Inference on autoregulation in gene expression with variance-to-mean ratio

Yue Wang[1,2,*] and Siqi He[3]

[1]Department of Computational Medicine, University of California, Los Angeles, California, United States of America
[2]Institut des Hautes Études Scientifiques, Bures-sur-Yvette, Essonne, France
[3]Simons Center for Geometry and Physics, Stony Brook University, Stony Brook, New York, United States of America
[*]E-mail address: yuew@g.ucla.edu (Y. W.). ORCID: 0000-0001-5918-7525

Abstract

Some genes can promote or repress their own expressions, which is called autoregulation. Although gene regulation is a central topic in biology, autoregulation is much less studied. In general, it is extremely difficult to determine the existence of autoregulation with direct biochemical approaches. Nevertheless, some papers have observed that certain types of autoregulations are linked to noise levels in gene expression. We generalize these results by two propositions on discrete-state continuous-time Markov chains. These two propositions form a simple but robust method to infer the existence of autoregulation from gene expression data. This method only needs to compare the mean and variance of the gene expression level. Compared to other methods for inferring autoregulation, our method only requires non-interventional one-time data, and does not need to estimate parameters. Besides, our method has few restrictions on the model. We apply this method to four groups of experimental data and find some genes that might have autoregulation. Some inferred autoregulations have been verified by experiments or other theoretical works.

Keywords.
inference; gene expression; autoregulation; Markov chain.

Frequently used abbreviations:
GRN: gene regulatory network.
VMR: variance-to-mean ratio

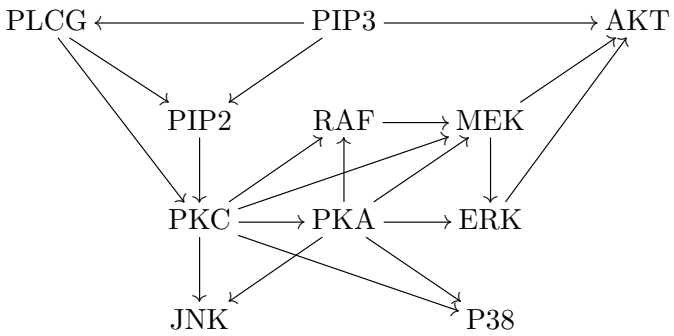

Figure 1: An example of a GRN in human T cells [85]. Each vertex is a gene. Each arrow is a regulatory relationship. Notice that it has no directed cycle.

1 Introduction

In general, genes are transcribed to mRNAs and then translated to proteins. We can use the abundance of mRNA or protein to represent the expression levels of genes. Both the synthesis and degradation of mRNAs/proteins can be affected (activated or inhibited) by the expression levels of other genes [42], which is called (mutual) gene regulation. Genes and their regulatory relations form a gene regulatory network (GRN) [16], generally represented as a directed graph: each vertex is a gene, and each directed edge is a regulatory relationship. See Fig. 1 for an example of a GRN.

The expression of one gene could promote/repress its own expression, which is called positive/negative autoregulation [11]. Autoregulation is very common in *E. coli* [63]. Positive autoregulation is also called autocatalysis or autoactivation, and negative autoregulation is also called autorepression [4, 21]. For instance, HOX proteins form and maintain spatially inhomogeneous expression of HOX genes [64]. For genes with position-specific expressions during development, it is common that the increase of one gene can further increase or decrease its level [78]. Autoregulation has the effect of stabilizing transposons in genomes [6], which can affect cell behavior [41]. Autoregulation can also stabilize the cell phenotype [2], which is related to cancer development [90, 52, 15].

While countless works infer the regulatory relationships between different genes (the GRN structure) [83], determining the existence of autoregulation is an equally important yet less-studied field. Due to technical limitations, it is difficult and sometimes impossible to directly detect autoregulation in experiments. Instead, we can measure gene expression profiles and infer the existence of autoregulation. In this paper, we consider a specific data type: measure the expression levels of certain genes without intervention for a single cell (which reaches stationarity) at a single time point, and repeat for many different cells to obtain a probability distribution for expression levels. Such single-cell non-interventional one-time gene expression data can be obtained with a relatively low cost [48].

With such single-cell level data for one gene V, we can calculate the ratio of variance

and mean of the expression level (mRNA or protein count). This quantity is called the variance-to-mean ratio (VMR) or the Fano factor. Many papers that study gene expression systems with autoregulations have found that negative autoregulation can decrease noise (smaller VMR), and positive autoregulation can increase noise (larger VMR) [70, 68, 30, 51, 25, 18, 17]. This means VMR can be used to infer the existence of autoregulation.

We generalize the above observation and develop two mathematical results that use VMR to determine the existence of autoregulation. They apply to some genes that have autoregulation. For genes without autoregulation, these results cannot determine that autoregulation does not exist. We apply these results to four experimental gene expression data sets and detect some genes that might have autoregulation.

We start with some setup and introduce our main results (Section 2). Then we cite some previous works on this topic and compare them with our results (Section 3). For a single gene that is not regulated by other genes (Section 4) and multiple genes that regulate each other (Section 5), we develop mathematical results to identify the existence of autoregulation. These two mathematical sections can be skipped. We summarize the procedure of our method and apply it to experimental data (Section 6). We finish with some conclusions and discussions (Section 7).

2 Setup and main results

One possible mechanism of "the increase of one gene's expression level further increases its expression level" is a positive feedback loop between two genes [31]. Here V_1 and V_2 promote each other, so that the increase of V_1 increases V_2, which in return further increases V_1. We should not regard this feedback loop as autoregulation. When we define autoregulation for a gene V, we should fix environmental factors and other genes that regulate V, and observe whether the expression level of V can affect itself. If V is in a feedback loop that contains other genes, then those genes (which regulate V and are regulated by V) cannot be fixed when we change V. Therefore, it is essentially difficult to determine whether V has autoregulation in this scenario. In the following, we need to assume that V is not contained in a feedback loop that involves other genes.

The actual gene expression mechanism might be complicated. Besides other genes/factors that can regulate a gene, for a gene V itself, it might switch between inactivated (off) and activated (on) states [10]. These states correspond to different transcription rates to produce mRNAs. When mRNAs are translated into proteins, those proteins might affect the transition of gene activation states, which forms autoregulation [23]. See Fig. 2 for an illustration. Therefore, for a gene V, we should regard the gene activation state, mRNA count, and protein count as a triplet of random variables (G, M, P), which depend on each other.

When we fix environmental factors and other genes that affect V, the triplet (G, M, P) should follow a continuous-time Markov chain. A possible state is the gene activation

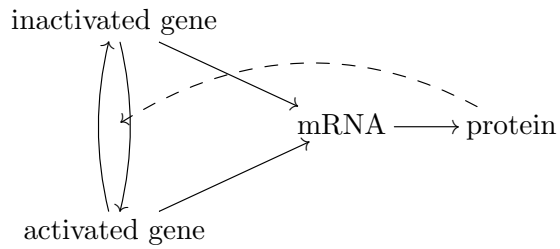

Figure 2: The mechanism of gene expression. A gene might switch between inactivated state and activated state, which correspond to different transcription rates. Gene is transcribed into mRNAs, which are translated into proteins. Proteins might (auto)regulate the state transition of the corresponding gene.

state on/off (for G), the mRNA count on $\mathbb{Z}$ (for M), and the protein count on $\mathbb{Z}$ (for P). Thus the total space is $\{0,1\} \times \mathbb{Z} \times \mathbb{Z}$. When we consider the expression level M or P (but have no access to the value of G), sometimes itself is Markovian (its dynamics can be fully determined by itself, without the knowledge of G), and we call this scenario "**autonomous**". In other cases, M or P itself is no longer Markovian (its dynamics explicitly depends on G), and we call this scenario "**non-autonomous**". We need to consider the triplet (G, M, P) in the non-autonomous scenario. This is similar to a hidden Markov model, where a two-dimensional Markov chain is no longer Markovian if projected to one dimension (since this dimension depends on the other dimension).

For the autonomous scenario, we can fully classify autoregulation for a gene V. Assume environmental factors and other genes that affect the expression of V are kept at constants. Define the expression level (mRNA count for example) of one cell to be $X = n$, the mRNA synthesis rate at $X = n - 1$ to be f_n, and the degradation rate for each mRNA molecule at $X = n$ to be g_n. This is a standard continuous-time Markov chain on $\mathbb{Z}$ with transition rates

$$\frac{1}{\Delta t}\mathbb{P}[X(t+\Delta t) = n \mid X(t) = n-1] = f_n,$$
$$\frac{1}{\Delta t}\mathbb{P}[X(t+\Delta t) = n-1 \mid X(t) = n] = ng_n.$$

Define the relative growth rate $h_n = f_n/g_n$. If there is **no autoregulation**, then h_n is a constant. **Positive autoregulation** means $h_n > h_{n-1}$ for some n, so that $f_n > f_{n-1}$ and/or $g_n < g_{n-1}$; **negative autoregulation** means $h_n < h_{n-1}$ for some n, so that $f_n < f_{n-1}$ and/or $g_n > g_{n-1}$. Notice that we can have $h_n > h_{n-1}$ for some n and $h_{n'} < h_{n'-1}$ for some other n', meaning that positive autoregulation and negative autoregulation can both exist for the same gene, but occur at different expression levels. Such non-monotonicity in regulating gene expression often appear in reality [1].

For the non-autonomous scenario, we can still define autoregulation. Consider the expression level X of V (mRNA count or protein count) and its interior factor I. If X is

the mRNA count, then I is the gene state; if X is the protein count, then I is the gene state and the mRNA count. If there is **no autoregulation**, then X cannot affect I, and for each value of I, the relative growth rate h_n of X is a constant. If X can affect I, or h_n is not a constant, then there is **autoregulation**. When X can affect I, there is a directed cycle ($X \to I \to X$), and the change of X can affect itself through I. In this case, it is not always easy to distinguish between positive autoregulation and negative autoregulation.

Quantitatively, for the autonomous scenario, when we fix other factors that might regulate this gene V, if V has no autoregulation, then $h_n = f_n/g_n$ is a constant h for all n. In this case, the stationary distribution of V satisfies $\mathbb{P}(X = n)/\mathbb{P}(X = n-1) = h/n$, meaning that the distribution is Poissonian with parameter h, $\mathbb{P}(X = n) = h^n e^{-h}/n!$, and VMR $= 1$. If there exists positive autoregulation of certain forms, VMR > 1; if there exists negative autoregulation of certain forms, VMR < 1. However, such results are derived by assuming that f_n, g_n take certain functional forms, such as linear functions [54, 57], quadratic functions [24], or Hill functions [67]. There are other papers that consider Markov chain models in gene expression/regulation [32, 61, 65, 14, 62, 45], but the role of VMR is not thoroughly studied.

In this paper, we generalize the above result of inferring autoregulation with VMR by dropping the restrictions on parameters. Consider a gene V in a known GRN, and assume it is not regulated by other genes, or assume other factors that regulate V are fixed. Assume we have the **autonomous** scenario, meaning that its expression level $X = n$ satisfies a general Markov chain with synthesis rate f_n and per molecule degradation rate g_n. We do not add any restrictions on f_n and g_n. Use the single-cell non-interventional one-time gene expression data to calculate the VMR of V. Proposition 1 states that VMR > 1 or VMR < 1 means the existence of positive/negative autoregulation.

Nevertheless, the autonomous condition requires some assumptions, and often does not hold in reality [5, 39, 33, 32]. Consider a gene V that is not regulated by other genes, and has no autoregulation. The mRNA count or the protein count is regulated by the gene activation state (an interior factor), which cannot be fixed. Due to this non-controllable factor, there might be transcriptional bursting [60, 19] or translational bursting [8], where transcription or translation can occur in bursts, and we have VMR > 1. This does not mean that Proposition 1 is wrong. Instead, it means that the expression level itself is not Markovian, and the scenario is non-autonomous. In this scenario, we should apply Proposition 2, described below, which states that no autoregulation means VMR ≥ 1.

