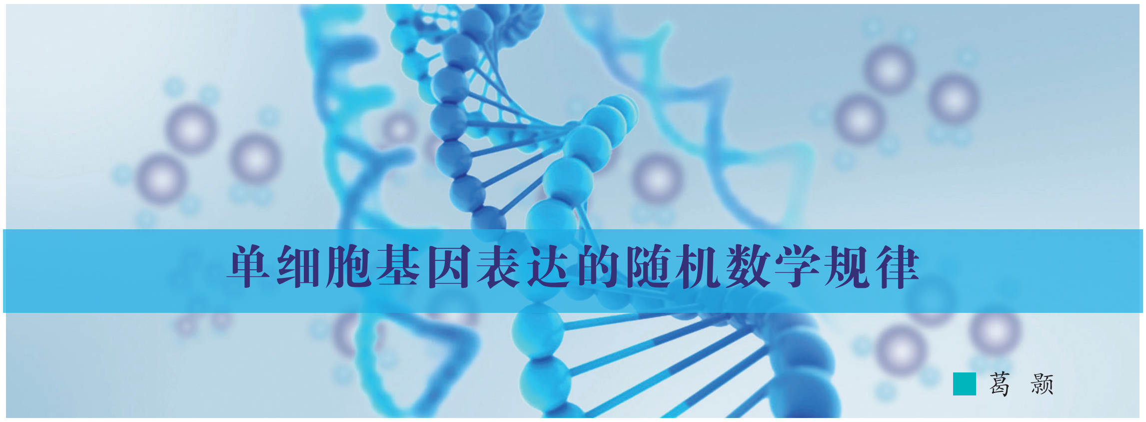

单细胞基因表达的随机数学规律

葛 颢

概率论和随机过程是数学体系中的一个另类，它的最早起源既不是对于数与形的探究，也不是对于物理问题的建模和分析。概率论最早起源于人们对于赌博游戏中随机规律的好奇，不过到了差不多十九世纪，人们也逐渐发现概率论和随机过程的知识可以被很好地用来刻画真实物理世界中的随机现象，其中首推布朗运动。

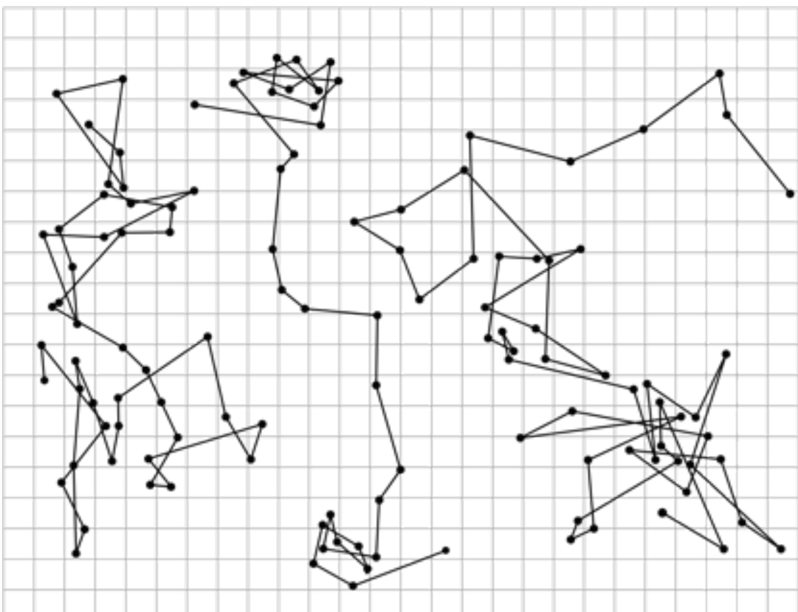

图 1. 二维布朗运动轨道 (from the book of Jean Baptiste Perrin, Les Atomes)

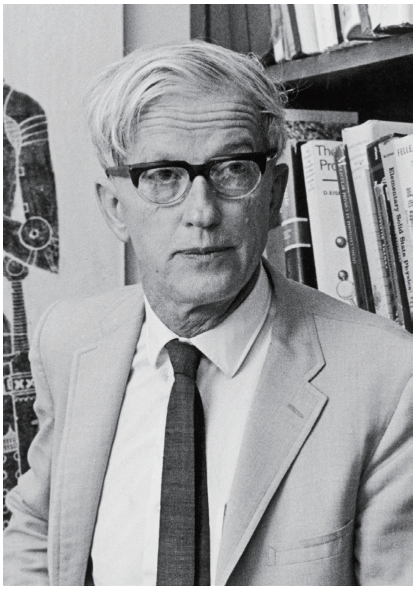

图 2. 德尔布吕克
（Max Delbrück, 1906-1981）

苏格兰植物学家罗伯特 • 布朗在 1827 年首先观察到了悬浮于水中的花粉迸出的微粒所做的无规则状运动，后人将之命名为布朗运动。后来直到 1905 年左右，才由爱因斯坦等推导出了第一个定量刻画布朗运动的数学规律，并被法国物理学家佩林首次在实验上证实，最终确认了分子和原子的存在。布朗运动，首先由生物学家发现，而后对物理学和数学都产生了深远的影响，这也许就是历史上最早的生物学和物理学以及数学的学科交叉。

从爱因斯坦的布朗运动理论开始，真正意义上的随机过程及其严格数学理论，在二十世纪上半叶被慢慢建立了起来。与此同时，随机过程理论也在物理学乃至化学领域里逐渐找到了用武之地。比如，物理学家朗之万所提出的牛顿定律在随机力下的修正，即朗之万方程，至今仍然是研究溶液里大分子或者胶体运动的经典模型，物理化学家克莱默就基于该模型给出了著名的化学过渡态理论的数学分析，与此相关的亚稳态理论至今仍然是随机过程数学理论研究中的热点。又比如，高分子聚合物在溶液里的构象，一直以来都是用随

机过程来建模的，不仅如此，化学家们在为高分子聚合物建模的过程中，还提出了不少新的随机过程模型，其中一些模型的严格数学理论已经成为现在随机过程理论中最核心的问题，2006 年和 2010 年都有数学家因为解决了其中的某些重要问题而获得了菲尔兹奖。再比如，物理学家德尔布吕克（1969 年诺贝尔生理学和医学奖得主）第一次用随机过程里的马尔可夫跳过程模型描述了自催化化学反应系统中化学物质分子数的随机涨落，这一类模型现在被称为化学主方程模型，得到了广泛的应用。除此之外，生物学中的遗传学其实很早就开始使用较为复杂的随机模型了，诞生了像哈代 - 温伯格（Hardy-Weinberg）平衡和费舍尔 - 怀特（Fisher-Wright）模型这样的著名理论。

