文章编号: 1000-5862(2016)01-0001-04

生化反应系统的二项矩和属性因子

周天寿
(中山大学数学与计算科学学院,广东 广州 510275)

摘要: 化学主方程对生化反应系统提供了一个建模框架,但它的分析与模拟一直是计算系统生物学的一个挑战. 另一方面,矩封闭方法对化学主方程提供了一种逼近,但普通的矩当其阶趋于无穷时并不趋于0,因此具有局限性. 这里,对概率密度函数引进二项矩,它具有 2 个突出的特点: 1) 当二项矩的阶充分大时二项矩趋于 0; 2) 二项矩能够方便地用来重构相应的概率密度函数. 基于二项矩,进一步引进反应物种的属性因子,它比普通的统计指标(如噪声强度、Fano 因子)具有某些优势. 此外,还给出了用二项矩来表示噪声强度和 Fano 因子的显式公式,并用简单的生物例子来说明二项矩的优势与 3 种统计指标的特征.

关键词: 二项矩; 属性因子; 噪声强度; Fano 因子; 反应系统
中图分类号: O 242; Q 332 文献标志码: A DOI: 10.16357/j.cnki.issn1000-5862.2016.01.01

0 引言

随机反应系统广泛存在于自然系统中,如化学领域的俄勒冈振子(Oregonator)[1]、生物学领域的基因调控网[2-3]、生态学领域的微观种群模型[4]等. 为了特征化这些微观系统的动力学行为,确定性方程(即 ODE 方程)在大多数情形无效,常常需要考虑反应系统的随机性方程或描述. 众所周知,化学主方程(描述反应系统中反应物种分子的联合概率密度的时间演化[5])在原理上对任何反应系统提供了一个数学建模框架. 然而,这种方程的分析与模拟一直是计算系统生物学的一个挑战,到目前为止并没有很好的方法来处理高维情形(即反应物种数目是很多的情形). 这是因为化学主方程关于反应物种数目是指数增长的,因此以前的数值方法,如著名的 Gillespie 随机模拟算法[5]和有限状态映射法[6],具有有限的应用前景,不能处理高维反应系统的随机行为.

作为化学主方程的一种近似,矩封闭方法[7-10]越来越来受到重视. 在大多数矩封闭方法中,仅考虑最初的几个低阶矩,如 1 阶矩和 2 阶矩(包括原点矩和中心矩). 这些低阶矩主要是用来计算反应系统中特定反应物种的统计量,如平均(或期望)、方差、噪声强度(被定义为标准差与平均之比)、Fano 因子(被定义为方差与平均之比)、分布的偏度(它特征化分布的非对称性的程度)、分布的峰度(它测量分布的尖峰的程度)等. 这些统计指标一方面可以用来简化随机分析和模拟,另一方面也常常被生物学家用来刻画某些感兴趣的化学物种的随机涨落程度,因此被广泛地使用. 然而,当分子数目比较少或随机涨落比较大时,这些指标具有局限性,甚至不能很好地刻画反应物种数目或浓度的随机性. 本质原因是普通的矩当其阶充分大时并不趋于 0(看下面的例子分析). 为克服普通矩的局限性,下面将引进二项矩,其最主要的优势是: 当二项矩的阶趋于无穷时二项矩趋于 0.

基于二项矩,将进一步(或自然地)引进一个定量化反应物种的随机张量程度的新因子,即属性因子,它被定义为 2 阶二项矩与 1 阶二项矩之比. 这种新因子的一个突出优点是: 若其值等于 1,则意味着相应的分布是泊松的(Poissonian); 若其值小于 1,则隐含着相应的分布是次泊松的(sub-Poissonian); 若其值大于 1,则隐含着相应的分布是超泊松的

收稿日期: 2015-12-10
基金项目: 国家自然科学基金/重大研究计划/重点支持项目(91230204)和科技部 973 项目子课题(2014CB964703)资助项目.
作者简介: 周天寿(1962-),男,江西南昌人,教授,博士生导师,主要从事系统生物学研究.

