise-deterministic Markov process). 文献 [56] 指出, 在稳态下, 该连续模型的蛋白数量服从如下的 Beta 分布:

$$p^{\mathrm{ss}}(x) = \frac{\Gamma(\beta)}{\Gamma(\alpha)\Gamma(\beta-\alpha)}s^{1-\beta}x^{\alpha-1}(s-x)^{\beta-\alpha-1}, \quad 0 < x < s.$$

对于相同的基因表达系统, 离散模型的平稳分布 p_n^{ss} 由方程 (2.5) 给出, 而连续模型的平稳分布 $p^{\mathrm{ss}}(x)$ 为 Beta 分布. 容易验证, 这两个分布也可以通过 Poisson 表示相连, 即方程 (4.1) 成立. 因此, 离散电报模型的平稳分布也称为 Beta-Poisson 分布 [58, 109].

通过上述两个例子可以看到, 对于很多随机基因表达系统, 其离散与连续模型可以通过 Poisson 表示相连. Dattani 和 Barahona [18] 证明了, 对于没有反馈、没有爆发的基因表达系统, 其离散与连续模型的平稳分布与瞬时分布一定可以通过 Poisson 表示相连. Wang 等 [112] 将其推广到了没有反馈、有爆发的多状态基因表达系统中. 对于有反馈的复杂基因调控网络, 文献 [112] 进一步证明了离散与连续模型的平稳分布与瞬时分布在蛋白分子数很大时一定可以通过 Poisson 表示相连. 该结果具有广泛的应用, 如果我们得到了某随机基因网络离散模型的解析解, 则可以利用 Poisson 表示自动得到连续模型的解析解.

5 随机基因调控网络

在真实的生物系统中, 多个基因往往会相互调控, 形成复杂的基因调控网络. 基本的基因调控包括正反馈与负反馈两种, 而多种反馈调控又可以相互叠加, 形成各种各样的网络模块. 其中最简单的网络模块应属自调控基因网络, 即某基因所产生的蛋白作为转录因子调控该基因自身的表达. 实验表明, 在大肠杆菌中, 超过 40% 的转录因子都会形成自调控基因网络 (参见文献 [91]), 而其中的大多数为负自调控 (参见文献 [89]).

经典的两状态电报模型并没有考虑基因的反馈调控. 最近, 大量的研究尝试将两状态模型推广到自调控基因网络中, 并对蛋白数量的平稳分布与瞬时分布进行解析求解. 在文献中, 最重要的 4 种自调控基因表达模型包括 Hornos 模型 [34]、Kumar 模型 [60]、Grima 模型 [29] 和修正 Kumar 模型 [41]. 具体地, 设 G 为基因的自由状态, G^* 为基因与蛋白的结合状态, P 为相应的蛋白. Hornos 模型的蛋白动力学由如下的反应所描述:

$$\begin{aligned} &G+P \xrightarrow{\sigma_b} G^*+P, \quad G^* \xrightarrow{\sigma_u} G, \\ &G \xrightarrow{\rho_u} G+P, \quad G^* \xrightarrow{\rho_b} G^*+P, \quad P \xrightarrow{d} \varnothing. \end{aligned}$$

这里蛋白的产生不具有爆发性, 即蛋白分子每次只产生一个. 该模型还假定基因在自由与结合状态下会以不同速率产生蛋白, 正反馈回路对应于结合状态生成速率大于自由状态生成速率的情形, 即 $\rho_b > \rho_u$; 而负反馈回路对应于前者小于后者的情形, 即 $\rho_b < \rho_u$. 该模型蛋白数量的平稳分布由文献 [34] 给出.

Kumar 模型由如下的反应所描述:

$$\begin{aligned} &G+P \xrightarrow{\sigma_b} G^*+P, \quad G^* \xrightarrow{\sigma_u} G, \\ &G \xrightarrow{\rho_u p^k(1-p)} G+kP, \quad G^* \xrightarrow{\rho_b p^k(1-p)} G^*+kP, \quad k \geqslant 1, \quad P \xrightarrow{d} \varnothing. \end{aligned}$$

与 Hornos 模型不同的是, 这里蛋白的产生具有爆发性, 爆发量服从以 p 为参数的几何分布. 该模型的平稳分布由文献 [60] 给出. 需要注意的是, 在上述两个模型中, 基因与蛋白之间的相互作用由催化反应 $G+P \longrightarrow G^*+P$ 所描述, 其中蛋白同时出现在反应的左右两端, 反应前后蛋白数量没有变化. 然而在真实的基因调控过程中, 基因与蛋白之间的相互作用应通过结合与解离实现, 结合发生时蛋白分子应减少一个, 而解离发生时蛋白分子应增加一个.

Grima 模型与修正 Kumar 模型考虑了基因与蛋白之间的结合与解离作用, 其中 Grima 模型由如下的反应所描述 (见图 4(a)):

$$\begin{aligned} &G+P \xrightarrow{\sigma_b} G^*, \quad G^* \xrightarrow{\sigma_u} G+P, \\ &G \xrightarrow{\rho_u} G+P, \quad G^* \xrightarrow{\rho_b} G^*+P, \quad P \xrightarrow{d} \varnothing. \end{aligned}$$