从二十世纪中叶开始，伴随着 DNA（核糖核酸）双螺旋结构和与此相关的一系列生物现象的发现，生物学进入了分子生物学和细胞生物学的崭新阶段；而由于实验数据的逐渐积累和实验手段的不断进步，对于定量的需求也越来越多，特别是一些系统层面的现象，并非是可以由若干个基因和蛋白质的存在与否或者突变与否等来完全刻画和解释，这就使得数学模型逐渐开始成为分析和解释这些现象的有力武器。从上世纪七十年代开始，用常微分方程和偏微分方程等为模型来解释生物现象的工作逐渐增多，特别是庞加莱开创于 19 世纪末的微分方程定性理论的广泛应用，使得人们可以从整体上认识一个生物系统的行为，包括多稳态、周期振荡和斑图的形成机制，等等。与此同时，一些先驱者也开始对单个细胞内部的 DNA 转录翻译调控过程建立随机模型，但是由于当时实验手段的限制，单细胞内部精细的现象还无法被直接观测到，因此当时这些工作并未得到足够的重视。

直到二十世纪八十年代末到九十年代，单分子的实验技术被发明了出来，人们终于可以跟踪单个分子的随机行为了，而溶液里单个分子或者几个分子之间的化学反应由于受到溶液分子无规则热运动的影响，本身就是随机的，对其动力学行为的定量刻画就必然需要建立随机模型。

这里就需要提到，利用随机模型对于自然现象进行建模和利用确定性的常微

分或者偏微分方程建模的本质区别。随机过程是有两种等价描述的，一种是刻画其分布函数或者密度函数是如何随着时间演化的，一般这要么是有限或者可数维的常微分方程，要么是偏微分方程，是确定性的，而且通常是线性的；另一种是刻画轨道的性质，比如对于离散状态的随机过程，我们需要描述的是如果已知当前轨道处于某一状态，那么还要等待多久才会跳跃到下一状态呢？这个等待时间以及下一状态所服从的联合概率分布是怎么样的。在单分子实验观测中，人们往往是观测到一条或者若干条轨道，因此对于随机轨道的数学分析和描述就显得格外的重要，而且有很多轨道层面的数学问题并不能够通过研究分布层面的常微分或者偏微分方程来解决，即使可以解决也异常地繁琐和困难，轨道分析有着其自身的特性和独特的技巧。另外，学过随机过程的都知道，这些具体的轨道都是不可重复的，而科学实验又讲究的就是可重复性。这就意味着我们需要从不可重复的具体轨道里提取出可以重复的统计规律，即分布规律，这就是随机轨道分析的目标。

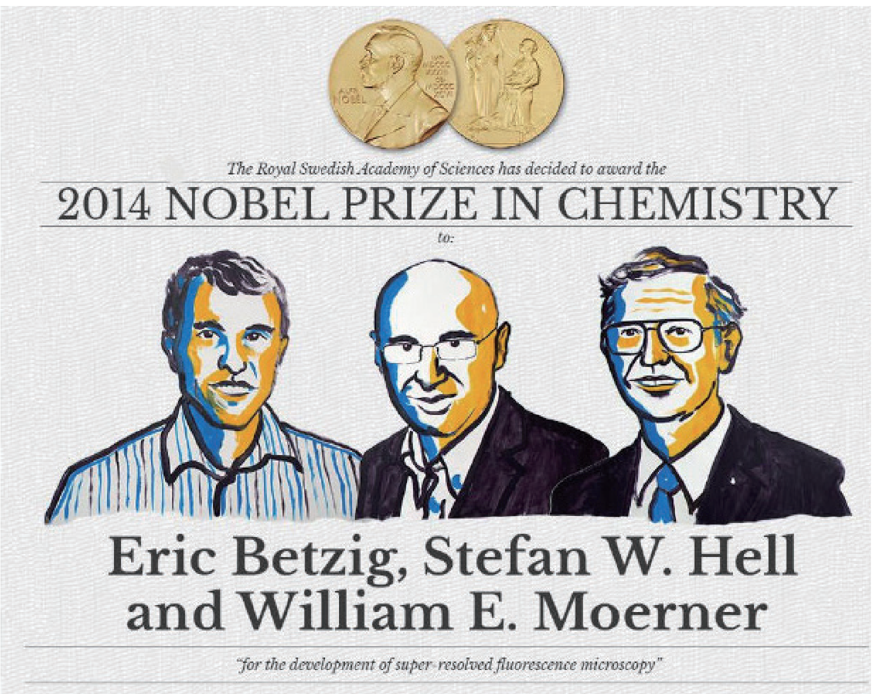

图 3. 2014 年诺贝尔奖授予了三位单分子观测技术方面的先驱

到了二十一世纪，随着单分子和单细胞实验技术的突飞猛进，人们已经可以定量追踪到单个细胞内部精细的随机现象了，特别是荧光标记实验，可以高分辨率地记录下单个细胞内部单个分子或者若干个分子数目随时间的变化。而单个细胞内部的化学反应所处的时间尺度大约在毫秒到微秒之间，而空间尺度大约就在微米到纳米之间，这正是随机规律起作用的时空尺度，比这个时空尺度再长再大的，其随机效应就不那么明显了，而比这个时空尺度再短再小的，也许量子效应就会起决定性的作用。于是原本还局限在物理化学领域的随机模型就开始慢慢地进入了生物学领域，围绕着单细胞中心法则给出了很多定量的，而又和实验十分吻合的精彩数学结果。

生物学中的中心法则，就是 DNA 会自我复制，也可以转录成信使 RNA (*m*RNA)，信使 RNA 又可以翻译成蛋白质的过程。因此最简单的中心法则随机模型（不包括 DNA 复制），就是把这个过程的每一步都简化成一个一步的化学反应来处理。早在二十世纪四十年代，克莱默就得出了热运动驱动下的单个基元化学反应的反应速率的表达式，同时人们也发现，这个反应的时间是随机的，近似地服从指数分布，于是最简单的一步化学反应所对应的随机模型就是等待一个指数分布的随机时间后完成该化学反应，该指数分布的平均值的倒数就被称为反应速率。于是最简单的中心法则随机模型就如图 4 所示，每一步化学反应上的参数就是该步化学反应的反应速率。

