文章编号：1006-3080(2025)01-0070-11 DOI：10.14135/j.cnki.1006-3080.20240227001

马尔可夫基因表达建模的神经网络矩闭合方法

顾冬洋， 姜庆超

（华东理工大学能源化工过程智能制造教育部重点实验室，上海 200237）

摘要：生物体的生命活动与基因表达密切相关，然而现有的基因表达矩闭合近似建模方法难以充分利用生化反应过程的潜在细节信息。本文提出了一种基于人工神经网络的矩闭合方法，其利用神经网络学习到矩方程组中高阶矩的低阶表示，使方程组实现闭合，并进一步结合常微分方程求解器对闭合后的方程组进行求解，最终获得精确的矩估计。实验结果表明，所提方法在预测精度和计算效率上显著优于传统方法。

关键词：基因表达建模；神经网络；矩闭合方法；随机模拟；最大熵原理

中图分类号：Q3 **文献标志码：**A

基因表达作为理解生物学现象的核心焦点之一，在生命科学领域的快速发展中占据着重要地位。基因表达是基因通过转录和翻译实现基因功能产物合成的过程，是生物体内调控和执行生命过程的关键步骤，通过细胞调控和执行基因的功能维持生物体的正常功能。了解基因表达机制不仅有助于理解生命的本质，还对揭示疾病发生、发展以及药物研发具有重要意义[1-3]。在生物学中概率主宰生物学，概率在噪声塑造生物系统行为方面起着至关重要的作用[4-7]。这里所述的“噪音”源自于活细胞内分子浓度的固有波动，主要是由生化反应的随机性引起的，尤其在低分子数量的生化反应中表现得尤为明显。因此，对于生化反应网络的建模主要倾向于对单个反应进行模拟来表现反应中分子数的随机波动[8]。基于这一见解，以化学主方程（Chemical Master Equation, CME）为基础的马尔可夫模型等低分子随机建模方法迅速流行[9-10]。同时，随机模拟算法（Stochastic Simulation Algorithm, SSA）也成为了解和获取基因表达动态过程的重要工具[10]。然而 SSA 的计算成本很高，适用性受到严重限制，难以应用于大型系统。

矩闭合近似方法（Moment Closure Approximations, MA）在研究基因表达网络的稳态和极限行为方面取得了重要成就。大多数矩闭合方法主要用于估计分布的矩[11-16]，从而得到关于所有 M 阶及以下联合分布矩的时间演化近似解[11-13]。进一步可以利用最终稳态时刻的近似矩值，使用最大熵原理重构相应的边缘概率分布[17-18]。通过分析系统的矩集合，这类方法能够从全局角度理解基因调控网络的整体行为。然而，传统矩闭合方法仍存在一些挑战和局限性。首先，传统矩闭合方法通常基于线性稳态分析，其在非线性系统中的适用性受限，而许多基因调控网络是非线性的，导致传统矩闭合方法无法充分捕捉基因表达网络潜在过程相互作用的复杂性。其次，传统方法往往局限于特定类型的生化反应网络，难以灵活处理不同生物体和细胞类型之间基因调控机制的差异。此外，在某些复杂生化反应网络中，即使采用更高阶的矩闭合方案，其精度也可能受到闭合方案和生化反应网络潜在物理过程复杂性的限制。总的来说，传统矩闭合方法在适用范围和精度方面存在不足。

人工神经网络（Artificial Neural Network, ANN）通常都是对自然界某种算法或者函数的逼近，也可能是对一种逻辑策略的表达[19]，近年来人工神经网络

收稿日期：2024-02-27

基金项目：国家自然科学基金（62322309）；上海市科技创新行动计划（23S41900500）

作者简介：顾冬洋（1996—），男，河南人，硕士生，主要研究方向为基因表达模型建模。E-mail: gudongyangg@163.com

通信联系人：姜庆超，E-mail: qchjiang@ecust.edu.cn

引用本文：顾冬洋，姜庆超. 马尔可夫基因表达建模的神经网络矩闭合方法 [J]. 华东理工大学学报（自然科学版），2025, 51(1): 70-80.

Citation: GU Dongyang, JIANG Qingchao. Neural Network Moment Closure Method for Markovian Gene Expression Modeling[J]. Journal of East China University of Science and Technology, 2025, 51(1): 70-80.

与其他学科领域联系日益紧密, 在各个领域得到广泛应用, 通过对神经网络层结构的探索和改进来解决不同领域的问题[20]。受此启发, 本文提出了一种基于神经网络的矩闭合方法, 称为神经网络矩闭合(Neural network moment closure)方法。该方法利用人工神经网络学习基因调控网络模型的矩方程组中高阶矩的低阶表示, 将未闭合的矩方程组闭合, 再通过线性常微分方程组求解获得估计的矩值。与传统矩闭合方法相比, 神经网络矩闭合方法无需对系统进行额外分布假设, 更能充分利用生化反应网络模型中的未知潜在特性, 捕捉背后复杂的物理相互作用。一旦神经网络学习到这种潜在相互作用, 说明所提出的方法能够学习到生化反应模型中的物理行为, 使矩闭合结果更加真实可信和准确。神经网络矩闭合方法不仅提供了一种获取矩闭合方法的新途径, 而且弥补了传统方法在生化反应网络系统模型近似中的不足。本研究有望推动基因表达建模领域的发展, 为深入理解基因调控网络的动态行为提供新的视角和方法。

1　预备知识

1.1　随机模拟算法

CME 所描述的随机过程本质上是一个连续时间马尔可夫过程, 其中连续反应事件之间的时间间隔服从指数分布[21]。由于从指数分布中抽样相对简单, 因此模拟生化反应的发生非常便捷且直接。SSA 算法基于概率分布的数值抽样, 可以模拟底层随机过程的精确样本路径, 从而提取准确样本, 是一种在状态空间中生成随机轨迹集合的动力学蒙特卡罗方法。这使得 SSA 能够在分子层面上捕获化学反应的随机性质, 提供精确的分子轨迹, 并且适用于广泛的化学反应网络。

