Abstract

基因网络是理解细胞调控反应的关键，是单细胞转录动力学实验观察的基础。虽然基因网络的信息编码在 RNA 表达数据中，但现有的计算框架目前无法从此类数据推断基因网络。相反，基因网络（由基因状态、其连通性和相关参数组成）目前是通过在学习相关速率参数之前预先指定基因状态数和连通性来推断的。因此，无法独立评估基因网络的正确性，这可能导致强烈的偏差。相比之下，我们在这里提出了一种方法，可以同时自洽地从单分子水平的 RNA 计数中学习基因状态、状态连通性和相关速率参数的完整分布。值得注意的是，我们的方法通过将网络本身视为随机变量，将源自波动的 RNA 计数的噪声传播到由数据保证的网络上。我们通过在贝叶斯非参数范式内操作来实现这一点。我们在大肠杆菌细胞中的 lacZ 通路、酿酒酵母细胞中的 STL1 通路上证明了我们的方法，并在合成数据上验证了其稳健性。

1 Introduction

RNA 动力学的定量测量在固定细胞群体和单个活细胞中一致地揭示了 RNA 计数和行为在空间、时间以及个体细胞中的复杂分布，甚至对于克隆细胞群体也是如此 1。广义上，“基因表达变异性”这一术语被用来解释这些普遍存在的、变异的和复杂的 RNA 表达分布。单细胞生物学的一个核心目标是理解基因表达变异性的分子来源及其下游后果。例如，最近的研究表明，某些稀有细胞，仅通过与克隆姐妹细胞相比，RNA 含量的瞬时波动可以驱动耐药性癌症或维持驱动发育的祖细胞 2,3,4。尽管实验方法能够识别稀有细胞，但从细胞间离散的 RNA 计数中稳健地确定基因网络仍然是一个开放的问题。

图 1 展示了通过基因状态数量、状态连接性和相关速率参数定义的基因网络示例。

在提供离散 RNA 计数的现有实验方法中，我们专注于单分子 RNA 荧光原位杂交（smFISH） 5,6。具体而言，smFISH 提供了独立荧光成像测定的快照数据，这些测定在固定样品上于离散的时间点上进行，通常是外部刺激后进行。这些测定提供了单个细胞中有限数量的 RNA 物种的转录本数量和位置，因此直接反映了细胞或组织在固定时的分子状态。

为了帮助突出计算推断基因网络从快照 smFISH 数据中所面临的主要挑战，我们考虑最简单的基因网络，如图 1 所示，该网络由一个单一的基因状态组成。单一状态模型预测每个时间间隔转录的 RNA 数量服从泊松分布 7,8。然而，泊松统计期望常常与这一观察结果不符 9,16，因此长期以来存在一种猜测，即采用二状态模型（图 1 a ii） 10,11,13,21。二状态或电报模型允许基因调控网络在非活跃状态和活跃状态之间转换。尽管二状态模型所产生的行为范围令人惊讶 21，但它常常未能描述 smFISH 快照数据中的观测分布。这反过来提示可能存在超出两种具有中间 RNA 生产速率的情况中，挑战立即显现，原因在于有多种方式可以连接 $N \geq 3$ 个基因状态（见图 1b 中的一些示例）。额外的反应路径可能更好地拟合现有数据，但也可能牺牲预测能力。事实上，模型推断，即在严格平衡数据描述与预测能力的基础上进行推断，目前还未在此问题中实现。

在许多先前的尝试中，采用了多种度量标准，包括泊松指标 7,9,11,16,22,25、交叉验证 26、非参数回归 27,33，或信息度量，这些方法通过比较一组截断的可能模型来证明引入额外基因状态及其网络结构的合理性（图 1a iii）。然而，所有这些方法都分别进行模型选择和参数推断，因此无法将 RNA 计数的内在随机性误差传播到基因网络推断中。因此，这些方法未能以统计上严谨的方式平衡描述能力和预测能力，且在给定数据的条件下，基因状态及其相关参数（从而是基因状态的连通性）每个提出的网络的相对概率仍然未知。

在进一步简化的过程中，这些方法中的一些完全忽视 RNA 计数的内在随机性，而偏向于使用质量作用公式，后者预测每个细胞中 RNA 分子数的时间演变 7,10,11,16,24,36,37。质量作用公式对于 smFISH 数据基本上是不足够的，因为 RNA 拷贝数可能很低，从而使得基因拷贝数波动的信息对于提取动力学参数至关重要 21。即使是使用前向 Kolmogorov 方程（化学主方程或 CME）解决这一问题的方法，预测单细胞 RNA 计数的概率性时间演变 8,35,37,43，仍然缺乏稳健的方法来学习基因状态。