We extend the idea of inferring autoregulation with VMR to a gene that is regulated by other genes, or with non-autonomous expression. Consider a gene V' in a known GRN. Denote other genes that regulate V' and the interior factors (gene state and/or mRNA count) of V' by $\boldsymbol{F}$. Denote the values of $V', \boldsymbol{F}$ as $X, \boldsymbol{Y}$. Assume V' is not contained in a feedback loop, and assume g_n, the per molecule degradation rate of V' at $X = n$, is not regulated by other genes or its interior factors (gene state and/or mRNA count). We do not add any restrictions on the synthesis rate f_n. Proposition 2 states that if V' has no autoregulation, then VMR$(X) \geq 1$. Therefore, VMR$(X) < 1$ means autoregulation for V'.

Proposition 2 is derived in a "one-step" Markov chain model, where at one time point, only transitions to the nearest neighbors are allowed: $(X = n, \boldsymbol{Y} = \boldsymbol{a}) \to (X = n+1, \boldsymbol{Y} = \boldsymbol{a})$, $(X = n, \boldsymbol{Y} = \boldsymbol{a}) \to (X = n-1, \boldsymbol{Y} = \boldsymbol{a})$, and $(X = n, \boldsymbol{Y} = \boldsymbol{a}) \to (X = n, \boldsymbol{Y} = \boldsymbol{a}')$. This one-step Markov chain model is the most common approach in stochastic representations of gene regulation [70, 30, 54, 51, 17]. Recently, there are some studies that consider "multi-step" Markov chain models, where at one time point, the change of mRNA/protein count can be accompanied with the change of other factors, such as the gene state [7, 43, 74]. For example, the following transition is allowed: $(G^*, M = n) \to (G, M = n+1)$. In this multi-step model, Proposition 2 is no longer valid: even without autoregulation, it is possible that $\mathrm{VMR}(X) < 1$. Consider an example that the production of one mRNA molecule needs many steps of gene state transition, and the gene returns to the initial step after producing one mRNA molecule: $G_1 \to G_2 \to \cdots \to G_k \to G_1 + M$, $M \to \emptyset$. Since there are many steps, the total time for one cycle of $G_1 \to \cdots \to G_k \to G_1 + M$ can be highly deterministic, such as 1 second. Assume the degradation probability for each mRNA molecule in 1 second is 0.01. Then the mRNA count is highly concentrated near 100, and $\mathrm{VMR}(X) < 1$ (close to 0.5 in numerical simulations).

Since multi-step models allow more transitions, they are more general than one-step models. However, it is still a question that whether such generalizations are necessary, since one-step models have good fitting with experimental data [38, 18, 9]. Proposition 2 provides a method to verify this problem: If a gene has $\mathrm{VMR}(X) < 1$, but we use other methods to determine that it has no autoregulation, then Proposition 2 states that one-step models deviate from reality, and multi-step models should be adopted. Therefore, when one-step models hold, Proposition 2 is a valid method to determine the existence of autoregulation; when one-step models do not hold, combined with other methods to determine autoregulation, Proposition 2 can detect the failure of one-step models.

In the scenario that Proposition 2 may apply, if $\mathrm{VMR} \geq 1$, Proposition 2 cannot determine whether autoregulation exists. In fact, with VMR, or even the full probability distribution, we might not distinguish a non-autonomous system with autoregulation from a non-autonomous system without autoregulation, which both have $\mathrm{VMR} \geq 1$ [9]. In the non-autonomous scenario, we only focus on the less complicated case of $\mathrm{VMR} < 1$, and derive Proposition 2 that firmly links VMR and autoregulation.

In reality, Proposition 1 and Proposition 2 can only apply to a few genes (which are not regulated by other genes or have $\mathrm{VMR} < 1$), and they cannot determine negative results. Thus the inference results about autoregulation are a few "yes" and many "we do not know". Besides, for the results inferred by Proposition 1, especially those with $\mathrm{VMR} > 1$ (positive autoregulation), we cannot verify whether their expression is autonomous, and the inference results are less reliable.

Current experimental methods can hardly determine the existence of autoregulation, and to determine that a gene does not have autoregulation is even more difficult. Therefore, about whether genes in a GRN have autoregulation, experimentally, we do not have "yes" or "no", but a few "yes" and many "we do not know". Thus there is no gold standard

to thoroughly evaluate the performance of our inference results. We can only report that some genes inferred by our method to have autoregulation are also verified by experiments or other inference methods to have autoregulation. Besides, if the result by Proposition 2 does not match with other methods, it is possible that the one-step model fails. Instead, in Appendix A, we test our methods with numerical simulations, and the performances of both Propositions are satisfactory.

3 Related works

There have been some results of inferring autoregulation with VMR [70, 68, 30, 51, 25, 18, 17]. However, these VMR-based methods have various restrictions on the model, and some of them are derived by applying linear noise approximations, which are not always reliable in gene regulatory networks [72].

Besides VMR-based methods, there are other mathematical approaches to infer the existence of autoregulation in gene expression [59, 86, 22, 73, 37, 91, 35, 34]. We introduce some works and compare them with our method. (**A**) Sanchez-Castillo et al. [59] considered an autoregressive model for multiple genes. This method (1) needs time series data; (2) requires the dynamics to be linear; (3) estimates a group of parameters. (**B**) Xing et al. [86] applied causal inference to a complicated gene expression model. This method (1) needs promoter sequences and information on transcription factor binding sites; (2) requires linearity for certain steps; (3) estimates a group of parameters. (**C**) Feigelman et al. [22] applied a Bayesian method for model selection. This method (1) needs time series data; (2) estimates a group of parameters. (**D**) Veerman et al. [73] considered the probability-generating function of a propagator model. This method (1) needs time series data; (2) estimates a group of parameters; (3) needs to approximate a Cauchy integral. (**E**) Jia et al. [37] compared the relaxation rate with degradation rate. This method (1) needs interventional data; (2) only works for a single gene that is not regulated by other genes; (3) requires that the per molecule degradation rate is a constant.

Compared to other more complicated methods, VMR-based methods (including ours) have two advantages: (1) VMR-based methods use non-interventional one-time data. Time series data require measuring the same cell multiple times without killing it, and interventional data require some techniques to interfere with gene expression, such as gene knockdown. Therefore, non-interventional one-time data used in VMR-based methods are much easier and cheaper to obtain. (2) VMR-based methods do not estimate parameters, and only calculate the mean and variance of the expression level. Some other methods need to estimate many parameters or approximate some complicated quantities, meaning that they need large data size and high data accuracy. Therefore, our method is easy to calculate, and need lower data accuracy and smaller data size.

Compared to other VMR-based methods, our method has few restrictions on the model, making them applicable to various scenarios with different dynamics. Besides, our deriva-

tions do not use any approximations.

In sum, compared to other VMR-based methods, our method is universal. Compared to other non-VMR-based methods, our method is simple, and has lower requirements on data quality.

Compared to other non-VMR-based methods, our method has some disadvantages: (1) The GRN structure needs to be known. (2) Our method does not work for certain genes, depending on regulatory relationships. Proposition 1 only works for a gene that is not regulated by other genes, and we require its expression to be autonomous; Proposition 2 only works for a gene that is not in a feedback loop. (3) Proposition 2 requires the per molecule degradation rate to be a constant, and it cannot provide information about autoregulation if VMR ≥ 1. (4) Our method only works for cells at equilibrium. Thus time series data that contain time-specific information cannot be utilized other than treated as one-time data. With just the stationary distribution, sometimes it is impossible to build the causal relationship (including autoregulation) [82]. Thus with this data type, some disadvantages are inevitable.

4 Scenario of a single isolated gene

4.1 Setup

We first consider the expression level (e.g., mRNA count) of one gene V in a single cell. At the single-cell level, gene expression is essentially stochastic, and we do not further consider dynamical system approaches with deterministic [79] or stochastic [89] operators. We use a random variable X to represent the mRNA count of V. We assume V is not in a feedback loop. We also assume all environmental factors and other genes that can affect X are kept at constant levels, so that we can focus on V alone. This can be achieved if no other genes point to gene V in the GRN, such as PIP3 in Fig. 1. Then we assume that the expression of V is autonomous, thus X satisfies a time-homogeneous Markov chain defined on $\mathbb{Z}^*$.

Assume that the mRNA synthesis rate at $X(t) = n-1$, namely the transition rate from $X = n-1$ to $X = n$, is $f_n \geq 0$. Assume that with n mRNA molecules, the degradation rate for each mRNA molecule is $g_n > 0$. Then the overall degradation rate at $X(t) = n$, namely the transition rate from $X = n$ to $X = n-1$, is $g_n n$. The associated master equation is

$$\begin{aligned}\frac{\mathrm{d}\mathbb{P}[X(t)=n]}{\mathrm{d}t} =& \mathbb{P}[X(t)=n+1]g_{n+1}(n+1) + \mathbb{P}[X(t)=n-1]f_n \\ &- \mathbb{P}[X(t)=n](f_{n+1}+g_n n).\end{aligned} \tag{1}$$

When f_n, g_n take specific forms, this master equation also corresponds to a branching process, so that related techniques can be applied [40]. Define the relative growth rate $h_n = f_n/g_n$. We assume that as time tends to infinity, the process reaches equilibrium, where (1) the stationary probability distribution $P_n = \lim_{t\to\infty} \mathbb{P}[X(t) = n]$ exists, and $P_n = P_{n-1}h_n/n$; (2) the mean $\lim_{t\to\infty} \mathbb{E}[X(t)]$ and the variance $\lim_{t\to\infty} \sigma^2[X(t)]$ are finite.

Such requirements can be satisfied under simple assumptions, such as assuming h_n has a finite upper bound [53, 80].

If $h_n > h_{n-1}$ for some n, then there exists positive autoregulation. If $h_n < h_{n-1}$ for some n, then there exists negative autoregulation. If there is no autoregulation, then h_n is a constant h, and the stationary distribution is Poissonian with parameter h. In this setting, positive autoregulation and negative autoregulation might coexist, meaning that $h_{n+1} < h_n$ for some n and $h_{n'+1} > h_{n'}$ for some n'.

4.2 Theoretical results

With single-cell non-interventional one-time gene expression data for one gene, we have the stationary distribution of the Markov chain X. We can infer the existence of autoregulation with the VMR of X, defined as $\mathrm{VMR}(X) = \sigma^2(X)/\mathbb{E}(X)$. The idea is that if we let f_n increase/decrease with n, and control g_n to make $\mathbb{E}(X)$ invariant, then the variance $\sigma^2(X)$ increases/decreases [76, Section 2.5.1]. We shall prove that $\mathrm{VMR} > 1$ implies the occurrence of positive autoregulation, and $\mathrm{VMR} < 1$ implies the occurrence of negative autoregulation. Notice that $\mathrm{VMR} > 1$ does not exclude the possibility that negative autoregulation exists for some expression level. This also applies to $\mathrm{VMR} < 1$ and positive autoregulation.

We can illustrate this result with a linear model:

Example 1. *Consider a Markov chain that satisfies Eq. 1, and set $f_n = k + b(n-1)$, $g_n = c$. Here b (can be positive or negative) is the strength of autoregulation, and c satisfies $c > 0$ and $c - b > 0$. We can calculate that $VMR = 1 + b/(c-b)$ (see Appendix B.1 for details). Therefore, $VMR > 1$ means positive autoregulation, $b > 0$; $VMR < 1$ means negative autoregulation, $b < 0$; $VMR = 1$ means no autoregulation, $b = 0$.*

Lemma 1. *Consider a Markov chain $X(t)$ that follows Eq. 1 with general transition coefficients f_n, g_n. Here $X(t)$ models the mRNA/protein count of one gene whose expression is autonomous. Calculate $VMR(X)$ at stationarity. (1) Assume $h_{n+1} \geq h_n$ for all n. We have $VMR(X) \geq 1$; moreover, $VMR(X) = 1$ if and only if $h_{n+1} = h_n$ for all n. (2) Assume $h_{n+1} \leq h_n$ for all n. We have $VMR(X) \leq 1$; moreover, $VMR(X) = 1$ if and only if $h_{n+1} = h_n$ for all n.*

We can take negation of Lemma 1 to obtain the following proposition.