这里蛋白的产生不具有爆发性, 该模型的平稳分布由文献 [29] 给出. 修正 Kumar 模型由如下的反应所描述 (见图 4(b)):

$$G+P \xrightarrow{\sigma_b} G^*, \quad G^* \xrightarrow{\sigma_u} G+P,$$

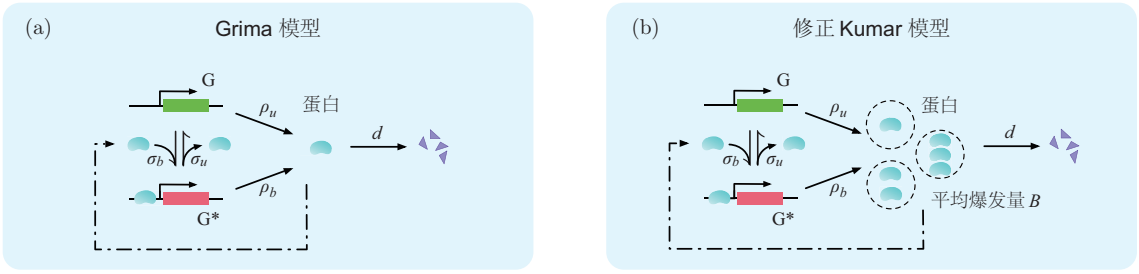

图 4 (网络版彩图) 自调控反馈回路的动力学模型 (参见文献 [41]). **(a) Grima 模型, 其中蛋白的产生不具有爆发性; (b) 修正 Kumar 模型, 其中蛋白的产生具有爆发性**

$$G \xrightarrow{\rho_u p^k(1-p)} G + kP, \quad G^* \xrightarrow{\rho_b p^k(1-p)} G^* + kP, \quad k \geqslant 1, \quad P \xrightarrow{d} \varnothing.$$

这里蛋白的产生具有爆发性, 该模型的平稳分布由文献 [29] 给出. 在上述两个模型中, 基因与蛋白的结合由反应 $G + P \longrightarrow G^*$ 所描述, 而解离由反应 $G^* \longrightarrow G + P$ 所描述, 其中蛋白只出现在反应的一端, 因此真正起到转录因子 (而非催化剂) 的作用.

在上述自调控基因网络模型中, Hornos 模型与 Grima 模型没有考虑蛋白的爆发性, 而 Hornos 模型与 Kumar 模型没有考虑基因与蛋白之间的结合与解离所造成的涨落. 修正 Kumar 模型同时考虑了这两个重要因素, 因此是 4 种模型中最精确的. 修正 Kumar 模型蛋白数量的平稳分布为 (参见文献 [29])

$$\begin{aligned} p_n = {} & K \sum_{k=0}^{n} \frac{(\alpha_1)_k (\alpha_2)_k (s)_{n-k}}{(\beta)_k (1)_k (1)_{n-k}} {}_2F_1(\alpha_1 + k, \alpha_2 + k; \beta + k; -wz_0) w^k p^{n-k} \\ & + KA \sum_{k=0}^{n} \frac{(\alpha_1 + 1)_k (\alpha_2 + 1)_k (s)_{n-k}}{(\beta + 1)_k (1)_k (1)_{n-k}} {}_2F_1(\alpha_1 + 1 + k, \alpha_2 + 1 + k; \beta + 1 + k; -wz_0) w^k p^{n-k} \\ & - KAa \sum_{k=0}^{n-1} \frac{(\alpha_1 + 1)_k (\alpha_2 + 1)_k (s)_{n-1-k}}{(\beta + 1)_k (1)_k (1)_{n-1-k}} {}_2F_1(\alpha_1 + 1 + k, \alpha_2 + 1 + k; \beta + 1 + k; -wz_0) w^k p^{n-1-k}, \end{aligned}$$

其中

$$\begin{aligned} & \alpha_1 + \alpha_2 = \frac{\rho_u + \sigma_u}{d + \sigma_b} - \frac{\rho_b}{d} - 1, \quad \alpha_1 \alpha_2 = \frac{\rho_b}{d} - \frac{(\rho_b - \rho_u)\sigma_u + \rho_u d}{d(d + \sigma_b)}, \\ & \beta = \frac{\sigma_u}{d + \sigma_b} + \frac{\rho_u \sigma_b p}{(d + \sigma_b)[d(1-p) + \sigma_b]}, \quad A = \frac{[(\rho_b - \rho_u)d + \rho_u \sigma_b]p}{d[d(1-p) + \sigma_b]\beta}, \\ & a = \frac{(\rho_b - \rho_u)d + \rho_b \sigma_b}{(\rho_b - \rho_u)d + \rho_u \sigma_b}, \quad w = \frac{(d + \sigma_b)(1-p)}{d(1-p) + \sigma_b}, \quad z_0 = \frac{d}{d + \sigma_b}, \end{aligned}$$

且 $K = (1-p)^s [{}_2F_1(\alpha_1, \alpha_2; \beta; w(1-z_0)) + A(1-a){}_2F_1(\alpha_1 + 1, \alpha_2 + 1; \beta + 1; w(1-z_0))]^{-1}$ 为归一化常数. 此外, Liu 等 [66] 还考虑了同时具有正反馈与负反馈的极小耦合基因网络, 并对蛋白数量的平稳分布进行了解析求解. Jia 等 [50] 将该结果推广到了具有蛋白爆发性的耦合基因网络中, 并研究了该模型的各种宏观极限.

在瞬时分布方面, Ramos 等 [86] 对 Hornos 模型的瞬时分布进行了解析求解, 然而他们给出的解是错误的. 该系统 Markov 跳过程模型的转移速率矩阵的特征值可能是复数, 然而文献 [86] 断言其所有特征值均为实数. Wu 等 [113] 对上述工作进行了修正, 利用谱分解方法得到了 Hornos 模型与 Kumar 模型瞬时分布的解析解以及转移速率矩阵的完整谱刻画, 其中非零特征值所对应的特征函数是具有四个正则奇点的 Heun 函数 (超几何函数只有三个正则奇点), 而所有非零特征值满足一个非平凡的连分数方程. 值得注意的是, 自调控基因网络的概率模型属于非对称 Markov 过程的范畴, 其生成元矩阵为无界非自伴算子. 刻画无界非自伴算子的完整谱结构在泛函分析中是极端困难的, 因此该工作可能对无界非自伴算子谱理论的发展有所启发. 此外, Jia 和 Grima [40] 利用复变函数方法, 在快速基因切换情形求解了修正 Kumar 模型蛋白数量的瞬时分布, 然而一般情况的瞬时分布求解至今仍未解决.

