第 56 卷　第 3 期
2020 年 5 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 56, No. 3
May, 2020

DOI:10.13232/j.cnki.jnju.2020.03.010

基因转录爆发的建模研究

李佳云,吴人杰*
(南京大学物理学院,南京,210093)

摘　要:基因转录是细胞最重要的活动之一,涉及众多分子事件,且不同基因间存在显著差异性.建立基因转录的模型有助于理解复杂的转录动力学和调控机制.如何构建合适的转录模型依然具有大的挑战性.近年来的实验发现,转录爆发是一类普遍的转录模式,揭示其特征(如转录爆发的频率和大小以及激活态和沉默态的持续时间等)和调控机制是当前的研究热点.人们相继提出两态模型和多态模型来理解转录现象.有些模型不再是简单的唯象模型,而是考虑了转录的分子过程,能够深入研究转录的内在机理.结合最近的实验和理论研究,综述不同转录模型的特点、合理性及其适用范围,特别比较了各个模型的优缺点,有助于在研究中选取合适的转录模型.随着单细胞实验技术的发展,构建基因转录的定量模型将起到越来越重要的作用.

关键词:基因转录,转录爆发,两态模型,多态模型,适用范围
中图分类号:Q615　　**文献标识码:**A

Modeling of transcriptional bursting

Li Jiayun, Wu Renjie*
(School of Physics, Nanjing University, Nanjing, 210093, China)

Abstract: Gene transcription is one of the most important cellular activities, involving various molecular events and exhibiting great variability among genes. Modeling of gene transcription can promote our understanding of the complex mechanisms for transcriptional kinetics and regulation. It is still challenging to construct suitable models under different conditions. It is established that transcriptional bursting has been a ubiquitous mode; it is essential to unravel the features of transcriptional bursting (such as burst frequency and size, as well as the duration of active and inactive gene states) and underlying regulatory mechanisms. Two-state and multi-state models have been proposed to investigate transcriptional bursting. Some models are no longer simple phenomenological ones; instead, they take into account molecular events involved in transcription and thus can be used to explore the intrinsic mechanisms for transcription. Integrating recent experimental and theoretical studies, the current work reviews widely used models of transcriptional bursting in the literature, including the two-state, continuum, multi-scale, and Wang-Liu-Wang (WLW) models. We analyze the essential features, rationality and applicability of models. Specifically, we list the advantages and disadvantages of these models to facilitate choosing an appropriate model in a special situation. With the advancement of single-cell technology, building quantitative models of gene transcription will play an increasingly important role.

Key words:gene transcription, transcriptional bursting, two-state model, multi-state model, scope of application

遗传信息自DNA流向RNA的转录过程是细胞最重要的生命活动之一,转录过程高度受控,涉及转录起始与延伸[1-2]、转录因子和聚合酶的募集[3-4]、染色质重构[4-5]、组蛋白修饰[6-7]等.原

基金项目:国家自然科学基金(11874209)
收稿日期:2020-05-14
*通讯联系人,E-mail:15995920872@163.com

核和真核基因的转录调控机制有很大的不同，真核基因转录涉及由RNA聚合酶、通用转录因子、媒介子、转录激活子等组成的转录机器的运转．尽管转录机器的基本架构已大致知晓，但其运转机制依然很不清楚．转录起始的关键步骤是如何完成和被调控的？转录机器是如何感知时变的信号，以合适的速率起始信使RNA（mRNA）的合成？分子的无规运动与基因表达的精确性是如何协调的？定量刻画mRNA数目随时间的变化是研究上述问题的基础．

传统观点认为基因转录是泊松过程[8]，这与许多传统实验的结果相符（mRNA或者蛋白质数量在一个稳定值附近变动）．但随着新技术的发展，人们发现从细菌到高等哺乳动物普遍存在着转录爆发（Transcriptional bursting）现象[9−12]：转录激活信号促进基因从沉默态转换到激活态、快速起始转录，在短时间内生成大量mRNA，然后再回到沉默态，上述过程重复进行，直到激活信号消失．转录爆发在细胞的信号转导过程中起到了重要作用[13]．

人们提出不少理论模型来解释转录爆发现象．比如，两态模型[14]认为，在激活态的快速转录、激活态与沉默态之间的转换都是马尔科夫过程（不受先前状态的影响，只由当前状态决定）．因其简单明了，两态模型得到了广泛应用．但单细胞技术的发展催生了大量新数据，挑战了原有的理论模型，比如两态模型无法解释不应期时长的单峰分布和mRNA丰度的多峰分布等特征．因此，针对具体的调控机制，Zhang et al[15−17]首次提出了多态模型．

不同的分子机制决定了不同的转录动力学，如转录爆发的频率调控和幅度调控等，导致转录产物丰度和持续时间的分布更加多样化．这更需要准确的模型给出定量化的解释，给出新的可供实验检验的理论预言．本文综述近年来的转录爆发模型，分析各模型的优缺点．

1　转录爆发与不应期

先简介基因的转录爆发现象．

转录过程曾被认为是一个平稳的过程，即单位时间内产生的mRNA数量是在一个平均值附近的小幅扰动，因此转录速率是常数，mRNA的产生是个泊松过程．传统的实验是针对细胞群体的测量，只保留了转录中mRNA的平均信息，丢失了单细胞的动态信息．随着单细胞技术的发展[18−19]，尤其是单分子荧光原位杂交（Single-molecule Fluorescence in situ Hybridization，SFSH）技术的出现，新生成的mRNA可以被追踪．人们发现，除了管家型基因[20]，很多基因的转录其实是间歇性的爆发过程，而非传统认为的泊松过程[11]．基因开始转录时会从沉默态（无转录活性）进入一个相对短暂的激活态，快速并大量地生成mRNA，再重新回到沉默态；该过程重复出现，直到转录信号消失或转录过程被阻遏．经过剪切加工，初生的mRNA变为成熟的mRNA、出核，而细胞核内的pre-mRNA也会被降解，这些都导致核内mRNA的减少．从低等的原核生物到高等的哺乳动物[9−12]，这一现象广泛存在，说明转录爆发是基因表达的一种基本模式．图1表达了单细胞中的转录爆发过程（基因Prl2c2的实验数据和模拟结果），整个过程只有一个稳定的转录信号，没有考虑对mRNA降解速率的调控．图中蓝线和绿线分别是实验得到的Prl2c2转录活性和mRNA数量随时间的演化曲线[12]；红线和黑线是基于两态模型和Gillespie算法[21]模拟得到的（$k_{\mathrm{ON}}=0.026\,\mathrm{min}^{-1}$，$k_{\mathrm{OFF}}=0.06\,\mathrm{min}^{-1}$，$k_{\mathrm{m}}=4\,\mathrm{min}^{-1}$，$\delta=0.0125\,\mathrm{min}^{-1}$），分别表示基因的活性和mRNA数量的变化．和简单的泊松过程相比，转录爆发包含了更多的动态信息（爆发大小、爆发频率和各个状态的持续时间等）．

