bioRxiv preprint doi: https://doi.org/10.1101/2022.03.31.486623; this version posted April 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Relative stability of mRNA and protein severely limits inference of gene networks from single-cell mRNA measurements

Tarun Mahajan[1], Michael Saint-Antoine[2], Roy D. Dar[3], Abhyudai Singh[4]

***Abstract*— Inference of gene regulatory networks from single-cell expression data, such as single-cell RNA sequencing, is a popular problem in computational biology. Despite diverse methods spanning information theory, machine learning, and statistics, it is unsolved. This shortcoming can be attributed to measurement errors, lack of perturbation data, or difficulty in causal inference. Yet, it is not known if kinetic properties of gene expression also cause an issue. We show how the relative stability of mRNA and protein hampers inference. Available inference methods perform benchmarking on synthetic data lacking protein species, which is biologically incorrect. We use a simple model of gene expression, incorporating both mRNA and protein, to show that a more stable protein than mRNA can cause loss in correlation between the mRNA of a transcription factor and its target gene. This can also happen when mRNA and protein are on the same timescale. The relative difference in timescales affects true interactions more strongly than false positives, which may not be suppressed. Besides correlation, we find that information-theoretic nonlinear measures are also prone to this problem. Finally, we demonstrate these principles in real single-cell RNA sequencing data for over 1700 yeast genes.**

I. Introduction

According to the "central dogma" of molecular biology [1], genes on DNA are transcribed into messenger RNA (mRNA), which are translated into proteins. The proteins carry out various functions within the cell including regulation of expression of other genes. These regulatory relationships form gene regulatory networks (GRNs), which control the complexity of cellular life [2], [3], and malfunctions in GRNs can lead to diseases like cancer [4].

Understanding GRN function and structure is essential for cell biologists, and the inference of their topology from static transcriptomic data is important [5]. Researchers have used statistical relationships between levels of mRNA to identify the underlying GRN. These statistical methods include correlation [6], regression [7]–[9], information-theoretic techniques [10]–[14], Bayesian techniques [15], [16], and more [17]–[23]. Benchmarking studies have assessed the comparative performance of different methods [24]. An excellent review of GRN inference methods can be found in [25].

[1]Tarun Mahajan is a PhD candidate with the Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA tarunm3@illinois.edu [2]Michael Saint-Antoine is a PhD candidate with the Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE 19716, USA mikest@udel.edu [3]Roy D. Dar is with the Faculty of Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA roydar@illinois.edu [4]Abhyudai Singh is with the Faculty of Electrical and Computer Engineering, Biomedical Engineering and Mathematical Sciences, University of Delaware, Newark, DE 19716, USA absingh@udel.edu

Many GRN inference techniques make the assumption that mRNA and protein counts for a given gene are correlated, and use mRNA transcript abundance as a proxy for protein abundance, which is difficult to measure in a high-throughput manner. However, in experiments the correlation between mRNA and corresponding protein counts can be weak [26]. In this paper, we use a model of gene expression to explore the impact of relative stability of mRNA and protein on the correlation between their abundances. We also explore information-theoretic measures–mutual information (MI) [27] and the phi-mixing coefficient [28]–using stochastic simulations [29]. MI quantifies the uncertainty in individual and joint distributions of random variables [27]. The phi-mixing coefficient is a measure of the statistical dependence of two random variables based on the difference between their conditional and unconditional probability distributions [28]. Both MI ([10]–[14]) and the phi-mixing coefficient ([30]) have been used for GRN inference. Finally, we demonstrate the established principles on a real single-cell RNA sequencing (scRNA-seq) dataset for yeast, *S. cerevisiae*.

We show that when protein is more stable than mRNA, GRN cannot be reliably inferred at the single-cell mRNA level. Even when the true GRN edges are identifiable, the false positive edges will dominate in correlation values. Collectively, this establishes a protein-mRNA lifetime-dependent loss in GRN signature in single-cell data.

II. A simple model of gene expression

We start with a simple model of gene expression. mRNA M is produced and degraded via a 1-dimensional Poisson birth-death process. Uppercase and lowercase letters represent a species and its molecular count, respectively; M and m are the mRNA and its count, respectively. Production of M is a Poisson birth process with rate $k_m F(\beta)$, and each event creates mRNA in bursts of size β, which is distributed according to any arbitrary positive-valued probability distribution $F(\beta)$. Protein P is created and degraded in a 1-dimensional conditional Poisson birth-death process. Each M molecule is translated into a molecule of P via a conditional Poisson birth process at a rate k_p. M and P are degraded as Poisson processes at rates $1/\tau_m$ and $1/\tau_p$, respectively. τ_m and τ_p are the respective average lifetimes of M and P. For the rest of the paper, we assume that τ_m is constant, and $\tau_p = \tau\,\tau_m$. This allows us to change τ_p/τ_m by varying only

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.31.486623; this version posted April 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

τ. The chemical reaction network for this system is

$$\begin{aligned} \varnothing \xrightarrow{k_m F(\beta)} \beta \times M, \quad M \xrightarrow{1/\tau_m} \varnothing \\ M \xrightarrow{k_p} M + P, \quad P \xrightarrow{1/(\tau\,\tau_m)} \varnothing. \end{aligned} \tag{1}$$

The time evolution of the joint probability distribution for m and p in (1) is given by the following chemical master equation (CME) [31]:

$$\begin{aligned} \frac{\partial P(m,p;t)}{\partial t} &= \left(\sum_{\beta} F(\beta)\, k_m\, P(m-\beta,p;t) \right. \\ &\left. + \frac{m+1}{\tau_m} P(m+1,p;t) \right) + \left(k_p\, m P(m,p-1;t) \right. \\ &\left. + \frac{p+1}{\tau\,\tau_m} P(m,p+1;t) \right) - \left(k_m + \frac{m}{\tau_m} + k_p\, m + \frac{p}{\tau\,\tau_m} \right) P(m,p;t), \end{aligned} \tag{2}$$

where $P(m,p;t)$ is the joint probability distribution for m and p at time t. (2) can be solved exactly for the steady-state moments [31]–[33]. The first order moments are given by

$$\langle m \rangle = \langle \beta \rangle\, k_m\, \tau_m, \quad \langle p \rangle = \langle m \rangle\, k_p\, \tau_p = \langle m \rangle\, k_p\, \tau\, \tau_m, \tag{3}$$

where the angle brackets represent statistical expectation. The second order moments involving only one species can be written as

$$\underbrace{\eta_m^2}_{\text{total noise}} := \frac{\langle m^2 \rangle - \langle m \rangle^2}{\langle m \rangle^2} = \underbrace{\frac{\left(\langle \beta^2 \rangle / \langle \beta \rangle + 1 \right)/2}{\langle m \rangle}}_{\text{intrinsic noise}}, \tag{4}$$