假设一个生化反应网络系统是由 N 个不同的化学反应物 $\{n_1,\cdots,n_N\}$ 和 M 个分别对应反应通道 $\{r_1,\cdots,r_M\}$ 的反应组成。每个反应都有一个倾向函数 $f_r(\boldsymbol{n})$, 反应系统状态用 $\boldsymbol{n}=[n_1(t),\cdots,n_N(t)]^{\mathrm{T}}$ 表示, $n_i(t), i=1,\cdots N$ 表示反应物 n_i 在 t 时刻的分子数, $[\cdot]^{\mathrm{T}}$ 表示向量的转置。直接随机模拟算法的模拟过程如下: 首先对将要发生的反应所需的时间间隔步长 τ 进行采样, 然后对反应集合中的某个具体反应进行采样, 从而确定是哪个反应在什么时间完成[22]。具体而言, $p(\tau|\boldsymbol{n},t)$ 表示下一个反应在 $\tau+t$ 时发生的概率, 并且该反应在一个无限小的时间间隔 dt 内完成; $p(r|\boldsymbol{n},t)$ 表示下一个反应是反应 r 的概率。这两个概率可以通过相应的计算公式从 $f_r(\boldsymbol{n})\mathrm{d}t$ 获得, 如式(1)、(2)所示:

$$p(\tau|\boldsymbol{n},t)=\lambda\exp(-\tau\lambda),\lambda=\sum_{r=1}^{R}f_r(\boldsymbol{n}) \tag{1}$$

$$p(r|\boldsymbol{n},t)=\frac{f_r(\boldsymbol{n})}{\lambda} \tag{2}$$

其中, $f_r(\boldsymbol{n})\mathrm{d}t$ 代表无穷小的时间间隔 dt 内发生第 r 个反应的概率, λ 代表反应系统的总倾向。对这两个公式进行采样可以获得如下两个参数:

$$\tau=-\ln(u_1)/\lambda \tag{3}$$

$$r=\mathrm{SIS}\sum_{i=1}^{r}f_r(\boldsymbol{n})>u_2\lambda \tag{4}$$

其中, u_1 和 u_2 为 0 到 1 之间的均匀随机数, SIS 代表满足公式的最小整数。直接法首先根据式(3)对下一个反应事件的时间点进行采样, 然后根据式(4)对发生某一反应进行采样, 迭代更新随机模拟过程的状态向量和时间。

由于随机模拟算法模拟系统中的每个化学反应事件都是明确的, 即使对于反应物种类较少的系统, 随机模拟算法的计算成本也很高。这种高计算成本的情况在分子数波动很大或单位时间内发生大量反应的情况下尤为明显。在第 1 种情况下, 为了获得统计上准确的结果, 必须模拟大量样本。而在第 2 种情况下, 由于反应事件之间的时间变得更短, 单次模拟的计算成本也变得昂贵。因此, 随机模拟算法的适用性受到严重限制, 并且很快就无法适用于大型系统。为了克服这些挑战, 近几十年来, 研究人员投入了大量精力来发展化学主方程的近似方法, 并出现了多种不同的方法。其中一种称为 Tau跳跃的方法(Tau-leaping)是一种模拟生化反应的近似方法, 它的主要目标是提供比 SSA 更高效的性能[23]。该方法的核心理念在于通过时间上的离散“跳跃”, 跨越多个反应事件, 从而避免了对每个单独反应事件进行模拟的需要。这允许系统在有限的时间段内经历多个反应, 大幅度减少了必须处理的事件总数, 加快了模拟的速度。除了 Tau 跳跃, 还有其他近似方法被提出来, 这些方法的共同目标是高效地近似 CME 的解, 以此降低计算的复杂性和成本。

1.2　近似方法

CME 有很多近似方法, 其中 3 种最常见的近似方法分别是化学朗之万方程(Chemical Langevin Equation, CLE)、系统尺寸展开(System Size Expansion, SSE)和 MA[24-25]。这 3 种方法易于实施, 无需对系统有任何预先的了解, 而且它们通常能够进行高效计

算,并提供精确近似。因此,它们已被成功应用于各种场合[26-30]。然而,这些方法在某些情况下的准确性可能大幅下降,尤其是当某些物种的拷贝数非常低时。如果关注的是过程的矩,CLE 通常比 SSE 和 MA 更为准确。但是,CLE在计算上的代价更高,因为它需要进行大量的随机模拟并集中平均来获取过程的矩,而其他方法只需求解一组有限的常微分方程。此外,当 CLE 定义为实值变量时,在零分子数处会遇到边界问题,实值修正又会引入新的不准确性[31]。通过将 CLE 扩展到复值变量可以解决边界问题,但会降低模拟的效率[32]。因此,如果只对过程的矩感兴趣,使用系统大小扩展或矩闭合近似似乎是更合适的选择。

另一方面,系统尺寸展开是基于小参数的系统扩展,而矩闭合近似是一种特定的近似方法。系统尺寸展开在大系统容量下可以保证准确性,因此在大规模系统下它更具吸引力。对于矩闭合近似,通常不期望能够在所有情况下保持同样的准确度。另外,系统大小扩展不适用于某些确定性具有多稳态的系统,这是矩闭合方法不具有的限制[33]。更进一步地,系统大小扩展仅在均值上高于线性噪声近似两个阶,在协方差上高一个阶[34],系统大小扩展的高阶矩修正比矩闭合方法更难以推导和实现;而矩闭合近似则可以推广到各种阶数[35-36]。CLE、系统大小扩展和矩闭合近似通常作为基础构建模块,为开发高级建模策略提供了框架。比如,有限状态投影算法(Finite State Projection Algorithm, FSP)的思想是将状态空间截断为有限子空间,并使用矩阵幂运算求出该子空间上分布的近似值[37]。鉴于这些因素,选择哪种方法更为合适,将取决于具体问题的细节。

在对比 CLE 和 SSE 的基础上,本文选择聚焦于 MA 中的矩闭合技术。矩闭合方法在操作性上提供了广泛的灵活性,近年来,多领域的专家和学者在人工智能技术的研究和应用中取得了突破性进展[38]。

对于线性反应系统,CME 方程可以在一定的期望阶数上进行数值求解。然而,对于非线性系统,低阶与高阶方程相互耦合,导致矩方程的无限耦合层次,因此不能直接求解。矩闭合方法通过一种近似的方式截断了这个无限阶方程组,常用的矩闭合近似就是通过将所有高于 M 阶的矩表示为低阶矩的函数来闭合矩方程。为了实现这个目标,一种方法是假设系统分布具有特定的函数形式,比如正态分布。这样的假设将 M 阶矩方程与高阶矩解耦,从而得到一组有限的解耦合的常微分方程组。数值求解这组闭合的方程就可以获得所需的矩估计值。这样的矩闭合方法称为“ M 阶矩闭合”。

$$\begin{aligned} y_{i_1,\cdots,i_k} &= \langle n_{i_1},\cdots,n_{i_k}\rangle \\ z_{i_1,\cdots,i_k} &= \begin{cases} \langle (n_{i_1}-y_{i_1}),\cdots,(n_{i_k}-y_{i_k})\rangle, & \text{if } k\geqslant 2 \\ y_{i_1}, & \text{if } k=2 \end{cases} \\ c_{i_1,\cdots,i_k} &= \partial_{s_{i_1}},\cdots,\partial_{s_{i_k}} g(s_1,\cdots,s_N)|_{s_1,\cdots,s_N=0} \end{aligned} \tag{5}$$

上面 3 个公式分别表示 k 阶的原始矩、中心矩和累积量。式(5)中的函数 $g(s)$ 为累积量生成函数,定义为:

$$g(s_1,\cdots,s_N)=\lg\langle\exp(s_1n_1+,\cdots,+s_Nn_N)\rangle \tag{6}$$

一些常见的矩闭合方法如下:

(1)正态分布矩闭合(Normal moment closure)方法(在文献中也称为累积量忽略矩闭合)[39-40]:将 M 阶以上的所有累积量均设为零,即:

$$c_{i_1,\ldots,i_k}=0,\ \text{for}\ k>M \tag{7}$$

值得注意的是,对于系统状态为正态分布的生化反应系统,高于二阶的累积量为零,因此称为“正态分布矩闭合方法”,将二阶和三阶的法向矩闭合近似分别称为“ 2MA ”和“ 3MA ”。

(2)忽略中心矩矩闭合(Central-moment-neglect moment closure)方法(在文献中也称为低色散矩闭合方法)[41]:将所有 M 阶以上的中心矩均设为零:

$$z_{i_1,\ldots,i_k}=0,\ \text{for}\ k>M \tag{8}$$

(3)微分匹配(Derivative matching):这种方法的核心思想是使用低阶矩来表示高于 M 阶的矩,以使闭合系统的矩的时间导数近似等于精确系统在某个初始时间点上 M 阶及以下的矩的时间导数。在文献 [42] 中,提出了一种具体的方法来生成相应的表达式。

2 神经网络获取矩闭合方法过程

本文提出的神经网络矩闭合方法的核心是假设有限数量的矩能够捕捉到所有必要的系统信息,通过神经网络学习到生化反应系统未闭合的矩方程组中高阶矩的低阶矩表示函数,就可以将矩方程组闭合,随后通过解闭合的微分方程组来获取矩估计值。

图 1 示出了整个实验流程。实验首先要构造所需的特定生化反应模型和输入数据集。虽然流程图中描绘的是一个基因调控网络模型,但方法同样适用于构建更广泛类型的生化反应模型。针对研究需要的生化反应模型,需要生成大量的随机参数组 $\boldsymbol{x}(i)$ 作为模型的输入,其中每个参数组代表生化反应模型的不同倾向函数的反应过程。为了让神经网络能

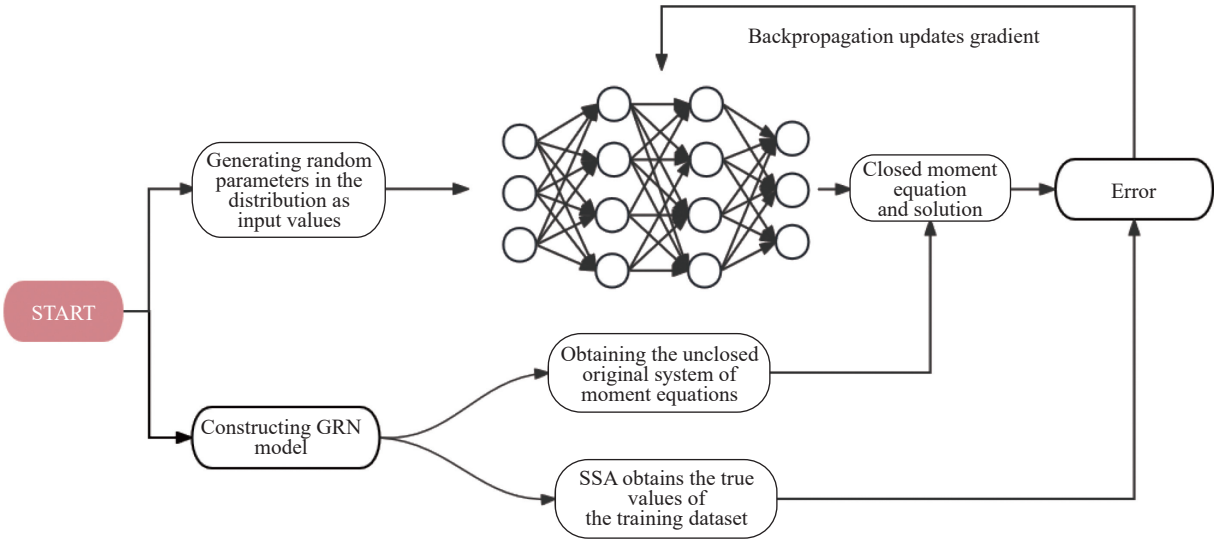

图 1　实验流程图

Fig. 1　Diagram of experimental workflow

够捕获生化反应模型的底层特性, 需要足够数量具有广泛性和代表性的倾向函数随机参数集。这些参数集的数量和范围可能需要根据实验结果进行进一步调整。

利用生成的有效数据集, 一方面, 需要获取生化反应系统的原始未闭合矩方程组(Raw moment equations)。这些方程组将运用神经网络学习到的矩闭合方案并求解闭合方程。另一方面, 运用 SSA 随机模拟并进行集中平均, 以获得生化反应模型的矩真实值 $\boldsymbol{y}(i)$, 此值将作为神经网络训练数据集的参考真值。神经网络的输出 $\boldsymbol{v}(i)$ 是高阶矩的低阶矩表示, 为了实现这一点, 需要针对不同生化反应网络构造不同的向量表达方式。将神经网络的输出 $\boldsymbol{v}(i)$ 代入到原始矩方程组中, 即可成功实现方程组的闭合, 这为常微分方程组的求解提供了便利, 进而获得了矩的估计值 $\hat{\boldsymbol{y}}(i)$。通过将求解得到的矩估计值与 SSA 得到的矩真实值进行比较, 得到模型的偏差, 利用偏差对神经网络进行反向传播更新梯度值, 直至满足预期的性能指标。

神经网络的训练过程遵循标准的训练算法, 如算法 2.1 所示。

算法 2.1　神经网络训练算法

1　加载数据集并归一化处理;
2　设置学习率 $\alpha = 0.1$, 正则化系数;
3　随机初始化神经网络权重和偏差 W, b;
4　repeat
5　　训练集样本进行随机排序;
6　　for $n \in$ train set do
7　　正向传播得到神经网络输出 $\boldsymbol{v}(i)$;
8　　闭合矩方程组, 使用常微分方程求解得到估计矩值 $\hat{\boldsymbol{y}}(i)$, 并求出目标函数;
9　　反向传播, 计算每一层的误差和导数;
10　　更新网络参数;
11　　end for
12　until 神经网络在测试集上的错误率不再下降
13　输出神经网络模型的参数 W, b

值得注意的是, 经过一轮训练后, 根据神经网络学习到的矩闭合效果, 可能需要对参数进行调整, 或者对网络结构进行优化, 以实现更精确的估算结果。

3　实验结果分析

3.1　基因调控网络模型及数据集介绍

3.1.1　基因调控网络模型　本文实验对象采用的是生化反应中极具代表性的基因调控网络(Gene Regulatory Network, GRN)模型。这种反应网络模型是一个用于描述细胞内或一个特定基因组内基因间相互作用的抽象模型, 在众多相互作用关系之中, 侧重于基因调控机制的相互作用。基因调控网络是生物体内控制基因表达的关键机制, 它涉及基因的转录和信使核糖核酸(mRNA)的翻译过程。图 2 示出了 GRN 模型示意图[43]。

在此模型中, 基因存在于“活跃” G 和“不活跃” G^* 两种状态, 并且可以通过启动子实现状态转换, 切换速率分别为 σ_u 和 σ_b, 每种状态下基因的蛋白质产生速率不同, 分别对应 ρ_u 和 ρ_b。基因从 G 到 G^* 的状态切换是通过蛋白质分子 P 与基因 G 的结合实现, 从 G^* 到 G 的状态切换是自发的, 并伴随产生一