通过追踪单细胞中mRNA数量的变化可以得到激活态和沉默态（不应期）的持续时间分布[12]，激活过程和沉默过程都涉及多个状态间的转变．转录爆发的激活过程往往包含多个相同或相似的状态转换过程[22]（生成单个mRNA的过程），而沉默过程多数情况下是多步骤的，且持续时间更长．激活过程和沉默过程的交替出现，说明必然存在一系列分隔两种过程的子状态（σ_i，$i=1,2,\cdots,n$），它们既可能进入激活过程，也可能进入沉默过程．假设基因从σ_i态经由激活态到达σ_j态且中间不经过其他σ状态所需时间的概率

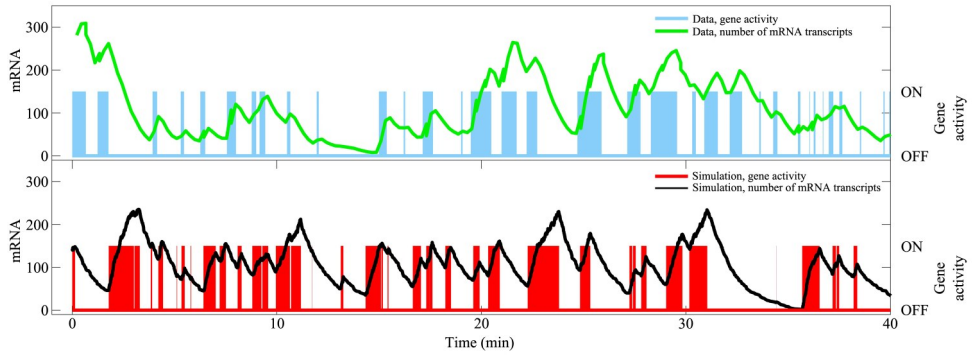


图 1　单细胞中的转录爆发

Fig. 1　Transcriptional bursting

密度函数为 $f_{ij}(t)$($\int_0^\infty f_{ij}(t)\mathrm{d}t=1, f_{ij}(t)\geqslant 0$)，经由该路径的概率为 p_{ij}($\sum_j p_{ij}=1-p_{i1}$，p_{i1} 为从 σ_i 态进入沉默态的概率)，而从沉默态进入 σ_i 的概率为 p_i($\sum_i p_i=1$)．所以，激活态的时间分布函数(f_A)为：

$$f_A(t)=\sum_{i=1}^{n}\frac{1}{1-p_{i1}}\sum_{j=1}^{n}p_{ij}f_{ij}(t)*\left[p_{j1}\delta(t)+\sum_{k=1}^{n}p_{jk}f_{jk}(t)*\cdots\right] \quad (1)$$

其中，* 代表卷积：$f(t)*g(t)=\int_0^\infty f(t)g(t-\tau)\mathrm{d}\tau$，“…”代表对卷积括号中的式子不断迭代取卷积的过程(需要改变对应的下标)．

考虑最简单的情况：只有一个 σ 状态($n=1$)或者一群难以区分的 σ 状态占主导，脚手架结构不断招募 Pol II，导致 $f_{ij}(t)\approx f_{\mathrm{mRNA}}(t)$，$p_{i1}=p$，$p_{ij}=\frac{1-p}{n}$，所以：

$$f_A(t)\approx p(1-p)^{i-1}\delta(t)\left[*f_{\mathrm{mRNA}}(t)\right]^i \quad (2)$$

同理可以得到 f_I：

$$f_I(t)\approx\sum_{i=1}^{\infty}(1-p)p^{i-1}\delta(t)\left[*f_S(t)\right]^i \quad (3)$$

$f_{\mathrm{mRNA}}(t)$ 为生成单个 mRNA 所需时间的分布函数，呈单峰分布(峰值不在 0 处)，代表一个多步骤过程．$f_S(t)$ 代表基因从沉默态开始激活的主要步骤，一般呈单峰分布，只有当某个限速步骤的持续时间占主导时，它才趋于指数分布．p 代表基因从激活态进入沉默态的概率，决定了爆发的大小($b=\frac{1}{p}$)．多数情况下 p 一般较小，导致 $f_A(t)$ 的 CV 接近 1，激活态时长的分布由峰值不在 0 的单峰分布逐渐接近指数分布(图 2a)，而不应期的时长则呈衰减分布(图 2b)或峰值不在 0 的单峰分布(图 2c)[12]．图 2a 和图 2b 是基于两态模型得到的($P(t_A)=k_{\mathrm{OFF}}e^{-k_{\mathrm{OFF}}t_A}$，$k_{\mathrm{OFF}}=1\,\mathrm{min}^{-1}$，$P(t_I)=k_{\mathrm{ON}}e^{-k_{\mathrm{ON}}t_I}$，$k_{\mathrm{ON}}=0.1\,\mathrm{min}^{-1}$)，而图 2c 是基于含不应期的多态模型得到的($P(t_I)=\frac{k_{1\to2}k_{\mathrm{OFF}}}{k_{1\to2}-k_{\mathrm{OFF}}}\left(e^{-k_{\mathrm{OFF}}t_I}-e^{-k_{1\to2}t_I}\right)$，$k_{\mathrm{OFF}}=1\,\mathrm{min}^{-1}$，$k_{1\to2}=0.2\,\mathrm{min}^{-1}$，$k_{\mathrm{ON}}=0.2\,\mathrm{min}^{-1}$)．