随机基因表达领域的另一个重要问题是研究基因表达噪声的来源与影响因素, 这里的噪声通常由基因产物分子数的变异系数或 Fano 因子所刻画. Paulsson 和 Swain 在这方面有一系列开创性工作. Paulsson [78]、Pedraza 和 Paulsson [82] 基于经典的基因表达模型, 根据不同的生物物理起源对基因表达噪声进行了分解, 并在一些反馈网络中给出了噪声的下界 [64]. Bowsher 和 Swain [10] 及 Swain 等 [98]

对复杂基因表达模型进行了噪声分解, 明确了基因表达的内噪声与外噪声. Lei [63] 进一步考虑了由于外部环境因素造成的速率参数的涨落对基因表达噪声的影响. Liu 等 [67] 和 Jia 等 [51] 分别将上述结果推广到了自调控基因网络中, 给出了基因表达噪声的两种不同的分解, 明确了反馈拓扑对基因表达噪声的影响. 当基因切换速率很快时, 正反馈会增大基因表达噪声, 而负反馈会减少基因表达噪声 (参见文献 [51]). 此外, Rosenfeld 等 [89] 研究了自调控基因网络的响应速度 (定义为系统演化到平稳均值的一半所需要的时间), 发现正反馈会减慢系统的响应速度, 而负反馈会加快系统的响应速度. 文献 [46] 将该结果推广到了自调控基因网络的弛豫速度 (定义为生成元第一非零特征值的模的倒数), 并发现了类似的现象.

上述研究基本局限于自调控基因网络, 但现实中基因调控网络的拓扑结构可能非常复杂. 早在 1981 年, 著名生物学家 Thomas [107] 提出了基因调控网络的两个重要猜想, 其中第一猜想断言网络中存在正反馈回路是系统产生多稳态 (multistability) 的必要条件, 第二猜想断言网络中存在负反馈回路是系统产生持续振荡的必要条件. Thomas [108] 随后于 1999 年提出了基因网络的第三猜想, 该猜想断言网络中同时存在正反馈与负反馈回路是系统产生混沌的必要条件. 对于基因网络的确定性模型, 多稳态即系统存在多个稳定的不动点, 持续振荡即系统存在稳定的极限环. 然而对于基因网络的随机模型, 多稳态应理解为基因产物分子数的多峰分布, 而持续振荡通常由基因产物分子数的自相关函数与功率谱所刻画. 这里功率谱定义为自相关函数的 Fourier 变换, 系统存在持续振荡当且仅当功率谱非单调 [85]. 更有意思的是, 在 Hopf 分支点附近, 基因网络可能存在噪声诱导振荡 (noised-induced oscillations) 现象, 即确定性模型只存在稳定不动点, 但随机模型可能存在由噪声所诱导的非单调功率谱 [69, 106]. 在解析理论方面, 文献 [47] 在自调控基因网络中给出了蛋白数量自相关函数与功率谱的解析表达式, 并研究了系统从随机爆发到随机振荡演化过程中的随机分岔现象.

如果基因之间的调控关系非常复杂, 我们通常无法对相应基因表达模型的概率分布、自相关函数、功率谱等重要统计量进行精确的解析求解. 最近, 大量的工作尝试对复杂随机基因网络的重要统计量进行近似的解析或半解析求解, 其中重要的近似方法包括多尺度方法 [70, 84, 104]、线性噪声近似 [69, 106]、矩封闭方法 [28, 61, 93]、Padé近似 [31]、线性映射近似 [13] 和 Holimap 方法 [44] 等.

关于随机基因网络的能量景观 (energy landscape) 也有大量的研究工作. 粗略地讲, 基因表达的概率分布与能量景观之间可以通过 Boltzmann 分布相联系. Ge 等 [24] 和 Zhang 等 [119] 研究了细胞周期的随机布尔网络模型 (也称为 Boltzmann 机) 及其能量景观, 发现细胞周期的相关信号通路在很大的噪声范围内仍能保持稳定. Zhu 等 [122] 建立了 λ 噬菌体基因开关动力学的随机微分方程模型及其能量景观理论. Assaf 等 [3]、Lv 等 [68] 和 Ge 等 [25] 利用随机过程的大偏差理论, 分别在不同的时间尺度假设下求解了自调控基因网络能量景观的解析表达式. 此外, Li 和 Wang [65] 及 Wang 等 [110] 基于生化网络的随机微分方程模型, 提出了基因调控网络的具有一般性的能量景观理论. Jia 等 [48] 研究了细菌表型切换与风险对冲的随机数学模型, 并求解了该模型的能量景观.

6 基因表达、细胞体积、细胞周期的耦合随机模型

截至目前, 我们介绍的所有模型都没有明确考虑细胞的生长与分裂, 因此只适用于非生长细胞, 而不适用于生长细胞. 具体地, 真实的生长细胞都要经历细胞生长、细胞分裂、基因复制等重要的细胞周期事件. 例如, 在真核生物中, 细胞周期被分为 G1 期、S 期、G2 期、M 期这四个阶段, 其中基因复制发生在 S 期, 细胞分裂发生在 M 期, 而在这四个阶段中细胞体积是不断增大的. 在基因复制的过程

中, 基因的拷贝数会增加一倍; 而在细胞分裂的过程中, 基因以及基因产物的拷贝数会减少一半, 如此往复. 这些重要的细胞周期事件都没有包含在经典的基因表达模型中.