(sup-Poissonian). 这些特性可类比于 Fano 因子[11]. 属性因子的另一个优点是它能很好刻画基因爆发表达的特征,这是普通的噪声指标所不能办到的.

1 二项矩的引入

为方便,这里仅考虑 1 维情形. 众所周知,一个概率分布 $P(n)$ 的普通矩在数学上被定义为(这里只考虑离散变量情形,完全类似地可考虑连续变量情形)

$$\langle n^k\rangle = \sum_{n=0}^{\infty} n^k P(n) \text{ (原点矩)},$$

$$\langle (n-\langle n\rangle)^k\rangle = \sum_{n=0}^{\infty} (n-\langle n\rangle)^k P(n) \text{ (中心矩)}.$$

这里整数 k 代表矩的阶. 假如对于 $k \geqslant 3$,所有的高级中心矩都等于 0,则前 2 个矩能够用来重构此分布(这是因为在连续随机变量情形时的分布是高斯分布,它完全由最初的 2 个矩决定). 然而,2 阶以上的矩一般并不等于 0,而是当 $k \to \infty$时可能会发散到无穷. 为此,考察下列简单的生化例子,其反应式为

$$\varnothing \xrightarrow{g} X, 2X \xrightarrow{d} \varnothing, \tag{1}$$

其中反应比率 g 代表反应物种 X 的生成率,d 代表 X 的降解率. 让 n 代表物种 X 在时刻 t 的分子数目,$P(n;t)$ 代表相应的概率密度函数. 假设相应的生化过程是马氏的,以便化学主方程可以应用,且相应的主方程为

$$\frac{\partial}{\partial t}P(n;t) = g[P(n-1;t) - P(n;t)] + d[(n+1)P(n;t) - n(n-1)P(n;t)].$$

感兴趣于静态概率分布,记为 $P(n)$. 不难导出相应的 k 阶原点矩和 k 阶中心矩分别满足下列迭代形式:

$$\langle n^{k+1}\rangle = \frac{1}{k}\sum_{i=0}^{k-1}\left[S\binom{k}{i}\langle n^i\rangle + (-1)^{k-i+1}\binom{k+1}{i}\langle n^{i+1}\rangle\right],$$

$$\langle (n-\langle n\rangle)^{k+1}\rangle = \sum_{i=0}^{k+1}(-1)^{k+1-i}\binom{k+1}{i}\langle n\rangle^{k+1-i}\langle n^i\rangle,$$

其中 $k = 0,1,2,\cdots$,$S = g/d$ 代表系统的静态. 用数学归纳法,不难证明 $\langle n^{k+1}\rangle > \langle n^i\rangle$,蕴含着当 $k \to \infty$时原点矩并不趋于 0. 数值模拟证实了:当 $k \to \infty$时中心矩也不趋于 0. 这些表明普通矩当它们的阶趋于无穷时并不趋于 0,见图 1.

由图 1 可得出:当矩的阶充分大时,普通矩(原点矩和中心矩)并不趋于 0,但二项矩趋于 0,这里参数值被设为 $g = 20$,$d = 1$.

由于普通矩的上述缺陷,故引进二项矩,它被定义为

$$b_k(t) = \sum_{N\geqslant k}\binom{N}{k}P(N;t),\ k = 0,1,2,\cdots,$$

这里 $\binom{N}{k}$ 代表普通的二项系数. 类似于普通矩的情形,k 叫做二项矩的阶. 这一定义并不奇怪,这是因为二项矩实际是相应于概率密度函数的母函数在特定点的泰勒展开式的系数. 事实上,假如 $G(z;t) = \sum_{n=0}^{\infty}P(n;t)z^n$ 是概率密度函数 $P(n;t)$ 的母函数,则并不困难地显示出

$$b_k(t) = \frac{1}{k!}\frac{\partial^k G(z;t)}{\partial z^k}\bigg|_{z=1},\ k = 0,1,2,\cdots$$

更为重要的是,能够用二项矩来重构相应的概率密度函数. 事实上,有重构公式

$$P(N;t) = \sum_{k\geqslant N}(-1)^k\binom{k}{N}b_k(t),\ N = 0,1,2,\cdots. \tag{2}$$

