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。

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

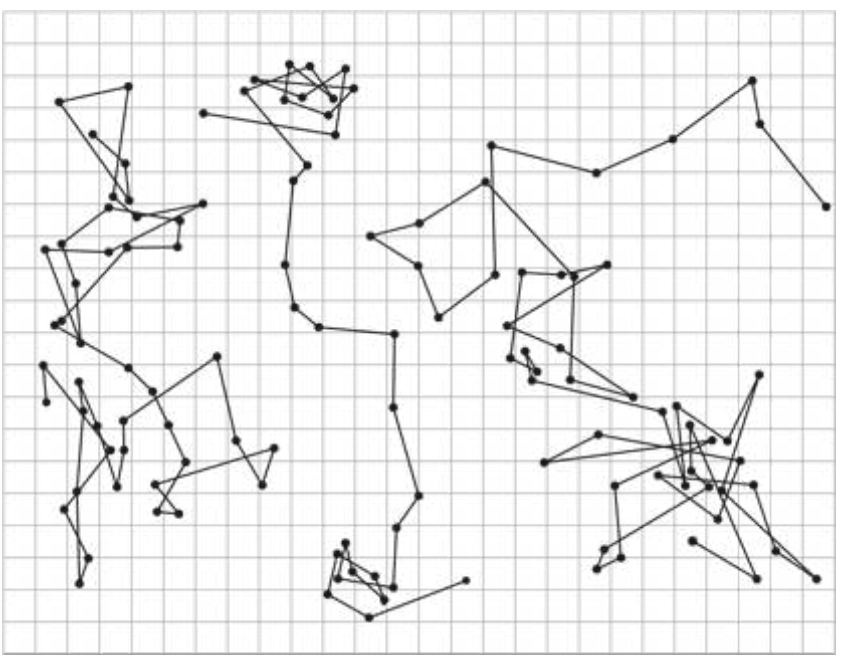


图 1. 二维布朗运动轨道 (from the book of Jean Baptiste Perrin, Les Atomes)

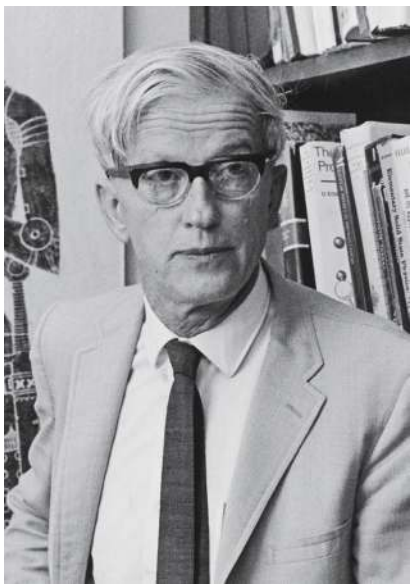


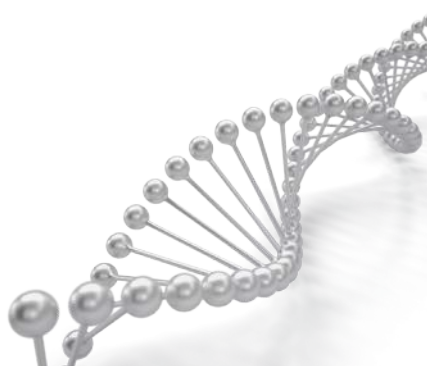
图 2. 德尔布吕克
（Max Delbrück, 1906-1981）

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。

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

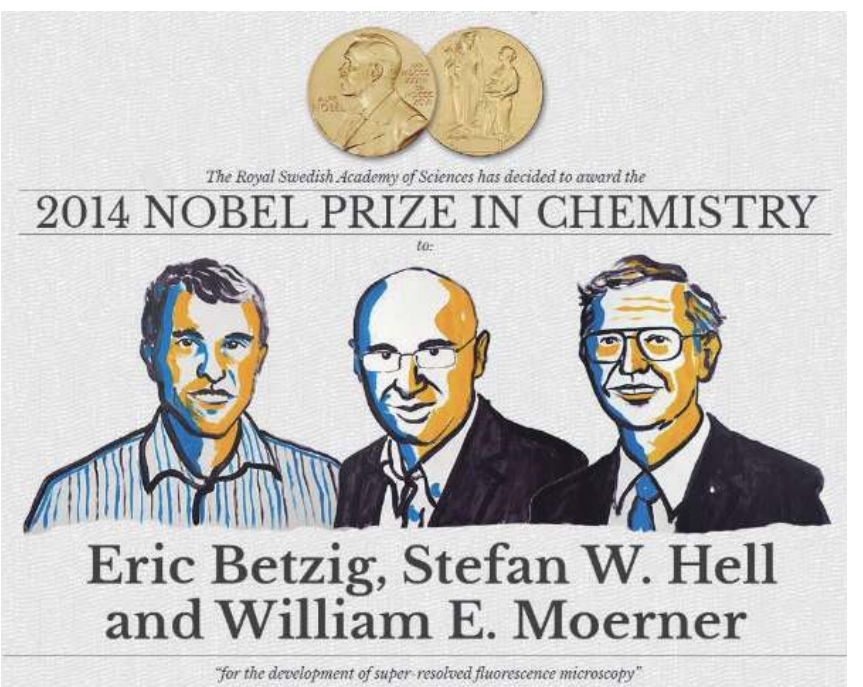


图 3. 2014 年诺贝尔奖授予了三位单分子观测技术方面的先驱

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。

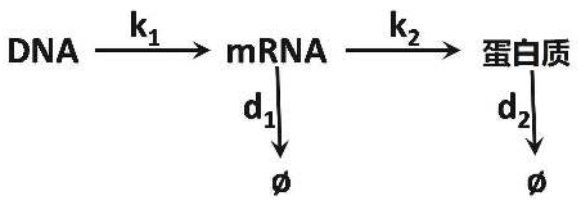


图 4. 最简化的单细胞中心法则随机模型的化学反应图

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。

我们可以在这个最简单的机制下建立随机模型，如下图所示。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