Proposition 1. *In the setting of Lemma 1, (1) If $VMR(X) > 1$, then there exists at least one value of n for which $h_{n+1} > h_n$; thus this gene has positive autoregulation. (2) If $VMR(X) < 1$, then there exists at least one value of n for which $h_{n+1} < h_n$; thus this gene has negative autoregulation. (3) If $VMR(X) = 1$, then either (A) $h_{n+1} = h_n$ for all n, meaning that this gene has no autoregulation; or (B) $h_{n+1} < h_n$ for one n and $h_{n'+1} > h_{n'}$ for another n', meaning that this gene has both positive and negative autoregulation (at different expression levels).*

Remark 1. *Results similar to Proposition 1 have been proven by Jia et al. in another model of expression for a single gene [38]. However, they require that $g_i = g_j$ for any i, j. Proposition 1 can handle arbitrary g_i, thus being novel.*

Proof of Lemma 1. Define $\lambda = -\log P_0$, so that $P_0 = \exp(-\lambda)$. Define $d_n = \prod_{i=1}^{n} h_i > 0$ and stipulate that $d_0 = 1$. We can see that

$$\frac{d_n d_{n+2}}{d_{n+1}^2} = \frac{h_{n+2}}{h_{n+1}}.$$

Also,

$$P_n = P_{n-1} f_n/(g_n n) = P_{n-1} h_n/n = \cdots = P_0(\prod_{i=1}^{n} h_i)/n! \ = e^{-\lambda}\frac{d_n}{n!}.$$

Then

$$\begin{aligned}\mathbb{E}(X^2) - \mathbb{E}(X) &= \sum_{n=1}^{\infty}(n^2 - n)P_n = e^{-\lambda}\sum_{n=1}^{\infty}(n^2 - n)\frac{d_n}{n!} \\ &= e^{-\lambda}\sum_{n=2}^{\infty}\frac{d_n}{(n-2)!} = e^{-\lambda}\sum_{n=0}^{\infty}\frac{d_{n+2}}{n!},\end{aligned}$$

$$[\mathbb{E}(X)]^2 = \left(\sum_{n=1}^{\infty} nP_n\right)^2 = e^{-2\lambda}\left(\sum_{n=1}^{\infty} n\frac{d_n}{n!}\right)^2 = e^{-2\lambda}\left(\sum_{n=0}^{\infty}\frac{d_{n+1}}{n!}\right)^2.$$

Besides,

$$1 = \sum_{n=0}^{\infty} P_n = e^{-\lambda}\sum_{n=0}^{\infty}\frac{d_n}{n!}.$$

Now we have

$$\mathbb{E}(X^2) - \mathbb{E}(X) - [\mathbb{E}(X)]^2 = e^{-2\lambda}\left(\sum_{n=0}^{\infty}\frac{d_n}{n!}\right)\left(\sum_{n=0}^{\infty}\frac{d_{n+2}}{n!}\right) - e^{-2\lambda}\left(\sum_{n=0}^{\infty}\frac{d_{n+1}}{n!}\right)^2.$$

(1) Assume $h_{n+1} \geq h_n$ for all n. Then

$$\begin{aligned}&\mathbb{E}(X^2) - \mathbb{E}(X) - [\mathbb{E}(X)]^2 \\ \geq &e^{-2\lambda}\left(\sum_{n=0}^{\infty}\frac{\sqrt{d_n d_{n+2}}}{n!}\right)^2 - e^{-2\lambda}\left(\sum_{n=0}^{\infty}\frac{d_{n+1}}{n!}\right)^2 \geq 0.\end{aligned} \tag{2}$$

Here the first inequality is from the Cauchy-Schwarz inequality, and the second inequality is from $d_n d_{n+2} \geq d_{n+1}^2$ for all n. Then $\mathrm{VMR}(X) = \{\mathbb{E}(X^2) - [\mathbb{E}(X)]^2\}/\mathbb{E}(X) \geq 1$.

Equality holds if and only if $d_n/d_{n+2} = d_{n+1}/d_{n+3}$ for all n (the first inequality of Eq. 2) and $d_n d_{n+2} = d_{n+1}^2$ for all n (the second inequality of Eq. 2). The equality condition is equivalent to $h_{n+1} = h_n$ for all n.

(2) Assume $h_{n+1} \le h_n$ for all n. Then $d_{n+2}/d_{n+1} \le d_{n+1}/d_n$, and $d_n \le h_1^n$ for all n. Define

$$H(t) = \sum_{n=0}^{\infty} \frac{d_n}{n!} t^n.$$

Since $0 < d_n \le h_1^n$, this series converges for all $t \in \mathbb{C}$, so that $H(t)$ is a well-defined analytical function on $\mathbb{C}$, and

$$H'(t) = \sum_{n=0}^{\infty} \frac{d_{n+1}}{n!} t^n, \quad \text{and} \quad H''(t) = \sum_{n=0}^{\infty} \frac{d_{n+2}}{n!} t^n.$$

In the following, we only consider $H(t), H'(t), H''(t)$ as real functions for $t \in \mathbb{R}$.

To prove $\mathrm{VMR}(X) \le 1$, we just need to prove $\mathbb{E}(X^2) - \mathbb{E}(X) - [\mathbb{E}(X)]^2 = e^{-2\lambda}\{H(1)H''(1) - [H'(1)]^2\} \le 0$. However, we shall prove $H''(t)H(t) \le [H'(t)]^2$ for all $t \in \mathfrak{I}$, where $\mathfrak{I} = (a, b)$ is a fixed interval in $\mathbb{R}$ with $0 < a < 1$ and $1 < b < \infty$. Thus $t = 1$ is an interior point of $\mathfrak{I}$. Since $H(t), H'(t), H''(t)$ have positive lower bounds on $\mathfrak{I}$, the following statements are obviously equivalent: (i) $H''(t)H(t) \le [H'(t)]^2$ for all $t \in \mathfrak{I}$; (ii) $\{\log[H'(t)/H(t)]\}' \le 0$ for all $t \in \mathfrak{I}$; (iii) $\log[H'(t)/H(t)]$ is non-increasing on $\mathfrak{I}$; (iv) $H'(t)/H(t)$ is non-increasing on $\mathfrak{I}$. To prove (i), we just need to prove (iv).

Consider any $t_1, t_2 \in \mathfrak{I}$ with $t_1 \le t_2$ and any $p, q \in \mathbb{N}$ with $p \ge q$. Since $d_{p+1}/d_p \le d_{q+1}/d_q$, and $t_1^{p-q} \le t_2^{p-q}$, we have

$$d_p d_q t_1^q t_2^q \left(\frac{d_{p+1}}{d_p} - \frac{d_{q+1}}{d_q}\right)(t_1^{p-q} - t_2^{p-q}) \ge 0,$$

which means

$$d_{p+1} d_q t_1^p t_2^q + d_{q+1} d_p t_1^q t_2^p \ge d_{p+1} d_q t_2^p t_1^q + d_{q+1} d_p t_2^q t_1^p.$$

Sum over all $p, q \in \mathbb{N}$ with $p \ge q$ to obtain

$$\begin{aligned} H'(t_1)H(t_2) &= \left(\sum_{n=0}^{\infty} \frac{d_{n+1}}{n!} t_1^n\right)\left(\sum_{n=0}^{\infty} \frac{d_n}{n!} t_2^n\right) \\ &\ge \left(\sum_{n=0}^{\infty} \frac{d_{n+1}}{n!} t_2^n\right)\left(\sum_{n=0}^{\infty} \frac{d_n}{n!} t_1^n\right) = H'(t_2)H(t_1). \end{aligned}$$

Thus $H'(t_1)/H(t_1) \ge H'(t_2)/H(t_2)$ for all $t_1, t_2 \in \mathfrak{I}$ with $t_1 \le t_2$. This means $H''(t)H(t) \le [H'(t)]^2$ for all $t \in \mathfrak{I}$, and $\mathrm{VMR}(X) \le 1$.

About the condition for the equality to hold, assume $h_{n'+1} < h_{n'}$ for a given n'. Then

$$d_{n'} d_{n'-1} t_1^{n'-1} t_2^{n'-1} \left(\frac{d_{n'+1}}{d_{n'}} - \frac{d_{n'}}{d_{n'-1}}\right)(t_1 - t_2) \ge C(t_2 - t_1)$$

for all $t_1, t_2 \in \mathfrak{I}$ with $t_1 \leq t_2$ and a constant C that does not depend on t_1, t_2. Therefore,

$$\begin{aligned}
&[H'(t_1)/H(t_1) - H'(t_2)/H(t_2)] \cdot [H(t_1)H(t_2)] \\
=&(\sum_{n=0}^{\infty} \frac{d_{n+1}}{n!} t_1^n)(\sum_{n=0}^{\infty} \frac{d_n}{n!} t_2^n) - (\sum_{n=0}^{\infty} \frac{d_{n+1}}{n!} t_2^n)(\sum_{n=0}^{\infty} \frac{d_n}{n!} t_1^n) \\
\geq& d_{n'} d_{n'-1} t_1^{n'-1} t_2^{n'-1} (\frac{d_{n'+1}}{d_{n'}} - \frac{d_{n'}}{d_{n'-1}})(t_1 - t_2) \\
\geq& C(t_2 - t_1).
\end{aligned}$$

Since $H(t)$ has a finite positive upper bound A and a positive lower bound B on $\mathfrak{I}$, we have

$$H'(t_1)/H(t_1) - H'(t_2)/H(t_2) \geq C(t_2 - t_1)/A^2,$$

meaning that

$$\forall t \in \mathfrak{I},\ [H'(t)/H(t)]' = \{H(t)H''(t) - [H'(t)]^2\}/[H(t)]^2 \leq -C/A^2,$$

and thus

$$\forall t \in \mathfrak{I},\ H(t)H''(t) - [H'(t)]^2 \leq -CB^2/A^2 < 0.$$

Therefore, $\mathbb{E}(X^2) - \mathbb{E}(X) - [\mathbb{E}(X)]^2 = e^{-2\lambda}\{H(1)H''(1) - [H'(1)]^2\} < 0$, and $\mathrm{VMR}(X) < 1$.

We have proved in (1) that if $h_{n+1} = h_n$ for all n, then $\mathrm{VMR}(X) = 1$. Thus when $h_{n+1} \leq h_n$ for all n, $\mathrm{VMR}(X) = 1$ if and only if $h_{n+1} = h_n$ for all n. □

In sum, for the Markov chain model of one gene (by assuming the expression to be autonomous), when we have the stationary distribution from single-cell non-interventional one-time gene expression data, we can calculate the VMR of X. $\mathrm{VMR}(X) > 1$ means the existence of positive autoregulation (while negative autoregulation might still be possible at different expression levels), and $\mathrm{VMR}(X) < 1$ means the existence of negative autoregulation (while positive autoregulation might still be possible at different expression levels). $\mathrm{VMR}(X) = 1$ means either (1) no autoregulation exists; or (2) both positive autoregulation and negative autoregulation exist (at different expression levels). In reality, many genes are non-autonomous, and transcriptional/translational bursting can make the VMR to be larger than 100 [54]. Since Proposition 1 does not apply to non-autonomous cases, such genes might not have autoregulations.

5 Scenario of multiple entangled genes

5.1 Setup

We consider m genes $V_1, \ldots, V_m$ for a single cell. Denote their expression levels by random variables $X_1, \ldots, X_m$. The change of X_i can depend on X_j (mutual regulation) and X_i

itself (autoregulation). Since these genes regulate each other, and their expression levels are not fixed, we cannot consider them separately. If the expression of gene V_k is non-autonomous, we also need to add its interior factors (gene state and/or mRNA count) into $X_1, \ldots, X_m$.