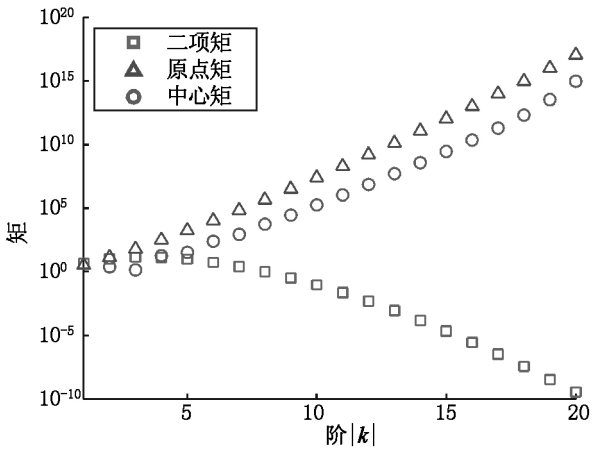

图 1 一个简单生化例子

下面考察前面的例子,即生化系统(1). 基于二项矩的定义并结合化学主方程,并不困难地导出下列二项矩方程:

$$\frac{\mathrm{d}b_k}{\mathrm{d}t} = gb_{k-1} - 2\mathrm{d}k(k+1)b_{k+1} - \mathrm{d}k(k-1)b_k,$$

其中 $k = 0,1,2,\cdots$. 注意到:由于概率的保守性,有 $b_0 = 1$. 有趣的是,对于这一例子,静态二项矩当其阶趋于无穷时趋于 0,见图 1. 此外,能够显示出下列关系:

$$b_{k+1} = \frac{1}{(k+1)!}\left[\langle n^{k+1}\rangle + \frac{(k+1)(k+2)}{2}\langle n^k\rangle + \cdots + (k+1)!\right].$$

更一般地,能够显示出 k 阶中心矩,记为 $\mu_k(t)$,与二项矩具有下列关系:

$$\mu_k(t) = (-b_1(t))^k + \sum_{i=0}^{k-1}\sum_{j=1}^{k-i} R(k,i,j)\cdot (j!)(b_1(t))^i b_j(t),$$

其中 $R(k,i,j) = (-1)^i\binom{k}{i}S(k-i,j)$, $S(n,k) = \sum_{i=0}^{k}(-1)^{k-i}\binom{k}{i}i^n$ 是第 2 类 Stirling 数[11]. 故利用上述关系,假如预先知道二项矩,则容易给出分布的偏度(记为 $\gamma_1(t)$)与峰度(记为 $\gamma_2(t)$)的计算公式. 事实上,有

$$\gamma_1(t) = \frac{\mu_3(t)}{\mu_2^{3/2}(t)}, \gamma_2(t) = \frac{\mu_4(t)}{\mu_2^2(t)} - 3.$$

2　属性因子及其应用

首先,能够用二项矩来表示普通的统计指标. 事实上,根据噪声强度(记为 NI)与 Fano 因子(记为 FF)的定义,能够显示出

$$NI^2 = (2b_2 + b_1 - b_1^2)/b_1^2,$$
$$FF = (2b_2 + b_1 - b_1^2)/b_1.$$

这里 b_1 和 b_2 分别代表 1 阶二项矩和 2 阶二项矩. 这 2 个统计指标已经广泛用于生化反应系统的随机分析. 类比于噪声强度或 Fano 因子的定义,自然地引入下列定义:

$$AF = 2b_2/b_1^2,$$

它代表 2 阶二项矩的 2 倍与 1 阶二项矩的平方之比,称 AF 为属性因子. 显然,上面定义的 3 个统计量均能够刻画反应物种的随机涨落程度.