$$\underbrace{\eta_p^2}_{\text{total noise}} := \frac{\langle p^2 \rangle - \langle p \rangle^2}{\langle p \rangle^2} = \underbrace{\frac{1}{\langle p \rangle}}_{\text{intrinsic noise}} + \underbrace{\frac{1}{1+\tau} \eta_m^2}_{\text{extrinsic noise}}, \tag{5}$$

where η_m^2 and η_p^2 are the total noise or expression fluctuation, for M and P, respectively [32], [33]. Intrinsic noise is the fluctuation in m or p caused by the discrete birth-death events for M or P, respectively. Intrinsic noise for any mRNA species is given by (4) [32], [33]. Intrinsic noise for any protein species is given by the first term on the right-hand size (RHS) in (5) [32], [33]. Extrinsic noise in (5) is the noise propagated from m to p. The decomposition of total noise into intrinsic and extrinsic in (5) is fairly standard in noisy expression research [32]–[36]. The numerator for intrinsic noise in (4),

$$B := \frac{\langle \beta^2 \rangle / \langle \beta \rangle + 1}{2}, \tag{6}$$

is the average contribution to noise from birth and death events for M [34], [35]. We propose that

$$B \propto \langle \beta \rangle + o(1), \tag{7}$$

where o is the Little-o notation. If β is deterministic, $B = (\langle \beta \rangle + 1)/2$. If β is distributed geometrically, as is known for many genes in different species [37], [38], $B = \langle \beta \rangle$. B is an estimate of mean-independent intrinsic noise for mRNA, and we modulate noise by varying B. We vary B by changing $\langle \beta \rangle$ ((7)). Whenever $\langle \beta \rangle$ is increased/decreased by a factor, we decrease/increase k_m by the same factor to keep first order moments constant at fixed τ.

We are interested in the dependence of steady-state species correlations on B, and τ. Since the moments in (3)- (5) are dependent on B and τ, we obtain the expression for correlation between m and p, $\mathrm{cor}(m,\, p)$, as a function of these moments.

Proposition 1 (**mRNA-protein correlation in a single gene**) *For a single gene* (1)*, steady-state correlation between m and p is*

$$\mathrm{cor}(m,\ p) = \frac{1}{1+\tau} \sqrt{\frac{\eta_m^2}{\eta_p^2}}. \tag{8}$$

(8) is obtained by solving (2) [31]–[33]. We examine the behavior of $\mathrm{cor}(m,\, p)$ when τ is fixed, and B is variable, and establish the following upper bound:

Theorem 2 (**Upper bound on mRNA-protein correlation in a single gene**) *For a single gene* (1)*, steady-state correlation between m and p is bounded from above by*

$$\mathrm{cor}(m,\ p) = \frac{1}{\sqrt{1+\tau}}, \textit{ for } B \gg \frac{1+\tau}{\tau} \frac{1}{k_p \tau_m} \tag{9}$$

$\mathrm{cor}(m,\, p)$ is an increasing function of η_m^2, which is an increasing function of B (see (4) and (6)). Consequently, $\mathrm{cor}(m,\, p)$ is an increasing function of B, and reaches its maximum value in (9) when the second term on the right hand side (RHS) in (5) is much larger than the first term. This gives us the constraint on B in (9) while using (3)-(4). (9) is depicted by the green curve in Fig. 1b. There is a τ-mediated loss in correlation as protein becomes more stable than mRNA. For extremely stable proteins, $\mathrm{cor}(m,\, p)$ might completely vanish. Actual $\mathrm{cor}(m,\, p)$ values will always be lower than (9). The gap between the upper bound and the true $\mathrm{cor}(m,\, p)$ is governed by the relative amounts of intrinsic and extrinsic noises in p. This is evident from (8) and (9). $\mathrm{cor}(m,\, p)$ reaches the upper bound only when B is large enough. However, if B is low or moderate, $\mathrm{cor}(m,\, p)$ can deviate significantly. $\mathrm{cor}(m,\, p)$'s dependence on τ and B is shown in Fig. 1b. Further, we have also validated the analytical result using exact stochastic simulations in Fig. 1c. Dependence of $\mathrm{cor}(m,\, p)$ on τ and B motivates the central theme of the paper. We next explore whether this dependence propagates to downstream genes in small GRN topologies.

III. Loss of Correlation for Simple GRNs

Count correlations in a two-gene cascade

For a two-gene cascade (Fig. (2)a), gene 1 has mRNA M_1 and protein P_1, and gene 2 has mRNA M_2 and protein P_2. P_1 regulates the production of M_2. All kinetic parameters are identical for the two genes. M_1 is created and degraded via a 1-dimensional Poisson birth-death process. All the other species are created and degraded via 1-dimensional conditional Poisson birth-death processes. The chemical reaction

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.31.486623; this version posted April 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

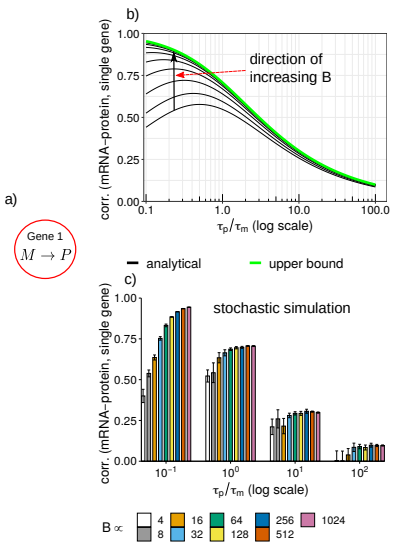

Fig. 1: **Plot of mRNA-protein correlation, cor(*m*, *p*),** for a **single gene** in (1) as a function of B, and τ. (a) GRN being considered–a single gene with its mRNA, M, and protein, P. (b) Analytical curves (black curves) obtained from (8) (one curve per B value). Nine different values were used for B: $B \propto \langle\beta\rangle \in 4\times\{2^0,2^1,\ldots,2^8\}$. The upper bound in (9) is depicted by the green curve. (c) $\mathrm{cor}(m,\,p)$ as a function of B and τ using exact stochastic simulations [29]. B was varied over the same range as before. The other kinetic parameters are: $k_m = 0.0282$, $\tau_m = 1/0.0025$, $k_p = 1.2/\tau_m$. k_m and τ_m are for the nanog gene from mouse embryonic stem cells ([39]). We assume that β is deterministic. Error bars show one standard deviation.