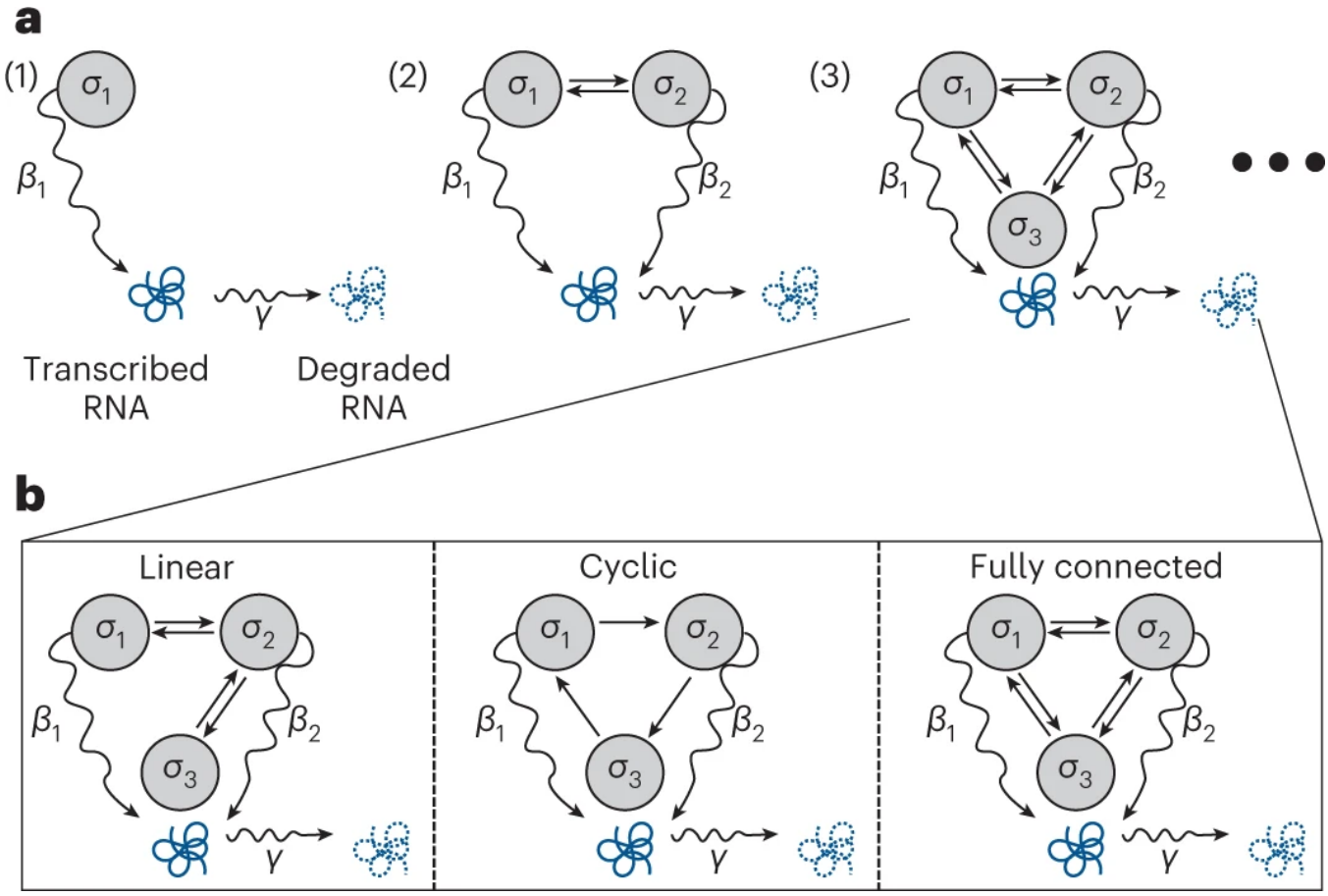

图 1: **基因状态模型示意图**。这里我们展示了一个、两个和三个状态基因模型的示意图（面板 a）。每个灰色圆圈表示一个基因可能占据的 RNA 生产状态，其独特的生产速率加以区分。直箭头表示基因状态之间的可能转变，曲箭头表示 RNA 转录（速率为 β）或降解（速率为 γ）。面板 b 展示了省略了一些转变的模型。正如我们很快将看到的，我们将提出一种从数据中直接推断基因网络（包括基因状态数量和相关速率）的方法。

在此，我们提出了一种方法，通过有意义地将固有噪声传播到模型估计中，基于 smFISH 快照数据集中的离散 RNA 计数，同时推断基因状态及其相关参数，从而得到基因状态的概率分布，并推断基因状态的连通性。我们在贝叶斯框架内实现这一目标，允许我们从后验概率分布中对基因网络进行采样。

为了构造这些后验分布，我们需要对基因状态数量进行先验假设，这需要使用贝叶斯非参数（BNP）方法 **??**，并使用基于 CME 的模型参数计算数据的似然函数。有了似然函数和非参数先验，我们可以同时且自洽地估计模型结构（即基因状态的数量），以及相关的速率参数，从而推断基因状态之间的连通性，这一切都由每个细胞观察到的 RNA 计数所决定。

我们通过首先在合成数据上测试该方法，然后在两个非常不同的实验系统中验证其适用性：大肠杆菌（*Escherichia coli*，*E. coli*）细胞中的 *lacZ* 通路 49和酿酒酵母（*Saccharomyces cerevisiae*，*S. cerevisiae*）细胞中的 *STL1* 通路 50。

2 Result

我们假设可以获得包含每个细胞 RNA 计数的快照 smFISH 数据，记作 m_j^i，这些数据在时间点 $t_{1:K}$ 收集，细胞索引为 $j = 1, ..., J_k$。为了简便起见，我们将所有细胞在所有时间点的 RNA 计数记作 $\tilde{m} = \{\{m_j^i\}_{j=1:J_k}\}_{k=1:K}$。利用这些信息，我们的目标是预测转录基因输出，即我们推断基因状态（即执行模型选择），同时推断相关的速率参数，从而得到基因状态的连通性（即参数估计）。

在这里，我们展示了我们的方法在使用实验数据和合成数据进行基因网络的模型选择和参数推断方面的能力。在贝叶斯非参数（BNP）方法中，模型结构（在本例中为基因状态的数量）被视为一个参数，因此与所有其他参数一起进行推断。其余的感兴趣的参数包括：每个基因状态的生产速率 β_l；不同基因状态之间的转变速率 $k_{l\to l'}$（对于 $l \neq l'$）；以及 RNA 降解的拷贝速率 γ。由于我们在 BNP 框架内工作，我们的参数估计是从速率以及基因状态的完全联合后验概率分布中提取的，这些速率和基因状态是同时学习并自洽的。下文将以直方图的形式展示这些后验样本。这些直方图自然地提供了对每个量值的完整评估，并展示了它们各自的不确定性。