图 2d 至图 2g 显示了 mRNA 丰度的分布图(基于含不应期的多态模型)，分布函数为：

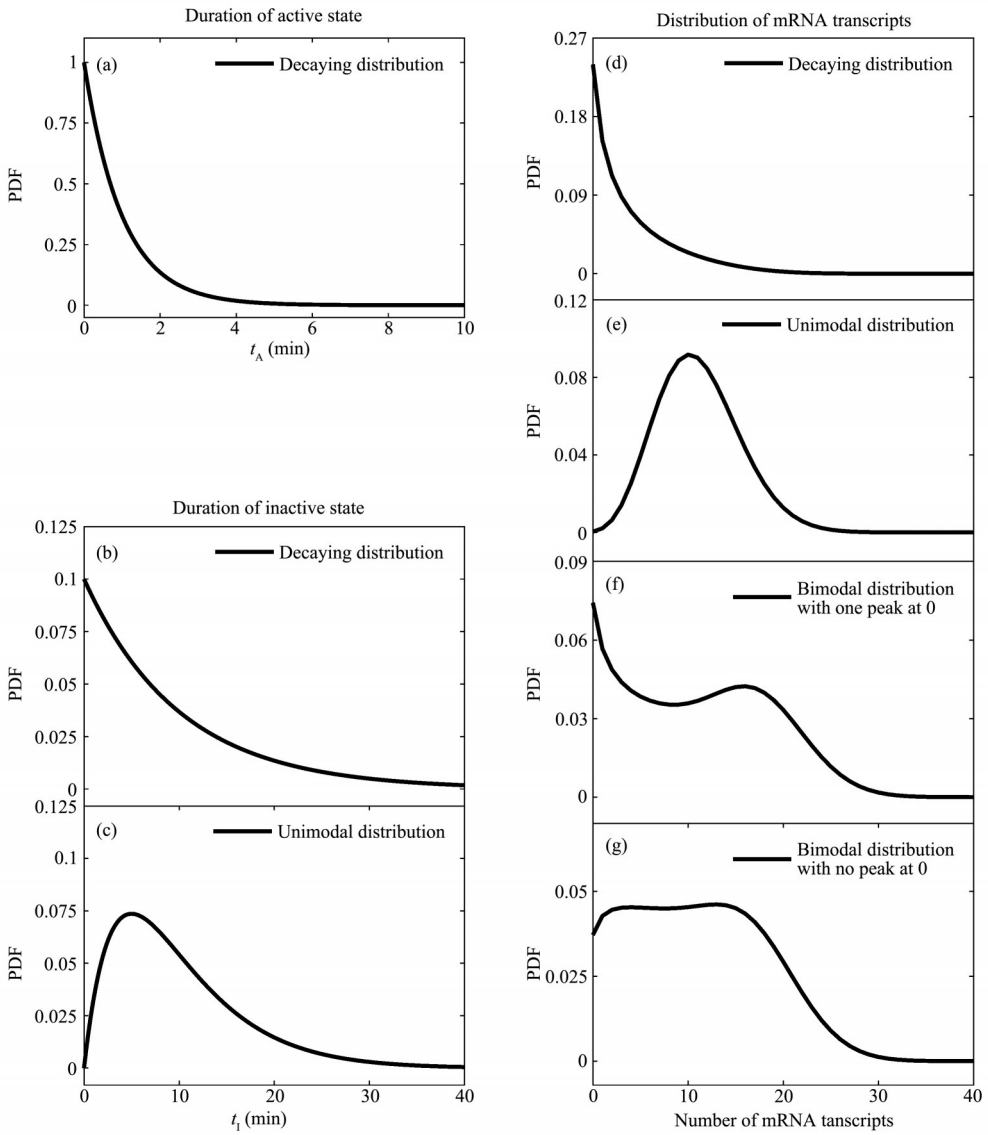
$$P(m)=\frac{\prod_{i=0}^{m-1}(a_1+i)(a_1+i)}{\prod_{i=0}^{m-1}(b_1+i)(b_2+i)}\frac{\mu^m}{m!}{}_2F_2(a_1+m,a_2+m;b_1+m,b_2+m;-\mu)$$

$$a_1=\frac{k_{1\to2}}{\delta},a_2=\frac{k_{\mathrm{ON}}}{\delta}$$

$$b_1=\frac{k_{\mathrm{OFF}}+k_{1\to2}+k_{\mathrm{ON}}+\sqrt{(k_{1\to2}+k_{\mathrm{ON}}-k_{\mathrm{OFF}})^2-4k_{1\to2}k_{\mathrm{ON}}}}{2\delta}$$

$$b_2=\frac{k_{\mathrm{OFF}}+k_{1\to 2}+k_{\mathrm{ON}}-\sqrt{\left(k_{1\to 2}+k_{\mathrm{ON}}-k_{\mathrm{OFF}}\right)^2-4k_{1\to 2}k_{\mathrm{ON}}}}{2\delta}$$

$\mu=\frac{k_{\mathrm{m}}}{\delta}$，${}_pF_q(a_1,\cdots,a_p;b_1,\cdots,b_q;z)$是广义超几何函数．当不应期的时长呈单峰分布时，mRNA丰度可能呈现多峰分布[20,22]（图2f至图2g），表明mRNA的生成涉及多个反应步骤，单分子事件是非独立的．图2d中mRNA的数量呈指数（衰减）分布[23−25]（$k_{\mathrm{OFF}}=0.2\,\mathrm{min}^{-1}$，$k_{1\to 2}=0.1\,\mathrm{min}^{-1}$，$k_{\mathrm{ON}}=0.1\mathrm{min}^{-1}$，$k_{\mathrm{m}}=2\,\mathrm{min}^{-1}$，$\delta=0.1\,\mathrm{min}^{-1}$），代表在mRNA的时序图中转录爆发的谷底为0或者接近0，意味着$\ln k_{\mathrm{m}}\tau_{\mathrm{A}}\lesssim\delta\tau_{\mathrm{I}}$（$k_{\mathrm{m}}$为转录速率常数，$\tau_{\mathrm{A}}$和$\tau_{\mathrm{I}}$分别为激活态和沉默态的平均持续时间，$\delta$为mRNA的降解速率常数）．图2e中mRNA的数量呈单峰分布，峰值不在0[23−26]（$k_{\mathrm{OFF}}=0.4\,\mathrm{min}^{-1}$，$k_{1\to 2}=1\,\mathrm{min}^{-1}$，$k_{\mathrm{ON}}=1\,\mathrm{min}^{-1}$，$k_{\mathrm{m}}=$



Schematically shown are the distributions of duration of active and silent gene states (Fig.2a～2c) and of the number of mRNA transcripts (Fig.2d～2g) under different conditions. PDF refers to probability density function.

图2　转录爆发的典型特征

Fig. 2　Typical features of transcriptional bursting

$2\,\mathrm{min}^{-1}$, $\delta=0.1\,\mathrm{min}^{-1}$),一般有 $\ln k_m\tau_A \geqslant \delta\tau_I$ 且 $\delta(\tau_A+\tau_I)<1$. 图 2f和图 2g则呈现双峰分布,只不过图 2f 中有一个峰在0处[27]($k_{OFF}=0.04\,\mathrm{min}^{-1}$, $k_{1\to2}=0.1\,\mathrm{min}^{-1}$, $k_{ON}=0.1\,\mathrm{min}^{-1}$, $k_m=2\,\mathrm{min}^{-1}$, $\delta=0.1\,\mathrm{min}^{-1}$),而图 2g 中两个峰都在非 0 处[25,27]($k_{OFF}=0.06\,\mathrm{min}^{-1}$, $k_{1\to2}=0.15\,\mathrm{min}^{-1}$, $k_{ON}=0.15\,\mathrm{min}^{-1}$, $k_m=2\,\mathrm{min}^{-1}$, $\delta=0.1\,\mathrm{min}^{-1}$);它们都要求 $\ln k_m\tau_A \geqslant \delta\tau_I$ 且 $\delta(\tau_A+\tau_I)>1$,但图 2f 中的 $\delta(\tau_A+\tau_I)$更大. 相比于指数分布和单峰分布,双峰分布意味着一个输入对应两个主要的输出,相对噪声较大,不确定性[27-28]更大. 而单峰分布相比于指数分布,相对噪声更小,有更多的信息(来自转录信号)传递给 mRNA 生成. 所以,基因最多遵循的是图 2e[23-26],其次是图 2d[23-25],最少的是图 2f和图 2g[25,27].