We can use a continuous-time one-step Markov chain on $(\mathbb{Z}^*)^m$ to describe the dynamics. Each state of this Markov chain, $(X_1 = n_1, \ldots, X_i = n_i, \ldots, X_m = n_m)$, can be abbreviated as $\boldsymbol{n} = (n_1, \ldots, n_i, \ldots, n_m)$. For gene V_i, the transition rate of $n_i - 1 \to n_i$ is $f_i(\boldsymbol{n})$, and the transition rate of $n_i \to n_i - 1$ is $g_i(\boldsymbol{n})n_i$. Transitions with more than one step are not allowed. The master equation of this process is

$$\begin{aligned}\frac{\mathrm{d}\mathbb{P}(\boldsymbol{n})}{\mathrm{d}t} =& \sum_i \mathbb{P}(n_1, \ldots, n_i + 1, \ldots, n_m) g_i(n_1, \ldots, n_i + 1, \ldots, n_m)(n_i + 1) \\ &+ \sum_i \mathbb{P}(n_1, \ldots, n_i - 1, \ldots, n_m) f_i(\boldsymbol{n}) \\ &- \mathbb{P}(\boldsymbol{n}) \sum_i [f_i(n_1, \ldots, n_i + 1, \ldots, n_m) + g_i(\boldsymbol{n})n_i].\end{aligned} \tag{3}$$

Define $\boldsymbol{n}_{\bar{i}} = (n_1, \ldots, n_{i-1}, n_{i+1}, \ldots, n_m)$. Define $h_i(\boldsymbol{n}) = f_i(\boldsymbol{n})/g_i(\boldsymbol{n})$ to be the relative growth rate of gene V_i. Autoregulation means for some fixed $\boldsymbol{n}_{\bar{i}}$, $h_i(\boldsymbol{n})$ is (locally) increasing/decreasing with n_i, thus $f_i(\boldsymbol{n})$ increases/decreases and/or $g_i(\boldsymbol{n})$ decreases/increases with n_i. For the non-autonomous scenario, another possibility for autoregulation is that V_i can affect its interior factors (gene state and/or mRNA count).

5.2 Theoretical results

With expression data for multiple genes, there are various methods to infer the regulatory relationships between different genes, so that the GRN can be reconstructed [83]. In the GRN, if there is a directed path from gene V_i to gene V_j, meaning that V_i can directly or indirectly regulate V_j, then V_i is an ancestor of V_j, and V_j is a descendant of V_i.

Fix a gene V_k in a GRN. We consider a simple case that V_k is not contained in any directed cycle (feedback loop), which means no gene is both an ancestor and a descendant of V_k, such as PIP2 in Fig. 1. This means V_k itself is a strongly connected component of the GRN. This condition is automatically satisfied if the GRN has no directed cycle. If the expression of V_k is non-autonomous, we need to add the interior factors (gene state and/or mRNA count) of V_k into $V_1, \ldots, V_m$, and it is acceptable that V_k regulates its interior factors. In this case, if the one-step model holds, we can prove that if V_k does not regulate itself, meaning that $h_k(\boldsymbol{n})$ is a constant for fixed $\boldsymbol{n}_{\bar{k}}$ and different n_k, and X_k does not affect its interior factors (if non-autonomous), then $\mathrm{VMR}(X_k) \geq 1$. The reason is that $\mathrm{VMR} < 1$ requires either a feedback loop or autoregulation. Certainly, $\mathrm{VMR} < 1$ might also mean that the one-step model fails. One intuition is to assume the transitions of $V_{\bar{k}}$ are extremely slow, so that V_k is approximately the average of many Poisson variables. It is

easy to verify that the average of Poisson variables has VMR ≥ 1. We need to assume that the per molecule degradation rate $g_k(\cdot)$ for V_k is not affected by $V_1, \ldots, V_m$, which is not always true in reality [42]. With this result, when VMR < 1, there might be autoregulation.

Proposition 2. *Consider the one-step Markov chain model for multiple genes, described by Eq. 3. Assume the GRN has no directed cycle, or at least there is no directed cycle that contains gene V_k. Assume $g_k(\cdot)$ is a constant for all $\boldsymbol{n}$. If V_k has no autoregulation, meaning that $h_k(\cdot)$ and $f_k(\cdot)$ do not depend on n_k, and V_k does not regulate its interior factors (gene state and/or mRNA count), then V_k has VMR ≥ 1. Therefore, V_k has VMR < 1 means V_k has autoregulation, or the one-step model fails.*

Paulsson et al. study a similar problem [29, 87], and they state Proposition 2 in an unpublished work [personal communication from Dr. Jiawei Yan in Jan. 2022]. Proposition 2 also appears in a preprint by Mahajan et al. [49], but the proof is based on a linear noise approximation, which requires that $f_k(\cdot)$ is linear with $\boldsymbol{n}_{\bar{k}}$. We propose a rigorous proof independently.

Proof. Denote the expression level of V_k by W. Assume the ancestors of V_k are $V_1, \ldots, V_l$. For simplicity, denote the expression levels of $V_1, \ldots, V_l$ by a (high-dimensional) random variable Y. Assume V_k has no autoregulation. Since V_k does not regulate $V_1, \ldots, V_l$, W does not affect Y. Denote the transition rate from $Y = i$ to $Y = j$ by $q_{ij} \geq 0$. Stipulate that $q_{ii} = -\sum_{j \neq i} q_{ij}$. When $Y = i$, the transition rate from $W = n$ to $W = n + 1$ is F_i (does not depend on n), and the transition rate from $W = n$ to $W = n - 1$ is G.

The master equation of this process is

$$\begin{aligned}
&\frac{\mathrm{d}\mathbb{P}[W(t) = n, Y(t) = i]}{\mathrm{d}t} \\
=&\mathbb{P}[W(t) = n-1, Y(t) = i]F_i + \mathbb{P}[W(t) = n+1, Y(t) = i]G(n+1) \\
&+ \sum_{j \neq i} \mathbb{P}[W(t) = n, Y(t) = j]q_{ji} - \mathbb{P}[W(t) = n, Y(t) = i](F_i + Gn + \sum_{j \neq i} q_{ij}).
\end{aligned}$$

Assume there is a unique stationary probability distribution $P_{n,i} = \lim_{t \to \infty} \mathbb{P}[W(t) = n, Y(t) = i]$. This can be guaranteed by assuming the process to be irreducible. Then we have

$$P_{n,i}\Big[F_i + Gn + \sum_j q_{ij}\Big] = P_{n-1,i}F_i + P_{n+1,i}G(n+1) + \sum_j P_{n,j}q_{ji}. \tag{4}$$

Define $P_i = \sum_n P_{n,i}$. Sum over n for Eq. 4 to obtain

$$P_i \sum_j q_{ij} = \sum_j P_j q_{ji}, \tag{5}$$

meaning that P_i is the stationary probability distribution of Y.

Define W_i to be W conditioned on $Y = i$ at stationarity. Then $\mathbb{P}(W_i = n) = \mathbb{P}(W = n \mid Y = i) = P_{n,i}/P_i$, and $\mathbb{E}(W_i) = \sum_n nP_{n,i}/P_i$. Multiply Eq. 4 by n and sum over n to obtain

$$\Big(G + \sum_j q_{ij}\Big) P_i \mathbb{E}(W_i) = F_i P_i + \sum_j q_{ji} P_j \mathbb{E}(W_j). \tag{6}$$

Here and in the following, we repeatedly apply the tricks of splitting n and shifting the index of summation. For example,

$$\begin{aligned}
&\sum_{n=1}^{\infty} P_{n-1,i} F_i n - \sum_{n=1}^{\infty} P_{n,i} F_i n \\
=&\sum_{n=1}^{\infty} P_{n-1,i} F_i (n-1) + \sum_{n=1}^{\infty} P_{n-1,i} F_i - \sum_{n=1}^{\infty} P_{n,i} F_i n \\
=&\sum_{n-1=0}^{\infty} P_{n-1,i} F_i (n-1) + \sum_{n-1=0}^{\infty} P_{n-1,i} F_i - \sum_{n=0}^{\infty} P_{n,i} F_i n \\
=&\sum_{n=0}^{\infty} P_{n,i} F_i n + F_i \sum_{n=0}^{\infty} P_{n,i} - \sum_{n=0}^{\infty} P_{n,i} F_i n = F_i P_i.
\end{aligned}$$

Sum over i for Eq. 6 to obtain

$$G \sum_i P_i \mathbb{E}(W_i) = \sum_i F_i P_i. \tag{7}$$

Multiply Eq. 4 by n^2 and sum over n to obtain

$$\Big(2G + \sum_j q_{ij}\Big) P_i \mathbb{E}(W_i^2) = F_i P_i + (2F_i + G) P_i \mathbb{E}(W_i) + \sum_j q_{ji} P_j \mathbb{E}(W_j^2). \tag{8}$$

Sum over i for Eq. 8 to obtain

$$2G \sum_i P_i \mathbb{E}(W_i^2) = \sum_i F_i P_i + 2 \sum_i F_i P_i \mathbb{E}(W_i) + G \sum_i P_i \mathbb{E}(W_i). \tag{9}$$

Multiply Eq. 6 by $\mathbb{E}(W_i)$ and sum over i to obtain

$$\begin{aligned}
&G \sum_i P_i [\mathbb{E}(W_i)]^2 + \sum_{i,j} P_i q_{ij} [\mathbb{E}(W_i)]^2 \\
=&\sum_i F_i P_i \mathbb{E}(W_i) + \sum_{i,j} P_j q_{ji} \mathbb{E}(W_i) \mathbb{E}(W_j).
\end{aligned} \tag{10}$$

Then we have

$$
\begin{aligned}
&\sum_i F_i P_i \mathbb{E}(W_i) - G\sum_i P_i[\mathbb{E}(W_i)]^2 \\
=&\sum_{i,j} P_i q_{ij}[\mathbb{E}(W_i)]^2 - \sum_{i,j} P_j q_{ji}\mathbb{E}(W_i)\mathbb{E}(W_j) \\
=&\frac{1}{2}\Big\{\sum_{i,j} P_i q_{ij}[\mathbb{E}(W_i)]^2 + \sum_i [\mathbb{E}(W_i)]^2 \sum_j P_i q_{ij} - 2\sum_{i,j} P_i q_{ij}\mathbb{E}(W_i)\mathbb{E}(W_j)\Big\} \\
=&\frac{1}{2}\Big\{\sum_{i,j} P_i q_{ij}[\mathbb{E}(W_i)]^2 + \sum_i [\mathbb{E}(W_i)]^2 \sum_j P_j q_{ji} - 2\sum_{i,j} P_i q_{ij}\mathbb{E}(W_i)\mathbb{E}(W_j)\Big\} \\
=&\frac{1}{2}\Big\{\sum_{i,j} P_i q_{ij}[\mathbb{E}(W_i)]^2 + \sum_{i,j} P_i q_{ij}[\mathbb{E}(W_j)]^2 - 2\sum_{i,j} P_i q_{ij}\mathbb{E}(W_i)\mathbb{E}(W_j)\Big\} \\
=&\frac{1}{2}\sum_{i,j} P_i q_{ij}[\mathbb{E}(W_i) - \mathbb{E}(W_j)]^2 \geq 0.
\end{aligned}
\tag{11}
$$

Here the first equality is from Eq. 10, the third equality is from Eq. 5, and other equalities are equivalent transformations.

Now we have

$$
\begin{aligned}
&\mathbb{E}(W^2) - \mathbb{E}(W) - [\mathbb{E}(W)]^2 \\
=&\sum_i P_i \mathbb{E}(W_i^2) - \sum_i P_i \mathbb{E}(W_i) - \Big[\sum_i P_i \mathbb{E}(W_i)\Big]^2 \\
=&\frac{1}{G}\sum_i F_i P_i \mathbb{E}(W_i) + \sum_i P_i \mathbb{E}(W_i) - \sum_i P_i \mathbb{E}(W_i) - \Big[\sum_i P_i \mathbb{E}(W_i)\Big]^2 \\
\geq&\sum_i P_i[\mathbb{E}(W_i)]^2 - \Big[\sum_i P_i \mathbb{E}(W_i)\Big]^2 \\
=&\Big(\sum_i P_i\Big)\sum_i P_i[\mathbb{E}(W_i)]^2 - \Big[\sum_i P_i \mathbb{E}(W_i)\Big]^2 \geq 0,
\end{aligned}
\tag{12}
$$

where the first equality is by definition, the second equality is from Eqs. 7,9, the first inequality is from Eq. 11, the third equality is from $\sum_i P_i = 1$, and the second inequality is the Cauchy-Schwarz inequality.