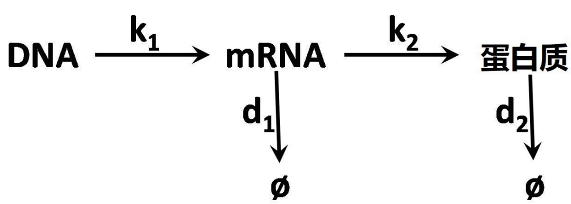

图 4. 最简化的单细胞中心法则随机模型的化学反应图

当然，这是个极其简化的模型，因为中心法则的每一步本质上都是由很多很多步化学反应构成的，即使在原核生物里亦是如此，那么为什么人们还可以

接受这样一个简化模型呢？这就要从建模方法论的角度来探讨了。大体来说，数学模型有两类：一类是定量的统计模型，可以准确拟合和预测实验的结果，例如开普勒三大天体运动定律以及生物信息学模型等；另一类是定性的机制模型，虽然不一定能从量上给予非常准确的预测，却可以从物理化学的理论角度给出更深层次的认识，提出某种合理的机制解释。二者结合得最好的当属牛顿的万有引力定律。 遗憾的是，对于生命体这样复杂的系统，极少有既在定性又在定量上都很完美的例子。因此，对于生物系统建模，需要在这二者之间寻找某种平衡。统计学家乔治•博克斯有句名言“所有的模型都是错误的，但是有些是有用的”(Every model is wrong, but some are useful)。所以，即使是如此简化的模型，其在细节上的确无法完美地刻画中心法则的过程，但是该模型整体上的一些定性结果，甚至是定量结果，却可以很好地符合实验。

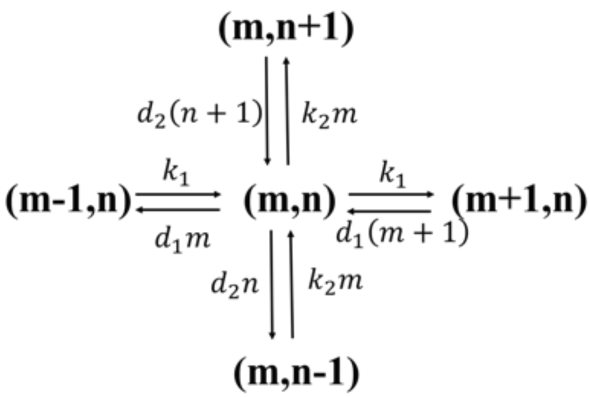

图 5. 最简单随机模型的状态跃迁图

我们可以在这个最简单的机制下建立随机模型，如上图所示。一个细胞的状态就由此时此刻细胞中这种 *m*RNA（信使核糖核酸分子）和蛋白质的个数 (m, n) 来描述，然后随机模型刻画的就是该细胞状态在二维非负整数格点上的随机跳跃，即从每个状态 (m, n) 出发，一共有四条边可以跳跃，其中 $(m, n) \geq (m+1, n)$ 对应着 DNA 的转录，即 DNA 合成了一个 *m*RNA 分子，$(m, n) \geq (m, n+1)$

对应着翻译过程，即 mRNA 生成了一个蛋白质分子，$(m, n) \geq (m-1, n)$ 和 $(m, n) \geq (m, n-1)$ 分别对应着 mRNA 和蛋白质分子的降解。可以想象成，每条边上都放有一个闹钟，该闹钟在等待一个服从对应参数指数分布的随机时间后就会闹响。然后真正的等待时间和跳跃就是选择这四个闹钟里最先闹的那一个来执行。这里讲的是该随机模型轨道层面的描述，如果写出对应的概率分布层面的演化方程，那就被称为化学主方程，因为这里是为化学反应系统在建模。该模型的化学主方程如下，其中 $p(m, n, t)$ 表示的是单细胞在 t 时刻恰好有 m 个 mRNA 分子和 n 个蛋白质分子的概率：

$$\begin{aligned}\frac{\mathrm{d}p(m,n,t)}{\mathrm{d}t} = & k_1 p(m-1,n,t) - k_1 p(m,n,t) + d_1(m+1)p(m+1,n,t) \\ & - d_1 mp(m,n,t) + k_2 mp(m,n-1,t) - k_2 mp(m,n,t) \\ & + d_2(n+1)p(m,n+1,t) - d_2 np(m,n,t).\end{aligned}$$

定义 mRNA 个数的平均值为 $\langle m\rangle(t) = \sum_{m,n=0}^{\infty} mp(m,n;t)$，蛋白质个数的平均值为 $\langle n\rangle(t) = \sum_{m,n=0}^{\infty} np(m,n;t)$，然后根据化学主方程可以得出

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle m\rangle(t) &= k_1 - d_1\langle m\rangle(t), \\ \frac{\mathrm{d}}{\mathrm{d}t}\langle n\rangle(t) &= k_2\langle m\rangle(t) - d_2\langle n\rangle(t).\end{aligned}$$

同样地，可以定义

$$\begin{aligned}\langle m^2\rangle(t) &= \sum_{m,n=0}^{\infty} m^2 p(m,n;t), \\ \langle n^2\rangle(t) &= \sum_{m,n=0}^{\infty} n^2 p(m,n;t), \\ \langle mn\rangle(t) &= \sum_{m,n=0}^{\infty} mnp(m,n;t),\end{aligned}$$

并得到

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle m^2\rangle(t)=k_1[2\langle m\rangle_t+1]-d_1[2\langle m^2\rangle(t)-\langle m\rangle(t)],$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\langle n^2\rangle(t)=k_2[2\langle mn\rangle(t)+\langle m\rangle(t)]-d_2[2\langle n^2\rangle(t)-\langle n\rangle(t)],$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\langle mn\rangle(t)=k_1\langle n\rangle(t)-d_1\langle mn\rangle(t)+k_2\langle m^2\rangle(t)-d_2\langle mn\rangle(t).$$