2 转录模型

转录过程的分子机制异常复杂,因此人们构建不同复杂程度的理论模型,定量刻画转录过程的某些特征,方便研究细胞的信号转导等过程. 模型构建后,可写出化学主方程,再基于 Gillespie 算法[21]开展数值计算. 下面逐一介绍当前文献中的主要模型(如图 3 所示).

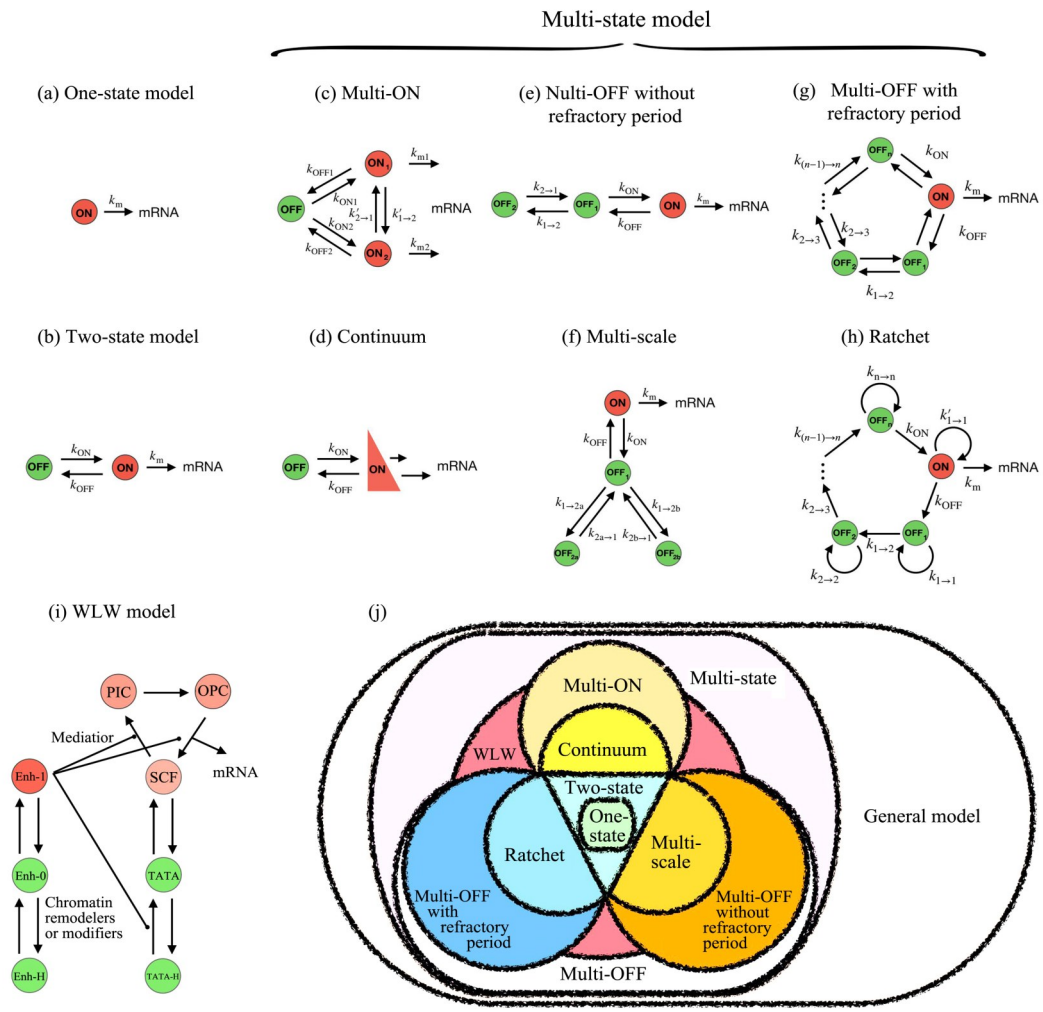
2.1 单态模型 在单态模型(图 3a)中,转录是泊松过程[8]. 转录速率是常数,其数值依赖于调控信号的强度,常用希尔函数来刻画[29]. 这一简化模型只保留了 mRNA 在细胞群体水平的平均信息,包含两个参数:转录速率常数 k_m 和降解速率常数 δ. 当研究 mRNA 或者蛋白质水平且它们的相对噪声较小(CV 较小)时,可以采用单态模型. 比如,当蛋白质的降解速率较慢,蛋白质丰度的 CV 较小,就常采用这一模型,甚至将转录和翻译过程合并在一起考虑. 该模型简单、参数少,适合用于研究复杂的细胞信号转导网络;缺点是太过简单,失去了基因转录的绝大部分信息.

2.2 两态模型 两态模型(又称 ON-OFF 模型,图 3b)向下兼容单态模型,向上又被多态模型所兼容,是当前广为使用的唯象模型[14,30]. 两态指基因启动子的两个状态:激活态(ON)和沉默态(OFF). 模型包含四个参数:从沉默态到激活态的转换速率常数 k_{ON}、从激活态到沉默态的转换速率常数 k_{OFF}、转录速率常数 k_m(激活态 $k_m>0$,沉默态 $k_m=0$)、mRNA 的降解速率常数 δ,直接决定转录爆发关键的四个特征:爆发频率(k_{ON})、大小($\frac{k_m}{k_{OFF}}$)和持续时间(激活态和沉默态的平均持续时间分别为 $\frac{1}{k_{OFF}}$ 和 $\frac{1}{k_{ON}}$,都呈指数分布)[31]、mRNA 的水平 $\left(\langle m\rangle\frac{k_m k_{ON}}{\delta(k_{ON}+k_{OFF})}\right)$ 和噪声 $\left(\sigma^2=\langle m\rangle\left[1-\langle m\rangle+\frac{k_m(k_{ON}+\delta)}{(k_{ON}+\delta)(k_{OFF}+\delta)-k_{ON}k_{OFF}}\right]\right)$. 调控信号对基因转录的影响,可分为调幅、调频和混合三种模式;在两态模型中,分别通过调节 k_m, k_{ON} 和 k_{OFF} 来实现这三种调节模式.