Since $\mathbb{E}(W^2) - [\mathbb{E}(W)]^2 \geq \mathbb{E}(W)$, $\mathrm{VMR}(W) = \{\mathbb{E}(W^2) - [\mathbb{E}(W)]^2\}/\mathbb{E}(W) \geq 1$. □

Remark 2. *In gene expression, the total noise ($\sigma^2(X)/(\mathbb{E}X)^2$) can be decomposed into intrinsic (cellular) noise and extrinsic (environmental) noise [3, 71, 27, 47, 75]. Inspired by that, we can decompose the VMR into intrinsic and extrinsic components. Denote intrinsic and extrinsic stochastic factors as I, E, and the expression level X is a deterministic*

function of these factors: $X = X(I, E)$. Then

$$VMR_{int} = \frac{\mathbb{E}_E(\mathbb{E}_{I|E}X^2) - \mathbb{E}_E(\mathbb{E}_{I|E}X)^2}{\mathbb{E}X},$$

$$VMR_{ext} = \frac{\mathbb{E}_E(\mathbb{E}_{I|E}X)^2 - [\mathbb{E}_E(\mathbb{E}_{I|E}X)]^2}{\mathbb{E}X},$$

where $\mathbb{E}_{I|E}$ is the expectation conditioned on E. This decomposition might lead to further understanding of Proposition 2.

We hypothesize that the requirement for $g_k(\cdot)$ in Proposition 2 can be dropped:

Conjecture 1. *Assume V_k is not contained in a directed cycle in the GRN, and V_k does not regulate its interior factors (gene state and/or mRNA count). If V_k has no autoregulation, meaning that $h_k(\cdot)$ does not depend on n_k (but might depend on $\boldsymbol{n}_{\bar{k}}$), then V_k has VMR ≥ 1.*

The main obstacle of proving this conjecture is that the second equality in Eq. 12 does not hold. The reason is that G_i cannot be extracted from the summation, and we cannot link $\sum_i P_i\mathbb{E}(W_i^2)$ and $\sum_i G_iP_i\mathbb{E}(W_i^2)$.

If the GRN has directed cycles, there is a result by Paulsson et al. [29, 87], which is proved under first-order approximations of covariances. The general case (when the approximations do not apply) has been numerically verified but not proved yet:

Conjecture 2. *Assume for each V_i, $g_i(\cdot)$ does not depend on $\boldsymbol{n}$, and $f_i(\cdot)$ does not depend on n_i (no autoregulation). Then for at least one gene V_j, we have VMR ≥ 1 [29, 87].*

Due to the existence of directed cycles, one gene can affect itself through other genes, and we cannot study them separately.

Notice that Conjecture 2 does not hold if g_i depends on $\boldsymbol{n}_{\bar{i}}$:

Example 2. *Consider a one-step Markov chain that satisfies Eq. 3, where $m = 2$, $f_1(n_2) = g_1(n_2) = 1$ for $n_2 = 2$, $f_1(n_2) = g_1(n_2) = 0$ for $n_2 \neq 2$, and $f_2(n_1) = g_2(n_1) = 1$ for $n_1 = 2$, $f_2(n_1) = g_2(n_1) = 0$ for $n_1 \neq 2$. The initial state is $(n_1 = 2, n_2 = 2)$. Then VMR $= 2e/(4e-1) \approx 0.55$ for both genes (see Appendix B.2 for details).*

Assume Conjecture 2 is correct. For m genes, if we find that VMR for each gene is less than 1, then we can infer that autoregulation exists, although we do not know which gene has autoregulation. Another possibility is that the one-step model fails.

6 Applying theoretical results to experimental data

We summarize our theoretical results into Algorithm 1. Proposition 1 applies to a gene that has no ancestor in the GRN. However, it requires the corresponding gene has autonomous expression (or the transition rates of gene states are high enough, so that the

1. **Input**

   Single-cell non-interventional one-time expression data for genes $V_1, \ldots, V_m$

   The structure of the GRN that contains $V_1, \ldots, V_m$

2. **Calculate** the VMR of each V_k

3. **If** V_k is not in a directed cycle (like PIP2 in Fig. 1) and VMR < 1

   Output V_k has autoregulation

   // Assume the degradation of V_k is not regulated by $V_1, \ldots, V_m$

   Else

   If V_k has no ancestor in the GRN (like PIP3 in Fig. 1) and VMR > 1

   Output V_k has autoregulation

   //Assume the expression of V_k is autonomous

   Else

   Output We cannot determine whether V_k has autoregulation

   End of if

   End of if

Algorithm 1: Detailed workflow of inferring autoregulation with gene expression data.

non-autonomous process is close to an autonomous process), which is difficult to validate and often does not hold in reality. Thus the inference result by Proposition 1 for VMR > 1 (positive autoregulation) is not very reliable. When VMR < 1 and Proposition 1 could apply, we should instead apply Proposition 2 to determine the existence of autoregulation, since Proposition 2 does not require the expression to be autonomous, thus being much more reliable, although it may fail if the one-step model does not hold. Proposition 2 applies when the gene is not in a feedback loop and has VMR < 1. Notice that our result cannot determine that a gene has no autoregulation.

For a given gene without autoregulation, its expression level satisfies a Poisson distribution, and VMR is 1. If we have n samples of its expression level, then the sample VMR (sample variance divided by sample mean) asymptotically satisfies a Gamma distribution $\Gamma[(n-1)/2, 2/(n-1)]$, and we can determine the confidence interval of sample VMR [20]. If the sample VMR is out of this confidence interval, then we know that VMR is significantly different from 1, and Propositions 1,2 might apply.

We apply our method to four groups of single-cell non-interventional one-time gene expression data from experiments, where the corresponding GRNs are known. Notice that we need to convert indirect measurements into protein/mRNA count. See Table 1 for our inference results and theoretical/experimental evidence that partially validates our results. See Appendix C for details. There are 186 genes in these four data sets, and we can only determine that 12 genes have autoregulation (7 genes determined by Proposition 1, and 5 genes determined by Proposition 2). Not every VMR is less than 1, so that Conjecture 2 does not apply. For the other 174 genes, (1) some of them are not contained in the known GRN, and we cannot determine if they are in directed cycles; (2) some of them are in directed cycles; (3) some of them have ancestors, and we cannot reject the hypothesis that VMR ≥ 1; (4) some of them have no ancestors, and we cannot reject the hypothesis that VMR $= 1$. Therefore, Proposition 1 and Proposition 2 do not apply, and we do not know whether they have autoregulation.

In some cases, we have experimental evidence that some genes have autoregulation, so that we can partially validate our inference results. Nevertheless, as discussed in the Introduction, there is no gold standard to evaluate our inference results. Besides, Proposition 2 requires that the one-step model holds, which we cannot verify.

In the data set by Guo et al. [26], Sanchez-Castillo et al. [59] inferred that 17 of 39 genes have autoregulation, and 22 genes do not have autoregulation. We infer that 5 genes have autoregulation, and 34 genes cannot be determined. Here 3 genes are shared by both inference results to have autoregulation. Consider a random classifier that randomly picks 5 genes and claims they have autoregulation. Using Sanchez-Castillo et al. as the standard, this random classifier has probability 62.55% to be worse than our result, and 10.17% to be better than our result. Thus our inference result is better than a random classifier, but the advantage is not substantial.

Source	Proposition 1	Proposition 2	Theory	Experiment
Guo et al. [26]	FN1 **HNF4A**	**TCFAP2C** **BMP4** CREB312	BMP4 [59] HNF4A [59] TCFAP2C [59]	BMP4 [55] HNF4A [12] TCFAP2C [44]
Psaila et al. [56]	BIM CCND1 **ECT2** PFKP			ECT2 [28]
Moignard et al. [50]		EIF2B1 HOXD8		
Sachs et al. [58]	PIP3			

Table 1: The autoregulation inference results by our method on four data sets. Source column is the paper that contains this data set. Proposition 1 column is the genes that can be only inferred by Proposition 1 to have autoregulation. Proposition 2 column is the genes that can be inferred by Proposition 2 to have autoregulation. Theory column is the genes inferred by both our method and other theoretical works to have autoregulation. Experiment column is the genes inferred by both our method and other experimental works to have autoregulation. **Bold** font means the inferred gene with autoregulation is validated by other results. Details can be found in Appendix C.

7 Conclusions

For a single gene that is not affected by other genes, or a group of genes that form a connected GRN, we develop rigorous theoretical results (without applying approximations) to determine the existence of autoregulation. These results generalize known relationships between autoregulation and VMR by dropping restrictions on parameters. Our results only depend on VMR, which is easy to compute and more robust than other complicated statistics. We also apply our method to experimental data and detect some genes that might have autoregulation.

Our method requires **independent** and **identically** distributed samples from the **exact stationary** distribution of a **fully observed** Markov chain, plus a known **GRN**. Proposition 1 requires that the expression is **autonomous**. Proposition 2 requires that the Markov chain model is **one-step**, the GRN has **no** directed **cycle**, and **degradation** is **not regulated**. If our inference fails, then some requirements are not met: (1) cells might affect each other, making the samples dependent; (2) cells are heterogeneous; (3) the measurements have extra errors; (4) the cells are not at stationarity; (5) there exist unobserved variables that affect gene expression; (6) the GRN is inferred by a theoretical method, which can be interfered by the existence of autoregulation; (7) the expression is non-autonomous; (8) the Markov chain is multi-step; (9) the GRN has unknown directed cycles; (10) the degradation rate is regulated by other genes. Such situations, especially the unobserved variables, are unavoidable. Therefore, current data might not satisfy these requirements, and our inference results should be interpreted as informative findings, not ground truths.

There are some known methods that overcome the above obstacles, and there are also some possible solutions that might appear in the future. (1) The dependency can be solved by better measurements for isolated cells that do not affect each other. In fact, the relationship between autoregulation and cell-cell interaction has been studied [46]. (2) About cell heterogeneity, we prove a result in Appendix D that if several cell types have VMR ≥ 1, then for a mixed population of such cell types, we still have VMR ≥ 1. Therefore, cell heterogeneity does not fail Proposition 2, since VMR < 1 for the mixture of several cell types means VMR < 1 for at least one cell type. (3) With the development of experimental technologies, we expect that the measurement error can decrease. (4) Some works study autoregulation in non-stationary situations [10, 69, 66, 36]. (5) Since hidden variables hurt any mechanism-based models, we can develop methods (especially with machine learning tools) that determine autoregulation based on similarities between gene expression profiles [84, 88, 77]. (6) Some GRN inference methods can also determine the existence of autoregulation [59]. (7) Many methods (including our Proposition 2) work in non-autonomous situations. (8) Some works study multi-step models [7, 43, 74]. (9) We expect the appearance of more advanced GRN inference methods. (10) If probabilists can prove Conjecture 1, then the restriction on degradation rate can be lifted.

In fact, other theoretical works that determine gene autoregulation, or general gene

	VMR < 1	VMR $= 1$ (true)	VMR > 1
$n = 100$	2.5%	95.3%	2.2%
$n = 1000$	2.3%	95.2%	2.5%
$n = 10000$	2.5%	95.1%	2.4%

Table 2: Success rates for determining VMR in Simulation 1. For different sample sizes n, calculate the rate that VMR is different from 1.

regulation, also need various assumptions and might fail. Nevertheless, with the development of experimental technologies and theoretical results, we believe that some obstacles will be lifted, and our method will be more applicable in the future. Besides, our method can be further developed and combined with other methods.