network is

$$\begin{aligned}
\{1\}\ \varnothing \xrightarrow{k_m F(\beta_1)} \beta_1\times \mathrm{M}_1, \quad \{2\}\ \mathrm{M}_1 \xrightarrow{1/\tau_m} \varnothing\\
\{3\}\ \mathrm{M}_1 \xrightarrow{k_p} \mathrm{M}_1+\mathrm{P}_1, \quad \{4\}\ \mathrm{P}_1 \xrightarrow{1/(\tau\,\tau_m)} \varnothing.\\
\{5\}\ \varnothing \xrightarrow{k_m F(\beta_2)\,h(p_1)} \beta_2\times \mathrm{M}_2, \quad \{6\}\ \mathrm{M}_2 \xrightarrow{1/\tau_m} \varnothing\\
\{7\}\ \mathrm{M}_2 \xrightarrow{k_p} \mathrm{M}_2+\mathrm{P}_2, \quad \{8\}\ \mathrm{P}_2 \xrightarrow{(\tau\,\tau_m)} \varnothing,
\end{aligned} \tag{10}$$

where β_1 and β_2 are the respective burst sizes for M_1 and M_2, P_1 regulates M_2 via $h(p_1)$, and numbers (in curly braces) represent an ordering on the reactions. The CME for (10) in compact form is [31]

$$\frac{\partial P(\boldsymbol{x};t)}{\partial t} = \sum_{r=1}^{R}\sum_{\beta_1,\beta_2} F(\beta_1)F(\beta_2)\Big(f_r(\boldsymbol{x}-\boldsymbol{d}_r)P(\boldsymbol{x}-\boldsymbol{d}_r;t) - f_r(\boldsymbol{x})P(\boldsymbol{x};t)\Big), \tag{11}$$

where $R=8$ is the total number of reactions in (10), $\boldsymbol{x} = (m_1,p_1,m_2,p_2)^T$, $f_r(\boldsymbol{x})$ is the propensity for reaction r, $\boldsymbol{d}_r = (\Delta m_1^r,\Delta p_1^r,\Delta m_2^r,\Delta p_2^r)^T$ is a vector containing the jump in species count because of reaction r, and $P(\boldsymbol{x};t)$ is the joint probability distribution for $\boldsymbol{x}$. β_1 and β_2 independent. We define

$$B_1 := \frac{\langle\beta_1^2\rangle/\langle\beta_1\rangle+1}{2},\ B_2 := \frac{\langle\beta_2^2\rangle/\langle\beta_2\rangle+1}{2}. \tag{12}$$

as the average contributions to noise from birth and death processes for M_1 and M_2, respectively ([32]–[35]):

(11) can be solved exactly for steady-state moments if $h(p_1)$ is a linear function [31]–[33]. For nonlinear $h(p_1)$, we solve (11) approximately using the linear noise approximation (LNA) approach [31]–[35], [40], [41]. For LNA, we linearize $h(p_1)$ around a deterministic concentration [32], [33], [41]. Then, evolution of the first-order moments is obtained using the extended moment generator from [42]:

$$d\langle\boldsymbol{x}\rangle/dt = -S\langle\boldsymbol{x}\rangle + \boldsymbol{g},$$

where S is a diagonal matrix of inverse lifetimes, and $\mathbf{g}$ is the vector of species production rates. Time evolution of the mean-normalized covariance matrix Σ, with entries $\Sigma_{ij} = (\langle x_i x_j\rangle - \langle x_i\rangle\langle x_j\rangle)/(\langle x_i\rangle\langle x_j\rangle)$ is given by ([31], [34], [35], [40], [41])

$$\frac{d\Sigma}{dt} = A\Sigma + \Sigma A^T + D, \tag{13}$$

where A and D are the mean-normalized jacobian and diffusion matrices, respectively. Ordering for the rows and columns of S, A, D and Σ corresponds to the species order in $\boldsymbol{x}$. The non-zero entries of A encode the structure of the expanded GRN including both mRNA and protein; there exists a regulation edge from species j to i iff $A_{ij}\neq 0$ ([31]–[35], [40], [41]). At steady state, first-order moments are given by

$$S\langle\boldsymbol{x}\rangle = \boldsymbol{g}, \tag{14}$$

and the Σ is given by the following Lyapunov equation ([31]–[35]):

$$A\Sigma + \Sigma A^T + D = 0. \tag{15}$$

(15) gives all the second order moments ([31]–[35], [41]). We are interested in the dependence of steady-state correlation between m_1 and m_2, $\mathrm{cor}(m_1,\,m_2)$, on B_1, B_2 and τ.

Proposition 3 (**mRNA-mRNA correlation in a two-gene cascade**) *For the two-gene cascade in* (10)*, correlation between m_1 and m_2 is*

$$\mathrm{cor}(m_1,\,m_2) = \frac{\theta_{m_2 p_1}}{2(1+\tau)}\sqrt{\frac{\eta_{m_1}^2}{\eta_{m_2}^2}}, \tag{16}$$

where $\eta_{m_1}^2$ and $\eta_{m_2}^2$ are the total noises in m_1 and m_2, respectively, which are obtained from (15) ([31]–[35], [41]). $\theta_{m_2 p_1}$ is the log-sensitivity of m_2 to changes in p_1 at steady state ([32]–[35]). Assume that $h(p_1)$ is saturating and $h(\langle p_1\rangle)$ is independent of $\langle p_1\rangle$. Then, $\theta_{m_2 p_1}$ is also independent of $\langle p_1\rangle$ at steady-state [32]–[35]. (16) is obtained by solving (15) ([31]–[33], [41]). We next establish the following upper bound on $\mathrm{cor}(m_1,\,m_2)$:

Theorem 4 (**Upper bound on mRNA-mRNA correlation in a two-gene cascade**) *For the two-gene cascade in* (10)*, correlation between m_1 and m_2 is bounded from above by*

$$\mathrm{cor}(m_1,\,m_2) = \frac{1}{\sqrt{2(1+2\tau)}}, \tag{17}$$

when

$$\frac{\theta_{m_2 p_1}^2(1+2\tau)}{2(1+\tau)^2}B_1 \gg B_2\frac{\langle m_1\rangle}{\langle m_2\rangle} + \frac{\theta_{m_2 p_1}^2\,\tau}{(1+\tau)}\frac{\langle m_1\rangle}{\langle p_1\rangle}, \tag{18}$$

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.31.486623; this version posted April 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

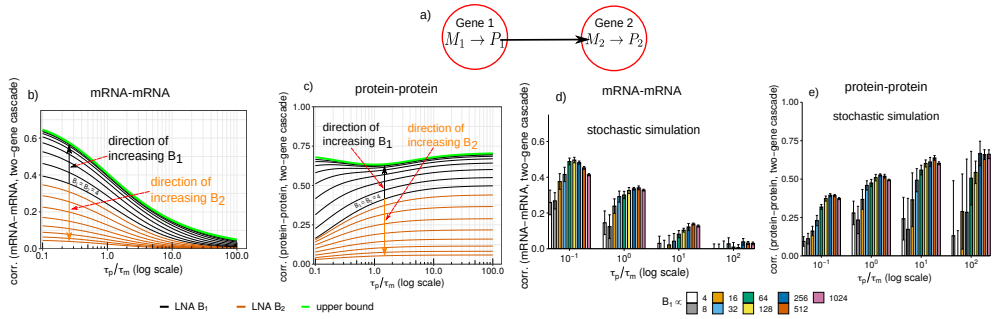

Fig. 2: **Plot of mRNA-mRNA, cor(m_1, m_2), and protein-protein, cor(p_1, p_2), correlation**, for a **two-gene cascade** in (10) as a function of B_1, B_2, and τ. (a) GRN being considered–a two gene cascade. Analytical curves (black (one curve per value of B_1, while $B_2 \propto \langle \beta_2 \rangle = 4$) and yellow (one curve per value of B_2, while $B_1 \propto \langle \beta_1 \rangle = 4$) curves) for cor($m_1$, m_2) and cor(p_1, p_2) are shown in (b) and (c), respectively. The upper bounds in (17) and (22) are depicted by the green curves in (b) and (c), respectively. cor(m_1, m_2) and cor(p_1, p_2) as functions of B_1 and τ using exact stochastic simulations [29] are shown in (d) and (e), respectively. B_1 was varied while B_2 was held constant (both in the same ranges as before). The values of other kinetic parameters are the same as mentioned in the caption for Fig. 1. Further, we assume that $h(p_1) = p_1^4/\left(p_1^4 + k_h^4\right)$, where k_h was selected such that at steady state $h(\langle p_1 \rangle) = 0.5$ for all values of the kinetic parameters. Error bars show one standard deviation.