令上式右边为零，则可以得到 mRNA 和蛋白质分子个数在平稳分布下的均值、方差等：

$$\langle m\rangle_{ss}=\frac{k_1}{d_1},\qquad \langle n\rangle_{ss}=\frac{k_1k_2}{d_1d_2},$$
$$\mathrm{var}[m]_{ss}=\langle m\rangle\frac{k_1}{d_1},\qquad \mathrm{var}[n]_{ss}=\frac{k_1k_2}{d_1d_2}\left(1+\frac{k_2}{d_1+d_2}\right),$$
$$\mathrm{cov}(m,n)_{ss}=\frac{k_1k_2}{d_1(d_1+d_2)}.$$

2004 年，瑞典生物学家约翰 • 鲍尔森 (他本科是数学和生物的双学位) 提出令该化学主方程的右端为零，即假设已达到平稳状态，并求解平稳情况下的 mRNA 和蛋白质个数的均值和方差，然后用变异系数的平方，也就是分子数概率分布的方差除以均值的平方，来作为刻划随机性大小的量，并以此厘清了当时互相矛盾的一些实验结果，指导实验学家如何正确的处理数据[1]。在这项工作之前，很多实验学家是使用方差除以均值来度量随机性的，这被称为 Fano factor。现在我们知道，Fano factor 可以用来刻画一个分布和泊松分布的差距 (泊松分布的 Fano factor 永远是 1)，但是由于它是有量纲的，不适合作为随机性大小的度量。

[1] Paulsson, J., Summing up the noise in gene networks. Nature, 2004. 427(6973).

[2] Taniguchi, Y., et al., Quantifying E-coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. Science, 2010. 329(5991).

除了均值和方差以外，这个模型的化学主方程还可以计算出单细胞中 mRNA 个数和蛋白质个数的相关系数为

$$\rho=\frac{\left\langle(m-\langle m\rangle_{ss})(n-\langle n\rangle_{ss})\right\rangle_{ss}}{\sqrt{\operatorname{var}(m)_{ss}}\cdot\sqrt{\operatorname{var}(n)_{ss}}}=\sqrt{\frac{d_2}{d_1+d_2}\frac{1}{1+\frac{d_1+d_2}{k_2}}}.$$

2010 年，哈佛大学谢晓亮实验室测量了大肠杆菌单细胞内一千多种蛋白质及其对应的 mRNA 的数目，并计算了相关系数，发现相关系数都非常小[2]。根据以上这样一个最简单模型得出的相关系数的表达式，就可以推断，该现象很可能是因为 mRNA 的降解速率 (d_1) 很大造成的，使得同一时刻测到的蛋白质和 mRNA 数目之间并不存在对应关系。

对于如此简单的模型，本以为我们还可以得到更多，但其实我们也只能得到其分布的均值、方差、协方差这些数字特征的表达式，其概率分布的表达式中含有超几何函数，使得具体应用起来十分困难[3]。单细胞内的单分子精度的实验测到的都是真正的概率分布，其信息远比若干个数字特征要多得多。因此人们就迫切希望得到其概率分布的近似简单表达式，至少在某些符合实验现象的合理假设下希望可以得到。在这件事上，随机数学模型的建立和分析就发挥了关键的作用。

不过，历史并不是沿着复杂模型加上合理假设推出简化模型这条路线来走的，而是反其道而行之，首先是根据实验观察建立了基于现象的简化模型，后来才慢慢意识到，直到最近才真正证明了这个简化模型其实就是更复杂模型在一定条件下的极限。

[3] Shahrezaei, V. and P.S. Swain, Analytical distributions for stochastic gene expression. Proceedings of the National Academy of Sciences of the United States of America, 2008. 105(45).

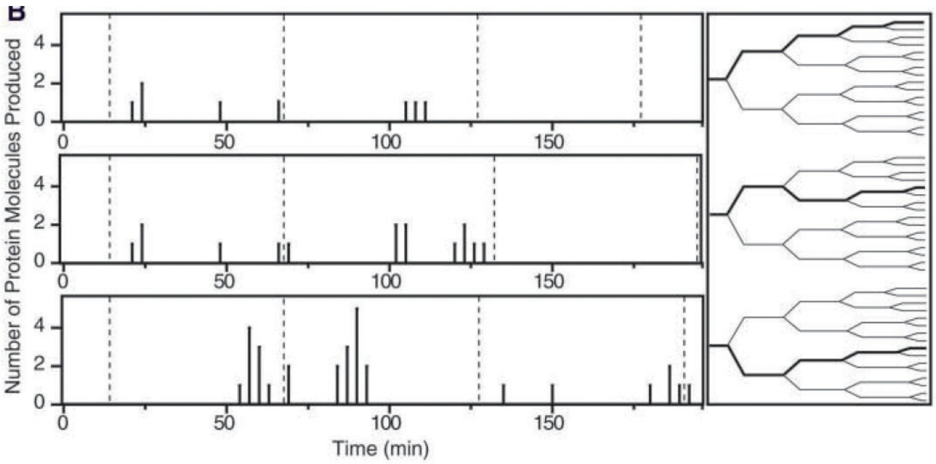

图 6. 蛋白质生成过程的爆发型动力学 . 来自文献[5]

早在 2006 年，又是哈佛大学的谢晓亮实验室首次观察到了单细胞中单个蛋白质分子合成过程中的爆发现象，即在每等待一个随机的时间间隔后，该单细胞内部会在很短的时间内密集地合成若干个该蛋白质分子，然后再次进入等待状态[4,5]。经过精巧的实验测量，实验学家发现每次蛋白质合成的爆发事件背后所对应的 *m*RNA 分子个数，至少在细菌中近似为 1。而且每个 *m*RNA 分子的生存时间相比较于相邻两个 *m*RNA 生成之间的时间间隔是很短的，因此每一次的爆发强度，其实就是每一个 *m*RNA 分子在其被降解之前所能产生的蛋白质分子数目，而相邻两次爆发之间的时间间隔，就是相邻两个 *m*RNA 分子合成的时间间隔。

有了这些准备，我们就可以建立一个只基于这些现象而且只关心蛋白质的生成与降解的随机模型，即延续最简单模型里的假设，假设每隔一个参数为 k_1 的指数分布的随机时间，DNA 会合成出一个 *m*RNA 分子，但这个 *m*RNA 分

[4] Cai, L., N. Friedman, and X.S. Xie, Stochastic protein expression in individual cells at the single molecule level. Nature, 2006. 440(7082).