首先，我们展示了该方法在合成数据上的稳健性，这些数据模仿了类似实验数据集的动力学，但涵盖了比实验数据更广泛的场景。随后，我们展示了在大肠杆菌（*E. coli*）细胞中的 *lacZ* 通路实验结果 49，然后是酿酒酵母（*S. cerevisiae*）细胞中的 *STL1* 通路实验结果 50。

2.1 Robustness Analysis

2.1.1 Number of states

与当前基因表达文献中的研究一致 **??**，我们在由一、二、三基因状态组成的三种不同模型上测试了我们的方法。在二状态和三状态模型中，其中一个状态的生产速率为零。

图 2 展示了该方法在一、二、三基因状态模型上的结果。由于我们使用的是合成数据且真实值已知，因此我们可以确认该方法成功地在接近真实值的参数上放置了显著的后验概率。

2.1.2 数据量

或许出人意料的是，对于最简单的基因网络（最容易推断的网络），基因状态的后验分布更为广泛。原因是微妙的：估计更多基因状态的模型可以通过为每个基因状态设定几乎相同的生产速率来近似一个单状态模型（但反之则不然）。然而，由于所有可能的基因状态的生产速率并非完全相同，我们的算法仍然倾向于选择真实的状态数量。

在图 3 中，我们测试了该方法对数据集大小的稳健性。对于具有两个基因状态的网络，该方法在每个时间点提取的细胞数量范围从三个数量级中做出了准确的推断。如图 3 所示，尽管我们发现模型的估计精度（后验分布的广度）确实随着数据量的增加而增加，但其准确性并没有变化。实际上，即使只提供少量细胞的数据，该方法仍能对二状态基因模型做出准确的推断，后验概率的广度反映了数据量较少的情况。

对于我们方法其余稳健性分析的详细概述，读者请参考 S 1.7 节。

2.1.3 合成数据生成

第 3.1 节中使用的合成数据是基于 Gillespie 随机模拟算法 52通过计算机模拟生成的。模型的详细信息在 S 1.1 节中概述，所有参数的推断过程在 S 2.1 节中详细描述。

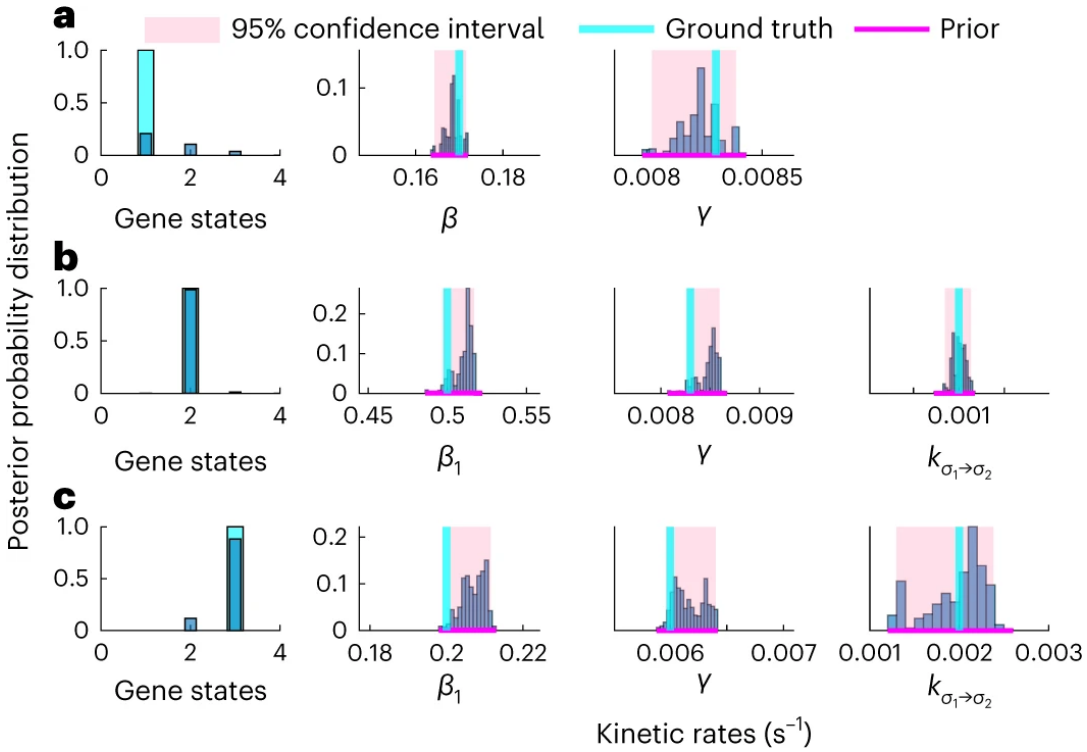

图 2: 从合成数据中对多种基因网络的准确推断。这里我们展示了对以下参数的后验分布：基因状态（第一列），生产速率 β_1（第二列），降解速率 γ（第三列），以及转变速率 $k_{1\to\sigma_1}$（第四列）。在第一行中，我们展示了一个单基因状态模型的分布，即在单一速率 β 下的生产和速率 γ 下的降解，没有转变速率。如预期，我们的后验最大值与真实值非常接近。我们在第二行和第三行中获得了类似的结果，分别表示二状态和三状态基因模型。每个合成数据集由 2000 个细胞在每个时间点观察，数据集包含 20 个时间点，均匀分布在 1 小时内 50。每个数据点通过 Gillespie 随机模拟算法生成 **??**。未在此展示的速率直方图可见于图 S12。对于此图和后续图，我们使用了公开可用的 MATLAB 实现的 Munkres 分配算法 54。该代码作为后处理步骤使用，仅用于在 MCMC 迭代中适当分配基因状态标签，仅影响我们的图形展示。