From (15), we obtain ([32], [33]),

$$\eta^2_{m_2} = \Sigma_{33} = \frac{\langle m_2^2 \rangle - \langle m_2 \rangle^2}{\langle m_2 \rangle^2} = \underbrace{\eta^2_{m_2\circ}}_{\text{intrinsic noise, } m_2} + \underbrace{\frac{\theta^2_{m_2 p_1}\,\tau}{1+\tau}\,\underbrace{\eta^2_{p_1\circ}}_{\text{intrinsic noise, } p_1} + \frac{\theta^2_{m_2 p_1}\,(1+2\tau)}{2\,(1+\tau)^2}\,\underbrace{\eta^2_{m_1\circ}}_{\text{intrinsic noise, } m_1}}_{\text{extrinsic noise, } m_2}. \tag{19}$$

From (16), (19) and $\eta^2_{m_1} = \eta^2_{m_1\circ}$, cor($m_1$, m_2) is an increasing function of $\eta^2_{m_1\circ}$, and achieves the upper bound in (17) when the third term on the RHS in (19) is much larger than the first two terms. (17) is depicted by the green curve in Fig. 2b. There is a τ-mediated loss in cor(m_1, m_2). As protein becomes more stable than mRNA, cor(m_1, m_2) decays much faster than cor(m_1, p_1). For extremely stable proteins, cor(m_1, m_2) will completely vanish. Actual cor(m_1, p_1) could be much less than (17) (Fig. 2).

(18) defines a tug-of-war between B_1 and B_2. Their relative magintudes dictate the gap between (17) and the actual cor(m_1, m_2). If B_1 is much higher than B_2, (18), then cor(m_1, m_2) will reach its upper bound (17) (see the black curves in Fig. 2b). However, if B_1 is much smaller than B_2, cor(m_1, m_2) can vanish even when mRNA is more stable than protein, and $\tau < 1$ (check the yellow curves in Fig. 2b). We also performed exact stochastic simulations ([29]) to verify (16) and (17), Fig. 2d. Next, we examine the steady-state correlation between p_1 and p_2.

Proposition 5 (**protein-protein correlation in a two-gene cascade**) *For the two-gene cascade* (10)*, correlation between p_1 and p_2 is*

$$\text{cor}(p_1,\ p_2) = \frac{\theta_{m_2 p_1}}{2\,(1+\tau)}\sqrt{\frac{\left(\eta^2_{m_1\circ} + \tau\,\eta^2_{p_1\circ}\right)^2}{\eta^2_{p_1}\,\eta^2_{p_2}}}, \tag{20}$$

$\eta^2_{p_1}$ is obtained from (5) by replacing η^2_p and η^2_m with $\eta^2_{p_1}$ and $\eta^2_{m_1}$, respectively, and from (15) ([32], [33])

$$\eta^2_{p_2} = \Sigma_{44} = \frac{\langle p_2^2 \rangle - \langle p_2 \rangle^2}{\langle p_2 \rangle^2} = \underbrace{\eta^2_{p_2\circ}}_{\text{intrinsic noise, } p_2} + \underbrace{\frac{1}{1+\tau}\,\eta^2_{m_2\circ} + \frac{\theta^2_{m_2 p_1}\,\tau\,(\tau+2)}{2\,(1+\tau)^2}\,\eta^2_{p_1\circ} + \frac{\theta^2_{m_2 p_1}\left(\tau^2+3\tau+1\right)}{2\,(1+\tau)^3}\,\eta^2_{m_1\circ}}_{\text{extrinsic noise, } p_2}. \tag{21}$$

(20) is obtained from (15) [31]–[33]. cor(p_1, p_2) is an increasing function of $\eta^2_{m_1\circ}$. Consequently, cor(p_1, p_2) is an increasing function of B_1, which establishes the following upper bound on cor(p_1, p_2):

Theorem 6 (**Upper bound on protein-protein correlation in a two-gene cascade**) *For the two-gene cascade* (10)*, correlation between p_1 and p_2 is bounded from above by*

$$\text{cor}(p_1,\ p_2) = \sqrt{\frac{(\tau+1)^2}{2\,(\tau^2+3\,\tau+1)}}, \tag{22}$$

when

$$B_2 \ll \frac{\theta^2_{m_2 p_1}}{2}\,B_1\,\frac{\langle m_2 \rangle}{\langle m_1 \rangle} + o(1). \tag{23}$$

Theorem 6 can be proved by substituting intrinsic noise terms in (20) to show that the upper bound is reached when the fourth term in the RHS of (21) is much larger than the first three terms. When B_1 is large to satisfy (23), cor(p_1, p_2) will reach its upper bound in (22) (see the black curves in Fig. 2c). For all values of τ, cor(p_1, p_2) is greater than or equal to 80% of its maximum of $1/\sqrt{2}$ (green curve in Fig. 2c). This is in contadiction to (17) where the upper bound was a monotonically decreasing function of τ. The upper bound for cor(p_1, p_2) does not exhibit a τ-mediated loss in correlation. cor(p_1, p_2) also exhibits a tug-of-war between B_1 and B_2 (yellow curves in Fig. 2c). This is evident from (23) and the fact that (20) is a decreasing function of B_2.

We also performed stochastic simulations [29] to demonstrate the dependence of cor(p_1, p_2) on B_1, and τ (Fig. 2e).