个蛋白质分子 P。此外，蛋白质分子 P 以固定的倾向函数速率 1 进行降解。该模型可以视为一个基本的反馈环路：如果 $\rho_u > \rho_b$，则形成负反馈，抑制自身的蛋白质表达；反之，则构成正反馈环，激励自身的蛋白质表达。为简化模型，这里没有明确地建模 mRNA，但这不会对实验结论产生影响。

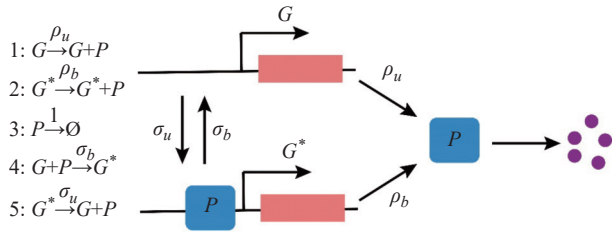

图 2 调控网络模型示意图

Fig. 2 Schematic of regulatory network model

3.1.2 基因调控网络数据集 为了实施图 1 所描述的基因调控网络模型的神经网络学习矩闭合方法，需要构建数据集 $D=\{\boldsymbol{X}(i),\boldsymbol{Y}(i)\},i=1,\cdots,M$，其中 $\boldsymbol{x}(i)=[\rho_b,\rho_u,\sigma_b,\sigma_u]$ 是神经网络的输入，即反应方程组的倾向函数组成的向量，$\boldsymbol{y}(i)$ 是模型经过 30 000 次 SSA 随机模拟并进行集中平均得到的精确矩值。由于本模型是双变量，所以用 μ_{ij} 分别代表基因和蛋白质 i 阶和 j 阶时的矩值，针对本文的模型将 $\boldsymbol{y}(i)$ 表示成 $\boldsymbol{y}(i)=[\mu_{10},\mu_{01},\mu_{11},\mu_{02},\mu_{12},\mu_{03}]$。数据集大小 M 设置为 4000 个，然后按照 9∶1 划分为训练集和测试集。

数据集生成方式如下：(1) 首先从特定的正态分布中生成 4 个参数，构成参数向量 $\boldsymbol{x}(i)=[\rho_b,\rho_u,\sigma_b,\sigma_u]$。具体地，分别从期望 75、方差 10 的正态分布中生成 ρ_b，从期望 15、方差为 5 的正态分布中生成 ρ_u，从期望 0.2、方差为 0.05 的正态分布中生成 σ_b 和 σ_u。如果生成的参数值小于等于 0，则将其丢弃并重新生成。重复这个过程，直到生成 $M=4\,000$ 组有效的参数向量，表示为 $\boldsymbol{X}(i)=[\boldsymbol{x}(1),\cdots,\boldsymbol{x}(M)]$。(2) 对于每一组参数 $\boldsymbol{x}(i)=[\rho_b,\rho_u,\sigma_b,\sigma_u],\ i=1,\cdots,M$，进行 30 000 次 SSA 随机模拟，并对模拟结果进行集合平均，得到该组参数的精确矩值 $\boldsymbol{y}(i)=[\mu_{10},\mu_{01},\mu_{11},\mu_{02},\mu_{12},\mu_{03}]$。对于 4000 组参数，可以得到精确矩值的集合 $\boldsymbol{Y}(i)=[\boldsymbol{y}(1),\cdots,\boldsymbol{y}(M)]$。(3) 将生成的 4000 组神经网络输入参数 $\boldsymbol{X}(i)$ 和对应的精确矩值 $\boldsymbol{Y}(i)$ 合并，形成最终的数据集 $D=\{\boldsymbol{X}(i),\boldsymbol{Y}(i)\},i=1,\cdots,M$。通过以上步骤，就可以构建一个含有输入参数和相应精确矩值的数据集，该数据集用于神经网络的训练过程。

3.1.3 数据集分析 为了确保后续实验结果的可靠性和有效性，本文从生成的数据集中选择 4 个不同的参数组合，并采用图形方式呈现了这些参数组合所对应的概率分布实例，如图 3 所示。图 3(a)描绘了 4 组参数通过 30 000 次随机模拟得到的系统状态记录，并对蛋白质数量随时间变化进行了平均处理的结果。从图中可以清晰看到，随着模拟时间的推移，蛋白质数量呈现出稳定趋势，这一现象表明所选用的数据集在随机模拟过程中已经达到了稳态。图 3(b)进一步示出了在基因调控网络模型达到稳态后，4 组参数下蛋白质数量的概率分布。通过这些概率分布曲线可以观察到蛋白质数量稳态均值覆盖了 50~90 的范围。

图 3 中展现的趋势和分布情况不仅揭示了蛋白质数量随时间的动态演变，而且也体现了在达到稳态时各个状态的概率分布。通过分析，可以确认数据集中的矩闭合值是在稳态条件下计算的，这一点对于验证数据集的精确性至关重要。此外，还可以观察到数据集具有广泛的代表性，这种特性对于保障数据集在模拟各类生化反应网络时的通用性和适用性极为关键，确保了模拟实验结果的稳定性和可重复性。通过选取覆盖多种可能情境的不同参数组合，确保数据集能够覆盖大范围的数据空间，这进一步证明所选数据集在适用性和可靠性方面的优势。

需要注意的是，本文所采用的基因调控网络模型，虽然是一种简化的抽象表达形式，它对于理解更为复杂的生化反应系统的动态行为提供了初始的框架。然而，对于那些对高度复杂生物过程的建模感兴趣的研究者来说，使用生成的模拟数据集之前，对其可信度进行细致的评估是必不可少的。为了确保所生成的模拟数据集能够准确地反映真实世界的数据特性，需要使用一系列细致的量化指标和对比分析方法：

(1) 统计一致性：包括对模拟数据集与真实数据集的平均值、中位数、方差等核心描述性统计指标进行比较，并利用 Kolmogorov-Smirnov 检验和 Q-Q 图等方式来详细对比数据分布的相似度；