近十年来, 随着荧光显微法、微流控装置、母机装置等实验技术的突破性进展, 人们可以在单细胞水平上获得基因表达的高通量时间序列数据 (参见文献 [76, 102, 111]). 具体地, 人们可以追踪单细胞内基因表达随时间的变化, 有的实验甚至可以持续追踪数十乃至数百个细胞周期. 在细胞分裂发生后, 母细胞分裂为两个子细胞, 该实验通常选取其中的某个子细胞进行继续追踪, 从而保证在任意时刻只有一个细胞被追踪, 因此所有被追踪的细胞都来源于同一个母细胞, 形成一个细胞谱系. 这种实验被称为单细胞谱系测量. 然而很多其他实验技术 (如流式细胞术、单细胞 RNA 测序等) 则是测量很多个细胞在同一个时间点上的基因表达, 这种实验被称为单细胞群体测量. 最近的研究表明, 基因表达的谱系测量与群体测量可能产生不同的涨落行为 [22, 103].

近年来的研究热点之一是, 以经典的随机基因表达模型为基础, 结合细胞周期中的关键生物学因素, 发展更加贴近生物学实际的基因表达动力学与细胞周期事件的耦合随机模型, 解释单细胞谱系测量与群体测量所呈现的全新动力学现象. 事实上, Berg [5] 早在 1978 年就发表了该领域的开创性工作, 该工作同时考虑了带爆发的基因表达动力学与细胞分裂, 并求解了该模型蛋白数量的瞬时分布与稳态群体分布. 然而该工作假定蛋白的降解速率可以被忽略, 该假设对于原核细胞是近似成立的, 然而在真核细胞中具有较大的偏差 (参见文献 [42]). Beentjes 等 [4] 和 Perez-Carrasco [83] 在上述模型的基础上进一步考虑了基因复制与基因产物的降解, 并对基因产物数量的瞬时分布、稳态谱系分布和稳态群体分布进行了解析求解.

需要指出的是, 以上工作没有明确考虑基因状态切换. Sun 等 [95] 改进了上述模型, 考虑了两状态电报模型、基因复制、细胞分裂的耦合随机动力学, 并求解了基因产物分子数的均值与方差. 文献 [43] 进一步给出了基因产物分子数的瞬时分布、稳态谱系分布和稳态群体分布的解析解. Jia 和 Grima [42] 研究了基因表达由细胞周期所诱导的随机振荡行为, 求解了基因产物分子数的自相关函数与功率谱, 给出了一系列理论预测并得到了单细胞谱系实验的验证.

在经典的基因表达模型中, 转录速率通常假定为常数. 然而最近的研究 [7, 77, 96] 表明, 在真核细胞中, 很多基因的 mRNA 与蛋白的浓度在每个细胞周期内近似保持不变, 该现象称为基因产物的浓度恒定性 (concentration homeostasis). 这里的恒定性指在一定外部环境范围内, 生物体通过器官与器官之间的协调联系, 得以维持体系内环境处于相对不变状态的这种特性. 由于浓度恒定性, 基因产物分子数正比于细胞体积, 因此该基因的转录速率应该也正比于细胞体积. 转录速率与细胞体积成正比的现象在生物学中被称为平衡生物合成 (balanced biosynthesis) [96]. 由于细胞体积在细胞周期内是不断增大的, 转录速率应该也是不断增大的. 因此假定转录速率为常数不足以描述真核细胞中真实的转录动力学.

为了解释浓度恒定性, 我们需要发展基因表达动力学、细胞体积动力学、细胞周期事件的耦合随机模型. 现有的动力学模型通常假定细胞体积在每个细胞周期内是指数生长的, 且当细胞分裂发生时, 细胞体积变为分裂前的一半. 细胞体积的指数增长在很多细胞类型中都得到了实验验证 (参见文献 [11, 94]). Cao 和 Grima [14] 考虑了平衡生物合成与确定性的细胞体积动力学, 并在上述假设下求解了基因产物数量的瞬时分布、稳态谱系分布和稳态群体分布. 然而该工作并没有考虑细胞体积的随机性和基因状态切换. Thomas 和 Shahrezaei [105] 考虑了基因状态切换, 并对该模型进行了解析求解. 然而该工作没有考虑基因复制. 文献 [43, 49] 鉴于以上模型的不足提出了一个较为精细的基因表达、细胞体积、细胞周期的耦合随机模型 (见图 5), 并对该模型进行了解析求解. 由于上述模型过于复杂, 具体的数学细节不再陈述.