Acknowledgments

Y.W. would like to thank Jiawei Yan for fruitful discussions, and Xiangting Li, Zikun Wang, Mingtao Xia for helpful comments.

Declarations

- Funding: This research was partially supported by NIH grant R01HL146552 (Y.W.).
- Competing interests: The authors declare no conflict of interest.
- Data and code availability: All code files are available in https://github.com/YueWangMathbio/Autoregulation.

A Simulation results

A.1 Test Proposition 1 without autoregulation

Simulation 1: Consider the Markov chain in Example 1 with $k = 1, b = 0, c = 1$. The stationary distribution is Poissonian with parameter 1. This process has no autoregulation with the true VMR $= 1$. For sample sizes $n = 100$, $n = 1000$, $n = 10000$, we repeat the experiment for 10000 times and calculate the rate that the sample VMR falls in the 95% confidence interval. See Table 2 for results. Proposition 1 has about 95% probability to produce the correct result that VMR $= 1$ (no autoregulation), since the confidence interval is 95%.

	VMR < 1	VMR $= 1$	VMR > 1 (true)
$n = 100$	0%	1.9%	98.1%
$n = 1000$	0%	0%	100%
$n = 10000$	0%	0%	100%

Table 3: Success rates for determining VMR in Simulation 2. For different sample sizes n, calculate the rate that VMR is different from 1.

	VMR < 1	VMR $= 1$	VMR > 1 (true)
$n = 100$	0.2%	81.6%	18.2%
$n = 1000$	0%	7.7%	92.3%
$n = 10000$	0%	0%	100%

Table 4: Success rates for determining VMR in Simulation 3. For different sample sizes n, calculate the rate that VMR is different from 1.

A.2 Test Proposition 1 with autoregulation

Simulation 2: Consider the Markov chain in Example 1 with $k = 1, b = 1, c = 2$. The stationary distribution is geometric with parameter 0.5. This process has positive autoregulation with the true VMR $= 2$. For sample sizes $n = 100$, $n = 1000$, $n = 10000$, we repeat the experiment for 10000 times and calculate the rate that the sample VMR falls in the 95% confidence interval. See Table 3 for results. When n is not too small, Proposition 1 always produces the correct result that VMR > 1 (positive autoregulation).

A.3 Test Proposition 2 without autoregulation

Simulation 3: Consider a Markov chain (G, M) that satisfies Eq. 3. G can take 0 and 1, and M can take values in $\mathbb{Z}$. G does not depend on M, and transition rates are both 10^{-10} for $G = 0 \to G = 1$ and $G = 1 \to G = 0$. Restricted on $G = 0$, M is the same as Example 1 with $k = 1, b = 0, c = 1$. Restricted on $G = 1$, M is the same as Example 1 with $k = 2, b = 0, c = 1$. The stationary distribution is the average of two Poisson distributions with parameters 1 and 2. This process has no autoregulation with the true VMR $= 1.167$. For sample sizes $n = 100$, $n = 1000$, $n = 10000$, we repeat the experiment for 10000 times and calculate the rate that the sample VMR falls in the 95% confidence interval. See Table 4 for results. When n increases, we are very likely to obtain the correct VMR > 1, but Proposition 2 cannot determine whether autoregulation exists.

A.4 Test Proposition 2 with negative autoregulation

Simulation 4: Consider a Markov chain (G, M) that satisfies Eq. 3. G can take 0 and 1, and M can take values in $\mathbb{Z}$. G does not depend on M, and transition rates are both 10^{-10} for $G = 0 \to G = 1$ and $G = 1 \to G = 0$. Restricted on $G = 0$, M is the same as

	VMR < 1 (true)	VMR $= 1$	VMR > 1
$n = 100$	56.6%	43.4%	0%
$n = 1000$	100%	0%	0%
$n = 10000$	100%	0%	0%

Table 5: Success rates for determining VMR in Simulation 4. For different sample sizes n, calculate the rate that VMR is different from 1.

	VMR < 1	VMR $= 1$	VMR > 1 (true)
$n = 100$	0%	51.8%	48.2%
$n = 1000$	0%	0%	100%
$n = 10000$	0%	0%	100%

Table 6: Success rates for determining VMR in Simulation 5. For different sample sizes n, calculate the rate that VMR is different from 1.

Example 1 with $k = 10, b = -1, c = 1$. Restricted on $G = 1$, M is the same as Example 1 with $k = 10, b = -1, c = 2$. This process has (negative) autoregulation with the true VMR $= 0.733$. For sample sizes $n = 100$, $n = 1000$, $n = 10000$, we repeat the experiment for 10000 times and calculate the rate that the sample VMR falls in the 95% confidence interval. See Table 5 for results. When n is not too small, Proposition 2 always produces the correct result that VMR < 1 (autoregulation).

A.5 Test Proposition 2 with positive autoregulation

Simulation 5: Consider a Markov chain (G, M) that satisfies Eq. 3. G can take 0 and 1, and M can take values in $\mathbb{Z}$. G does not depend on M, and transition rates are both 10^{-10} for $G = 0 \to G = 1$ and $G = 1 \to G = 0$. Restricted on $G = 0$, M is the same as Example 1 with $k = 10, b = 1, c = 4$. Restricted on $G = 1$, M is the same as Example 1 with $k = 10, b = 1, c = 5$. This process has (positive) autoregulation with the true VMR $= 1.357$. For sample sizes $n = 100$, $n = 1000$, $n = 10000$, we repeat the experiment for 10000 times and calculate the rate that the sample VMR falls in the 95% confidence interval. See Table 6 for results. When n is not too small, we can always obtain the correct result that VMR > 1, but Proposition 2 cannot determine whether autoregulation exists.

B Details of examples

B.1 Details of Example 1

In Example 1, the stationary distribution P_n exists, and satisfies

$$[b(n-1) + k]P_{n-1} = cnP_n. \tag{13}$$

Taking summation for Eq. (13), we have

$$\begin{aligned}
&\sum_{n=1}^{\infty}[b(n-1)+k]P_{n-1} = \sum_{n=1}^{\infty} cnP_n, \\
\Rightarrow &b\sum_{n=0}^{\infty} nP_n + k\sum_{n=0}^{\infty} P_n = c\sum_{n=0}^{\infty} nP_n, \\
\Rightarrow &b\mathbb{E}X + k = c\mathbb{E}X.
\end{aligned}$$

Thus $\mathbb{E}X = k/(c-b)$.

Also, multiplying Eq. (13) by $n-1$ and taking summation, we have

$$\begin{aligned}
&\sum_{n=1}^{\infty}[b(n-1)^2+k(n-1)]P_{n-1} = \sum_{n=1}^{\infty} cn^2P_n - \sum_{n=1}^{\infty} cnP_n, \\
\Rightarrow &b\sum_{n=0}^{\infty} n^2P_n + k\sum_{n=0}^{\infty} nP_n = c\sum_{n=0}^{\infty} n^2P_n - c\sum_{n=0}^{\infty} nP_n, \\
\Rightarrow &b\mathbb{E}(X^2) + k\mathbb{E}X = c\mathbb{E}(X^2) - c\mathbb{E}X.
\end{aligned}$$

Thus $\mathbb{E}(X^2) = (c+k)k/(c-b)^2$, and the variance $\sigma^2(X) = ck/(c-b)^2$. Then $\mathrm{VMR}(X) = 1 + b/(c-b)$.

B.2 Details of Example 2

In Example 2, the process is restricted to two lines: $n_1 = 2$ and $n_2 = 2$. Since this process has no cycle, it is detailed balanced [81], and the stationary distribution $P(n_1, n_2)$ satisfies

$$P(k,2) = (k+1)P(k+1,2)$$

and

$$P(2,k) = (k+1)P(2,k+1).$$

Restricted on $n_1 = 2$ or $n_2 = 2$, the stationary distribution is Poissonian, $P(k,2) = c/(ek!)$. After normalization, we find that $c = 2e/(4e-1)$. Thus

$$P(k,2) = P(2,k) = \frac{2}{k!(4e-1)}.$$

Besides,

$$P(2,\cdot) = \sum_{k=0}^{\infty} P(2,k) = \sum_{k=0}^{\infty} \frac{2}{k!(4e-1)} = \frac{2e}{4e-1}.$$

Then for the first gene X, we have

$$\begin{aligned}\mathbb{E}X &= \sum_{k\neq 2} kP(k,2) + 2P(2,\cdot) = \sum_{k=0}^{\infty} kP(k,2) + 2[P(2,\cdot) - P(2,2)] \\ &= \sum_{k=0}^{\infty} \frac{2k}{k!(4e-1)} + \frac{4e-2}{4e-1} = \frac{6e-2}{4e-1},\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}(X^2) &= \sum_{k\neq 2} k^2P(k,2) + 4P(2,\cdot) = \sum_{k=0}^{\infty} k^2P(k,2) + 4[P(2,\cdot) - P(2,2)] \\ &= \sum_{k=0}^{\infty} \frac{2k(k-1)}{k!(4e-1)} + \sum_{k=0}^{\infty} \frac{2k}{k!(4e-1)} + \frac{8e-4}{4e-1} = \frac{12e-4}{4e-1}.\end{aligned}$$

Thus

$$\mathrm{VMR}(X) = \frac{\mathbb{E}(X^2)}{\mathbb{E}X} - \mathbb{E}X = 2 - \frac{6e-2}{4e-1} = \frac{2}{4e-1} \approx 0.55.$$

Due to symmetry, the other gene also has $\mathrm{VMR}(Y) \approx 0.55$.

C Details of applications on experimental data

In experiments, the expression levels of genes are not directly measured as mRNA or protein counts. Rather, they are measured as cycle threshold (Ct) values or fluorescence intensity values. Such indirect measurements need to be converted. Related details can be found in other papers [38].

Guo et al. [26] measured the expression (mRNA) levels of 48 genes for mouse embryo cells at different developmental stages. We consider three groups (16-cell stage, 32-cell stage, 64-cell stage) that have more than 50 samples. Sanchez-Castillo et al. [59] used such data to infer the GRN structure, including autoregulation, but the inferred GRN only contains 39 genes. We cannot guarantee that the other 9 genes have no ancestors in the true GRN (to apply Proposition 1) or these genes are not contained in directed cycles (to apply Proposition 2). Thus we ignore those 9 genes not in this GRN. In the inferred GRN, genes BMP4, CREB312, and TCFAP2C are not contained in directed cycles. In the 16-cell stage group with 75 samples, if there is no autoregulation, then the 95% confidence interval of VMR is $[0.7041, 1.3470]$. BMP4 (VMR $= 0.2139$), CREB312 (VMR $= 0.1971$), and TCFAP2C (VMR $= 0.3468$) have significantly small VMR, and we can apply Proposition 2 to infer that BMP4, CREB312, and TCFAP2C might have autoregulation. In the other two groups, these genes do not have VMR < 1, and the results are relatively weak. Besides, in the inferred GRN, genes FN1 and HNF4A have no ancestors. For the 16-cell stage with 75

samples, the VMR of FN1 and HNF4A are 3.4522 and 1.3599, outside of the 95% confidence interval $[0.7041, 1.3470]$; for the 32-cell stage with 113 samples, the VMR of FN1 and HNF4A are 93.1070 and 46.7688, outside of the 95% confidence interval $[0.7554, 1.2784]$; for the 64-cell stage with 159 samples, the VMR of FN1 and HNF4A are 117.3059 and 93.9589, outside of the 95% confidence interval $[0.7917, 1.2322]$. Thus we can apply Proposition 1 to infer that FN1 and HNF4A (VMR > 1 for all three cell groups) might have positive autoregulation. Nevertheless, it is more likely that the expressions of FN1 and HNF4A are non-autonomous, and there is no autoregulation. Sanchez-Castillo et al. [59] inferred that BMP4, HNF4A, TCFAP2C have autoregulation. Besides, there is experimental evidence that BMP4 [55], HNF4A [12], TCFAP2C [44] have autoregulation. Therefore, our inference results are partially validated.