(2) 时间序列分析：分析模拟数据集和真实数据集分子数量随时间变化的行为模式，确保模拟数据能够精确地再现真实生物系统的动态特性；

(3) 再现性测试：对于每组参数多次运行模拟过程，并检查结果的再现性和变异性，有助于验证模拟过程的稳定性。

在实际实施中，需要充分考虑到研究目的的具体性和所使用数据集的独特性质，以便选取最适合的评估工具和方法。

3.2 神经网络训练结果

本文构建的人工神经网络旨在学习基因调控网

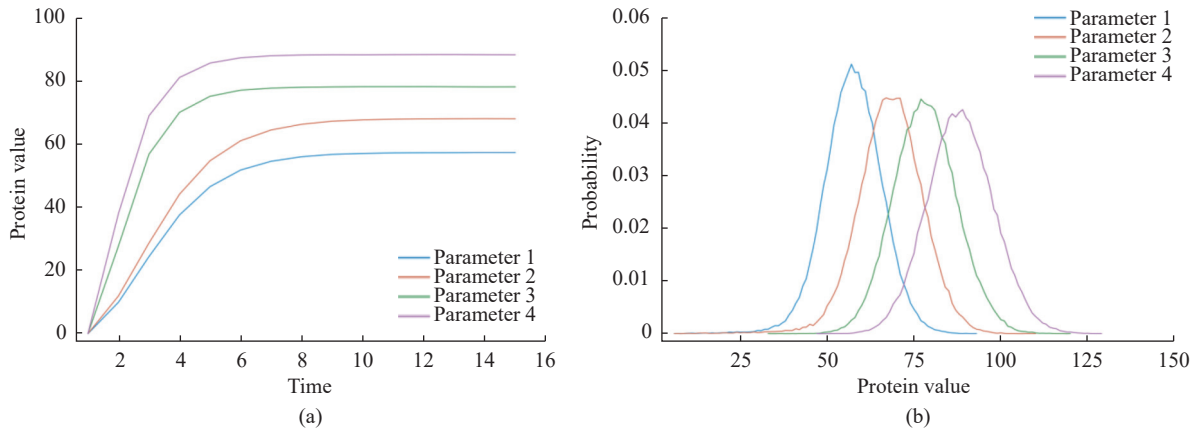

图 3　数据集分布示意图

Fig. 3　Schematic diagram of dataset

络模型中的内在反应特性, 因此神经网络设计相对灵活, 允许多种修改和实验, 只要能够有效捕捉生化反应模型的关键特征即可。具体而言, 针对本文的研究对象所构建的神经网络包括: (1)一个由 4 个神经元组成的输入层; (2)两个隐藏层, 每层各含 10 个神经元; (3)包含 7 个神经元的输出层。网络中输入层与隐藏层之间采用 ReLU 函数作为激活函数。在训练过程中, 采用 ADAM 优化器推荐的标准对学习率进行自适应调整。针对不同的反应网络需要构建不同的神经网络的输出层, 如下所示:

$$\hat{\mu}_{13} = b\mu_{10}^{v_{10}}\mu_{01}^{v_{01}}\mu_{11}^{v_{11}}\mu_{02}^{v_{02}}\mu_{12}^{v_{12}}\mu_{03}^{v_{03}} \tag{9}$$

将式(9)代入到未闭合方程组中三阶及以下部分, 就成功实现了方程组的闭合。对闭合后的方程组进行常微分求解得到最终的矩估计值 $\hat{\boldsymbol{y}}(i) = [\hat{\mu}_{10},\hat{\mu}_{01},\hat{\mu}_{11},\hat{\mu}_{02},\hat{\mu}_{12},\hat{\mu}_{03}]$。有了矩估计值 $\hat{\boldsymbol{y}}(i)$ 和数据集中的矩真实值 $\boldsymbol{y}(i)$, 就可以设置目标函数, 目标函数旨在最小化估计值和真实值之间的差异, 具体形式如下:

$$\begin{aligned} L =& \lg(\hat{\boldsymbol{y}}_i/\boldsymbol{y}_i) + (v_{10}+v_{11}+v_{12}-1)^2 + \\ & (v_{01}+v_{11}+2v_{02}+2v_{12}+3v_{03}-3)^2 \end{aligned} \tag{10}$$

自定义的损失函数可能需要根据训练效果的不同进行权重调整和优化, 以便更精细地学习生化反应的潜在细节。

神经网络的训练使用标准反向传播算法来进行权重更新和训练。为了衡量训练的有效性, 本文追踪了损失函数的变化, 并通过训练周期的演进来评估模型性能(图 4)。如图 4 所示, 损失函数在训练初期迅速下降, 表明模型从初始状态迅速学习并调整参数以最小化损失。随着训练的深入, 损失函数下降的速度减慢, 并最终趋于稳定。定义成功的收敛标准为, 若损失函数在连续 20 个训练周期内保持在一个特定的范围内波动, 便认为模型已经稳定学习到了数据的特征。在本实验中, 损失函数在后续 30 个周期内保持稳定, 由此可以判断模型已经成功收敛。

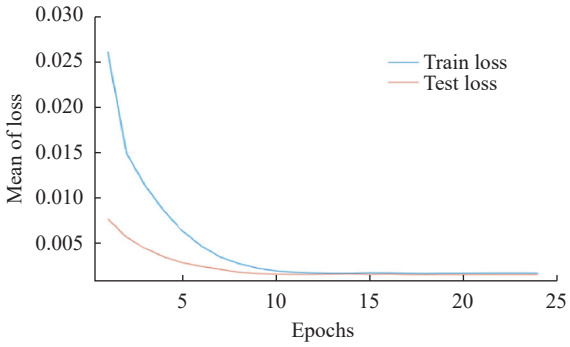

图 4　训练过程 loss 图

Fig. 4　Training process loss

3.3　结果准确性

由于本文实验采用的基因调控网络模型最终得出 6 个矩估计值, 因此评估结果也集中在这 6 个矩值上。图 5 示出了估计值的不同方法箱型图。图 5 中的箱型图对比了基于神经网络的矩闭合方法、SSA 和传统矩闭合方法在所考虑的基因调控网络模型中的准确度表现。图中的 SSA 方法表示模型经过 2000 次 SSA 随机模拟到达稳态后计算出的三阶矩以下矩值, 低数量模拟的 SSA 方法由于其固有的随机性, 准确度会受到部分限制。图中的“Normal”和“DM”分布代表传统矩闭合方法, 分别对应于第 1.2 节中的正态分布矩闭合方法和微分匹配矩闭合方法。

从图 5 中的结果来看, 神经网络矩闭合方法在准确性方面明显超越了低数量 SSA 模拟计算得到的矩估计值。尽管这是基于较少数量的随机模拟得出的结论, 但依然能展示神经网络矩闭合方法的相对准确性, 从侧面说明了 SSA 方法在获得精确的矩估