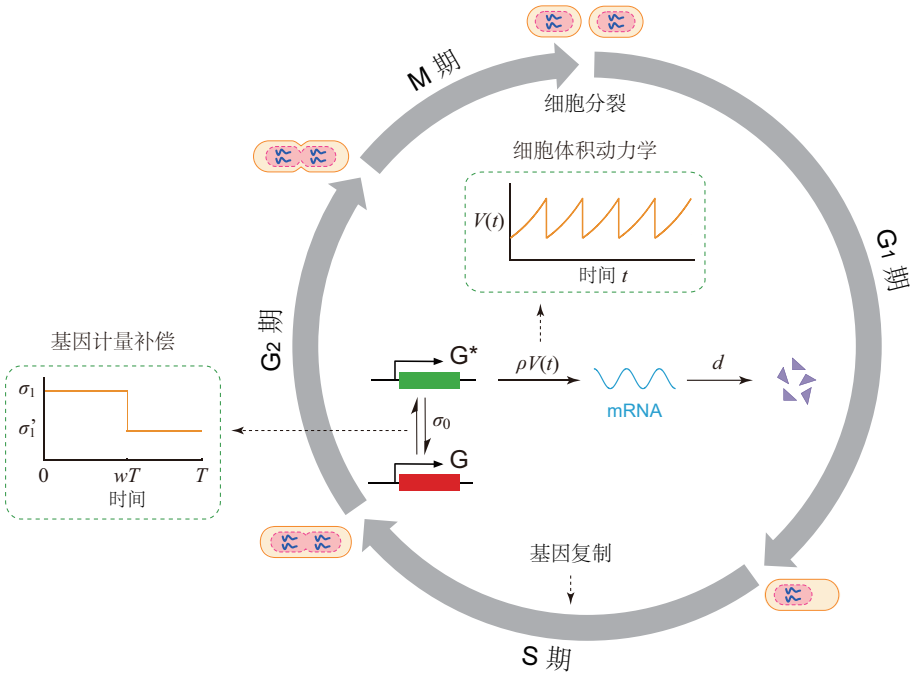

图 5 (网络版彩图) 基因表达、细胞体积、细胞周期的耦合动力学模型 (参见文献 [43]). 该模型假定细胞体积 $\boldsymbol{V(t)}$ 在每个细胞周期内是指数生长的. 由于平衡生物合成, 基因产物的合成速率与细胞体积成正比. 由于基因计量补偿效应, 基因的激活速率在基因复制后会产生下降. 当细胞分裂发生时, 母细胞平分为两个体积相同的子细胞, 且每个基因产物分子以相同的概率分配到两个子细胞中

7 总结与讨论

近二十年来, 基因表达随机动力学的理论与实验都取得了突破性进展. 本文对单细胞基因表达与基因调控的经典数学模型及其理论结果进行了较为全面的梳理, 包括经典一状态与两状态基因表达模型、经典两阶段与三阶段基因表达模型、多状态基因表达模型、自调控基因网络模型、离散与连续基因表达模型等. 随机过程作为研究随机现象的数学工具, 在基因表达动力学的研究中扮演至关重要的作用. 对于离散基因表达模型, 其动力学可以由无穷状态空间上的 Markov 跳过程所描述. 对于连续基因表达模型, 其动力学可以由 (带切换的杂交) 常微分方程或随机微分方程所描述, 其中常微分方程对应于基因产物不具有爆发性的情形, 随机微分方程对应于基因产物具有爆发性的情形. 特别地, 概率论中的很多经典概率分布都可以看作不同基因表达模型的基因产物分布, 如 Poisson 分布、几何分布、指数分布、负二项分布、Gamma 分布和 Beta 分布等. 而且该领域与特殊函数理论有着深刻的联系, 复杂基因表达模型的基因产物分布通常带有合流超几何函数、Gauss 超几何函数、广义超几何函数和 Heun 函数等. 对不同基因表达模型的基因产物分布进行解析求解, 对深入理解细胞的表型异质性与基因调控机制, 以及精确推断基因表达动力学的模型与参数都起到至关重要的作用. 鉴于本文作者知识面的局限性, 单细胞随机基因表达领域仍有许多研究方向并未在本文中系统性阐述, 如时滞与非 Markov 基因表达模型 [20,100,118]、基因表达动力学的参数推断与模型选择 [53,57,74] 等. 希望本文可以帮助读者快速了解随机基因表达与基因调控动力学的经典结果, 迅速进入该领域的研究前沿.

致谢 作者衷心感谢钱敏平教授长期以来的栽培、关心和帮助.

参考文献

1 Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol, 2010, 11: 1–12
2 Anderson D F, Kurtz T G. Stochastic Analysis of Biochemical Systems. Berlin: Springer, 2015
3 Assaf M, Roberts E, Luthey-Schulten Z. Determining the stability of genetic switches: Explicitly accounting for mRNA noise. Phys Rev Lett, 2011, 106: 248102
4 Beentjes C H L, Perez-Carrasco R, Grima R. Exact solution of stochastic gene expression models with bursting, cell cycle and replication dynamics. Phys Rev E, 2020, 101: 032403
5 Berg O G. A model for the statistical fluctuations of protein numbers in a microbial population. J Theoret Biol, 1978, 71: 587–603
6 Bernstein J A, Lin P H, Cohen S N, et al. Global analysis of *Escherichia coli* RNA degradosome function using DNA microarrays. Proc Natl Acad Sci USA, 2004, 101: 2758–2763
7 Berry S, Pelkmans L. Mechanisms of cellular mRNA transcript homeostasis. Trends Cell Biol, 2022, 32: 655–668
8 Bo S, Celani A. Multiple-scale stochastic processes: Decimation, averaging and beyond. Phys Rep, 2017, 670: 1–59
9 Bokes P, King J R, Wood A T A, et al. Exact and approximate distributions of protein and mRNA levels in the low-copy regime of gene expression. J Math Biol, 2012, 64: 829–854
10 Bowsher C G, Swain P S. Identifying sources of variation and the flow of information in biochemical networks. Proc Natl Acad Sci USA, 2012, 109: E1320–E1328
11 Cadart C, Monnier S, Grilli J, et al. Size control in mammalian cells involves modulation of both growth rate and cell cycle duration. Nat Commun, 2018, 9: 3275
12 Cai L, Friedman N, Xie X S. Stochastic protein expression in individual cells at the single molecule level. Nature, 2006, 440: 358–362
13 Cao Z, Grima R. Linear mapping approximation of gene regulatory networks with stochastic dynamics. Nat Commun, 2018, 9: 3305
14 Cao Z, Grima R. Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. Proc Natl Acad Sci USA, 2020, 117: 4682–4692
15 Chaturvedi S, Gardiner C W. The poisson representation. II Two-time correlation functions. J Stat Phys, 1978, 18: 501–522
16 Chen J, Jiao F. A novel approach for calculating exact forms of mRNA distribution in single-cell measurements. Mathematics, 2021, 10: 27
17 Chen M, Luo S, Cao M, et al. Exact distributions for stochastic gene expression models with arbitrary promoter architecture and translational bursting. Phys Rev E, 2022, 105: 014405
18 Dattani J, Barahona M. Stochastic models of gene transcription with upstream drives: Exact solution and sample path characterization. J R Soc Interface, 2017, 14: 20160833
19 Delbrück M. Statistical fluctuations in autocatalytic reactions. J Chem Phys, 1940, 8: 120–124
20 Fralix B, Holmes M, Löpker A. A Markovian arrival stream approach to stochastic gene expression in cells. J Math Biol, 2023, 86: 79
21 Friedman N, Cai L, Xie X S. Linking stochastic dynamics to population distribution: An analytical framework of gene expression. Phys Rev Lett, 2006, 97: 168302
22 García-García R, Genthon A, Lacoste D. Linking lineage and population observables in biological branching processes. Phys Rev E, 2019, 99: 042413
23 Gardiner C W, Chaturvedi S. The Poisson representation. I. A new technique for chemical master equations. J Stat Phys, 1977, 17: 429–468
24 Ge H, Qian H, Qian M. Synchronized dynamics and non-equilibrium steady states in a stochastic yeast cell-cycle network. Math Biosci, 2008, 211: 132–152
25 Ge H, Qian H, Xie X S. Stochastic phenotype transition of a single cell in an intermediate region of gene state switching. Phys Rev Lett, 2015, 114: 078101
26 Gillespie D T. Exact stochastic simulation of coupled chemical reactions. J Phys Chem, 1977, 81: 2340–2361
27 Golding I, Paulsson J, Zawilski S M, et al. Real-time kinetics of gene activity in individual bacteria. Cell, 2005, 123: 1025–1036
28 Grima R. A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. J Chem Phys, 2012, 136: 154105
29 Grima R, Schmidt D R, Newman T J. Steady-state fluctuations of a genetic feedback loop: An exact solution. J Chem Phys, 2012, 137: 035104
30 Grün D, Kester L, Van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods, 2014, 11: 637–640