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.31.486623; this version posted April 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Count correlations in a three-gene fanout motif

Next, we show that the false positive correlation between two genes having a common upstream TF, but not regulating each other, is less susceptible to a τ-mediated loss of correlation. This false positive correlation is most of the time greater than the true positive correlation in a two-gene cascade. For this, we study a fanout network, which has three genes. Genes 1 and 2 satisfy (10). Gene 3 satistifes the following additional reaction channels:

$$\begin{aligned} \{9\}\ \varnothing \xrightarrow{k_m F(\beta_3) h(p_1)} \beta_3 \times \mathrm{M}_3, \quad \{10\}\ \mathrm{M}_3 \xrightarrow{1/\tau_m} \varnothing \\ \{11\}\ \mathrm{M}_3 \xrightarrow{k_p} \mathrm{M}_3 + \mathrm{P}_3, \quad \{12\}\ \mathrm{P}_3 \xrightarrow{1/(\tau\,\tau_m)} \varnothing, \end{aligned} \tag{24}$$

where reaction numbering has been continued from (24), M_3 and P_3 are the mRNA and protein for gene 3, respectively, β_3 is the burst size for m_3. We assume that β_2 and β_3 are identically distributed. We also define

$$B_3 = B_2, \tag{25}$$

where B_2 is given in (12), as the average contribution of the birth and death events in (24) to noise in m_3. Genes 2 and 3 are identical, and do not regulate each other. They have a common regulator, P_1. Correlation between the species of genes 2 and 3 defines false positive correlation ($\mathrm{cor}(m_2,\ m_3)$ and $\mathrm{cor}(p_2,\ p_3)$). Now, $\boldsymbol{x} = (m_1, p_1, m_2, p_2, m_3, p_3)^T$. All moments upto the second-order can be obtained by solving (14) and (15) with additional species, M_3 and P_3. All moments for m_3 amd p_3 can be obtained from the moments of m_2 and p_2, respectively. Consequently, $\theta_{m_3 p_1} = \theta_{m_2 p_1}$. Next, we find $\mathrm{cor}(m_2,\ m_3)$.

Proposition 7 (**mRNA-mRNA correlation in a three-gene fanout**) *For the three-gene fanout* (10) *and* (24)*, correlation between m_2 and m_3 is given by*

$$\mathrm{cor}(m_2,\ m_3) = \frac{\theta^2_{m_2 p_1}\ \frac{\tau}{\tau+1}\ \eta^2_{p_1\circ} + \theta^2_{m_2 p_1}\ \frac{2\tau+1}{2(\tau+1)^2}\ \eta^2_{m_1\circ}}{\eta^2_{m_2}} \tag{26}$$

$\theta_{m_2 p_1}$ and $\eta^2_{m_2}$ in (26) are interchangeable with $\theta_{m_3 p_1}$ and $\eta^2_{m_3}$, respectively. Proposition 7 is proved by solving (15) [32], [33]. On substituting (19) in (26), we see that $\mathrm{cor}(m_2,\ m_3)$ is an increasing function of B_1, and establish the following upper bound on $\mathrm{cor}(m_2,\ m_3)$:

Theorem 8 (**Upper bound on mRNA-mRNA correlation in a three-gene fanout**) *For the three-gene fanout* (10) *and* (24)*, correlation between m_2 and m_3 is bounded from above by*

$$\begin{aligned} &\mathrm{cor}(m_2,\ m_3) = 1 \\ &\textit{for } B_1 \gg \frac{2\,(\tau+1)^2}{\theta^2_{m_2 p_1}\,(2\,\tau+1)}\,B_2\,\frac{\langle m_1\rangle}{\langle m_3\rangle} + \frac{2\,\tau\,(\tau+1)}{(2\,\tau+1)}\,\frac{\langle m_1\rangle}{\langle p_1\rangle}. \end{aligned} \tag{27}$$

$\mathrm{cor}(m_2,\ m_3)$ depends on B_1 through $\eta^2_{m_1\circ}$. For $\eta^2_{m_2}$ in the denominator in (26), substituting (19), we get (27). Interestingly, the upper bound (27) is independent of τ. This is in contrast to $\mathrm{cor}(m_1,\ m_2)$, where (17) which decays with τ. This implies that $\mathrm{cor}(m_2,\ m_3)$ will always dominate $\mathrm{cor}(m_1,\ m_2)$ and $\mathrm{cor}(m_1,\ m_3)$ when protein is more stable than mRNA, and can even dominate when mRNA is more stable (Fig. (3)b). Next, we directly compare $\mathrm{cor}(m_1,\ m_2)$ to $\mathrm{cor}(m_2,\ m_3)$.

Theorem 9 (**mRNA-mRNA correlation in two-gene cascade vs three-gene fanout**) *For the three-gene fanout* (10) *and* (24)

$$B_2 < \frac{\theta^2_{m_2 p_1}}{2}\,B_1\,\frac{\langle m_2\rangle}{\langle m_1\rangle} \implies \mathrm{cor}(m_2,\ m_3) > \mathrm{cor}(m_1,\ m_2). \tag{28}$$

On comparing (16) and (26), while using (19), we get (28). (28) demonstrates a tug-of-war between B_1 and B_2, where their relative magintudes dictate whether $\mathrm{cor}(m_1,\ m_2)$ or $\mathrm{cor}(m_2,\ m_3)$ will dominate. The relative magnitudes of B_1 and B_2 also determine whether $\mathrm{cor}(m_1,\ m_2)$ and $\mathrm{cor}(m_1,\ m_3)$ will achieve their upper bounds (see Theorems 4 and 8). When (18) and (27) are true, $\mathrm{cor}(m_1,\ m_2)$ and $\mathrm{cor}(m_2,\ m_3)$ achieve their upper bounds. However, at the same time, $\mathrm{cor}(m_2,\ m_3)$ will begin to dominate $\mathrm{cor}(m_1,\ m_2)$. This implies that kinetic conditions which allow inference of the GRN also confound inference via the false positive edges.

The dependence of $\mathrm{cor}(m_2,\ m_3)$ on B_1 (cyan curves), B_2 (dark-green curves) and τ based on (26) is shown in Fig. 3b. We have also shown the respective curves for $\mathrm{cor}(m_1,\ m_2)$ based on (16) there for the sake of comparison. We also performed exact stochastic simulation to verify these results as shown in Figs. 3d. Next, we study the correlation between p_2 and p_3, $\mathrm{cor}(p_2,\ p_3)$.

Proposition 10 (**protein-protein correlation in a three--gene fanout**) *For the three-gene fanout* (10) *and* (24)*, correlation between p_2 and p_3 is*

$$\mathrm{cor}(p_2,\ p_3) = \frac{\theta^2_{m_2 p_1}\ \frac{\tau(\tau+2)}{2(\tau+1)^2}\ \eta^2_{p_1\circ} + \theta^2_{m_2 p_1}\ \frac{\tau^2+3\tau+1}{2(\tau+1)^3}\ \eta^2_{m_1\circ}}{\eta^2_{p_2}} \tag{29}$$

$\theta_{m_2 p_1}$ and $\eta^2_{p_2}$ in (29) are interchangeable with $\theta_{m_3 p_1}$ and $\eta^2_{p_3}$, respectively. Proposition 10, is proved by solving (15) [32], [33]. Like before, we assume that B_1, B_2 and B_3 are varied in a manner which preserves the first order moments at fixed τ. On substituting (21) in (29), we see that $\mathrm{cor}(p_2,\ p_3)$ is an increasing function of B_1. Now, we establish the following upper bound on $\mathrm{cor}(p_2,\ p_3)$:

Theorem 11 (**Upper bound on protein-protein correlation in a three-gene fanout**) *For the three-gene fanout* (10) *and* (24)*, correlation between p_2 and p_3 is bounded from above by*

$$\mathrm{cor}(p_2,\ p_3) = 1, \textit{ for } B_2 \ll \frac{\theta^2_{m_2 p_1}}{2}\,B_1\,\frac{\langle m_2\rangle}{\langle m_1\rangle}, \tag{30}$$

$\mathrm{cor}(p_2,\ p_3)$ depends on B_1 through $\eta^2_{m_1\circ}$. For $\eta^2_{p_2}$ in the denominator in (29), substituting (21), we get (30) for reaching the upper bound. Similar to $\mathrm{cor}(m_2,\ m_3)$, the upper bound for $\mathrm{cor}(p_2,\ p_3)$ is independent of τ. Therefore, it is possible that $\mathrm{cor}(p_2,\ p_3)$ might dominate $\mathrm{cor}(p_1,\ p_2)$ and $\mathrm{cor}(p_1,\ p_3)$. Next, we directly compare $\mathrm{cor}(p_2,\ p_3)$ to $\mathrm{cor}(p_1,\ p_2)$:

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.31.486623; this version posted April 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

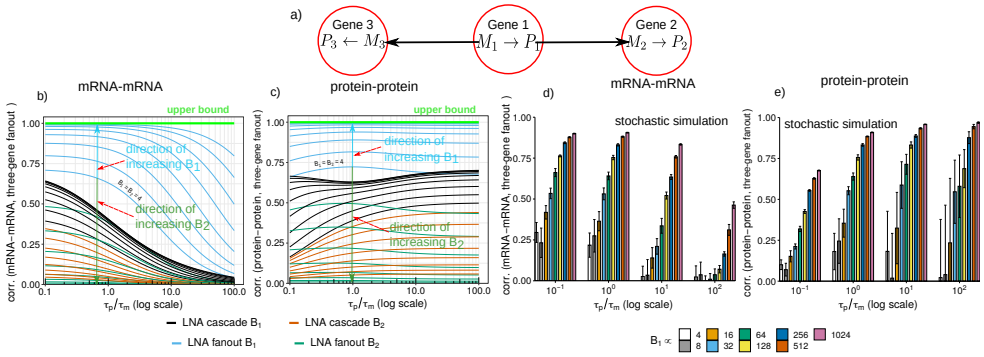

Fig. 3: **Plot of mRNA-mRNA, cor(m_2, m_3), and protein-protein, cor(p_2, p_3), correlation**, for a **three-gene fanout** in (10) and (24) jointly as a function of B_1, B_2, and τ. (a) GRN being considered–a three-gene fanout Analytical curves (cyan (one curve per value of B_1, while $B_2 \propto \langle \beta_2 \rangle = 4$) and dark-green (one curve per value of B_2, while $B_1 \propto \langle \beta_1 \rangle = 4$) curves) for cor($m_2$, m_3) and cor(p_2, p_3) are shown in (b) and (c), respectively. The analytical curves from Figs. 2a and 2b are also shown in black and yellow curves in (b) and (c), respectively, for comparison. The upper bound of 1 is shown in green curves in (b) and (c). cor(m_2, m_3) and cor(p_2, p_3) as functions of B_1 and τ using exact stochastic simulations [29] are shown in (d) and (e), respectively. B_1 was varied while B_2 was held constant (both in the same ranges as before). The values of other kinetic parameters are the same as mentioned in the caption for Fig. 1. The values of other kinetic parameters are the same as before (see caption of Fig. 1). For details on the regulation function $h(p_1)$, see caption of Fig. 2. Error bars show one standard deviation.

Theorem 12 (**Protein-protein correlation in two-gene cascade vs three-gene fanout**) *For the three-gene fanout* (10) *and* (24)*,*

$$B_2 < \frac{\theta_{m_2 p_1}^2}{2} B_1 \frac{\langle m_2 \rangle}{\langle m_1 \rangle} \implies \text{cor}(p_2,\ p_3) > \text{cor}(p_1,\ p_2). \quad (31)$$

On comparing (20) and (29), while using (19), we get (31). Similar to Theorem 9, Theorem 12 demonstrates a tug-of-war between B_1 and B_2, where their relative magintudes dictate whether the true positive correlation cor(p_1, p_2) (cor(p_1, p_3)) or the false positive correlation cor(p_2, p_3) will dominate in single-cell protein measurements. Strong bursting in M_1 relative to M_2 can make cor(p_2, p_3) dominate.

The kinetic regime which allows cor(p_1, p_2) and cor(p_1, p_3) to achieve their upper bounds also allows cor(p_2, p_3) to reach its upper bound (compare the constraints in Theorem 6 and Theorem 11). However, the upper bound of cor(p_2, p_3) is larger than that of cor(p_1, p_2) and cor(p_1, p_3): 1 vs $1/\sqrt{2}$. Even though protein-protein correlations are not subject to τ-mediated loss in correlation, yet the relative gap between the upper bounds for cor(p_2, p_3) and cor(p_1, p_2) or cor(p_1, p_3) will cause the false positive correlation cor(p_2, p_3) to dominate the true positive correlations cor(p_1, p_2) and cor(p_1, p_3). For GRN inference from single-cell protein measurements, this implies that true positive and false positives will always be observed together. Again, the kinetic conditions which allow inference of the network also confound inference via the false positive edges.

The dependence of cor(p_2, p_3) on B_1 (cyan curves), B_2 (dark-green curves) and τ based on (29) is shown in Fig. 3c. We have also shown the respective curves for cor(p_1, p_2) based on (20) there for the sake of comparison. We also performed exact stochastic simulation to verify these results as shown in Figs. 3e.

IV. Loss of Information-Theoretic Measures for Simple GRNs

We also study the behavior of MI and the phi-mixing coefficient as a function of mRNA bursting and τ for the single gene and the two-gene cascade topologies using stochastic simulations ([29]). For a single-gene, like correlation, we observe a τ-mediated loss in MI between m_1 and p_1 (Fig. 4, top). For the two-gene cascade, there is τ-mediated loss in MI between m_1 and m_2 as well (Fig. 4, middle). Similar to correlation, MI between m_1 and p_1 appears to be larger in magnitude compared to m_1 and m_2. Finally, MI between p_1 and p_2 for the two-gene cascade in (10) exhibits an opposite trend to MI between m_1 and m_2, and has a τ-mediated loss when τ decreases rather than increasing (Fig. 4, bottom). The phi-mixing coefficient exhibited a similar behavior (Fig. 4b).

V. Loss of Inference Accuracy for Real Yeast GRN in Single-Cell RNA Sequencing Data

We collected the experimentally inferred GRN for *S. cerevisiae* from the yeastract database [43]. We collected transcriptome- and proteome-wide mRNA and protein degradation rates from [44] and [45], respectively. We used scRNA-seq data generated in [46]. We only retained cells grown under complete medium conditions as degradation rate measurements were made under these conditions [44], [45]. Further, the scRNA-seq data has 12 different genotypes, including the wildtype, and we used all the genotypes.