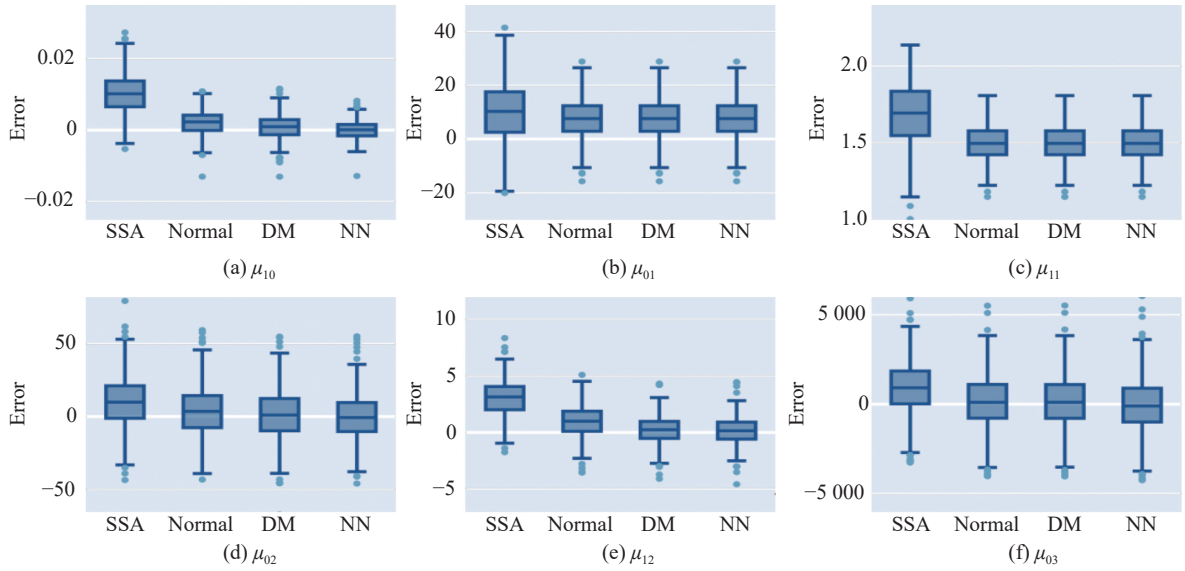

图5 估计值的不同方法箱型图

Fig. 5 Boxplot of moment estimates from different methods

计值时需要进行大量的计算平均,而这正是矩闭合方法的价值所在,它显著减少了计算量的需求。从图中还可以看到,神经网络矩闭合方法在 μ_{10} 、μ_{02} 、μ_{12} 、μ_{03} 这几个矩估计值上表现得优于传统矩闭合方法,直接证明了神经网络矩闭合方法在准确度方面相比于传统的矩闭合方法在基因调控网络模型具有显著优势。

R^2 是一个统计指标,用于衡量观测数据与拟合模型之间的吻合程度,取值范围从0到1,越接近1表示模型与观测数据的拟合度越高。图6示出了神经网络矩闭合方法得到的矩估计值的 R^2 拟合图,突显了这些矩值之间的高度相关性,以进一步验证本文方法在基因调控网络模型中的可靠性。从图中可以清晰地看出,每个矩值的 R^2 拟合值都接近1,表明神经网络矩闭合方法能够有效地捕捉到这些矩之间的紧密关联,进一步说明了神经网络矩闭合方法在揭示基因调控网络模型中生化反应动态内在规律性的能力。

神经网络矩闭合方法在灵活性上优于传统矩闭合技术,特别是在满足精度要求的可调整性方面。研究者不仅可以针对整体模型精度进行优化,还能够对特定参数进行细致的调校,这一切均通过修改训练阶段目标函数(参考式(10))中的权重实现,或者可以在目标函数中添加额外感兴趣的项以进一步细化。

3.4 结果快速性

表1所示为神经网络矩闭合方法与其他一些算法单次获得矩闭合估计值所需的平均计算时间对比结果。具体来说,对于数据集中一组数据,SSA方法和Tau-leaping方法的时间消耗包括了随机模拟过程和集合平均获取矩值;传统矩闭合方法时间消耗包括获取矩方程组、利用传统公式闭合矩方程组和求解闭合方程组获得矩估计;FSP方法包括计算系统的概率密度向量和计算矩值;而神经网络矩闭合方法的时间消耗则包括获取矩方程组、训练神经网络、利用神经网络输出闭合矩方程组合求解闭合方程组获得矩估计。平均计算时间基于本文4000组参数的数据集得出,该时间反映了求得最终矩估计值所需的平均时长。SSA方法,使用的是3.2节中选择的10000次模拟并作为真值的数据。Tau-leaping方法和SSA相同,也是进行了10000次模拟并集合平均。对于传统矩闭合方法,表中平均计算时间为正态分布矩闭合和微分匹配矩闭合两种方法的平均计算时间。

由结果清楚地显示,神经网络矩闭合方法在计算速度上明显优于SSA方法,并且随着生化反应模型复杂性的提升和模拟规模的扩大,这种速度优势将非常显著。与评估中的其他3种方法相比,神经网络矩闭合方法同样展现出了速度上的优越性。这强调了在进行复杂生化反应模拟时,利用神经网络进行矩闭合近似作为提高计算效率的有力工具,尤其在传统算法难以承受高计算负荷时更显其价值。图中神经网络矩闭合方法虽然在表中仅展示了整体的平均计算速度,但神经网络矩闭合方法中最耗时的环节预计为网络训练过程。后续分析将进一步探究数据量的增加对神经网络训练时间的影响。

图7示出了随着数据集样本量的增加,SSA、传统矩闭合方法和神经网络矩闭合方法在获得矩闭合

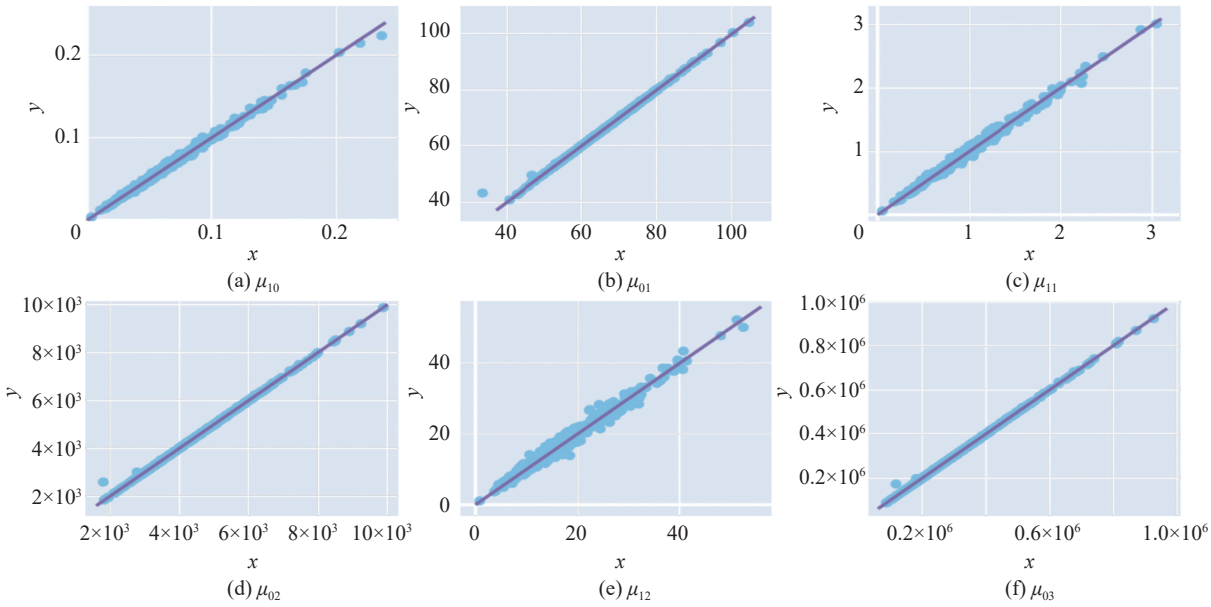

图 6　矩估计值的 *R*-squared 拟合图

Fig. 6　*R*-squared fitting chart for moment estimaties

表 1　计算时间对比

Table 1　Comparison of computation time

Method	Average computation time/s
SSA	2.588
Tau-leaping	1.146
Traditional moment closure	0.971
FSP	0.521
Neural network moment closure	0.479