当 mRNA 数目满足泊松分布、调控信号影响主要限速步骤(持续时间呈指数分布),两态模型可以很好地描述 mNRA 数量和基因活性的随时演化等,也适用于描述单态模型适用的情形,但计算量更大. 果蝇间隙(gap)基因中由统一的转录动力学机制导致的不同空间图样就可以用两态模型很好地描述[26]. 当 mRNA 数目呈多峰分布、转录受多种信号调控或者调控信号作用于非限速步骤(持续时间偏离指数分布)时,两态模型就不再适用. 两态模型是唯象模型,其参数包含了整体的均值信息,往往刻画的是关键的限速步骤,但忽略了许多细节.

2.3 多态模型 多态(multi-state)模型是在两态模型的基础上发展起来的. 它包含多个启动子活性状态,可描述持续时间偏离指数分布、呈单峰或双峰分布的情形,以及 mRNA 数目呈双峰分布的情形. 简单来说,就是激活态或沉默态是有记忆的. mRNA 的产生涉及多个反应步骤,单分子事件之间存在关联(记忆). 这涉及阻遏物、转录因子和中介物的复杂调控过程以及染色质重构或组蛋白修饰等,其中的许多分子机制还远不清楚.

从两态模型出发,可分别将 ON 态和 OFF 态细分,得到多 ON 态[15,32]和多 OFF 态[16,22-23,32]两大类唯象模型(图 3c 至图 3h). 各个状态受限速步骤的影响,但并不一一对应(可能是多个步骤的合并). 激活态的时长呈指数或近指数分布时,如果



Shown are the one-state (a), two-state (b), multi-ON (c), continuum (d), multi-OFF without necessary refractory period (e), multi-scale (f), multi-OFF with necessary refractory period (g), ratchet (h) and WLW (i) models. (j) shows model compatibility.

图3　转录模型

Fig. 3　Modeling of transcription

不考虑激活态中子状态的差异(或差异不大)，多ON态可以合并为单ON态，所以一般在考虑RNA聚合酶状态时才使用多ON态模型. 对许多基因，如与催乳素[33]基因相关的启动子表现出很强的记忆性，不应期时长呈单峰分布，多OFF态模型更常见. 模型的参数包括：从沉默态到激活态的转换速率常数k_{ON}、从激活态到沉默态的转换速率常数k_{OFF}、激活态之间的转换速率常数$k'_{i\to j}$、沉默态之间的转换速率常数$k_{i\to j}$、各个激活态的转录速率常数$k_{\mathrm{m}i}$以及mRNA的降解速率常数δ(i,j代表的是转换前后的状态).

多ON(multi-ON)[15,17]模型(图3c)考虑的是存在多个ON态的情形. 影响ON态数量的因素很多，如转录因子的空间分布、TAD(Topologically Assocaited Domain)和RNA聚合酶的相分离等. 当ON态是由大量受转录因子或表观遗传标记的特异性结合所定义的子状态组成，每个子状态具有不同的启动速率，并且在时间上紧密相邻，或者ON态中启动子受RNA聚合酶局部时变浓度影响、导致大量准连续的启动速率时，可以用连续性(continuum)模型[34](图3d)来刻画转录过程. 连续性模型主要考虑启始转录速率的波动，涉及多种因素，如RNA聚合酶的相分离、转录因子的空间分布等. 两个连续启动事件间隔几秒到

十多秒不等;这些间隔不服从单指数分布,而是服从大量的指数分布,其期望值是准连续分布的.因此,相应的起爆速度几乎跨越连续的区间.从本质上来讲,不管是考虑ON态中的子状态还是考虑RNA聚合酶的时空分布,连续性模型都应该归为多ON态模型,但由于ON态的持续时间变化不大,连续性模型又可以看成是参数可变的两态模型.连续性模型适合于研究由时空特异性导致、拥有不同分子结合速率的转录过程.

根据是否存在基因启动子重新激活前必须经历的不应期,多OFF态模型又分成无不应期[32](图3e)和含不应期[16,22-23](图3g)的两类模型.无不应期模型有更多可能的启动路径,很容易延伸发展为多尺度(multi-scale)模型[35](图3f).多尺度模型考虑的是基因在再次激活前,存在多条路径且耗费的时间长短不同.举例来说,考虑结合在启动子上的脚手架结构(Scaffold Complex,SCF)的不完全拆解和完全拆解,就可以导致不应期存在多个时间尺度.含不应期的multi-OFF模型存在很强的不可逆性;如果沉默态足够多,模型还会呈现出很强的周期性.考虑到转录过程存在可替代的路径(分岔)以及微弱的可逆反应,模型可进一步发展为棘轮模型[36-38](图3e).在OFF态除了沿着主要方向进行的反应,其余分岔和可逆反应都简化为反应常数为$k_{i\to i}$的简单反应.而在ON态中,反应常数为$k'_{i\to j}$的反应对应着依靠逆反应经由OFF态直接回到ON态的过程.在mRNA均值和持续时间均值都相同的情况下,多OFF态模型的噪声往往比两态模型要小.

相比两态模型,多态模型对mRNA分布的刻画更准确,更适用于研究信息的传递,能够更好地解释许多实验现象.Wang et al[39]提出的模型(简写为WLW模型)(图3i)是从基础的转录机制来解释转录爆发,揭示了转录的生物化学机制对转录动力学的影响.它将转录中启动子的变化分成三部分:mRNA生成、脚手架结构的装配与拆解、组蛋白的装配和修饰,并通过激活子的结合与解离来控制这三个部分.WLW模型考虑的是含TATA框的基因转录过程,其中最重要的就是增强子(enhancer)和TATA框上的状态变化.增强子可以是被组蛋白占据的状态(Enh-H)、裸露的状态(Enh)或者被转录因子占据的状态(Enh-1).结合在增强子上的激活子通过使脚手架结构中的媒介子(Mediator)异构化[40-41]来调控起始转录速率,将信息传递给Pol II.这意味着ON态的持续时间和转录速率是耦合在一起的,难以像其他的唯象模型将两者分离.当激活子结合在增强子上时,通过招募染色质重构酶[4-5]和修饰酶[6-7],使得占据TATA框的组蛋白很容易被清除,为脚手架结构的构建做好准备.转录因子是否占据增强子以及转录因子的种类则会影响TATA框上状态转变的速率,而转录因子的浓度越高,增强子被其占据的频率越高,导致转录速率、激活态和沉默态时长是转录因子浓度的函数.与增强子相对的是沉默子,能够产生相反的效果.如果没有TATA框,WLW模型的结构也可以用,只不过反应过程有所不同:没有TATA框,不再是TFIID中的TATA框结合蛋白(TATA-box binding protein,TBP)与TATA框结合,而是TFIID中其他亚基结合到启动子上.其实,可将WLW模型中TATA框的概念延申为代指或标识核心启动子区域.