31 Gupta A, Khammash M. Frequency spectra and the color of cellular noise. Nat Commun, 2022, 13: 4305
32 Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol, 2019, 20: 296
33 Harper C V, Finkenstädt B, Woodcock D J, et al. Dynamic analysis of stochastic transcription cycles. PLoS Biol, 2011, 9: e1000607
34 Hornos J E M, Schultz D, Innocentini G C P, et al. Self-regulating gene: An exact solution. Phys Rev E, 2005, 72: 051907
35 Iyer-Biswas S, Hayot F, Jayaprakash C. Stochasticity of gene products from transcriptional pulsing. Phys Rev E, 2009, 79: 031911
36 Jia C. Model simplification and loss of irreversibility. Phys Rev E, 2016, 93: 052149
37 Jia C. Reduction of Markov chains with two-time-scale state transitions. Stochastics, 2016, 88: 73–105
38 Jia C. Simplification of Markov chains with infinite state space and the mathematical theory of random gene expression bursts. Phys Rev E, 2017, 96: 032402
39 Jia C. Kinetic foundation of the zero-inflated negative binomial model for single-cell RNA sequencing data. SIAM J Appl Math, 2020, 80: 1336–1355
40 Jia C, Grima R. Dynamical phase diagram of an auto-regulating gene in fast switching conditions. J Chem Phys, 2020, 152: 174110
41 Jia C, Grima R. Small protein number effects in stochastic models of autoregulated bursty gene expression. J Chem Phys, 2020, 152: 084115
42 Jia C, Grima R. Frequency domain analysis of fluctuations of mRNA and protein copy numbers within a cell lineage: Theory and experimental validation. Phys Rev X, 2021, 11: 021032
43 Jia C, Grima R. Coupling gene expression dynamics to cell size dynamics and cell cycle events: Exact and approximate solutions of the extended telegraph model. Iscience, 2023, 26: 105746
44 Jia C, Grima R. Holimap: An accurate and efficient method for solving stochastic gene network dynamics. Nat Commun, 2024, 15: 6557
45 Jia C, Li Y. Analytical time-dependent distributions for gene expression models with complex promoter switching mechanisms. SIAM J Appl Math, 2023, 83: 1572–1602
46 Jia C, Qian H, Chen M, et al. Relaxation rates of gene expression kinetics reveal the feedback signs of autoregulatory gene networks. J Chem Phys, 2018, 148: 095102
47 Jia C, Qian H, Zhang M Q. Exact power spectrum in a minimal hybrid model of stochastic gene expression oscillations. SIAM J Appl Math, 2024, 84: 1204–1226
48 Jia C, Qian M, Kang Y, et al. Modeling stochastic phenotype switching and bet-hedging in bacteria: Stochastic nonlinear dynamics and critical state identification. Quant Biol, 2014, 2: 110–125
49 Jia C, Singh A, Grima R. Concentration fluctuations in growing and dividing cells: Insights into the emergence of concentration homeostasis. PLoS Comput Biol, 2022, 18: e1010574
50 Jia C, Wang L Y, Yin G G, et al. Single-cell stochastic gene expression kinetics with coupled positive-plus-negative feedback. Phys Rev E, 2019, 100: 052406
51 Jia C, Xie P, Chen M, et al. Stochastic fluctuations can reveal the feedback signs of gene regulatory networks at the single-molecule level. Sci Rep, 2017, 7: 16037
52 Jia C, Zhang M Q, Qian H. Emergent Lévy behavior in single-cell stochastic gene expression. Phys Rev E, 2017, 96: 040402
53 Jiao F, Li J, Liu T, et al. What can we learn when fitting a simple telegraph model to a complex gene expression model? PLoS Comput Biol, 2024, 20: e1012118
54 Jiao F, Sun Q, Tang M, et al. Distribution modes and their corresponding parameter regions in stochastic gene transcription. SIAM J Appl Math, 2015, 75: 2396–2420
55 Jiao F, Zhu C. Regulation of gene activation by competitive cross talking pathways. Biophys J, 2020, 119: 1204–1214
56 Karmakar R, Bose I. Graded and binary responses in stochastic gene expression. Phys Biol, 2004, 1: 197–204
57 Kilic Z, Schweiger M, Moyer C, et al. Gene expression model inference from snapshot RNA data using Bayesian non-parametrics. Nat Comput Sci, 2023, 3: 174–183
58 Kim J K, Marioni J C. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. Genome Biol, 2013, 14: 1–12
59 Ko M S H. A stochastic model for gene induction. J Theoret Biol, 1991, 153: 181–194
60 Kumar N, Platini T, Kulkarni R V. Exact distributions for stochastic gene expression models with bursting and feedback. Phys Rev Lett, 2014, 113: 268105
61 Lakatos E, Ale A, Kirk P D W, et al. Multivariate moment closure techniques for stochastic kinetic models. J Chem