Psaila et al. [56] measured the expression (mRNA) levels of 90 genes for human megakaryocyte-erythroid progenitor cells. Chan et al. [13] inferred the GRN structure (autoregulation not included). In the inferred GRN, genes BIM, CCND1, ECT2, PFKP have no ancestors. BIM has 214 effective samples, and VMR is 187.7, outside of the 95% confidence interval $[0.8191, 1.1987]$. CCND1 has 68 effective samples, and VMR is 111.3, outside of the 95% confidence interval $[0.6905, 1.3660]$. ECT2 has 56 effective samples, and VMR is 8.2, outside of the 95% confidence interval $[0.6618, 1.4069]$. PFKP has 134 effective samples, and VMR is 82.1, outside of the 95% confidence interval $[0.7742, 1.2543]$. Thus we can apply Proposition 1 to infer that BIM, CCND1, ECT2, PFKP might have positive autoregulation. Nevertheless, it is more likely that the expressions of these four genes are non-autonomous, and there is no autoregulation. There is experimental evidence that ECT2 has autoregulation [28], which partially validates our inference results. No other gene fits the requirement of Proposition 2.

Moignard et al. [50] measured the expression (mRNA) levels of 46 genes for mouse embryo cells. Chan et al. [13] inferred the GRN structure (autoregulation not included). Gene EIF2B1 has 3934 effective samples, and VMR is 0.66, outside of the 95% confidence interval $[0.9563, 1.0447]$. Gene EIF2B1 has 12 effective samples, and VMR is 0.24, outside of the 95% confidence interval $[0.3469, 1.9927]$. We can apply Proposition 2 to infer that EIF2B1 and HOXD8 might have autoregulation. No other gene fits the requirement of Proposition 1.

Sachs et al. [58] measured the expression (protein) levels of 11 genes in the RAF signaling pathway for human T cells. The measurements were repeated for 14 groups of cells under different interventions. Werhli et al. [85] inferred the GRN structure (autoregulation not included). In the inferred GRN (Fig. 1), PIP3 gene has no ancestor, and its VMRs in all 14 groups are larger than 5, while the 95% confidence intervals for all 14 groups are contained in $[0.8, 1.2]$. Therefore, we can apply Proposition 1 and infer that PIP3 might have positive autoregulation. Nevertheless, it is more likely that the expression of PIP3 is non-autonomous, and there is no autoregulation. No other gene fits the requirement of Proposition 2.

D Heterogeneity and VMR

Proposition 3. *Consider n independent random variables $X_1, \ldots, X_n$ and probabilities $p_1, \ldots, p_n$ with $\sum p_i = 1$. Consider an independent random variable R that equals i with probability p_i. Construct a random variable Z that equals X_i when $R = i$. If each X_i has VMR ≥ 1, then Z has VMR ≥ 1.*

Proof. We only need to prove this for $n = 2$. The case for general n can be proved by mixing two variables iteratively.

Consider random variables X, Y and construct Z that equals X or Y with probability p or $1-p$. Since $\mathrm{VMR}(X) \geq 1$, $\mathrm{VMR}(Y) \geq 1$, we have $\mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \geq \mathbb{E}(X)$ and $\mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 \geq \mathbb{E}(Y)$. Then

$$
\begin{aligned}
&\mathrm{VMR}(Z) \\
=&\frac{p\mathbb{E}(X^2) + (1-p)\mathbb{E}(Y^2)}{p\mathbb{E}(X) + (1-p)\mathbb{E}(Y)} \\
&+ \frac{-p^2[\mathbb{E}(X)]^2 - 2p(1-p)\mathbb{E}(X)\mathbb{E}(Y) - (1-p)^2[\mathbb{E}(Y)]^2}{p\mathbb{E}(X) + (1-p)\mathbb{E}(Y)} \\
=&\frac{p\mathbb{E}(X^2) - p[\mathbb{E}(X)]^2 + (1-p)\mathbb{E}(Y^2) - (1-p)[\mathbb{E}(Y)]^2}{p\mathbb{E}(X) + (1-p)\mathbb{E}(Y)} \\
&+ \frac{p(1-p)[\mathbb{E}(X)]^2 - 2p(1-p)\mathbb{E}(X)\mathbb{E}(Y) + p(1-p)[\mathbb{E}(Y)]^2}{p\mathbb{E}(X) + (1-p)\mathbb{E}(Y)} \\
\geq&\frac{p\mathbb{E}(X) + (1-p)\mathbb{E}(Y)}{p\mathbb{E}(X) + (1-p)\mathbb{E}(Y)} + \frac{p(1-p)[\mathbb{E}(X) - \mathbb{E}(Y)]^2}{p\mathbb{E}(X) + (1-p)\mathbb{E}(Y)} \\
\geq&1.
\end{aligned}
$$

□

References

[1] Angelini, E., Wang, Y., Zhou, J. X., Qian, H., and Huang, S. A model for the intrinsic limit of cancer therapy: Duality of treatment-induced cell death and treatment-induced stemness. *PLOS Comput. Biol. 18*, 7 (2022), e1010319.

[2] Barros, R., da Costa, L. T., Pinto-de Sousa, J., Duluc, I., Freund, J.-N., David, L., and Almeida, R. CDX2 autoregulation in human intestinal metaplasia of the stomach: impact on the stability of the phenotype. *Gut 60*, 3 (2011), 290–298.

[3] Baudrimont, A., Jaquet, V., Wallerich, S., Voegeli, S., and Becskei, A. Contribution of RNA degradation to intrinsic and extrinsic noise in gene expression. *Cell Rep. 26*, 13 (2019), 3752–3761.

[4] Baumdick, M., Gelléri, M., Uttamapinant, C., Beránek, V., Chin, J. W., and Bastiaens, P. I. A conformational sensor based on genetic code expansion reveals an autocatalytic component in EGFR activation. *Nat. Commun. 9*, 1 (2018), 1–13.

[5] Bokes, P., King, J. R., Wood, A. T., and Loose, M. Multiscale stochastic modelling of gene expression. *J. Math. Biol. 65*, 3 (2012), 493–520.

[6] Bouuaert, C. C., Lipkow, K., Andrews, S. S., Liu, D., and Chalmers, R. The autoregulation of a eukaryotic DNA transposon. *eLife 2* (2013).

[7] Braichenko, S., Holehouse, J., and Grima, R. Distinguishing between models of mammalian gene expression: telegraph-like models versus mechanistic models. *J. R. Soc. Interface 18*, 183 (2021), 20210510.

[8] Cagnetta, R., Wong, H. H.-W., Frese, C. K., Mallucci, G. R., Krijgsveld, J., and Holt, C. E. Noncanonical modulation of the eIF2 pathway controls an increase in local translation during neural wiring. *Mol. Cell 73*, 3 (2019), 474–489.

[9] Cao, Z., and Grima, R. Linear mapping approximation of gene regulatory networks with stochastic dynamics. *Nat. Commun. 9*, 1 (2018), 1–15.

[10] Cao, Z., and Grima, R. Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proc. Natl. Acad. Sci. U.S.A. 117*, 9 (2020), 4682–4692.

[11] Carrier, T. A., and Keasling, J. D. Investigating autocatalytic gene expression systems through mechanistic modeling. *J. Theor. Biol. 201*, 1 (1999), 25–36.

[12] Chahar, S., Gandhi, V., Yu, S., Desai, K., Cowper-Sal·lari, R., Kim, Y., Perekatt, A. O., Kumar, N., Thackray, J. K., Musolf, A., et al. Chromatin profiling reveals regulatory network shifts and a protective role for hepatocyte nuclear factor 4α during colitis. *Mol. Cell. Biol. 34*, 17 (2014), 3291–3304.

[13] Chan, T. E., Stumpf, M. P., and Babtie, A. C. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst. 5*, 3 (2017), 251–267.

[14] Chen, X., and Jia, C. Limit theorems for generalized density-dependent Markov chains and bursty stochastic gene regulatory networks. *J. Math. Biol. 80*, 4 (2020), 959–994.

[15] Chen, X., Wang, Y., Feng, T., Yi, M., Zhang, X., and Zhou, D. The overshoot and phenotypic equilibrium in characterizing cancer dynamics of reversible phenotypic plasticity. *J. Theor. Biol. 390* (2016), 40–49.

[16] Cunningham, T. J., and Duester, G. Mechanisms of retinoic acid signalling and its roles in organ and limb development. *Nat. Rev. Mol. Cell. Biol. 16*, 2 (2015), 110–123.

[17] Czuppon, P., and Pfaffelhuber, P. Limits of noise for autoregulated gene expression. *J. Math. Biol. 77*, 4 (2018), 1153–1191.

[18] Dessalles, R., Fromion, V., and Robert, P. A stochastic analysis of autoregulation of gene expression. *J. Math. Biol. 75*, 5 (2017), 1253–1283.

[19] Dobrinić, P., Szczurek, A. T., and Klose, R. J. PRC1 drives Polycomb-mediated gene repression by controlling transcription initiation and burst frequency. *Nat. Struct. Mol. Biol. 28*, 10 (2021), 811–824.

[20] Eden, U. T., and Kramer, M. A. Drawing inferences from Fano factor calculations. *J. Neurosci. Methods 190*, 1 (2010), 149–152.

[21] Fang, J., Ianni, A., Smolka, C., Vakhrusheva, O., Nolte, H., Krüger, M., Wietelmann, A., Simonet, N. G., Adrian-Segarra, J. M., Vaquero, A., et al. Sirt7 promotes adipogenesis in the mouse by inhibiting autocatalytic activation of Sirt1. *Proc. Natl. Acad. Sci. U.S.A. 114*, 40 (2017), E8352–E8361.

[22] Feigelman, J., Ganscha, S., Hastreiter, S., Schwarzfischer, M., Filipczyk, A., Schroeder, T., Theis, F. J., Marr, C., and Claassen, M. Analysis of cell lineage trees by exact bayesian inference identifies negative autoregulation of Nanog in mouse embryonic stem cells. *Cell Syst. 3*, 5 (2016), 480–490.

[23] Firman, T., Wedekind, S., McMorrow, T., and Ghosh, K. Maximum caliber can characterize genetic switches with multiple hidden species. *J. Phys. Chem. B 122*, 21 (2018), 5666–5677.

[24] Giovanini, G., Sabino, A. U., Barros, L. R., and Ramos, A. F. A comparative analysis of noise properties of stochastic binary models for a self-repressing and for an externally regulating gene. *Math. Biosci. Eng. 17*, 5 (2020), 5477–5503.

[25] Grönlund, A., Lötstedt, P., and Elf, J. Transcription factor binding kinetics constrain noise suppression via negative feedback. *Nat. Commun. 4*, 1 (2013), 1–5.

[26] Guo, G., Huss, M., Tong, G. Q., Wang, C., Sun, L. L., Clarke, N. D., and Robson, P. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell 18*, 4 (2010), 675–685.

[27] Ham, L., Brackston, R. D., and Stumpf, M. P. Extrinsic noise and heavy-tailed laws in gene expression. *Phys. Rev. Lett. 124*, 10 (2020), 108101.

[28] Hara, T., Abe, M., Inoue, H., Yu, L., Veenstra, T. D., Kang, Y., Lee, K., and Miki, T. Cytokinesis regulator ECT2 changes its conformation through phosphorylation at Thr-341 in G2/M phase. *Oncogene 25*, 4 (2006), 566–578.

[29] Hilfinger, A., Norman, T. M., Vinnicombe, G., and Paulsson, J. Constraints on fluctuations in sparsely characterized biological systems. *Phys Rev. Lett. 116*, 5 (2016), 058101.

[30] Hornos, J. E., Schultz, D., Innocentini, G. C., Wang, J., Walczak, A. M., Onuchic, J. N., and Wolynes, P. G. Self-regulating gene: an exact solution. *Phys. Rev. E 72*, 5 (2005), 051907.