总结一下,单态模型是最简单的模型,主要刻画mRNA的均值.它能发展为两态模型,为两态模型兼容(图3j).在合适参数下,两态模型可以退化为单态模型.两态模型能刻画转录爆发大小和频率的均值,以及部分呈指数分布的基因激活态和沉默态持续时间.研究mRNA的分布及其对信号的响应时,常常要运用多态模型.多态模型分为多ON态和多OFF态模型,它们兼容单态和两态模型.当研究具体的转录机制时,多ON态模型可发展为连续性模型.而多OFF态模型根据是否存在必须经历的不应期,分成两类;考虑到具体的分子机制,两者又可以分别发展为棘轮模型和多尺度模型.连续性模型侧重于刻画转录因子或聚合酶的局部时空变化对转录的影响,可用于研究相分离对转录的影响.棘轮模型主要研究不可逆反应和路径分岔的影响,适合于研究转录中的能耗.多尺度模型侧重于脚手架结构对转录的影响,适合研究温度、启动子序列、转录因子等对转录的影响.WLW模型考虑作用于增强子区域的转录因子,突出的是转录因子自身生物功

能对转录过程（对增强子、TATA框和Pol II的状态）的影响．只要把不断招募Pol II的过程合并为ON态，再把增强子、TATA框的状态向量与OFF态对应，WLW模型就可以转化为考虑基因状态的模型，归并到多态模型里，但是激活态时长与转录速率耦合在一起．事实上，WLW模型只要多考虑一些转录因子、染色质的重构与修饰反应，就能再现多态模型中的复杂动力学，兼容连续性模型、多尺度模型和棘轮模型．当然，受限于内在的转录机制，WLW模型不能模拟一些细菌基因的动力学．

此外，有外部信号（如转录因子，Transcription factor，TF）调控转录时，实际上有两组参数，分别对应有TF结合和没有TF结合的情形．假设TF的结合可以提高转录水平，那么在没有TF结合时，转录事件发生的频率很低，可以视为泊松过程．在保证模拟结果大致不变的情况下，对单态模型、两态模型和多态模型而言，没有TF结合的一组参数常常可以简化为一个参数（k_{m0}），即基因在没有被TF结合时的基本转录速率．多数情况下，k_{m0}很小而被忽略，或者直接等于0．

上述的单态、两态、多态模型，只要适当变化一下，与Zhang and Zhou[15]提出的一般模型是相容的（图3i），且参数个数和计算量不断增加．针对mRNA均值、单细胞mRNA的动力学和分布，用单态模型、两态模型以及多ON和多OFF模型来预测和解释现象就够了．当研究的问题与转录机制、能耗、具体的信号调控、信号转导过程有关时，就要运用连续性模型、棘轮模型、多尺度模型和WLW模型了．很多情况下，这些模型模拟的结果与多ON或多OFF模型的结果相差不大（在实验误差内），但是添加转录的分子机制势必会增加复杂性和计算量，所以往往只有在研究具体分子机制或者信息传输时，才采用这些模型．在保证模型能够表达研究所需转录信息的情况下，则一般选择更加简单的模型．当出现新的现象、原有模型无法解释时，先尝试能够兼容原模型的已有模型，失败后再尝试其他的现有模型．如果都失败，再基于两态模型，根据激活态和沉默态时间的分布、转录爆发大小以及mRNA数量分布随调控因子的变化，发展新模型．

3　数据分析与数值模拟

确定模型后便可根据实验数据确定模型参数、开展数值模拟、解释实验现象或预测新结果．

3.1　数据分析　实验给出的数据多种多样，可以是蛋白质或mRNA信息、核内或核外信息、群体或单细胞信息．这里主要介绍如何对单细胞中ON态时长、OFF态时长、mRNA数量等进行分析．如果这三组数据只是在稳态下测得的，那就已经丢失了部分转录信息，无法还原所有转录细节．用这三组数据可分别得到激活态和沉默态持续时间、mRNA数量的平均值，即$\langle t_A\rangle$，$\langle t_I\rangle$和$\langle m\rangle$．进而得到转录速率常数与降解速率常数之比$\frac{k_m}{\delta}$，再关联mRNA降解的实验数据就可以得到δ和k_m，构建两态模型．分析激活态（沉默态）持续时间的分布，如果是指数（衰减）分布，激活态（沉默态）就是由一个限速步骤或者一串相同或相似、可多次快速进行的反应所主导，一般直接采用单ON（OFF）态模型．当对应的mRNA分布出现多峰分布就要考虑连续性模型．如果激活态（沉默态）持续时间呈现单峰分布或者mRNA分布出现多峰，就要采用多ON（多OFF）模型．计算持续时间的变异系数CV，$\frac{1}{CV^2}$大致提示有多少等效的限速步骤，一般取大于$\frac{1}{CV^2}$的状态数来模拟．对于多尺度模型，$\frac{1}{CV^2}$一般不超过2．如果$\frac{1}{CV^2}$很大，一般会采用不可逆的多OFF态模型[23]（棘轮模型）．这些都是基于稳态数据得到的，如果还有暂态的数据，可以先用稳态的数据构建模型，再通过模拟与暂态比较．如果暂态中激活态持续时间、沉默态持续时间和mRNA数量的分布不在模型（由稳态数据构建）允许的情形中，就要去寻找能够兼容的模型，一般都类似WLW模型．

不妨以研究转录因子浓度对转录的影响为例来说明．首先，根据转录爆发大小的变化，判断由激活态进入沉默态的概率p的变化，判断TF对两态转换速率的影响；然后，根据激活态和沉默态持续时间分布的变化判断TF是否对两态子状态的

转换有影响以及是否作用在限速步骤上;再根据持续时间CV的变化,判断两态中有几个限速步骤,TF作用在哪几个限速步骤上;最后,结合分子机制建立模型,并解释新现象.