Phys, 2015, 143
62 Larsson A J M, Johnsson P, Hagemann-Jensen M, et al. Genomic encoding of transcriptional burst kinetics. Nature, 2019, 565: 251–254
63 Lei J. Stochasticity in single gene expression with both intrinsic noise and fluctuation in kinetic parameters. J Theoret Biol, 2009, 256: 485–492
64 Lestas I, Vinnicombe G, Paulsson J. Fundamental limits on the suppression of molecular fluctuations. Nature, 2010, 467: 174–178
65 Li C, Wang J. Landscape and flux reveal a new global view and physical quantification of mammalian cell cycle. Proc Natl Acad Sci USA, 2014, 111: 14130–14135
66 Liu P, Yuan Z, Huang L, et al. Roles of factorial noise in inducing bimodal gene expression. Phys Rev E, 2015, 91: 062706
67 Liu P, Yuan Z, Wang H, et al. Decomposition and tunability of expression noise in the presence of coupled feedbacks. Chaos, 2016, 26: 043108
68 Lv C, Li X, Li F, et al. Constructing the energy landscape for genetic switching system driven by intrinsic noise. PLoS One, 2014, 9: e88167
69 McKane A J, Nagy J D, Newman T J, et al. Amplified biochemical oscillations in cellular systems. J Stat Phys, 2007, 128: 165–191
70 Mélykúti B, Hespanha J P, Khammash M. Equilibrium distributions of simple biochemical reaction systems for time-scale separation in stochastic reaction networks. J R Soc Interface, 2014, 11: 20140054
71 Miao Z, Deng K, Wang X, et al. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. Bioinformatics, 2018, 34: 3223–3224
72 Moran M A, Satinsky B, Gifford S M, et al. Sizing up metatranscriptomics. ISME J, 2013, 7: 237–243
73 Munsky B, Khammash M. The finite state projection algorithm for the solution of the chemical master equation. J Chem Phys, 2006, 124: 044104
74 Munsky B, Li G, Fox Z R, et al. Distribution shapes govern the discovery of predictive models for gene regulation. Proc Natl Acad Sci USA, 2018, 115: 7533–7538
75 Munsky B, Neuert G, van Oudenaarden A. Using gene expression noise to understand gene regulation. Science, 2012, 336: 183–187
76 Nakaoka H, Wakamoto Y. Aging, mortality, and the fast growth trade-off of Schizosaccharomyces pombe. PLoS Biol, 2017, 15: e2001109
77 Padovan-Merhar O, Nair G P, Biaesch A G, et al. Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. Mol Cell, 2015, 58: 339–352
78 Paulsson J. Summing up the noise in gene networks. Nature, 2004, 427: 415–418
79 Paulsson J. Models of stochastic gene expression. Phys Life Rev, 2005, 2: 157–175
80 Paulsson J, Ehrenberg M. Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. Phys Rev Lett, 2000, 84: 5447–5450
81 Peccoud J, Ycart B. Markovian modeling of gene-product synthesis. Theor Population Biol, 1995, 48: 222–234
82 Pedraza J M, Paulsson J. Effects of molecular memory and bursting on fluctuations in gene expression. Science, 2008, 319: 339–343
83 Perez-Carrasco R, Beentjes C, Grima R. Effects of cell cycle variability on lineage and population measurements of messenger RNA abundance. J R Soc Interface, 2020, 17: 20200360
84 Popović N, Marr C, Swain P S. A geometric analysis of fast-slow models for stochastic gene expression. J Math Biol, 2016, 72: 87–122
85 Qian H, Qian M. Pumped biochemical reactions, nonequilibrium circulation, and stochastic resonance. Phys Rev Lett, 2000, 84: 2271–2274
86 Ramos A F, Innocentini G C P, Hornos J E M. Exact time-dependent solutions for a self-regulating gene. Phys Rev E, 2011, 83: 062902
87 Risso D, Perraudeau F, Gribkova S, et al. A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun, 2018, 9: 284
88 Robinson M D, McCarthy D J, Smyth G K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 2010, 26: 139–140
89 Rosenfeld N, Elowitz M B, Alon U. Negative autoregulation speeds the response times of transcription networks. J Mol Biol, 2002, 323: 785–793
90 Shahrezaei V, Swain P S. Analytical distributions for stochastic gene expression. Proc Natl Acad Sci USA, 2008, 105: 17256–17261

91 Shen-Orr S S, Milo R, Mangan S, et al. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet, 2002, 31: 64–68

92 Shi C, Jiang Y, Zhou T. Queuing models of gene expression: Analytical distributions and beyond. Biophys J, 2020, 119: 1606–1616