From the GRN, we extracted edges between master regulators (TFs without any incoming regulation), and their target genes. For these edges, we computed correlation between the mRNA counts of the TF and its target. These values are shown in red in Fig. 5a. We find that these true positive edges do not violate the upper bound on such correlations. Since the TF and the target genes can have different lifetimes, we

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.31.486623; this version posted April 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

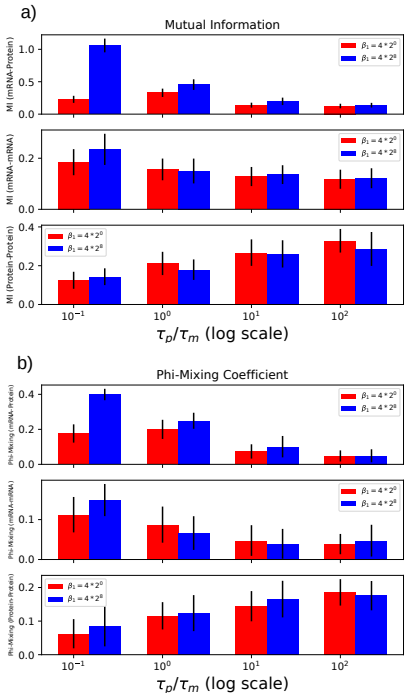

Fig. 4: **Plot of mutual information and phi-mixing coefficient for a single-gene and the two-gene cascade** as functions of B (B_1), and τ. (a) and (b) show results for mutual information (MI) and phi-mixing coefficient, respectively. (top) MI/phi-mixing coefficient between m_1 and p_1 for the single gene in (1) computed using exact stochastic simulations [29] for two values of B–one low $\propto \langle\beta\rangle = 4$ (red), and one high $\propto \langle\beta\rangle = 1024$ (blue), and four different values of τ. (middle) MI/phi-mixing coefficient between m_1 and m_2 for the two-gene cascade in (10). (bottom) MI/phi-mixing coefficient between p_1 and p_2 for the two-gene cascade in (10). The values of other kinetic parameters are the same as before (see caption of Fig. 1). For details on the regulation function $h(p_1)$, see caption of Fig. 2. Error bars show one standard deviation.

recompute the upper bound in (4), which becomes

$$\text{cor}(m_1,\, m_2) \leq \frac{1}{\sqrt{(1+\tau)}} \tag{32}$$

Interestingly, allowing different lifetimes, enables the upper bound to reach the maximum possible value of 1 when M_1 is much more stable than M_2. The upper bound in (32) is shown by the green curve in Figs. (5)a and (5)b.

Within the limit of statistical variability, the true positive correlations from yeast scRNA-seq data do not violate (32), and consequently face a τ-mediated loss (Fig. (5)a). The false positive edges (blue spheres in Fig. (5)a) are not constrained by the upper bound. For false positive edges, we calculate correlation between genes which do not regulate each other, but are regulated by the master regulators. This observation shows that the insights generated from small network topologies are valid for larger GRNs as well. Further, we observed a similar behavior when we calculated normalized mutual information [47] instead of correlation for the true positive and false positive edges in Fig. (5)b.

Discussion

We have established fundamental limits on inferrability of GRN topology from static single-cell mRNA and protein

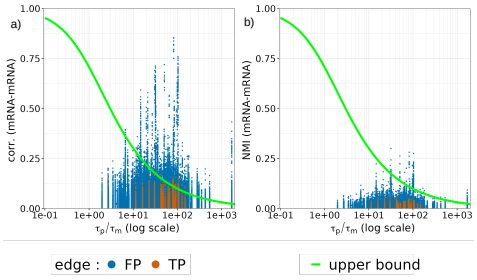

Fig. 5: **Loss in correlation and mutual information for yeast in single-cell RNA sequencing data.** (a) Comparision of correlation and normalized mutual information (NMI) between a TF and its target gene (red spheres) against correlation and normalized mutual information (NMI) between two genes with common regulators but no edge between them (blue spheres) are shown in (a) and (b), respectively. The green curves represent the upper bound on correlation between a TF and its target gene given in (32)

measurements. We find a relative stability-mediated loss in correlation and information-theoretic measures for mRNA species; when protein is more stable than mRNA, steady-state mRNA counts are not enough to infer the underlying GRN. This is exacerbated by the robustness of false positive correlations to this loss. The kinetic conditions which allow discovery of true positive correlations between a TF and its target gene also hinder GRN inference by amplification of false positive correlations.

We also found these constraints to be true for scRNA-seq data for yeast, *S. cerevisiae*, suggesting that the relative stability issue is true for real systems. This raises an important question on the limits of identifiability from static data. What about the dependence of other inference tasks, such as kinetic estimation, trajectory inference, clustering and differential expression, on relative stability of mRNA and protein and its propagation over GRN?

We used a simple model of gene expression, which does not incorporate complex processes such as post-transcriptional and post-transcriptional modifications. A future research direction is the establishment of constraints on GRN inferrability in more general settings.

scRNA-seq is not a static snapshot. Cells can be present in multiples states, and not not steady-state prior to sequencing. Can this be leveraged to circumvent the issues we have raised? This is an interesting problem to unpack.

GRN is essential for cellular functioning [2]–[4]. Consequently, a knowledge of its topology is important for understanding and controlling cellular functions. Experimental and computational efforts have been spent over the last two decades to unravel GRN topology for different species. However, the computational problem still remains unsolved. We have provided one explanation for this difficulty. This will motivate development of methods to circumvent the limitations we have unraveled.

Acknowledgment

References

[1] F. H. Crick, "On protein synthesis," *Symposia of the Society for Experimental Biology*, vol. 12, pp. 138–63, 1958.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.31.486623; this version posted April 1, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[2] M. Ptashne, "The chemistry of regulation of genes and other things," *Journal of Biological Chemistry*, vol. 289, no. 9, p. 5417–5435, 2014.

[3] G. Karlebach and R. Shamir, "Modelling and analysis of gene regulatory networks," *Nature Reviews Molecular Cell Biology*, vol. 9, no. 10, p. 770–780, 2008.

[4] P. K. Kreeger and D. A. Lauffenburger, "Cancer systems biology: A network modeling perspective," *Carcinogenesis*, vol. 31, no. 1, p. 2–8, 2009.

[5] M. M. Saint-Antoine and A. Singh, "Network inference in systems biology: Recent developments, challenges, and applications," *Current Opinion in Biotechnology*, vol. 63, p. 89–98, 2020.

[6] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, 2005.

[7] A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert, "TIGRESS: Trustful inference of gene regulation using stability selection," *BMC Systems Biology*, vol. 6, no. 1, 2012.

[8] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring regulatory networks from expression data using tree-based methods," *PLoS ONE*, vol. 5, no. 9, 2010.