3.2 数值模拟 对模型构建后的数值模拟,主要通过化学主方程来计算分布信息,用Gillespie算法[21]来模拟基因状态和mRNA的时变轨迹.所有模型适当变化后都可以转变为态模型.假设模型中共有N个状态,$P=(P_1,P_2,\cdots,P_N)^T$,$P_i(m;t)(i=1,2,\cdots,N)$是基因处于第i状态、存在m个mRNA的概率.K是$N\times N$矩阵,其元素k_{ij}是状态i到状态j的转变速率.$\Lambda=\mathrm{diag}(k_{m1},k_{m2},\cdots,k_{mN})$,$k_{mi}(i=1,2,\cdots,N)$是状态$i$的转录速率(如果是OFF态,则$k_{mi}=0$).$\delta$是mRNA的降解速率常数,$I$是单位算符,$E$是移位算符(算符作用后,变量增加一位).模型的化学主方程为:

$$\frac{\mathrm{d}P(m;t)}{\mathrm{d}t}=KP(m,t)+\delta(E-I)\left[mP(m,t)\right]+\Lambda(E^{-1}-I)\left[P(m;t)\right] \tag{4}$$

通过上式,便能模拟概率分布随时间的变化.如果要得到单细胞中mRNA的时序图,就要使用Gillespie算法.

4 总 结

本文介绍了刻画转录爆发的几个主要模型:单态、两态[14,30]和多态模型[15−16,22−23,34−35,39],涉及模型的结构、应用和优缺点比较(如表1所示).

表1 转录模型的适用性

Table 1 Applicability of the models of transcription

模型	单态模型(图3a)	两态模型(图3b)	多态模型(图3c至图3i)		
			多ON模型(含连续性模型)	多OFF模型(含棘轮模型和多尺度模型)	WLW模型
激活态的持续时间分布	/	图2a	主要时图2a,少部分情况下为峰值不为0的单峰分布,但是也接近图2a	图2a	主要时图2a,少部分情况下为峰值不为0的单峰分布,但是也接近图2a
沉默态的持续时间分布	/	图2c	图2c	图2b,图2c	图2b,图2c
mRNA数量分布	图2e	图2d至图2f	图2d至图2g	图2d至图2g	图2d至图2g
保留信息	细胞群体水平的平均mRNA信息				
		mRNA的噪声强度、爆发频率和大小、激活态和沉默态的平均持续时间			
			mRNA分布性质		
			激活态持续时间分布	沉默态持续时间分布	激活态和沉默态的持续时间分布
精确度	低	中等	较高	较高	高
参数	很少,可由实验测得	少,可由实验测得	中等,部分可由实验测得,部分需要假设且要与实验相符	中等,部分可由实验测得,部分需要假设且要与实验相符	多,主要步骤可由实验测得,部分未明确的机制需要假设
计算量	很小	小	中等	中等	大
适用情况	mRNA或蛋白质的CV较小,不考虑单细胞中mRNA的时变	研究单细胞转录产物数量变化,基因状态切换以及转录调控等	mRNA数量呈多峰分布、激活态时长呈单峰分布,或者研究转录调控和信号传导时采用多ON态模型。在研究相分离对转录的影响时可用连续性模型,侧重于转录因子或者聚合酶的局部时空变化对转录的影响	研究周期性和能耗时,可采用棘轮模型。研究温度、启动子序列、转录因子等对转录的影响时用多尺度,侧重于脚手架结构对转录的影响	研究转录因子自身生物功能对转录过程的影响

在单态模型中，转录是一个简单的泊松过程[8]；当mRNA或者蛋白质的CV较小时，可用于模拟细胞信号转导网络的动力学．两态模型比单态模型多了一个沉默态：沉默态没有转录，而在激活态，mRNA快速生成．它适用于研究单细胞的转录爆发动力学，能够有效描述爆发频率、爆发大小和平均持续时间等[28]，但无法解释持续时间的单峰分布[12]和mRNA数目的双峰分布[27－28]．多态模型是将激活态或沉默态由单态变成多态，使激活或沉默过程拥有记忆，从而预测和解释持续时间的单峰分布和mRNA的双峰分布等现象，并能更好地预测噪声在信号转导中的作用．限于当前的技术水平，多态模型的参数不易确定．

构建转录模型，主要是为了能够定量解释实验现象，提供可供实验检验的理论预言．当研究对象是细胞系或者受精卵时，考虑细胞分裂[42]、基因复制[43]和体积变化[44]时，可根据研究的需要选择不同的模型．构建转录模型受到研究对象所需模型层次的影响．转录调控模式、包含转录的信号传导等问题是当前的研究热点，往往需要精确的转录模型，更多地与转录机制相结合．随着更多蛋白质标签和有机染料的开发利用[45－47]，会有更多的单细胞数据出现，更多的蛋白质工作机理将被揭示，促进对转录机制的理解．构建转录模型，解释和预测转录爆发动力学仍将是一个重要的研究方向．构建更详细的转录模型十分依赖实验数据，依赖于科学技术．利用分子模拟、力化学技术、单分子荧光原位杂交技术（smFISH）、荧光成像（Fluorescence Imaging）以及二项分（Binomial Partitioning）[48]，能够得到单分子作用形式和单细胞中特定基因的转录轨迹．困难在于转录过程涉及太多的反应、分子作用机制多样、各种信号干扰，限制逆推模型结构．因此，设计能够精准调控单步反应的实验和发展有效处理含噪声数据的数学方法是必须的．此外，研究转录对于外界信号（随时空变化）的响应，也是一个重要的方向．它将阐明外界信息是如何通过存在不确定性的转录过程传递下去的，证明转录精巧性的一面．

参考文献

[1] Jonkers I, Kwak H, Lis J T. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. eLife, 2014, 3: e02407.

[2] Stasevich T J, Hayashi-Takanaka Y, Sato Y, et al. Regulation of RNA polymerase II activation by histone acetylation in single living cells. Nature, 2014, 516(7530): 272－275.

[3] Senecal A, Munsky B, Proux F, et al. Transcription factors modulate c-Fos transcriptional bursts. Cell Reports, 2014, 8(1): 75－83.

[4] Voss T C, Hager G L. Dynamic regulation of transcriptional states by chromatin and transcription factors. Nature Reviews Genetics, 2014, 15(2): 69－81.

[5] Brown C R, Mao C, Falkovskaia E, et al. Linking stochastic fluctuations in chromatin structure and gene expression. PLoS Biology, 2013, 11(8): e1001621.

[6] Nicolas D, Zoller B, Suter D M, et al. Modulation of transcriptional burst frequency by histone acetylation. Proceedings of the National Academy of Sciences of the United States of America, 2018, 115(27): 7153－7158.

[7] Muramoto T, Müller I, Thomas G, et al. Methylation of H3K4 is required for inheritance of active transcriptional states. Current Biology, 2010, 20(5): 397－406.

[8] Sanchez A, Choubey S, Kondev J. Stochastic models of transcription: from single molecules to single cells. Methods, 2013, 62(1): 13－25.

[9] Chubb J R, Trcek T, Shenoy S M, et al. Transcriptional pulsing of a developmental gene. Current Biology, 2006, 16(10): 1018－1025.