[5] Yu, J., et al., Probing gene expression in live cells, one protein molecule at a time. Science, 2006. 311(5767).

子的生存时间极短，可以忽略不计，且这个 mRNA 在其生存时间内会以一定的概率分布产生出一定量的蛋白质分子。那现在的问题就是单个 mRNA 分子在降解之前能产生出的蛋白质分子数目服从怎样的概率分布呢？一旦有了该分布类型，建立这个随机过程模型就算是万事俱备了。

这是一个典型的随机过程数学问题，即假设每个 mRNA 分子是在等待一个参数为 d_1 的指数分布随机时间后消亡，而在这段生存时间内，每过一个参数为 k_2 的指数分布的随机时间间隔，该 mRNA 会合成出一个蛋白质分子，这其实就是随机过程里经典的泊松过程，只是这里该过程的总时间也是随机的。计算一次爆发过程中产生 n 个蛋白质分子的概率，需要用到的就是概率论里著名的全概公式，其实也就是类似于二重积分，即先在条件固定的情况下计算条件概率，再对所有不同条件下计算出的条件概率进行概率加权平均。具体到这个问题，就是先计算在该 mRNA 分子的生存时间为 t 的条件下，该 mRNA 一共生成了 n 个蛋白质分子的概率，这是典型的泊松过程计算，结果是参数为 k_2t 的泊松分布。而又因为该 mRNA 分子的生存时间本身就是一个概率密度为 $d_1e^{-d_1t}$ 的指数分布随机变量，二者一结合，就可以计算出一次爆发过程中产生 n 个蛋白质分子的概率为

$$G_n \underline{\underline{\mathrm{def}}}\ p(n)=\int_0^{+\infty} d_1\mathrm{e}^{-d_1t}\cdot\mathrm{e}^{-k_2t}\frac{(k_2t)^n}{n!}\mathrm{d}t=q^n(1-q),$$

其中 $q=\dfrac{k_2}{k_2+d_1}$。

这是个标准的几何分布。其实，这个分布早在上世纪七十年代末，就由单细胞随机模型的先驱们给出了[6]。

[6] Berg, O.G., A model for the statistical fluctuations of protein numbers in a microbial population. J Theor Biol, 1978. 71(4).

至此，我们终于可以给出这样一个基于转录爆发现象而建立的随机模型的概率分布 $p(n, t)$，即在 t 时刻恰好有 n 个蛋白质分子的概率，随时间的演化方程

$$\frac{\mathrm{d}p(n,t)}{\mathrm{d}t}=k_1\left(\sum_{j=1}^{n}G_j p(n-j,t)-qp(n,t)\right)+d_2[(n+1)p(n+1,t)-np(n,t)],$$

其中 d_2 是单个蛋白质分子由于自身降解或者细胞分裂而导致的衰减速率。

上面我们说了，建立这样的只依赖于蛋白质分子数目的简化模型，就是为了得到蛋白质分子数目的概率分布，而不仅仅是均值、方差等数字特征。这个目的可以达到吗？也就是说，让上述方程的右端等于零，可不可以显式地写出平稳状态下，细胞内具有 n 个蛋白质分子的概率呢？

早在 2000 年，鲍尔森和埃伦伯格就在无任何实验启发的情况下构建了这个随机模型，并计算出了其长时间后的平稳分布，发现是负二项分布 [7]。该文并无任何推导过程。由于上述模型虽然状态空间是一维非负整数格点，但它并不是只能在紧邻的状态间跳跃，所以该模型还是具有相当复杂度的。不过幸运的是，只要我们令化学主方程的右边为零，且分别将 n 等于 0，1，2,⋯代入，就可以猜出平稳分布的表达式，再利用数学归纳法证明即可。当然，我们也可以利用母函数等方法证明，只是在这个问题上，还是猜测加数学归纳法最简便。

负二项分布本来是在概率论的赌博模型里被提出来的，从未想到有朝一日居然可以在某个自然现象中被观测到，让人只能概叹大自然的神奇。此平稳分布的表达式如下：

[7] Paulsson, J. and M. Ehrenberg, Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. Physical Review Letters, 2000. 84(23).

$$p^{ss}(n)=\frac{b^n}{(1+b)^{a+n}}\frac{\Gamma(a+n)}{\Gamma(a)n!},$$

其中参数 $a = k_1/d_2$ 表示一个细胞周期时间内的爆发次数，而 $b = k_2/d_1$ 则是每次爆发所生成的蛋白质分子数的平均值。这两个数字都是可以在实验中直接测量的，最后的平稳分布也是可以测量的。2006 年谢晓亮实验室的文章中，这三个量被分别测量到，发现它们与该理论符合得很好。当然，并不是说这就意味着该模型里的所有假设都是对的，但这的确表明，通过这样一个基本符合实验观测的简单模型，是可以抓住真正的复杂生命过程中的一些重要定量性质的。最近，贾晨和钱纮从数学上还证明了该模型正是最简单机制模型在一定