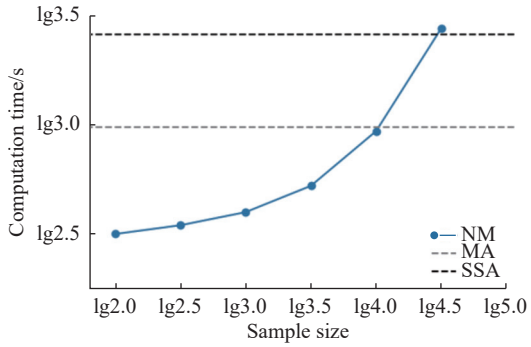

图 7　SSA 与 NN 计算时间对比

Fig. 7　Comparison of SSA and NN computation time

估计值时所需的平均计算时间的变化。对于 SSA 和传统矩闭合方法, 由于它们在获取矩值时采用了固定的实现途径, 因此这两种方法的平均计算时间保持不变, 不受数据集规模影响。这一点可以从图中的黑色虚线和浅灰色虚线观察得到。神经网络矩闭合方法的平均计算时间随着数据集样本量的增加而提升, 这是因为数据集规模的扩大导致了更长的网络训练时间。值得强调的是, 在数据集样本量为 1 000 时, 神经网络矩闭合方法已能达到 SSA 在进行 30 000 次随机模拟后的集合平均矩值精度。从图中可以明显看出, SSA 所需的计算时间大约是神经网络矩闭合方法的 6 倍, 而传统矩闭合方法所需时间则约为神经网络方法的两倍半。因此, 相较于 SSA 和传统矩闭合方法, 神经网络矩闭合方法在计算效率上具有显著优势。

这种计算效率的显著提升主要归功于神经网络矩闭合方法继承并强化了传统矩闭合方法在近似建模领域的优势, 同时规避了 SSA 在执行大规模随机模拟并集合平均过程中所固有的高计算需求。随着生化反应系统规模的扩张, SSA 的计算负担将急剧增加, 而矩闭合方法所需的计算资源几乎不受影响。此外, 矩闭合技术在求解微分方程组时能够运用先进的时间步长优化技术, 根据反应动力学的实际特性动态调整求解步长, 由此节约了不必要的计算资源。最关键的是, 神经网络矩闭合方法通过神经网络的学习能力, 实现了对高阶矩方程组中高阶矩的低阶近似表达, 在大量模拟的情况下有效避免了直接计算复杂高阶矩的需求。如果研究者需要对时间效率有极端的要求, 迫切需要快速执行大规模模拟时, 可以牺牲精度提升时间效率。通过选用较小的数据集合或限制迭代次数, 可以大幅缩短神经网络训练所需的时间。尽管这样做可能会影响结果的精细度, 但在特定的实验环境中, 这种方法仍能有效地满足对快速处理的需求。

4 结束语

在基因调控网络建模过程中,随机模拟算法在获取矩值时需进行大量的随机模拟并集合平均,导致计算量庞大和复杂性增加。而依赖于简化假设的传统矩闭合方法则无法充分描绘真实系统的复杂性,不能有效捕捉大量相互作用的生化反应模型系统的物理细节。因此,本文提出了一种新颖的神经网络矩闭合方法,它通过在整个生化反应网络中探索潜在关联,能够更全面地捕捉生化反应模型中的动态行为。实验证明,相较于传统方法,神经网络矩闭合方法在对基因表达模型的预测精度和时间效率上都表现出一定的优势,为基因表达建模研究提供了一种更准确和高效的分析工具。

尽管神经网络矩闭合方法在生化反应建模方面取得了显著的进展,但也存在着挑战和改进的空间。本文的实验验证主要局限于特定的基因调控网络模型,因此该方法在遇到未知情境时的泛化能力可能不足。此外,尽管本文在方法验证阶段使用的是模拟数据集,但与实际生物实验数据的结合是提升方法可靠性和应用实用性的关键。未来的研究应当着重于将神经网络矩闭合方法应用于更为广泛的生化反应模型,并提升模型可解释性,以改善用户对预测决策的理解。同时,与更多的反应类型的结合也将是增强方法鲁棒性和验证可行性的重要步骤。总而言之,通过解决现有问题并成功地将研究前景转化为实际成果,神经网络矩闭合方法有望在生化反应建模领域实现更广泛的应用。

参考文献:

[1] CHEN J L, YANG L, WANG Q, *et al*. Helix-sense-selective and enantiomer-selective living polymerization of phenyl isocyanide induced by reusable chiral lactide using achiral palladium initiator[J]. Macromolecules, 2015, 48(21): 7737-7746.

[2] WAN Y M, MU Q, KRZYSZTOŃ R, *et al*. Adaptive DNA amplification of synthetic gene circuit opens a way to overcome cancer chemoresistance[J]. Proceedings of the National Academy of Sciences, 2023, 120(49): e2303114120.

[3] 隋馨莹, 徐平, 段昌柱, 等. p62 蛋白的分子功能及其在疾病中的研究进展 [J]. 生物工程学报, 2023, 39(4): 1374-1389.

[4] 刘晓, 张学博, 陈大明, 等. 2022 年合成生物学发展态势 [J]. 生命科学, 2023, 35(1): 63-71.

[5] MCADAMS H H, ARKIN A. Stochastic mechanisms in gene expression[J]. Proceedings of the National Academy of Sciences, 1997, 94(3): 814-819.

[6] SWAIN P S, ELOWITZ M B, SIGGIA E D. Intrinsic and extrinsic contributions to stochasticity in gene expression[J]. Proceedings of the National Academy of Sciences, 2002, 99(20): 12795-12800.

[7] 李岩, 张绍武. 生物分子网络弹性研究进展 [J]. 生物化学与生物物理进展, 2022, 49(10): 1987-2000.

[8] PERKINS T J, SWAIN P S. Strategies for cellular decision-making[J]. Molecular Systems Biology, 2009, 5(1): 326.

[9] 周天寿, 唐云. 分子系统生物学的数学建模与分析 [J]. 数学建模及其应用, 2017, 6(1): 1-12.

[10] WILKINSON D J. Stochastic Modelling for Systems Biology[M]. [s.l.]: Chapman & Hall, 2006.

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

[12] ELF J, EHRENBERG M. Fast evaluation of fluctuations in biochemical networks with the linear noise approximation[J]. Genome Research, 2003, 13(11): 2475-2484.

[13] GRIMA R. An effective rate equation approach to reaction kinetics in small volumes: Theory and application to biochemical reactions in nonequilibrium steady-state conditions[J]. The Journal of Chemical Physics, 2010, 133(3): 035101.

[14] THOMAS P, MATUSCHEK H, GRIMA R. How reliable is the linear noise approximation of gene regulatory networks[J]. BMC Genomics, 2013, 14: 1-15.