[31] Hui, Z., Jiang, Z., Qiao, D., Bo, Z., Qiyuan, K., Shaohua, B., Wenbing, Y., Wei, L., Cheng, L., Shuangning, L., et al. Increased expression of LCN2 formed a positive feedback loop with activation of the ERK pathway in human kidney cells during kidney stone formation. *Sci. Rep. 10*, 1 (2020), 1–12.

[32] Jia, C. Simplification of Markov chains with infinite state space and the mathematical theory of random gene expression bursts. *Phys. Rev. E 96*, 3 (2017), 032402.

[33] Jia, C. Kinetic foundation of the zero-inflated negative binomial model for single-cell RNA sequencing data. *SIAM J. Appl. Math. 80*, 3 (2020), 1336–1355.

[34] Jia, C., and Grima, R. Dynamical phase diagram of an auto-regulating gene in fast switching conditions. *J. Chem. Phys. 152*, 17 (2020), 174110.

[35] Jia, C., and Grima, R. Small protein number effects in stochastic models of autoregulated bursty gene expression. *J. Chem. Phys. 152*, 8 (2020), 084115.

[36] Jia, C., and Grima, R. Frequency domain analysis of fluctuations of mRNA and protein copy numbers within a cell lineage: theory and experimental validation. *Phys. Rev. X 11*, 2 (2021), 021032.

[37] Jia, C., Qian, H., Chen, M., and Zhang, M. Q. Relaxation rates of gene expression kinetics reveal the feedback signs of autoregulatory gene networks. *J. Chem. Phys. 148*, 9 (2018), 095102.

[38] Jia, C., Xie, P., Chen, M., and Zhang, M. Q. Stochastic fluctuations can reveal the feedback signs of gene regulatory networks at the single-molecule level. *Sci. Rep. 7*, 1 (2017), 1–9.

[39] Jia, C., Zhang, M. Q., and Qian, H. Emergent Lévy behavior in single-cell stochastic gene expression. *Phys. Rev. E 96*, 4 (2017), 040402.

[40] Jiang, D.-Q., Wang, Y., and Zhou, D. Phenotypic equilibrium as probabilistic convergence in multi-phenotype cell population dynamics. *PLOS ONE 12*, 2 (2017), e0170916.

[41] Kang, Y., Gu, C., Yuan, L., Wang, Y., Zhu, Y., Li, X., Luo, Q., Xiao, J., Jiang, D., Qian, M., et al. Flexibility and symmetry of prokaryotic genome rearrangement reveal lineage-associated core-gene-defined genome organizational frameworks (vol 5, e01867, 2014). *MBIO 6*, 1 (2015).

[42] Karamyshev, A. L., and Karamysheva, Z. N. Lost in translation: ribosome-associated mRNA and protein quality controls. *Front. Genet. 9* (2018), 431.

[43] Karmakar, R., and Das, A. K. Effect of transcription reinitiation in stochastic gene expression. *J. Stat. Mech. Theory Exp. 2021*, 3 (2021), 033502.

[44] Kidder, B. L., and Palmer, S. Examination of transcriptional networks reveals an important role for TCFAP2C, SMARCA4, and EOMES in trophoblast stem cell maintenance. *Genome Res. 20*, 4 (2010), 458–472.

[45] Ko, Y., Kim, J., and Rodriguez-Zas, S. L. Markov chain Monte Carlo simulation of a Bayesian mixture model for gene network inference. *Genes Genom. 41*, 5 (2019), 547–555.

[46] Levenberg, S., Katz, B.-Z., Yamada, K. M., and Geiger, B. Long-range and selective autoregulation of cell-cell or cell-matrix adhesions by cadherin or integrin ligands. *J. Cell Sci. 111*, 3 (1998), 347–357.

[47] Lin, J., and Amir, A. Disentangling intrinsic and extrinsic gene expression noise in growing cells. *Phys. Rev. Lett. 126*, 7 (2021), 078101.

[48] Luecken, M. D., and Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol. 15*, 6 (2019), e8746.

[49] Mahajan, T., Singh, A., and Dar, R. Topological constraints on noise propagation in gene regulatory networks. *bioRxiv* (2021).

[50] Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A. J., Tanaka, Y., Wilkinson, A. C., Buettner, F., Macaulay, I. C., Jawaid, W., Diamanti, E., et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol. 33*, 3 (2015), 269–276.

[51] Munsky, B., Neuert, G., and Van Oudenaarden, A. Using gene expression noise to understand gene regulation. *Science 336*, 6078 (2012), 183–187.

[52] Niu, Y., Wang, Y., and Zhou, D. The phenotypic equilibrium of cancer cells: From average-level stability to path-wise convergence. *J. Theor. Biol. 386* (2015), 7–17.

[53] Norris, J. R. *Markov chains*. Cambridge university press, 1998.

[54] Paulsson, J. Models of stochastic gene expression. *Phys. Life Rev. 2*, 2 (2005), 157–175.

[55] Pramono, A., Zahabi, A., Morishima, T., Lan, D., Welte, K., and Skokowa, J. Thrombopoietin induces hematopoiesis from mouse ES cells via HIF-1α–dependent activation of a BMP4 autoregulatory loop. *Ann. N. Y. Acad. Sci. 1375*, 1 (2016), 38–51.

[56] Psaila, B., Barkas, N., Iskander, D., Roy, A., Anderson, S., Ashley, N., Caputo, V. S., Lichtenberg, J., Loaiza, S., Bodine, D. M., et al. Single-cell profiling of human megakaryocyte-erythroid progenitors identifies distinct megakaryocyte and erythroid differentiation pathways. *Genome Biol. 17*, 1 (2016), 1–19.

[57] Ramos, A. F., Hornos, J. E. M., and Reinitz, J. Gene regulation and noise reduction by coupling of stochastic processes. *Phys. Rev. E 91*, 2 (2015), 020701.

[58] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science 308*, 5721 (2005), 523–529.

[59] Sanchez-Castillo, M., Blanco, D., Tienda-Luna, I. M., Carrion, M., and Huang, Y. A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics 34*, 6 (2018), 964–970.

[60] Shahrezaei, V., and Swain, P. S. Analytical distributions for stochastic gene expression. *Proc. Natl. Acad. Sci. U.S.A. 105*, 45 (2008), 17256–17261.

[61] Sharma, A., and Adlakha, N. Markov chain model to study the gene expression. *Adv. Appl. Sci. Res. 5*, 2 (2014), 387–393.

[62] Shen, H., Huo, S., Yan, H., Park, J. H., and Sreeram, V. Distributed dissipative state estimation for Markov jump genetic regulatory networks subject to round-robin scheduling. *IEEE Trans. Neural Netw. Learn. Syst. 31*, 3 (2019), 762–771.

[63] Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. Network motifs in the transcriptional regulation network of escherichia coli. *Nat. Genet. 31*, 1 (2002), 64–68.

[64] Sheth, R., Bastida, M. F., Kmita, M., and Ros, M. "Self-regulation," a new facet of Hox genes' function. *Dev. Dyn. 243*, 1 (2014), 182–191.

[65] Shmulevich, I., Gluhovsky, I., Hashimoto, R. F., Dougherty, E. R., and Zhang, W. Steady-state analysis of genetic regulatory networks modelled by probabilistic boolean networks. *Comp. Funct. Genomics 4*, 6 (2003), 601–608.

[66] Skinner, S. O., Xu, H., Nagarkar-Jaiswal, S., Freire, P. R., Zwaka, T. P., and Golding, I. Single-cell analysis of transcription kinetics across the cell cycle. *Elife 5* (2016), e12175.

[67] Stewart, A. J., Seymour, R. M., Pomiankowski, A., and Reuter, M. Underdominance constrains the evolution of negative autoregulation in diploids. *PLOS Comput. Biol. 9*, 3 (2013), e1002992.

[68] Swain, P. S. Efficient attenuation of stochasticity in gene expression through post-transcriptional control. *J. Mol. Biol. 344*, 4 (2004), 965–976.

[69] Swain, P. S., Elowitz, M. B., and Siggia, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. U.S.A. 99*, 20 (2002), 12795–12800.

[70] Thattai, M., and Van Oudenaarden, A. Intrinsic noise in gene regulatory networks. *Proc. Natl. Acad. Sci. U.S.A. 98*, 15 (2001), 8614–8619.

[71] Thomas, P. Intrinsic and extrinsic noise of gene expression in lineage trees. *Sci. Rep. 9*, 1 (2019), 1–16.

[72] Thomas, P., Matuschek, H., and Grima, R. How reliable is the linear noise approximation of gene regulatory networks? *BMC Genom. 14*, 4 (2013), 1–15.

[73] Veerman, F., Popović, N., and Marr, C. Parameter inference with analytical propagators for stochastic models of autoregulated gene expression. *Int. J. Nonlinear Sci. Numer. Simul.* (2021).

[74] Voliotis, M., Cohen, N., Molina-París, C., and Liverpool, T. B. Fluctuations, pauses, and backtracking in DNA transcription. *Biophys. J. 94*, 2 (2008), 334–348.

[75] Wang, D.-G., Wang, S., Huang, B., and Liu, F. Roles of cellular heterogeneity, intrinsic and extrinsic noise in variability of p53 oscillation. *Sci. Rep. 9*, 1 (2019), 1–11.

[76] Wang, Y. *Some Problems in Stochastic Dynamics and Statistical Analysis of Single-Cell Biology of Cancer*. Ph.D. thesis, University of Washington, 2018.

[77] Wang, Y. Two metrics on rooted unordered trees with labels. *Algorithms Mol. Biol. 17*, 1 (2022), 1–17.

[78] Wang, Y., Kropp, J., and Morozova, N. Biological notion of positional information/value in morphogenesis theory. *Int. J. Dev. Biol. 64*, 10-11-12 (2020), 453–463.

[79] Wang, Y., Minarsky, A., Penner, R., Soulé, C., and Morozova, N. Model of morphogenesis. *J. Comput. Biol. 27*, 9 (2020), 1373–1383.

[80] Wang, Y., Mistry, B. A., and Chou, T. Discrete stochastic models of SELEX: Aptamer capture probabilities and protocol optimization. *J. Chem. Phys. 156*, 24 (2022), 244103.

[81] Wang, Y., and Qian, H. Mathematical representation of Clausius' and Kelvin's statements of the second law and irreversibility. *J. Stat. Phys. 179*, 3 (2020), 808–837.

[82] Wang, Y., and Wang, L. Causal inference in degenerate systems: An impossibility result. In *International Conference on Artificial Intelligence and Statistics* (2020), PMLR, pp. 3383–3392.

[83] Wang, Y., and Wang, Z. Inference on the structure of gene regulatory networks. *J. Theor. Biol. 539* (2022), 111055.

[84] Wang, Y., Zhang, B., Kropp, J., and Morozova, N. Inference on tissue transplantation experiments. *J. Theor. Biol. 520* (2021), 110645.

[85] Werhli, A. V., Grzegorczyk, M., and Husmeier, D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics 22*, 20 (2006), 2523–2531.

[86] Xing, B., and Van Der Laan, M. J. A causal inference approach for constructing transcriptional regulatory networks. *Bioinformatics 21*, 21 (2005), 4007–4013.

[87] Yan, J., Hilfinger, A., Vinnicombe, G., Paulsson, J., et al. Kinetic uncertainty relations for the control of stochastic reaction networks. *Phys. Rev. Lett. 123*, 10 (2019), 108101.

[88] Yang, W., Peng, L., Zhu, Y., and Hong, L. When machine learning meets multiscale modeling in chemical reactions. *J. Chem. Phys. 153*, 9 (2020), 094117.

[89] Ye, F. X.-F., Wang, Y., and Qian, H. Stochastic dynamics: Markov chains and random transformations. *Discrete Contin. Dyn. Syst. - B 21*, 7 (2016), 2337.

[90] Zhou, D., Wang, Y., and Wu, B. A multi-phenotypic cancer model with cell plasticity. *J. Theor. Biol. 357* (2014), 35–45.

[91] Zhou, T., and Zhang, J. Analytical results for a multistate gene model. *SIAM J. Appl. Math. 72*, 3 (2012), 789–818.