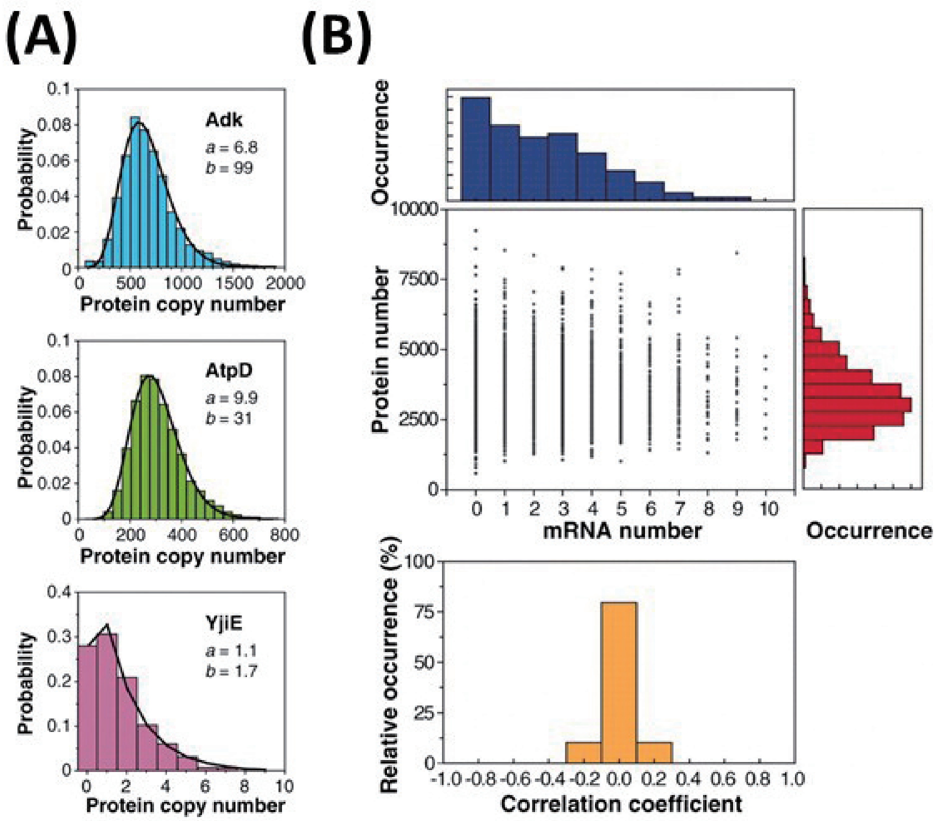

图 7. (A) 单细胞中蛋白质分子个数的平稳分布
(B) 单细胞中信使核糖核酸和蛋白质分子个数的相关系数 . 来自文献 [2]

条件下的近似[8]。

关于单细胞内的蛋白质分子数目的演化和分布，似乎已经被最简单的数学模型及其简化模型定量地刻画得很好了。那么是不是单细胞内的 *m*RNA 分子数目的演化和分布也可以被前面所述的最简单机制模型或者它的变形刻画得很好呢？答案是否定的。其实早在 2005 年，普林斯顿大学的考克斯实验室就在世界上首次观测到了单个细胞中的单个 *m*RNA 分子的生成和降解现象，并且发现了一种名为转录爆发的现象[9]。转录爆发，顾名思义，类似于蛋白质合成中的翻译爆发现象，说的是，DNA 似乎存在着某种神奇的机制，它会使得即使在完全不存在抑制蛋白结合的情况下，DNA 也会时不时地偷懒不工作，即转录停止，过一段时间后再重新开启。而这种 DNA 自身具有多种不同的转录状态的现象并未包括在前面所述的最简单的机制模型里。而且，2010 年谢晓亮实验室也发现，实际中 *m*RNA 分子个数的平稳分布的方差是明显大于均值的，这与最简单机制模型中预测的该分布会是方差等于均值的泊松分布，是很不一样的[2]。

早在上世纪九十年代，就有学者提出了具有两个不同的 DNA 转录状态的随机模型[10]，即俗称的两状态模型：

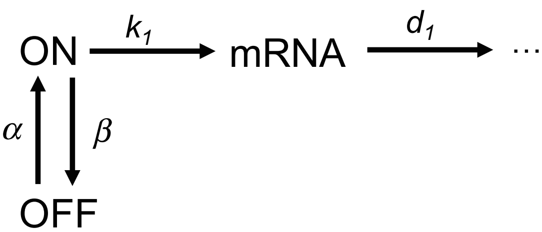

图 8. 两状态基因开关模型的化学反应图

[8] Jia, C., M.Q. Zhang, and H. Qian, Emergent Levy behavior in single-cell stochastic gene expression. Physical Review E, 2017. 96(4).

[9] Golding, I., et al., Real-time kinetics of gene activity in individual bacteria. Cell, 2005. 123(6).

[10] Peccoud, J. and B. Ycart, Markovian modeling of gene-product synthesis. Theoretical Population Biology, 1995. 48(2).

$$
\begin{tikzcd}[column sep=large, row sep=large]
(\mathbf{on,m\text{-}1}) \arrow[r, shift left, "k_1"] \arrow[d, shift right, "\beta"'] & (\mathbf{on,m}) \arrow[l, shift left, "d_1 m"] \arrow[r, shift left, "k_1"] \arrow[d, shift right, "\beta"'] & (\mathbf{on,m+1}) \arrow[l, shift left, "d_1(m+1)"] \arrow[d, shift right, "\beta"'] \\
(\mathbf{off,m\text{-}1}) \arrow[u, shift right, "\alpha"'] & (\mathbf{off,m}) \arrow[l, "d_1 m"] \arrow[u, shift right, "\alpha"'] & (\mathbf{off,m+1}) \arrow[l, "d_1(m+1)"] \arrow[u, shift right, "\alpha"']
\end{tikzcd}
$$

图 9. 两状态基因开关随机模型的状态跃迁图

从数学上来看，这个模型和前面的两个模型一样，也都是属于离散状态的马尔可夫跳过程。其状态概率 $p_{on}(m,\ t)$ 和 $p_{off}(m,\ t)$ 分别表示 t 时刻细胞内恰好有 m 个 mRNA 分子且 DNA 正好处于 ON 或者 OFF 状态的概率，它们随时间的演化方程为

$$
\begin{aligned}
\frac{\mathrm{d}p_{\text{on}}(m;t)}{\mathrm{d}t} &= k_1 p_{\text{on}}(m-1) + d_1(m+1)p_{\text{on}}(m+1) + \alpha p_{\text{off}}(m) \\
&\quad -(k_1 + d_1 m + \beta)p_{\text{on}}(m;t), \mathrm{K} \approx k_0 e^{-\frac{\Delta\Phi_{AB}}{\sigma}}, \\
\frac{\mathrm{d}p_{\text{off}}(m)}{\mathrm{d}t} &= d_1(m+1)p_{\text{off}}(m+1) + \beta p_{\text{on}}(m) - (d_1 m + \alpha)p_{\text{off}}(m).
\end{aligned}
$$

同样地，由于该模型里并没有二阶或者二阶以上的非线性化学反应，其得出的随机模型也是线性的，因此可以很容易算出平稳状态下的均值和方差，特别是 Fano factor 的表达式

$$
F = 1 + \frac{\alpha\beta k_1}{(\alpha+\beta)(\alpha+\beta+d_1)\beta}
$$

显然是大于 1 的。

因此，两状态模型可以很好地解释实验里观测到的 Fano factor 大于 1 的现象。但是两状态模型是个一般模型，具体到了转录爆发这件事情上，在没有抑制蛋白结合的前提下，DNA 产生这样两个 ON 和 OFF 状态的分子机制又是什么