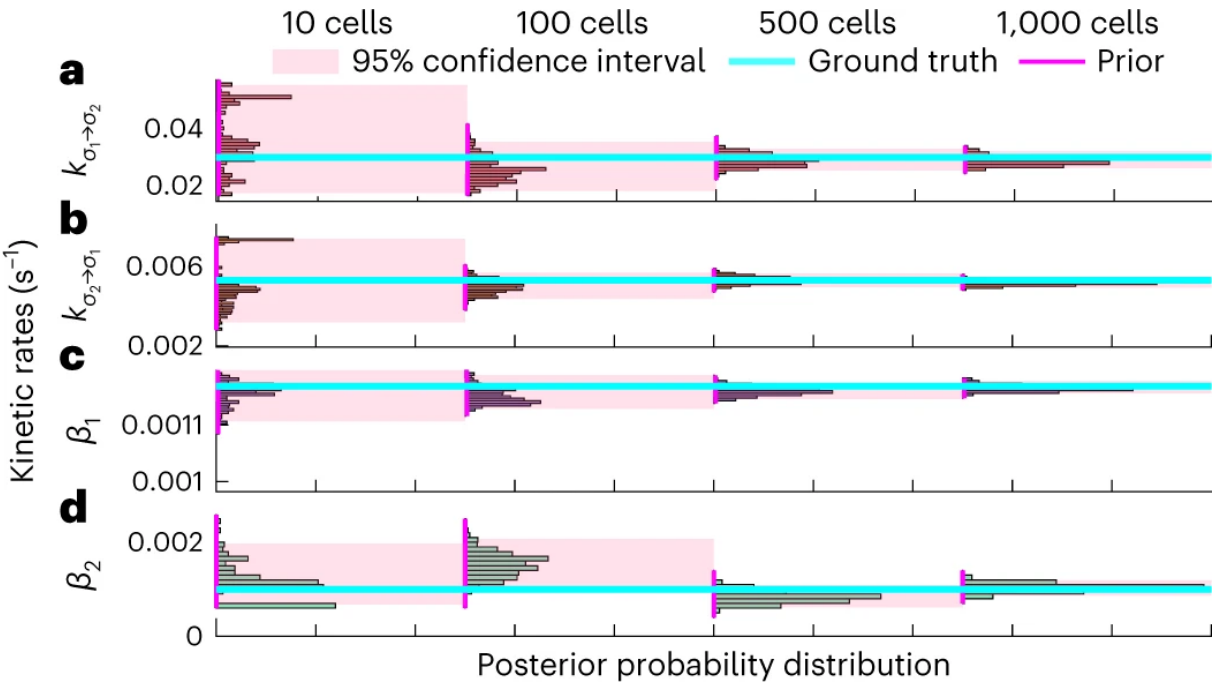

图 3: **灵敏度分析：数据量**。这里我们展示了生产速率 β_1 和转变速率 $k_{1\to\sigma_2}$、$k_{2\to\sigma_1}$ 的后验分布，这些是具有两个基因状态的网络的结果。每一行展示了不同动力学速率的后验分布，随着数据量的增加，其宽度逐渐变窄。因此，我们发现我们的后验最大值与真实值非常接近，并且随着数据量的增加，置信区间明显变窄。合成数据集由每个时间点观察的 10、50、500 或 1000 个细胞组成，且 20 个采集时间点均匀分布在 1 小时内，基于典型的实验程序 50。与之前一样，每个数据点是通过 Gillespie 随机模拟算法生成的 52。对于此图及后续图，我们省略了基因状态数量的后验分布，因为即使每个时间点只有 10 个细胞，它们也能完全确定正确的基因状态数量。

2.2 实验结果

我们分析了来自 *E. coli* 49和 *S. cerevisiae* 50细胞的 smFISH RNA 计数数据。为了简化，我们在此通过使用之前已建立的结果来校准 RNA 分子的降解速率 **??**，并在图 S 2 中展示了如预期的那样，校准一个速率对推断其他速率的不确定性有减少的净效应。

有关实验条件的更多详细信息，请参阅 49（*E. coli*）和 50（*S. cerevisiae*）。

2.2.1 E. coli

在这里，我们展示了对 *E. coli* 细胞中 *lacZ* 通路的基因状态数量和参数的同时推断，细胞在慢生长培养基中生长。我们的结果展示在图 4 中。在此图中，我们还展示了 49（称为 Wang et al.）通过最大似然法获得的点估计。需要明确的是， 49提出了一个模型（即预先指定基因状态），并在给定该模型的条件下，从数据中学习参数。

我们发现他们假定的状态与我们从数据中直接学习到的结果一致。在参数方面，我们也发现最低生产速率的结果大致一致。然而，在比较我们最大后验估计（MAP）与 49中报告的结果时，我们发现 $k_{\sigma_1\to\sigma_2}$、$k_{\sigma_2\to\sigma_1}$ 和 β_1 的参数分别存在 70%、40% 和 43% 的差异。这种分歧凸显了一个核心问题：即使仅学习速率（并手动假设状态），49也不能高效地对其高维后验进行采样。

为此，我们采用了哈密顿蒙特卡罗（HMC）和并行温度（PT）方法，使我们的算法避免陷入局部极大值。因此，我们发现通过比较似然函数，明显偏向于我们所收敛的参数（与 49中报告的参数相比）。事实上，我们发现我们的 MAP 估计（θ'）比 49的估计（θ）更可能，二者的对数似然比值为

$$\ln\left(\frac{P(\tilde{m}|\theta')}{P(\tilde{m}|\theta)}\right) \approx 83.$$

图 4 中可以看到我们的算法的对数似然曲线超越了 49的方法。

有趣的是，尽管 θ 和 θ' 之间有显著差异，时间点之间的 RNA 计数直方图看起来定性上相似；参见图 S 3。这是预期中的结果，因为静态直方图不包含我们在分析按时间顺序排列的快照数据时所利用的时间信息。同样，这也突显了通过比较时间点之间的 RNA 直方图来评估动力学速率和基因状态的局限性，而这种方法正是本文所提出的。

