Sparse identification of bacterial transcriptional regulation

Yu Fu[1], Ido Golding[1]

1. Department of Physics, College of Engineering, University of Illinois at Urbana-Champaign

Introduction

In bacteria, while the connection between a gene's regulatory architecture and its expression is well-established in the context of individual model gene circuits, a genome-scale perspective of such connection is lacking. A recent collaborative project from our lab aimed to develop a genome-wide classification of transcriptional regulation based on the response of each gene to its own replication. However, some genes exhibit transcriptional behavior that falls outside the current theoretical framework. Meanwhile, many data-driven machine learning algorithms aimed to construct ordinary differential equation models have emerged in recent years. These algorithms may be helpful in discovering hidden regulatory mechanisms beyond the current model. In this work, we aim to use machine learning-aided methods to perform a genome-wide classification of transcriptional behavior based on real *Escherichia coli* (*E.*) *coli* data. After considering several different algorithms, we now focus on the sparse identification of nonlinear dynamics (SINDy) algorithm, which is more practical due to its high computational efficiency. We reported some simple test cases of remarkable agreement between the simulation data generated from the mRNA transcription model and the ODE model learned by SINDy. We anticipate that our attempt will serve as a starting point for learning more complicated transcriptional regulation models using data-driven methods.

Background

Fundamental model of gene replication[1]

- Within one cell life cycle t_D, mRNA transcripts at a rate k_t.
- At time t_r, gene replicates and mRNA transcription doubles: $2k_t$.
- Meanwhile, mRNA will also degrade at a constant rate k_d

$$\frac{d\overline{m}(t)}{dt} = \begin{cases} k_t - k_d\overline{m} & 0 < t < t_r \\ 2k_t - k_d\overline{m} & t_r < t < t_D \end{cases}$$

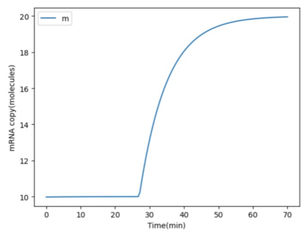


Figure 2: analytical solution of mRNA replication

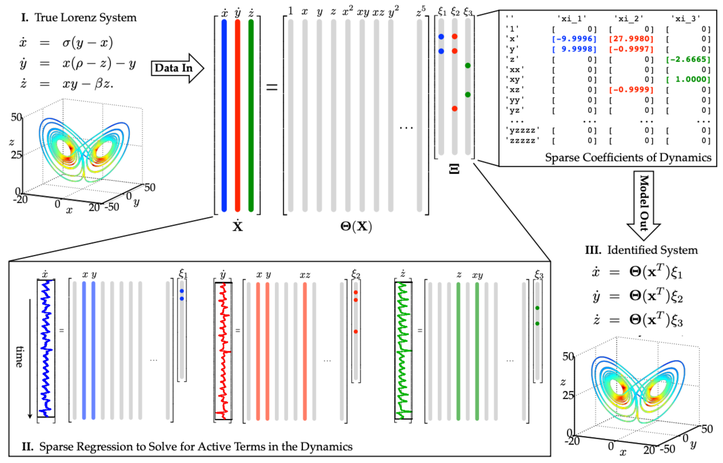
Reference

[1]Peterson, Joseph R., et al. "Effects of DNA replication on mRNA noise." *Proceedings of the National Academy of Sciences* 112.52 (2015): 15886-15891.

[2] Yanai, Itai et al. "Transcription-replication interactions reveal principles of bacterial genome regulation." *Research square* rs.3.rs-2724389. 31 Mar. 2023, doi:10.21203/rs.3.rs-2724389/v1. Preprint.

[3] SINDy-PI: A Robust Algorithm for Parallel Implicit Sparse Identification of Nonlinear Dynamics

Sparse identification of nonlinear dynamics (SINDy)



Choice of SINDy library: why polynomial library?

- We introduced simulated protein by ground model
- We use $\{m, 1, p, p^2, \ldots, p^n\}$ as the library.
- We assume protein would regulate transcription rate.

ground model

$$\frac{d}{dt}m(t) = k_0 - \delta_m m(t)$$
$$\frac{d}{dt}p(t) = k_1 m(t) - \delta_p p(t)$$

mechanism 1: Rates of transcription from regulated genes

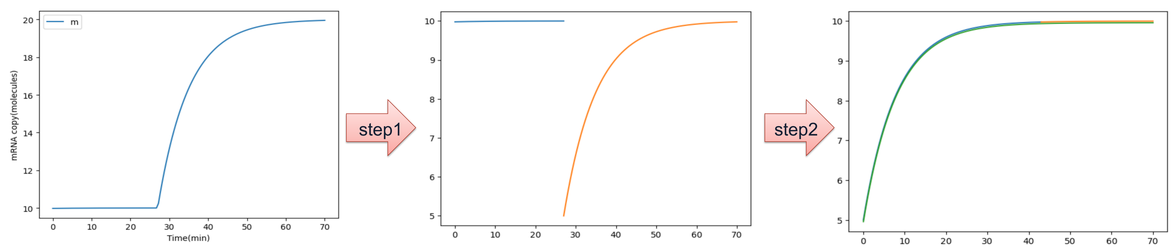
$$\text{rate of activated transcription} = \alpha_0 + \alpha\frac{[P]/K}{1+[P]/K} \sim \alpha_0 + \alpha\left(\frac{[P]}{K} - \left(\frac{[P]}{K}\right)^2\right)$$

$$\text{rate of repressible transcription} = \alpha_0 + \alpha\frac{1}{1+[P]/K} \sim \alpha_0 + \alpha\left(1 - \frac{[P]}{K}\right)$$