[15] ENGBLOM S. Computing the moments of high dimensional solutions of the master equation[J]. Applied Mathematics and Computation, 2006, 180(2): 498-515.

[16] GILLESPIE C S. Moment-closure approximations for mass-action models[J]. IET Systems Biology, 2009, 3(1): 52-58.

[17] ALE A, KIRK P, STUMPF M P. A general moment expansion method for stochastic kinetic models[J]. The Journal of Chemical Physics, 2013, 138(17): 174101.

[18] ANDREYCHENKO A, MIKEEV L, WOLF V. Model reconstruction for moment-based stochastic chemical kinetics[J]. ACM Transactions on Modeling and Computer Simulation, 2015, 25(2): 1-19.

[19] ANDREYCHENKO A, MIKEEV L, WOLF V. Reconstruction of multimodal distributions for hybrid moment-based chemical kinetics[J]. Journal of Coupled Systems and Multiscale Dynamics, 2015, 3(2): 156-163.

[20] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[21] 孙志军, 薛磊, 许阳明, 等. 深度学习研究综述 [J]. 计算机应用研究, 2012, 29(8): 2806-2810.

[22] GILLESPIE D T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics, 1976, 22(4): 403-434.

[23] GARDINER C W. Handbook of Stochastic Methods[M]. Berlin: Springer, 1985.

[24] GILLESPIE D T. Approximate accelerated stochastic simulation of chemically reacting systems[J]. The Journal of Chemical Physics, 2001, 115(4): 1716-1733.

[25] YIN S, DING S, XIE X, *et al*. A review on basic data-driven approaches for industrial process monitoring(Review)[J]. IEEE Transactions on Industrial Electronics, 2014, 11: 6414-6428.

[26] KAMPEN V, GODFRIED N. Stochastic Processes in Physics and Chemistry[M]. Third edition. Amsterdam: Elsevier, 1992.

[27] LIAO S, VEJCHODSKÝ T, ERBAN R. Tensor methods for parameter estimation and bifurcation analysis of stochastic reaction networks[J]. Journal of The Royal Society Interface, 2015, 12(108): 20150233.

[28] MCKANE A J, NAGY J D, NEWMAN T J, *et al*. Amplified biochemical oscillations in cellular systems[J]. Journal of Statistical Physics, 2007, 128: 165-191.

[29] 姜特, 陈志刚, 万永菁. 基于注意力机制的多任务 3D CNN-BLSTM 情感语音识别 [J]. 华东理工大学学报 (自然科学版), 2022, 48(4): 534-542.

[30] HEAVNER B D, SMALLBONE K, BARKER B, *et al*. Yeast 5: An expanded reconstruction of the Saccharomyces cerevisiae metabolic network[J]. BMC Systems Biology, 2012, 6: 55.

[31] CHALLENGER J D, MCKANE A J. Synchronization of stochastic oscillators in biochemical systems[J]. Physical Review E, 2013, 88(1): 012107.

[32] GRIMA R. A study of the accuracy of moment-closure approximations for stochastic chemical kinetics[J]. The Journal of Chemical Physics, 2012, 136(15): 154105.

[33] SCHNOERR D, SANGUINETTI G, GRIMA R. The complex chemical Langevin equation[J]. The Journal of Chemical Physics, 2014, 141(2): 024103.

[34] BIANCALANI T, DYSON L, MCKANE A J. Noise-induced bistable states and their mean switching time in foraging colonies[J]. Physical Review Letters, 2014, 112(3): 038101.

[35] THOMAS P, MATUSCHEK H, GRIMA R. Intrinsic noise analyzer: A software package for the exploration of stochastic biochemical kinetics using the system size expansion[J]. PloS One, 2012, 7(6): e38518.

[36] KAZEROONIAN A, FRÖHLICH F, RAUE A, *et al*. CERENA: Chemical reaction network analyzer: A toolbox for the simulation and analysis of stochastic chemical kinetics[J]. PloS One, 2016, 11(1): e0146732.

[37] SCHNOERR D, SANGUINETTI G, GRIMA R. Comparison of different moment-closure approximations for stochastic chemical kinetics[J]. The Journal of Chemical Physics, 2015, 143(18): 185101.

[38] MUNSKY B, KHAMMASH M. The finite state projection algorithm for the solution of the chemical master equation[J]. The Journal of Chemical Physics, 2006, 124(4): 044104.

[39] FAN S, GEISSMANN Q, LAKATOS E, *et al*. MEANS: Python package for moment expansion approximation, iNference and Simulation[J]. Bioinformatics, 2016, 32(18): 2863-2865.

[40] GOODMAN L A. Population growth of the sexes [J]. Biometrics, 1953, 9(2): 212-225.

[41] WHITTLE P. On the use of the normal approximation in the treatment of stochastic processes[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1957, 19(2): 268-281.

[42] HESPANHA J. Moment closure for biochemical networks [C]//3rd International Symposium on Communications, Control and Signal Processing. Saint Julian's, Malta: IEEE, 2008: 142-147.

[43] SINGH A, HESPANHA J P. Lognormal moment closures for biochemical reactions [C]//Proceedings of the 45th IEEE Conference on Decision and Control. San Diego, CA, USA: IEEE, 2006: 2063-2068.

Neural Network Moment Closure Method for Markovian Gene Expression Modeling

GU Dongyang, JIANG Qingchao

(Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China)

Abstract: Gene expression is pivotal in numerous biological processes, making the comprehension and analysis of its modeling critically important. The gene regulatory network modeling process often involves stochastic simulation algorithms, which necessitate extensive random simulations and ensemble averaging to determine moment values. This results in a considerable computational burden and added intricacy. Traditional moment closure approximations, based on oversimplified distribution assumptions, fall short in capturing the intricate nature of real-world systems and fail to accurately represent the nuances of biochemical reaction models with extensive interactions. Such methods typically neglect the full spectrum of possibilities inherent in biochemical reactions, characterized by complex interplays among numerous components.To overcome these obstacles, this study exploits the exceptional capabilities of artificial neural networks for regression analysis and introduces a novel moment closure approximation for gene regulatory networks that harnesses these networks. This innovative method employs neural networks to infer low-order moments representations of higher-order moments, subsequently utilizing ordinary differential equation solvers to compute the predicted moment values. This approach effectively resolves the limitations of traditional moment closure approximations, which do not adequately leverage the intricate details present in biochemical reaction models.The research utilizes simulated datasets, meticulously validated for integrity and reliability. A comparative analysis of the moment values predicted by the neural network-based method against those derived from traditional approaches demonstrates a marked increase in precision with the neural network method. Furthermore, when assessing computational time across varying sample data, the neural network moment closure method is shown to outperform both traditional moment closure and stochastic simulation algorithms in terms of efficiency.To summarize, the enhanced precision and computational efficiency of the neural network moment closure method not only underscore its validity but also introduce an innovative tool and methodology for advancing gene regulatory network research.

Key words: gene expression modeling; neural networks; moment closure method; stochastic simulation; principle of maximum entropy

（责任编辑：李娟）