[9] N. Singh and M. Vidyasagar, "bLARS: An algorithm to infer gene regulatory networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 2, p. 301–314, 2016.

[10] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano, "ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7, no. S1, 2006.

[11] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements," *Biocomputing 2000*, 1999.

[12] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles," *PLoS Biology*, vol. 5, no. 1, 2007.

[13] P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi, "Information-theoretic inference of large transcriptional regulatory networks," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, p. 1–9, 2007.

[14] T. E. Chan, M. P. Stumpf, and A. C. Babtie, "Gene regulatory network inference from single-cell data using multivariate information measures," *Cell Systems*, vol. 5, no. 3, 2017.

[15] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *Journal of Computational Biology*, vol. 7, no. 3-4, p. 601–620, 2000.

[16] J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink, and E. D. Jarvis, "Advances to Bayesian network inference for generating causal networks from observational biological data," *Bioinformatics*, vol. 20, no. 18, p. 3594–3603, 2004.

[17] T. Xu, L. Ou-Yang, X. Hu, and X.-F. Zhang, "Identifying gene network rewiring by integrating gene expression and gene network data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 6, p. 2079–2085, 2018.

[18] H. Zhao and Z.-H. Duan, "Cancer genetic network inference using Gaussian graphical models," *Bioinformatics and Biology Insights*, vol. 13, p. 117793221983940, 2019.

[19] M. Bansal, G. D. Gatta, and D. di Bernardo, "Inference of gene regulatory networks and compound mode of action from time course gene expression profiles," *Bioinformatics*, vol. 22, no. 7, p. 815–822, 2006.

[20] V. A. Huynh-Thu and G. Sanguinetti, "Combining tree-based and dynamical systems for the inference of gene regulatory networks," *Bioinformatics*, vol. 31, no. 10, p. 1614–1622, 2015.

[21] C. Biane, F. Delaplace, and T. Melliti, "Abductive network action inference for targeted therapy discovery," *Electronic Notes in Theoretical Computer Science*, vol. 335, p. 3–25, 2018.

[22] S. Barman and Y.-K. Kwon, "A Boolean network inference from time-series gene expression data using a genetic algorithm," *Bioinformatics*, vol. 34, no. 17, p. i927–i933, 2018.

[23] K. Kishan, R. Li, F. Cui, Q. Yu, and A. R. Haake, "GNE: A deep learning framework for gene network inference by aggregating biological information," *BMC Systems Biology*, vol. 13, no. S2, 2019.

[24] M. M. Saint-Antoine and A. Singh, "Evaluating pruning methods in gene network inference," in *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, 2019, pp. 1–7.

[25] V. A. Huynh-Thu and G. Sanguinetti, "Gene regulatory network inference: An introductory survey," *Methods in Molecular Biology*, p. 1–23, 2018.

[26] Y. Liu, A. Beyer, and R. Aebersold, "On the dependency of cellular protein levels on mRNA abundance," *Cell*, vol. 165, no. 3, p. 535–550, 2016.

[27] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 4, p. 623–656, 1948.

[28] I. A. Ibragimov, "Some limit theorems for stationary processes," *Theory of Probability and Its Applications*, vol. 7, no. 4, p. 349–382, 1962.

[29] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *J. Phys. Chem.*, vol. 81, no. 25, pp. 2340–2361, 1977-12-01.

[30] N. Singh, M. E. Ahsen, N. Challapalli, H.-S. Kim, M. A. White, and M. Vidyasagar, "Inferring genome-wide interaction networks using the phi-mixing coefficient, and applications to lung and breast cancer," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 4, no. 3, p. 123–139, 2018.

[31] D. Schnoerr, G. Sanguinetti, and R. Grima, "Approximation and inference methods for stochastic biochemical kinetics—a tutorial review," *J. Phys. A: Math. Theor.*, vol. 50, no. 9, p. 093001, 2017-01.

[32] A. Singh and M. Soltani, "Quantifying intrinsic and extrinsic variability in stochastic gene expression models," *Plos one*, vol. 8, no. 12, p. e84301, 2013.

[33] A. Singh, "Transient changes in intercellular protein variability identify sources of noise in gene expression," *Biophysical Journal*, vol. 107, no. 9, pp. 2214–2220, 2014.

[34] J. Paulsson, "Summing up the noise in gene networks," *Nature*, vol. 427, no. 6973, pp. 415–418, 2004.

[35] A. Hilfinger, T. M. Norman, G. Vinnicombe, and J. Paulsson, "Constraints on fluctuations in sparsely characterized biological systems," *Physical review letters*, vol. 116, no. 5, p. 058101, 2016.

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

[37] J. Peccoud and B. Ycart, "Markovian Modeling of Gene-Product Synthesis," *Theoretical Population Biology*, vol. 48, no. 2, pp. 222–234, 1995-10-01.

[38] V. Shahrezaei and P. S. Swain, "Analytical distributions for stochastic gene expression," *PNAS*, vol. 105, no. 45, pp. 17 256–17 261, 2008-11-11.

[39] H. Ochiai, T. Sugawara, T. Sakuma, and T. Yamamoto, "Stochastic promoter activation affects nanog expression variability in mouse embryonic stem cells," *Scientific reports*, vol. 4, no. 1, pp. 1–9, 2014.

[40] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry*. Elsevier, 1992, vol. 1.

[41] S. Modi, M. Soltani, and A. Singh, "Linear Noise Approximation for a Class of Piecewise Deterministic Markov Processes," in *2018 Annual American Control Conference (ACC)*, 2018-06, pp. 1993–1998.

[42] A. Singh and J. Hespanha, "Models for Multi-Specie Chemical Reactions Using Polynomial Stochastic Hybrid Systems," in *Proceedings of the 44th IEEE Conference on Decision and Control*, 2005-12, pp. 2969–2974.

[43] P. T. Monteiro, J. Oliveira, P. Pais, M. Antunes, M. Palma, M. Cavalheiro, M. Galocha, C. P. Godinho, L. C. Martins, N. Bourbon, *et al.*, "Yeastract+: a portal for cross-species comparative genomics of transcription regulation in yeasts," *Nucleic acids research*, vol. 48, no. D1, pp. D642–D649, 2020.

[44] B. Neymotin, R. Athanasiadou, and D. Gresham, "Determination of in vivo rna kinetics using rate-seq," *Rna*, vol. 20, no. 10, pp. 1645–1652, 2014.

[45] R. Christiano, N. Nagaraj, F. Fröhlich, and T. C. Walther, "Global proteome turnover analyses of the yeasts s. cerevisiae and s. pombe," *Cell reports*, vol. 9, no. 5, pp. 1959–1965, 2014.

[46] C. A. Jackson, D. M. Castro, G.-A. Saldi, R. Bonneau, and D. Gresham, "Gene regulatory network reconstruction using single-cell rna sequencing of barcoded genotypes in diverse environments," *elife*, vol. 9, p. e51254, 2020.

[47] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *Journal of machine learning research*, vol. 3, no. Dec, pp. 583–617, 2002.