图 5 比较了我们学习的模型与 49中估计的模型，该模型用于分析在甘油中培养的 *E. coli* 数据。

在快速生长培养基（葡萄糖在 $37°C$）中，我们的完整非参数方法自信地推断出三个基因状态，而与此相对，49假设只有两个状态。由于我们预测了不同的模型，因此在这里直接比较似然性比之前的情况更为困难。

然而，为了直接比较似然性，我们限制我们的算法，仅通过手动强加一个二状态基因表达模型（将一个生产速率固定为零），来推断参数（图 S 7）。我们发现与之前展示的差异相似：$k_{\sigma_1\to\sigma_2}$、$k_{\sigma_2\to\sigma_1}$ 和 β_1 的参数分别存在 78%、67% 和 18% 的差异。似然比再次偏向于我们的估计：

$$\ln\left(\frac{P(\tilde{m}|\theta')}{P(\tilde{m}|\theta)}\right) \approx 4.7\times 10^3.$$

这再次表明，即使假设一个较简单的具有较少状态的模型，局部极大值的存在可能导致错误的参数估计值。再次强调了像我们这样同时优化的方法的必要性。

2.2.2 *S. cerevisiae*

图 6 展示了在 *S. cerevisiae* 细胞中对 *STL1* 通路的模型推断结果。我们将我们的基因状态和参数推断结果与 39中仅估计基因参数（并假设基因状态）的结果进行了比较。我们的分析证实了先前方法在参数估计之前假设的 *S. cerevisiae* 中 *STL1* 基因的四个染色质重组状态 26,39。然而，如第 2 节和第 3.2.1 节所述，这种基因状态数量的预设可能导致仅对给定观察集局部最大化似然的参数估计。由于我们改进了对模型空间的探索，我们学习到的参数（θ'），如图 6 所示，计算得出比 39中的参数更可能，且可能性增加的倍数为

$$\ln\left(\frac{P(\tilde{m}|\theta')}{P(\tilde{m}|\theta)}\right) \approx 350 \text{ for } \textit{STL1} \text{ transcription.}$$

有关预测分布的比较，请参见图 S 8，类型参见第 3.2.1 节。

3 Discussion

推断给定观察到的快照 RNA 表达数据的最可能的调控网络结构，提出了独特的挑战，这些挑战一直妨碍着准确识别生物物理反应的数量和连通性及其组成参数，无论是单独还是同时进行。我们的方法实现了模型和参数的自洽和同时推断，并改善了其他方法的局限性，包括：1) 假设稳态动力学 63，以及 2) 基因状态数量的模型选择与参数推断的分离 22,23,34,35。

我们通过实验数据和模拟的快照 RNA 表达数据评估了我们方法的有效性。对于快速生长培养基中的 *E. coli*，我们的方法确定（图 5）三状态模型比先前使用的二状态模型更为可能。额外的状态是 *lacZ* 基因的一个中间生产状态，其生产速率介于先前分析中假设的“开”和“关”速率之间 49。对于 *S. cerevisiae* 中的 *STL1* 通路，我们的方法确认先前使用的四状态网络，具有多个生产状态，是最可能的模型 **??**。关键的是，这里详细描述的方法不预先假设基