- Other more complicated mechanism with multiple protein can always be represented by polynomial library with Taylor expansion.

Rescaling of mRNA data

- motivation: SINDy can only infer model with continuous parameter while in the fundamental model the transcription rate is a step function.
- step1: divide the mRNA number by 2 after gene replicates.
- step2: set the original gene replication time as starting time.



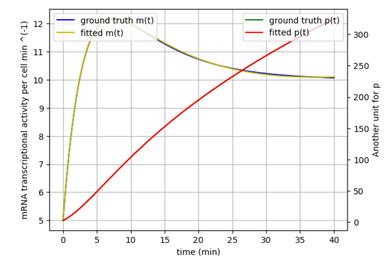
$$\frac{d\overline{m}(t)}{dt} = \begin{cases} k_t - k_d\overline{m} & 0 < t < t_r \\ 2k_t - k_d\overline{m} & t_r < t < t_D \end{cases} \quad \text{After rescaling} \quad \frac{d\overline{m}'(t)}{dt} = k_t - k_d\overline{m}'$$

Results on simulated data

Overshoot activation

Simulated model

$$\frac{dm}{dt} = k_t\left(1 + (\alpha - 1)e^{-k_b t}\right) - k_d m$$
$$k_b = 0.3\sim\text{min}^{-1}, k_t = 1.26\sim\text{min}^{-1},$$
$$k_d = 0.126\sim\text{min}^{-1}, t_D = 70\sim\text{min}, \ \alpha = 3$$
$$\frac{dm}{dt} = 1.26 + 2.52e^{-0.3t} - 0.126m$$
$$m(0) = 5$$
$$\frac{dp}{dt} = m - \frac{p}{70}$$
$$p(0) = 2.5$$

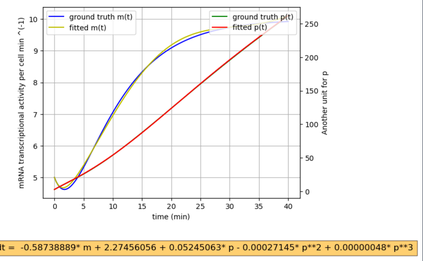


fitted model: dm/dt = -0.36071719* m + 5.00336425 - 0.00938148* p + 0.00001606* p**2

Drop repression

Simulated model

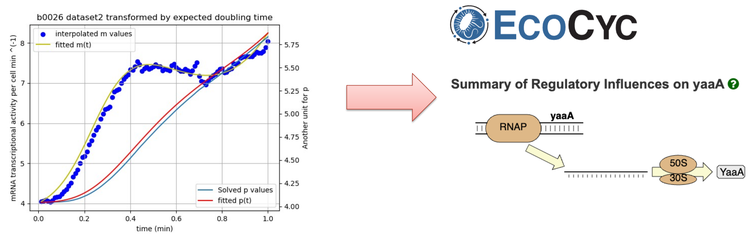
$$\frac{dm}{dt} = k_t\left(1 + (\alpha - 1)e^{-k_b t}\right) - k_d m$$
$$k_b = 0.3\sim\text{min}^{-1}, k_t = 1.26\sim\text{min}^{-1},$$
$$k_d = 0.126\sim\text{min}^{-1}, t_D = 70\sim\text{min}, \ \alpha = 0.1$$
$$\frac{dm}{dt} = 1.26 - 1.134e^{-0.3t} - 0.126m$$
$$m(0) = 5$$
$$\frac{dp}{dt} = m - \frac{p}{70}$$
$$p(0) = 2.5$$



ted model: dm/dt = -0.58738889* m + 2.27456056 + 0.05245063* p - 0.00027145* p**2 + 0.00000048* p**3

Future research

- Apply SINDy on experimental data of *Escherichia coli* (*E.*) *coli*.
- *Compare the modeling result with database such EcoCyc to see if the inferred model consists with recorded regulatory mechanism and if it can predict unknown mechanism.*



Conclusions

- With SINDy algorithm, we have shown that SINDy can correctly modeling activation and repressive effect from time-series data of mRNA. Number.

Acknowledgement

Thanks for Golding lab. Especially thanks for Prof. Ido Golding, Tianyou Yao, Kevin McDonald and Yuncong Geng for discussion and instruction.

ILLINOIS