为了帮助读者理解上面的定义,下面考察一个最简单的反应系统,即单物种的生灭过程: $\varnothing \xrightarrow{g} X \xrightarrow{d} \varnothing$. 让 n 表示反应物种 X 的分子数目, $P(n;t)$ 为随机变量 X 的概率密度函数,则相应的二项矩方程为

$$\frac{\mathrm{d}b_k}{\mathrm{d}t} = gb_{k-1} - \mathrm{d}kb_k, k = 0,1,2,\cdots.$$

由此,容易求得静态二项矩的分析表达: $b_k = \lambda^k/k!$, 其中 $\lambda = g/d$. 这样,利用重构公式(2)可获得静态分布,记为 $P(n)$, 其分析表达式为 $P(n) = \mathrm{e}^{-\lambda}\lambda^n/n!$, 它是一个 Poisson 分布,其特征参数为 λ. 由此,很容易计算出此分布的 2 阶二项矩和 1 阶二项矩,它们是: $b_1 = \lambda$, $b_2 = \lambda^2/2$. 这样, $NI = 1/\lambda$, $FF = AF = 1$. 回忆起下列事实[12]: 若 $FF < 1$, 则相应的分布是次泊松的; 若 $FF = 1$, 则相应的分布是泊松的; 若 $FF > 1$, 则相应的分布是超泊松的. 故得出: 假如 $AF < 1$, 则相应的分布是超泊松的; 假如 $AF = 1$, 则相应的分布是泊松的; 假如 $AF > 1$, 则相应的分布是超泊松的. 故表明属性因子与 Fano 因子具有同等的功能.

为了显示出属性因子比 Fano 因子具有更大优势,分析另外一个例子. 这一例子是一个 2 状态的基因表达模型[13-15], 其生化反应式包括

$$A \underset{g}{\overset{f}{\rightleftharpoons}} I, A \xrightarrow{\mu} A + X, X \xrightarrow{d} \varnothing, \quad (3)$$

其中 A 代表基因的活性状态, I 代表基因的非活性状态, X 代表基因产物(mRNA 或蛋白质), f 和 g 分别代表基因从活性状态到非活性状态、从非活性状态到活性状态的转移率, μ 代表基因产物的产生率, d 代表基因产物的降解率. 基于式(3)的静态主方程,容易计算出基因产物 X 的 3 个统计量:

$$NI^2 = \frac{1}{\langle n\rangle} + \frac{\langle\tau_{\mathrm{off}}\rangle^2}{(1+\langle\tau_{\mathrm{off}}\rangle)(1+\langle\tau_{\mathrm{on}}\rangle)-1},$$
$$FF = 1 + \frac{\langle\tau_{\mathrm{off}}\rangle^2}{(1+\langle\tau_{\mathrm{off}}\rangle)(1+\langle\tau_{\mathrm{on}}\rangle)-1}\langle n\rangle,$$
$$AF = 1 + \frac{\langle\tau_{\mathrm{off}}\rangle^2}{(1+\langle\tau_{\mathrm{off}}\rangle)(1+\langle\tau_{\mathrm{on}}\rangle)-1},$$

其中 $\langle n\rangle = (\tilde{\mu}\langle\tau_{\mathrm{on}}\rangle)/(\langle\tau_{\mathrm{on}}\rangle + \langle\tau_{\mathrm{off}}\rangle)$ 代表基因产物分子的平均数目(这里 $\tilde{\mu} = \mu/d$), $\langle\tau_{\mathrm{on}}\rangle$ 和 $\langle\tau_{\mathrm{off}}\rangle$ 分别代表基因处在活性和非活性状态的平均时间,即 $\langle\tau_{\mathrm{on}}\rangle = 1/f$, $\langle\tau_{\mathrm{off}}\rangle = 1/g$. 从上面的 3 个显式表达可看出: 1) Fano 因子和属性因子总是大于 1; 2) 转录率越大,则噪声强度越小, Fano 因子越大,但噪声强度属性因子与转录率无关; 3) 假如基因停留在非活性状态非常长,即 $\langle\tau_{\mathrm{off}}\rangle$ 很大,则噪声强度和属性因子均很大,但 Fano 因子趋于 $1 + \mu/(1 + \langle\tau_{\mathrm{on}}\rangle)$; 4) 假如基因停留在活性状态非常长,即 $\langle\tau_{\mathrm{on}}\rangle$ 很大,则噪声强度趋于 $1/\mu$, Fano 因子趋于 1,属性因子也趋于 1. 由文献[16-17]知,属性因子即为基因表达的爆发大小.