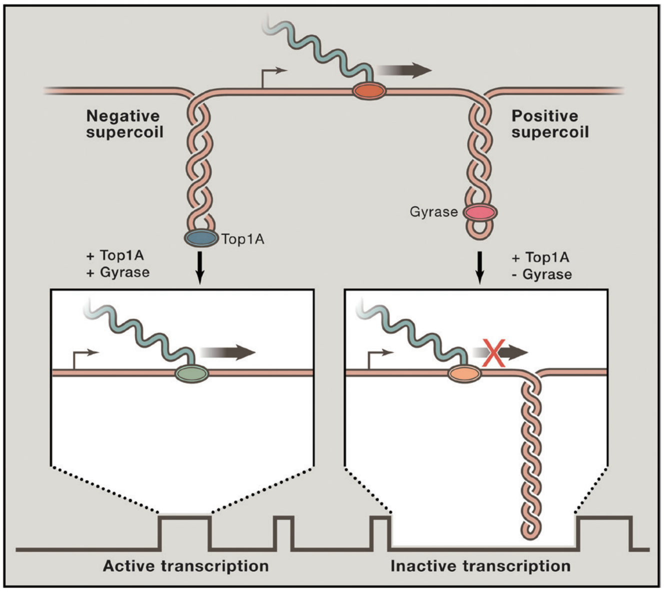

图 10. 大肠杆菌细胞里的 DNA 拓扑结构和转录爆发现象之间的关系 . 见文献 [11]

呢？在发现该分子机制的过程中，两状态模型又能发挥怎样的作用呢？这里要稍微展开说一下这个故事的来龙去脉。

在 2011 年左右，人们发现即使是在细菌中，DNA 分子也不是自由的，它们会被分成一段一段的锚定在一些大的蛋白质分子上。同时早在上世纪八十年代，人们就发现在 DNA 转录的过程中，处于 RNA 聚合酶前方的 DNA 链将聚集所谓正超螺旋（positive supercoiling），通俗的来讲就是 DNA 双螺旋结构原本是每 10.5 个碱基对转一圈（360 度）的，而现在被转的越来越紧，每转一圈的碱基对数目越来越少，DNA 形变越来越厉害了；与此相对应的，处于 RNA 聚合酶后方的 DNA 链将产生负超螺旋（negative supercoiling），即完成一圈的碱基对数目越来越多。这两种 DNA 形变将由于 DNA 被锚定在一些大分子上而无法相互抵消。因此细胞内需要有两种酶，拓朴异构酶

[11] Levens, D. and Larson, D.R.: A New Twist on Transcriptional Bursting, Cell, 158, 241-242, 2014

（Topoisomerase IA）专门负责在 RNA 聚合酶的后方释放负超螺旋，而旋转酶(gyrase)则专门负责在 RNA 聚合酶的前方释放正超螺旋。然而有研究表明，拓扑异构酶的活性是很高的，而旋转酶的活性是不高的，而且旋转酶在活细胞内的分子数也并不十分多，平均到每一段被铆定的 DNA 片段的话只有大约一个旋转酶分子。

2014 年，谢晓亮实验室发展了一套高通量的体外单分子荧光技术，可以实时地高分辨率地观测到单个分子上正在发生的转录过程。通过这一技术发现，转录过程在 RNA 聚合酶前段不断聚集起来的正超螺旋，会渐渐地减慢 RNA 聚合酶的延伸速度，并最终彻底阻止转录的起始。而旋转酶和 DNA 分子的结合又可以使得转录得以继续。因此，就可以合理猜测，这就是活细胞内部广泛存在的转录爆发现象的机制。

然而，事情并非如此简单。生物学是极其复杂的，体外的单分子实验技术自然是高超的，但是体外观察到的机制就一定是体内该现象产生的主要原因吗？而且棘手的是，体外如此高分辨率的实验由于实验技术所限，是无法在体内完成的，那么我们又如何来验证该机制是否在体内发挥着重要作用呢？

这个时候就需要数学模型的作用了，笔者也很荣幸地参与了此项工作。我们采用了上面所述的两状态基因开关模型。实验上虽然无法在体内完成十分高精度的单分子追踪实验，但是却可以精准地测出每一个单细胞中该 mRNA 分子的个数，即可以测量出平稳分布来。然而，和最简单的机制模型一样，两状态模型的平稳分布一般是没有简单的表达式的，而没有具体表达式，就无法和实验数据进行更为详细的对比和拟合。

不过我们注意到，实验上发现，旋转酶分子和 DNA 分子的结合和解离，相比于 DNA 处于转录活跃状态的转录速率来说，是很慢的，即模型中的 α 和 β 是很小的。因此我们可以借助应用数学上常用的渐近分析的办法，发现平稳分

布可以近似为一个在零点单独再加一个峰的泊松分布，被称为零点膨胀的泊松分布：

$$p(0)=\frac{\alpha}{\alpha+\beta}\mathrm{e}^{-\lambda_1}+\frac{\beta}{\alpha+\beta},\qquad p(m)=\frac{\alpha}{\alpha+\beta}\frac{\lambda_1^m}{m!}\mathrm{e}^{-\lambda_1}\ \ (m\geq 1),$$

其中 $\lambda = k_1/d_1$。

一旦有了平稳分布的具体表达式，我们就可以直接利用实验数据来推断 DNA 转录处于活跃和不活跃状态的平均时间的比值了，而如果体外观测到的机制是正确的，那么这个比例就会随着旋转酶的不同浓度，以及旋转酶和 DNA 结合的不同强度而发生明显的变化。经过分析实验数据，我们发现，该比值强烈依赖于细胞内的旋转酶浓度和活性，且其变化的趋势和模型的预测完全相符。在很多个基因上俱是如此。综合以上的实验和数学模型分析，间接但是严格的证明，细菌里的转录爆发机制，就是来源于旋转酶分子与 DNA 分子的不断随机结合与解离，从而导致该 DNA 片段的超螺旋情况发生改变[12]。

上面这些单细胞中心法则的随机模型，都是没有考虑到调控和反馈的，而基因调控才是生命体实现各种功能的根本。带有调控机制的随机数学模型在过去二十年也得到了长足的发展，但是由于生物学的复杂性，略微复杂一些的模型已经很难得出显式解了，大都依赖于数值模拟，且都需要具体问题具体分析，普适性的理论并不多。

尽管如此，数学家们还是在下面两个方向上取得了进展。一个是高维化学反应系统的随机模型在多尺度下的简化以及数值模拟的加速方法，即随机过程和计算数学的结合。另一个则是具有正反馈的生物化学反应系统会出现多个

[12] Chong, S.S., et al., Mechanism of Transcriptional Bursting in Bacteria. Cell, 2014. 158(2).