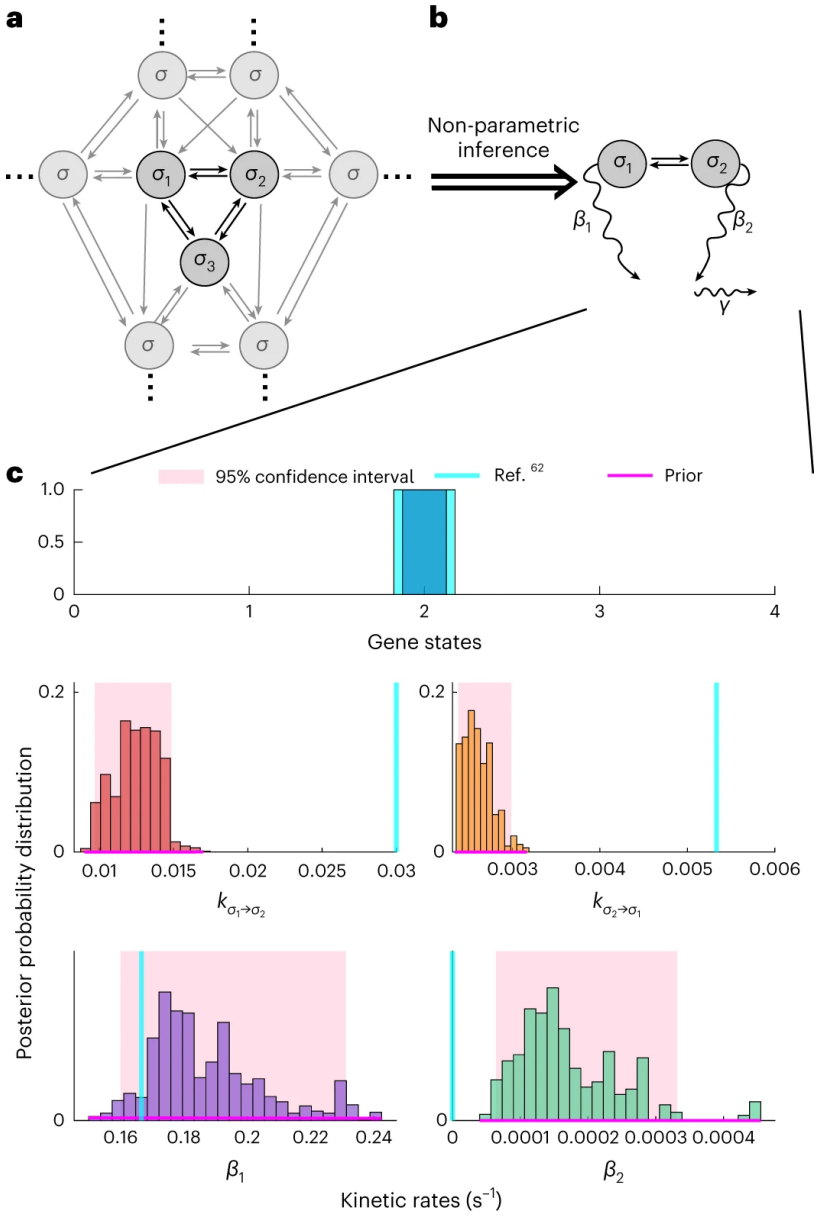

图 4: **对慢生长 *E. coli* 数据的推断**。在此图中，我们展示了在 30°C 下用甘油培养的 *E. coli* 中的 *lacZ* 通路分析结果。每个面板展示了不同模型参数的后验概率分布，并与 Wang 等 49的估计结果进行了比较，后者的估计值以垂直青色线表示。粉红色阴影区域表示 MCMC 样本中 95% 的区间。我们恢复了一个二基因状态模型，其参数与 49中的参数不同，具体内容如第 3.2.1 节所述。底部面板展示了我们方法的对数似然函数超过 49（以水平青色线表示）的痕迹。有关此处展示的速率的联合直方图，请参见图 S 9。

因状态的数量或连通性。最后，我们通过使用由模拟调控网络创建的合成快照 RNA 表达数据，展示了我们方法的稳健性，这些数据旨在挑战任何计算推断方法。这些结果展示了一种通用的、同时的、自洽的方法，用于从 smFISH 获得的快照 RNA 表达数据中推断基因调控模型和相关速率。

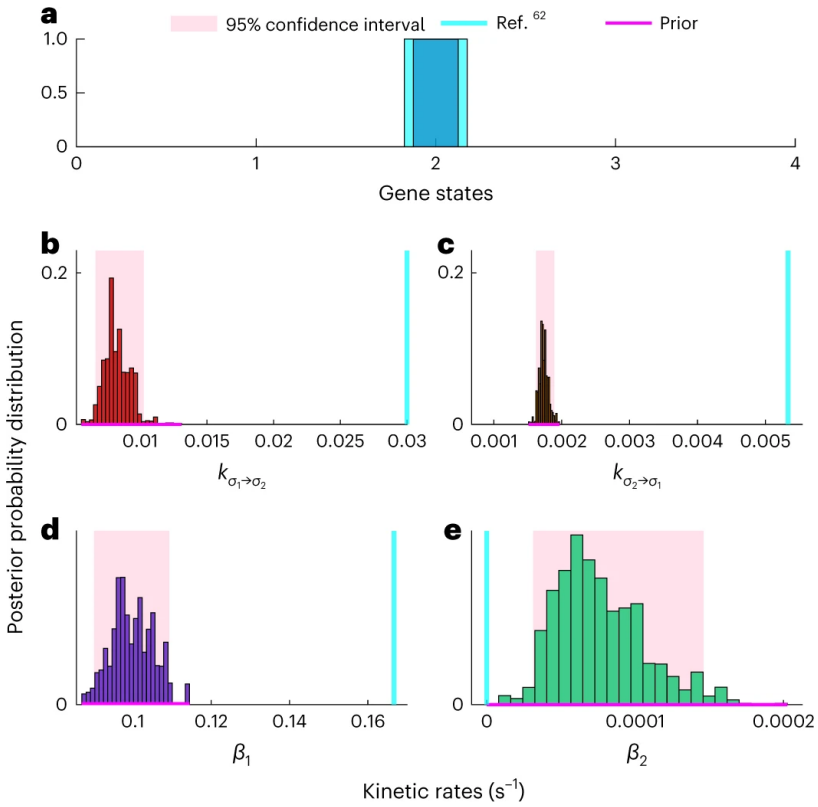

图 5: **对快速生长 *E. coli* 数据的非参数推断**。在这里，我们展示了在葡萄糖培养基中，37°C 下培养的 *E. coli* 中的 *lacZ* 通路推断出的速率的一个子集。我们的方法强烈支持三个基因状态。与后续的三基因状态动态结果一样，我们展示了与具有两个最大生产速率的状态相关的生产速率 β_1、β_2 和转变速率 $k_{\sigma_1\to\sigma_2}$、$k_{\sigma_2\to\sigma_1}$。除了基因状态的数量外，Wang 等 49推断的参数被省略，因为这些参数是基于一个假设的基因状态数量，而与我们推断的基因状态数量不同。图 S 5 进一步说明了直方图拟合可能不是估计速率的适当方法。有关此处展示的速率的联合直方图，请参见图 S 10。

我们可以对我们的框架进行一些附加扩展。首先，通过最小的修改，我们的方法可以利用 smFISH 量化的快照 RNA 表达数据中包含的空间信息，例如，用于确定 RNA 从细胞核到细胞质的转运速率。事实上，RNA 从细胞核到细胞质的转运这一额外约束以前已经改善了参数识别。