[10] Golding I, Paulsson J, Zawilski S M, et al. Real-time kinetics of gene activity in individual bacteria. Cell, 2005, 123(6): 1025－1036.

[11] Raj A, Peskin C S, Tranchina D, et al. Stochastic mRNA synthesis in mammalian cells. PLoS Biology, 2006, 4(10): e309.

[12] Suter D M, Molina N, Gatfield D, et al. Mammalian genes are transcribed with widely different bursting kinetics. Science, 2011, 332(6028): 472－474.

[13] Wang Y, Ni T, Wang W, et al. Gene transcription in bursting: a unified mode for realizing accuracy and stochasticity. Biological Reviews, 2019, 94(1): 248－258.

[14] Ko M S H. Induction mechanism of a single gene molecule: Stochastic or deterministic? BioEssays, 1992, 14(5): 341－346.

[15] Zhang J, Zhou T. Promoter-mediated tran-scriptional dynamics. Biophysical Journal, 2014, 106(2): 479－488.

[16] Zhang J, Chen L, Zhou T. Analytical distribution and tunability of noise in a model of promoter progress. Biophysical Journal, 2012, 102(6): 1247－1257.

[17] Zhou T, Zhang J. Analytical results for a multistate gene model. SIAM Journal on Applied Mathematics, 2012, 72(3): 789－818.

[18] Bertrand E, Chartrand P, Schaefer M, et al. Localization of ASH1 mRNA particles in living yeast. Molecular Cell, 1998, 2(4): 437－445.

[19] Femino A M, Fay F S, Fogarty K, et al. Visualization of single RNA transcripts in situ. Science, 1998, 280(5363): 585－590.

[20] Zenklusen D, Larson D R, Singer R H. Single-RNA counting reveals alternative modes of gene expression in yeast. Nature Structural & Molecular Biology, 2008, 15(12): 1263－1271.

[21] Gillespie D T. Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry, 1977, 81(25): 2340－2361.

[22] Pedraza J M, Paulsson J. Effects of molecular memory and bursting on fluctuations in gene expression. Science, 2008, 319(5861): 339－343.

[23] Zoller B, Nicolas D, Molina N, et al. Structure of silent transcription intervals and noise characteristics of mammalian genes. Molecular Systems Biology, 2015, 11(7): 823.

[24] Sepúlveda L A, Xu H, Zhang J, et al. Measurement of gene regulation in individual cells reveals rapid switching between promoter states. Science, 2016, 351(6278): 1218－1222.

[25] Fritzsch C, Baumgärtner S, Kuban M, et al. Estrogen－dependent control and cell-to-cell variability of transcriptional bursting. Molecular Systems Biology, 2018, 14(2): e7678.

[26] Zoller B, Little S C, Gregor T. Diverse spatial expression patterns emerge from unified kinetics of transcriptional bursting. Cell, 2018, 175(3): 835－847.e5.

[27] Ochab-Marcinek A, Tabaka M. Bimodal gene expression in noncooperative regulatory systems. Proceedings of the National Academy of Sciences of the United States of America, 2010, 107(51): 22096－22101.

[28] To T L, Maheshri N. Noise can induce bimodality in positive transcriptional feedback loops without bistability. Science, 2010, 327(5969): 1142－1145.

[29] Raj A, van Oudenaarden A. Nature, nurture, or chance: stochastic gene expression and its consequences. Cell, 2008, 135(2): 216－226.

[30] Peccoud J, Ycart B. Markovian modeling of gene-product synthesis. Theoretical Population Biology, 1995, 48(2): 222－234.

[31] Dar R D, Razooky B S, Singh A, et al. Transcriptional burst frequency and burst size are equally modulated across the human genome. Proceedings of the National Academy of Sciences of the United States of America, 2012, 109(43): 17454－17459.

[32] Rodriguez J, Ren G, Day C R, et al. Intrinsic dynamics of a human gene reveal the basis of expression heterogeneity. Cell, 2018, 176(1－2): 213－226.e18.

[33] Harper C V, Finkenstädt B, Woodcock D J, et al. Dynamic analysis of stochastic transcription cycles. PLoS Biology, 2011, 9(4): e1000607.

[34] Corrigan A M, Tunnacliffe E, Cannon D, et al. A continuum model of transcriptional bursting. eLife, 2016, 5: e13051.

[35] Tantale K, Mueller F, Kozulic-Pirher A, et al. A single-molecule view of transcription reveals convoys of RNA polymerases and multi-scale bursting. Nature Communications, 2016, 7: 12248.

[36] Krasnov A N, Mazina M Y, Nikolenko J V, et al. On the way of revealing coactivator complexes cross-talk during transcriptional activation. Cell & Bioscience, 2016, 6: 15.

[37] Lemaire V, Lee C F, Lei J, et al. Sequential recruitment and combinatorial assembling of

multiprotein complexes in transcriptional activation. Physical Review Letters，2006，96(19)：198102.

[38] Wang Y，Liu F，Li J，et al. Reconciling the concurrent fast and slow cycling of proteins on gene promoters. Journal of the Royal Society Interface，2014，11(96)：20140253.

[39] Wang Y，Liu F，Wang W. Dynamic mechanism for the transcription apparatus orchestrating reliable responses to activators. Scientific Reports，2012，2：422.

[40] Kornberg R D. Mediator and the mechanism of transcriptional activation. Trends in Biochemical Sciences，2005，30(5)：235－239.

[41] Malik S，Roeder R G. Dynamic regulation of pol II transcription by the mammalian Mediator complex. Trends in Biochemical Sciences，2005，30(5)：256－263.

[42] Huh D，Paulsson J. Random partitioning of molecules at cell division. Proceedings of the National Academy of Sciences of the United States of America，2011，108(36)：15004－15009.

[43] Peterson J R，Cole J A，Fei J，et al. Effects of DNA replication on mRNA noise. Proceedings of the National Academy of Sciences of the United States of America，2015，112(52)：15886－15891.

[44] 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. Molecular Cell，2015，58(2)：339－352.

[45] Chen J，Zhang Z，Li L，et al. Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. Cell，2014，156(6)：1274－1285.

[46] Paakinaho V，Presman D M，Ball D A，et al. Single-molecule analysis of steroid receptor and cofactor action in living cells. Nature Communications，2017，8：15896.

[47] Grimm J B，English B P，Chen J J，et al. A general method to improve fluorophores for live-cell and single－molecule microscopy. Nature Methods，2015，12(3)：244－250.

[48] Phillips R，Belliveau N M，Chure G，et al. Figure 1 theory meets figure 2 experiments in the study of gene expression. Annual Reviews of Biophysics，2019，48：121－163.

（责任编辑　杨可盛）