稳态，求解在不同稳态之间的跃迁速率。下面我们来简单介绍一下后一个方面的工作。

多个稳态的意思就是在基因组相同的情况下，单细胞在形态上或者分布上呈现出聚集成若干类的情形。比如 mRNA 或者蛋白质的分布呈现明显的双峰或者多峰的情况，而这样的每个峰都被称为是单细胞的一个表型或者稳态。多稳态的产生机制大体上有两大类。一类是类似前面提到的转录爆发机制，即是由于 DNA 不同状态间的转移特别慢而导致的，而这一类稳态之间的跃迁速率也很简单，就是由这个 DNA 不同状态间的转移速率所决定。另一类产生多稳态的机制则更为常见，是由于比较强的正反馈所形成的，这样形成的稳态一般都很稳定。这一类稳态之间的跃迁，可以在随机性相对比较小的时候得到比较简单的表达式，这就需要用到在二十世纪后半叶才建立起来的随机过程里著名的大偏差理论，特别是其中弗雷德林 - 温策尔（Freidlin-Wentzell）型大偏差理论[13,14]，其核心思想是如何刻画单条轨道的“概率密度”以及如何构造类似于能量函数的拟势函数。

过去二十年，人们得到了在不同的多尺度参数假设下，这类细胞表型（多稳态）间的跃迁速率[15,16,17,18]，它们都有一个统一的形式，可以类比于物理化

[13] Dembo, A. and O. Zeitouni, Large deviations techniques and applications. 2nd ed. Applications of mathematics. 1998, New York: Springer. xvi, 396 pp.

[14] Freidlin, M.I., A.D. Wentzell, and J. Szucs, Random perturbations of dynamical systems. Third ed. Grundlehren der mathematischen wissenschaften. 2012, Heidelberg: Springer. xxviii, 458 pp.

[15] Ge, H., H. Qian, and X.S. Xie, Stochastic phenotype transition of a single cell in an intermediate region of gene state switching. Phys Rev Lett, 2015. 114(7).

[16] Zhu, X.M., et al., Calculating biological behaviors of epigenetic states in the phage lambda life cycle. Funct Integr Genomics, 2004. 4(3).

[17] Fang, X.N., et al., Nonequilibrium physics in biology. Reviews of Modern Physics, 2019. 91(4).

[18] Lu, M.Y., J. Onuchic, and E. Ben-Jacob, Construction of an effective landscape for multistate genetic switches. Physical Review Letters, 2014. 113(7).

学中著名的 Kramers 速率公式，即首先可以证明存在一个类似能量函数的多势阱函数，被称为 landscape 函数，中文翻译为景观或者地貌函数，其实就是弗雷德林 - 温策尔理论中的拟势函数。它的每个势阱都对应于一个稳态，而要在两个稳态间跃迁，就必须要跨过势垒。设从稳态 A 到稳态 B 的势垒高度为 $\Delta\Phi AB$，其随机性大小可以由某个参数 σ 来度量，那么从 A 跃迁到 B 的速率 k 就可以在 σ 很小的情况下渐近地表示成

$$\mathrm{k} \approx k_0 e^{-\frac{\Delta\Phi_{AB}}{\sigma}},$$

其中的系数 k_0 一般是不可以由地貌函数来获得的，也没有一般表达式，这是和平衡态化学动力学里的 Kramers 速率公式最大的区别，也是生命体作为非平衡系统的特征之一。不过，现在的这个速率公式还是抓住了跃迁速率的最高阶，即指数阶的定量表达式。其实地貌函数不仅可以作为不同表型之间相对稳定性的定量刻画，而且还预言了在随机性很小的情况下，会出现非平衡态相变现象[19]。

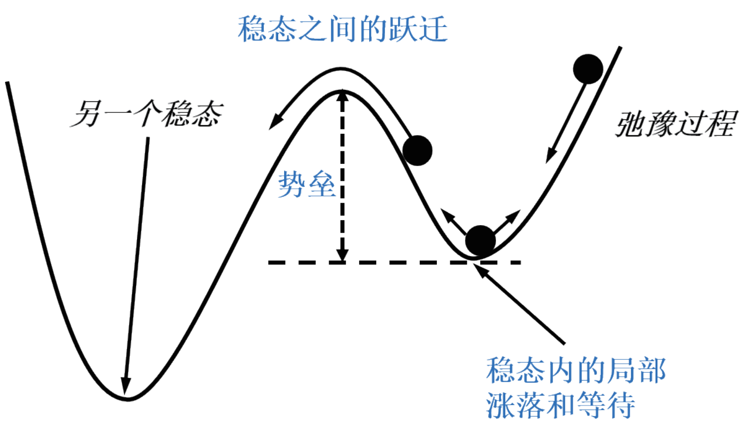

图 11. 双稳态系统的动力学

[19] Ge, H. and H. Qian, Thermodynamic limit of a nonequilibrium steady state: Maxwell-type construction for a bistable biochemical system. Physical Review Letters, 2009. 103(14)3.

对于大偏差理论及其在单细胞生物学中的应用，这里就不再继续展开。其实，这样一个地貌函数的图景，告诉我们的远不止上面提到的这些，比如我们来看任何一个在此地貌函数下运动的个体，它都会先进行一个弛豫过程，到达一个距离自己最近的稳态的底部，并在此进行漫长时间的等待和局部的涨落，由于随机性的驱动，它会突然在某一个时刻，快速地越过某个势垒，从而到达另一个稳态。这样一个动力学的图像，并不仅仅可以用来刻画单个细胞的运动，还可以用来刻画健康和疾病的关系，甚至可以用于描述进化过程中的种群突变，乃至描述一个社会历史的变迁。

作者简介：葛颢，北京国际数学研究中心长聘副教授，北京大学生物医学前沿创新中心特聘研究员，博士生导师。

此文源自《数学文化》2020 年第 4 期。感谢 Global Science Press 允许本通讯转载。