其次，在我们的框架中修改测量模型可能允许转录速率、基因状态转变和 RNA 降解的时间变化 66。最后，随着基因物种密度的增加，使用高度多重化的 smFISH 方法，我们方法的灵活网络连通性可能允许调控模型探索共变基因表达的最可能调控网络 67,68,69。

上述推广将为我们似然计算引入额外的复杂性，而这一部分已经是最耗时的推断步骤。额外的复杂性直接来源于状态数量的增加和连通图复杂度的提高。这两者都会改变生成矩阵 **A**（见 S 1.2 节），使其在状态数目较多时变得更大，或者在连通性更密集时变得更加稠密。如果生成矩阵保持稀疏，CME 解决方案的时间成本大致与 **A** 的大小成线性比例，基于 FSP 的 Krylov 子空间方法 **??**可能比这里使用的 CME 方法更为高效。**A** 的 CME 计算成本如何随密度变化则比状态数的变化更为复杂。在某些密度之上，最近提出的量化张量训练方法 72可能更高效，因为基于 FSP 的 Krylov 子空间方法采用增量时间步进，而不是直接跳到分析所需的时间。或者，已经有一些有前景的尝试使用神经网络求解 ODE 73。除了促进由于稠密 CME 生成矩阵所带来的困难外，神经网络方法可能进一步使得非马尔科夫基因转录模型的参数推断成为可能 74。

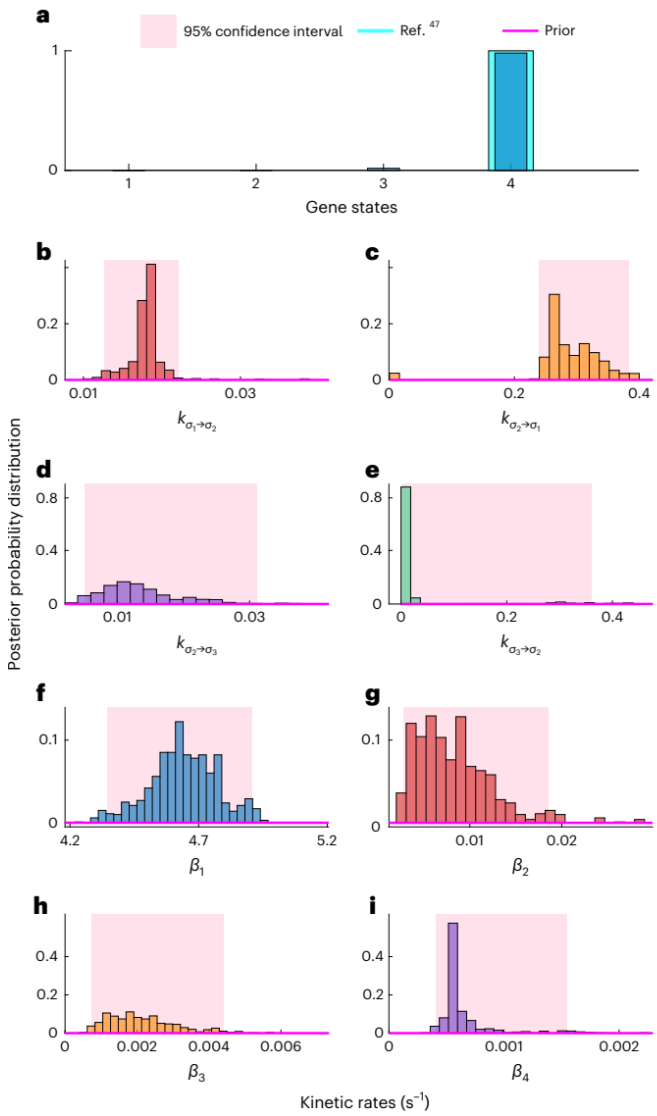

图 6: **对 *S. cerevisiae* 数据的推断**。在此图中，我们展示了与我们推断的四个基因状态之间的线性转换相关的速率的子集，这与 39的结果一致（图中引用为 Munsky 等）。我们发现我们的方法高可信度地推断出了一个具有四个基因状态的转录模型。此外，转变速率与所谓的“线性”切换非常吻合（即，基因可以从每个基因状态离开并进入最多两个其他基因状态）。这两个结果与先前假设的模型一致，该模型描述了 *STL1* 基因的染色质重塑 26。有关此处展示的速率的联合直方图，请参见图 S 11。

还可能通过直接的图像基因表达动态推断基因网络，这些动态在活细胞中的实时数据中获得 **??**。然而，这种方法通过实时获取有限数量分子的动力学，牺牲了更高的数据密度。而且，基因操作限制了获得对局部分子和生物物理相互作用的洞察。

通过预测建模理解基因表达的下游后果，直接推动了快照 RNA 表达数据的使用，尤其是在更高数据密度的情况下。虽然去除快照 RNA 表达数据中的时间相关性直接妨碍了获得基因调控动力学的直接洞察，我们可以通过增加快照 RNA 表达数据中的时间点、RNA 物种、细胞和刺激条件的密度来填补知识空白。事实上，本文展示了如何最大化从快照 RNA 表达数据中推断出的信息，并获得基因调控网络结构及其组成速率的概率分布。我们通过将基因调控

网络识别问题重新构建为贝叶斯非参数范式，并开发了推断基因状态、连通性和参数所需的工具，达到了这一目标。所提出方法的概率输出现在可能使我们能够学习反映任何给定快照 RNA 表达数据集支持的信心的网络。

4 Methods

我们假设可以获得包含每个细胞 RNA 计数的快照 smFISH 数据，记作 $m^j_{t_k}$，这些数据在时间点 $t_{1:K}$ 收集，细胞索引为 $j = 1, \ldots, J_k$。为了简便起见，我们将所有细胞在所有时间点的 RNA 计数记作 $\tilde{m} = \left\{m^j_{t_k}\right\}_{j=1:J_k, k=1:K}$。利用这些信息，我们的目标是推断基因表达模型，即同时推断基因状态及其相关速率参数（从而推断基因状态的连通性） 8,16,17,46,47,62,63,**??**。