93 Singh A, Hespanha J P. Approximate moment dynamics for chemically reacting systems. IEEE Trans Automat Control, 2010, 56: 414–418

94 Soifer I, Robert L, Amir A. Single-cell analysis of growth in budding yeast and bacteria reveals a common size regulation strategy. Curr Biol, 2016, 26: 356–361

95 Sun Q, Jiao F, Lin G, et al. The nonlinear dynamics and fluctuations of mRNA levels in cell cycle coupled transcription. PLoS Comput Biol, 2019, 15: e1007017

96 Sun X M, Bowman A, Priestman M, et al. Size-dependent increase in RNA polymerase II initiation rates mediates gene expression scaling with cell size. Curr Biol, 2020, 30: 1217–1230

97 Suter D M, Molina N, Gatfield D, et al. Mammalian genes are transcribed with widely different bursting kinetics. Science, 2011, 332: 472–474

98 Swain P S, Elowitz M B, Siggia E D. Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc Natl Acad Sci USA, 2002, 99: 12795–12800

99 Szavits-Nossan J, Grima R. Steady-state distributions of nascent RNA for general initiation mechanisms. Phys Rev Res, 2023, 5: 013064

100 Szavits-Nossan J, Grima R. Solving stochastic gene-expression models using queueing theory: A tutorial review. Biophys J, 2024, 123: 1034–1057

101 Taniguchi Y, Choi P J, Li G W, et al. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science, 2010, 329: 533–538

102 Tanouchi Y, Pai A, Park H, et al. A noisy linear map underlies oscillations in cell size and gene expression in bacteria. Nature, 2015, 523: 357–360

103 Thomas P. Analysis of cell size homeostasis at the single-cell and population level. Front Phys China, 2018, 6: 64

104 Thomas P, Popović N, Grima R. Phenotypic switching in gene regulatory networks. Proc Natl Acad Sci USA, 2014, 111: 6994–6999

105 Thomas P, Shahrezaei V. Coordination of gene expression noise with cell size: Extrinsic noise versus agent-based models of growing cell populations. J R Soc Interface, 2021, 18: 20210274

106 Thomas P, Straube A V, Timmer J, et al. Signatures of nonlinearity in single cell noise-induced oscillations. J Theoret Biol, 2013, 335: 222–234

107 Thomas R. On the relation between the logical structure of systems and their ability to generate multiple steady states or sustained oscillations In: Numerical Methods in the Study of Critical Phenomena. Berlin-Heidelberg: Springer, 1981: 180–193

108 Thomas R. Deterministic chaos seen in terms of feedback circuits: Analysis, synthesis, "labyrinth chaos". Internat J Bifur Chaos, 1999, 9: 1889–1905

109 Vu T N, Wills Q F, Kalari K R, et al. Beta-Poisson model for single-cell RNA-seq data analyses. Bioinformatics, 2016, 32: 2128–2135

110 Wang J, Xu L, Wang E. Potential landscape and flux framework of nonequilibrium networks: Robustness, dissipation, and coherence of biochemical oscillations. Proc Natl Acad Sci USA, 2008, 105: 12271–12276

111 Wang P, Robert L, Pelletier J, et al. Robust growth of *Escherichia coli*. Curr Biol, 2010, 20: 1099–1103

112 Wang X, Li Y, Jia C. Poisson representation: A bridge between discrete and continuous models of stochastic gene regulatory networks. J R Soc Interface, 2023, 20: 20230467

113 Wu B, Holehouse J, Grima R, et al. Solving the time-dependent protein distributions for autoregulated bursty gene expression using spectral decomposition. J Chem Phys, 2024, 160: 074105

114 Xu H, Skinner S O, Sokac A M, et al. Stochastic kinetics of nascent RNA. Phys Rev Lett, 2016, 117: 128101

115 Yin G G, Zhang Q. Continuous-Time Markov Chains and Applications: A Two-Time-Scale Approach. New York: Springer, 2012

116 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

117 Zhang J, Zhou T. Promoter-mediated transcriptional dynamics. Biophys J, 2014, 106: 479–488

118 Zhang J, Zhou T. Markovian approaches to modeling intracellular reaction processes with molecular memory. Proc Natl Acad Sci USA, 2019, 116: 23542–23550

119 Zhang Y, Qian M, Ouyang Q, et al. Stochastic model of yeast cell-cycle network. Phys D, 2006, 219: 35–39

120 Zhang Z, Deng Q, Wang Z, et al. Exact results for queuing models of stochastic transcription with memory and

crosstalk. Phys Rev E, 2021, 103: 062414
121 Zhou T, Zhang J. Analytical results for a multistate gene model. SIAM J Appl Math, 2012, 72: 789–818
122 Zhu X M, Yin L, Hood L, et al. Robustness, stability and efficiency of phage λ genetic switch: Dynamical structure analysis. J Bioinform Comput Biol, 2004, 2: 785–817

Single-cell stochastic gene expression dynamics: Mathematical models and analytical theory

Chen Jia

Abstract Gene expression is one of the most crucial dynamic processes within cells. Gene expression and gene regulation play decisive roles in cell differentiation, proliferation, apoptosis, signal transduction, stress response, and the onset and progression of various diseases. Over the past two decades, significant advancements have been made in both the theory and experiments of stochastic gene expression dynamics. In this review, we introduce the classical mathematical models and theoretical frameworks for single-cell stochastic gene expression, with a particular focus on the analytical methods for deriving the copy number distributions of mRNAs and proteins. Additionally, we provide a comprehensive review of key literature in this field.

Keywords **gene regulation, gene expression bursting, Markov process, hybrid differential equation, steady-state distribution, time-dependent distribution, power spectrum, special function**

MSC(2020) **60J27, 60J28, 34A38, 60H10, 92C40, 92B05**

doi: 10.1360/SSM-2024-0263