3　结论与讨论

本文对概率密度函数引入了二项矩以及对反应物种引进了一种新的统计指标(即属性因子). 由上面的简单例子分析可看出: 二项矩比普通的矩(如原点矩和中心矩)具有更大优势,属性因子也比普通的统计指标(噪声强度和 Fano 因子)在定量化噪声方面具有更大优势.

尽管上面的二项矩和属性因子的定义是在 1 维情形给出的,但这些定义亦容易扩充到高维情形. 例

如,一般地,有

$$b_k(t) = \sum_{N\geqslant k}\binom{N}{k}P(N;t),$$

其中 $\binom{N}{k} = \prod_{i=1}^{n}\binom{N_i}{k_i}$, $N = (N_1,\cdots,N_n)$, $k = (k_1,\cdots,k_n)$. 假如想要刻画某个反应物种的随机涨落程度,则可根据下列公式来计算相应的属性因子:

$$AF^{(i)} = 2b_2^{(i)}/(b_1^{(i)})^2.$$

最后,对于一般的生化反应网络,也能够导出二项矩的时间演化方程. 为此,假如一个反应系统包含 M 个反应物种(记为 X_i)和 J 个反应式. 让 $r\xrightarrow{c_r^s}s$ 代表发生在此反应系统中一个代表性的反应式: $\sum_{i=1}^{M}r_iX_i\xrightarrow{c_r^s}\sum_{i=1}^{M}s_iX_i$,其中化学计量系数是 r_i 和 s_i 为非负整数. 让向量 $N = (N_1,\cdots,N_M)$ 代表整个系统的微观状态,其中 N_i 是物种 X_i 的拷贝数,则相应的二项矩方程为[18]

$$\frac{\mathrm{d}b_k(t)}{\mathrm{d}t} = \sum_{r\to s}(r!)c_r^s\left[\sum_{i=0}^{k}\left(\binom{s}{i}\mathbf{1}^{s-i} - \binom{r}{i}\mathbf{1}^{r-i}\right)\cdot\binom{r+k-i}{r}b_{r+k-i}(t)\right], \quad (4)$$

这里规定:假如 $s = (s_1,\cdots,s_M)$ 中存在某个成分 s_i 小于 $r = (r_1,\cdots,r_M)$ 中相应的成分 r_i,即 $s_i < r_i$,则定义 $\mathbf{1}^{s-i} = 0$. 容易看出:1)方程(4)关于二项矩是一个线性方程组;2)低阶二项矩方程依赖于高阶二项矩,因此为了获得一个封闭系统,需要截取. 至于在什么条件下可以截取,需要进一步讨论,这里就不讨论了.

4 参考文献

[1] Zaikin A N, Zhabotinsky A M. Concentration wave propagation in two-dimensional liquid-phase self-oscillating system [J]. Nature, 1970, 225(5232): 535-537.

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

[3] 周天寿,胡长春. 三类基因振子和它们的基本动力学[J]. 江西师范大学学报:自然科学版, 2008, 32(1): 1-5.

[4] Lotka A J. Contribution to the theory of periodic reaction [J]. J Phys Chem, 1910, 14(3): 271-274.

[5] Gillespie D T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions [J]. J Comput Phys, 1976, 22(4): 403-434.

[6] Munsky B, Khammash M. The finite state projection algorithm for the solution of the chemical master equation [J]. J Chem Phys, 2006, 124(4): 044104.

[7] Ale A, Kirk P, Stumpf M P. A general moment expansion method for stochastic kinetic models [J]. J Chem Phys, 2013, 138(17): 3869-3876.

[8] Smadbeck P, Kaznessis Y N. A closure scheme for chemical master equations [J]. Proc Natl Acad Sci U S A, 2013, 110(35): 14261-14265.

[9] Zechner C, Ruess J, Krenn P, et al. Moment-based inference predicts bimodality in transient gene expression [J]. Proc Natl Acad Sci U S A, 2012, 109(21): 8340-8345.