在“结果”部分中使用的所有合成数据都是通过基于 Gillespie 的随机模拟算法 **??**的计算机模拟生成的。模型的详细信息在下文中概述，所有参数的推断过程在接下来的部分中描述。

4.1 Model formulation

在每个基因状态 σ_l 中，基因以速率 β_l 转录 RNA 拷贝。所有 RNA 会根据总速率 γ 随机降解。基因可以随机转变，例如从状态 σ_l 转变到状态 $\sigma_{l'}$，其转变速率为 $k_{\sigma_l \to \sigma_{l'}}$。为方便起见，我们将所有参数统称为

$$\theta = (\sigma_\star, k_{\sigma_l \to \sigma_{l'}}, \beta_1, \beta_2, \ldots, \gamma)$$

其中 $\sigma_1 = \sigma_\star$ 表示初始基因状态。

为了在贝叶斯框架内推断 θ，我们必须首先指定似然函数 $P(\tilde{m}|\theta)$。给定测量值 $\tilde{m}$，似然函数为

$$P(\tilde{m}|\theta) = \prod_{k=1}^{K} \prod_{j=1}^{J_k} \left(\sum_{l=1}^{N} P^\theta_l(\sigma_l, m^l_{jt_k}) \right).$$

其中 $P^\theta_t \equiv (P^\theta_t(\sigma_1, 1), \ldots, P^\theta_t(\sigma_1, M), P^\theta_t(\sigma_2, 1), \ldots, P^\theta_t(\sigma_2, M), \ldots, P^\theta_t(\sigma_N, 1), \ldots, P^\theta_t(\sigma_N, M))^T$ 满足化学主方程（CME），$\mathbf{P}^\theta_t = \mathbf{A} \cdot \mathbf{P}^\theta_t$，其中 $\mathbf{A}$ 是生成矩阵，其对 θ 的依赖关系在 S 1.2 节中详细描述。

4.2 Model Inference

为了使用我们的似然函数构建后验分布，我们需要对所有模型参数设定先验。θ 中各量的先验选择仅为计算上的便利，具体细节见 S 1.4 节。

这里我们仅扩展基因状态上使用的非参数先验。

在非参数公式中，我们必须理论上考虑无限多个基因状态，并允许数据将这些无限可能性缩小为由数据支持的状态。这类似于常规（参数化）贝叶斯方法，后者通常假设对参数有广泛的先验，并最终通过似然函数允许数据将参数估计变得更加精确（即，后验分布变得更加集中）。

为了计算上的便利，我们使用 Beta-Bernoulli 过程 58,59作为这些状态是否存在的正式先验。

简而言之，我们引入了无限多个中间的二元（伯努利）指标变量 b_l，称为负荷，当基因状态 σ_l 被数据认为是必要时，b_l 等于 1，否则为 0。为了使计算可行，我们引入了一个所谓的弱极限 L，设置了可能的基因状态数量的上限 58,59。我们将所有负荷 $\{b_l\}_{l=1:L}$ 统称为 $\mathbf{b}$。

Beta-Bernoulli 过程先验 58,59 对负荷的描述为：

$$q_l \sim \text{Beta}\left(\frac{\zeta}{L}, \frac{L-1}{L}\right)$$

$$b_l|q_l \sim \text{Bernoulli}(q_l),$$

其中 q_l 是描述负荷 b_l 被“激活”或等于 1 的成功概率的超参数，ζ 是一个超超参数。给定这个先验，我们可以从数据中学习哪些基因状态是必要的。

给定似然函数和所有先验分布，我们现在可以构建我们的后验概率分布的显式形式 $P(\mathbf{q}, \mathbf{b}, \theta|\tilde{m})$。由于我们的似然函数没有假设解析形式，我们使用自定义的马尔科夫链蒙特卡洛（MCMC）60,61,62采样方案，从 $P(\mathbf{q}, \mathbf{b}, \theta|\tilde{m})$ 中生成伪随机数。

重要的是，能够有效地探索后验分布，尤其是在推断基因状态的困难增加的情况下，这将使我们能够摆脱困扰其他方法（第 3.2.2 节）参数评估的陷阱（局部极大值）。

考虑到这一点，我们使用了整体的吉布斯采样方案来构建我们的马尔科夫链。在这个吉布斯采样方案中，我们可以直接从它们的联合边际后验分布中采样初始条件 σ_* 和负荷 $\mathbf{b}$。相比之下，成功概率 $\mathbf{q}$ 使用梅特罗波利斯-黑斯廷斯采样方案进行采样。由于以下事实：1) 我们同时学习离散（基因状态数量、初始条件）和连续（动力学速率）参数；2) 各个连续参数之间存在显著的尺度分离，我们可能会遇到在可能模型空间的大部分区域内没有特征的后验分布。

为了解决问题 1)，我们使用并行温度（PT）对所有参数进行采样，以便更好地探索离散参数。在我们的 PT 方案中，我们使用哈密顿蒙特卡罗（HMC）采样提出连续参数，从而解决问题 2)。这些采样方案首次联合使用，允许我们在合理的时间尺度内推断基因状态及其相关参数，避免了第 3.2.1 节中提到的局部极大值。

4.2.1 Reporting summary

有关研究设计的更多信息，请参阅与本文链接的《自然投资组合报告摘要》。

5 Data and Code availability

Extended data is available for this paper at https://doi.org/10.1038/ s43588-022-00392-0.

MatLab Code is available at https://zenodo.org/records/7425217.