[10] Grima R. A study of the accuracy of moment-closure approximations for stochastic chemical kinetics [J]. J Chem Phys, 2012, 136(15): 1591-1596.

[11] Balakrishnan N, Johnson N L, Kotz S. A note on relationships between moments, central moments and cumulants from multivariate distributions [J]. Stat Probabil Lett, 1998, 39(1): 49-54.

[12] He Yong, Barkai E. Super- and sub-poissonian photon statistics for single molecule spectroscopy [J]. J Chem Phys, 2004, 122(18): 184703.

[13] Friedman N, Cai Long, Xie X Sunny. Linking stochastic dynamics to population distribution: an analytical framework of gene expression [J]. Phys Rev Lett, 2006, 97(16): 168302.

[14] 周天寿. 基因表达系统的研究进展:概率分布[J]. 江西师范大学学报:自然科学版, 2012, 36(3): 221-229.

[15] Zhang Jiajun, Zhou Tianshou. Promoter-mediated transcriptional dynamics [J]. Biophys J, 2014, 106(2): 479-488.

[16] Singh A, Bokes P. Consequences of mRNA transport on stochastic variability in protein levels [J]. Biophys J, 2012, 103(5): 1087-1096.

[17] Wang Qianliang, Zhou Tianshou. Dynamical analysis of mCAT2 gene models with CTN-RNA nuclear retention [J]. Phys Biol, 2015, 12(1): 016010.

[18] Zhang Jiajun, Huang Lifang, Zhou Tianshou. Comment on 'binomial moment equations for chemical reaction networks' [J]. Phys Rev Lett, 2014, 112(8): 088901.

(下转第 21 页)

Science,2005,309(5743):2075-2078.
[27] Ochab-Marcinek A,Tabaka M. Bimodal gene expression in noncooperative regulatory systems [J]. Proc Natl Acad Sci USA,2010,107(51):22096-22101.

The miRNA Regulation-Induced Random Gain and Bimodal Expression

SHI Changhong
(School of Public Health,Guangzhou Medical University,Guangzhou Guangdong 510275,China)

Abstract: While miRNA often post-transcriptionally regulates gene expression through accelerating the degradation of mRNA,recent studies indicate that miRNA in some cells can regulate the expression of mRNA in a switching manner. Based on this,in this paper a model of gene expression at the transcription level is established,which considers miRNA post-transcriptional regulation of mRNA degradation. By analytical solution to and numerical simulations of the corresponding chemical master equation,the effect of noise in the miRNA regulation process on the mRNA expression is studied both quantitatively and qualitatively. Our analysis shows that this noise can not only raise the expression level of mRNA (such a phenomenon is called as random gain) but also can induce the bimodal expression of mRNA. These results indicate that miRNA post-transcriptional regulation is a mechanism of efficiently controlling gene expression.

Key words: miRNA; post-transcriptional regulation; random gain; bimodal expression; gene model

(责任编辑: 王金莲)

(上接第 4 页)

The Binomial Moments and Attribute Factors for Biochemical Reaction Systems

ZHOU Tianshou
(School of Mathematics and Computational Science,Sun Yet-Sen University,Guangzhou Guangdong 510275,China)

Abstract: Chemical master equations (CMEs) provide a framework for modeling of biochemical reaction systems, but its analysis and simulation are a challenge in computational systems biology. On the other hand,moment-closure methods provide approximations for CMEs but ordinary moments have shortcomings,e. g. ,they do not tend to zero as their orders go to infinity. Binomial moments for a distribution are introduced,which have two remarkable features: 1) binomial moments tend to zero as their orders go to infinity; 2) they can be conveniently used to reconstruction of the corresponding distribution. Based on binomial moments,it further introduces the attribution factor of a reactive species,which has more advantages than common statistical indices such as noise intensity and Fano factor. In addition,it gives explicit formulae for calculating common statistical indices,and uses simple biological examples to show advantages of binomial moments and characteristics of three statistics.

Key words: binomial moment; attribute factor; noise intensity; Fano factor; reaction system

(责任编辑: 王金莲)