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

I. Abstract

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.

II. Background

Fundamental model of gene replication[1]

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

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

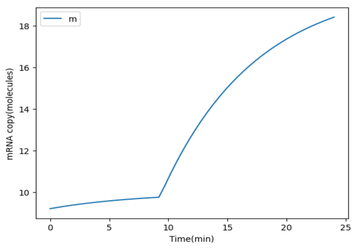

Fig 1. Analytical solution of the model above.

III. Sparse identification of nonlinear dynamics (SINDy)

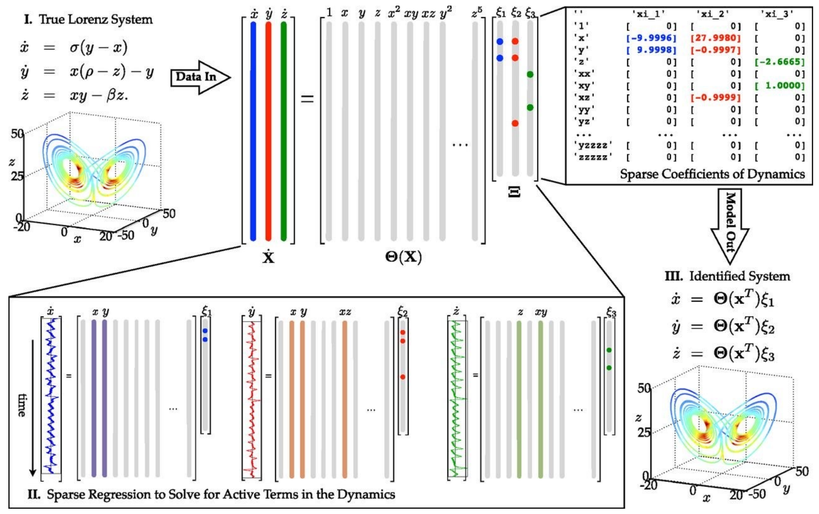

Fig 2. Schematic of the SINDy algorithm, demonstrated on the Lorenz equations. (Image from Ref [3].)

IV. Choice of SINDy library: Why polynomial library?

- We introduce a simulated protein by ground model
- We use $\{m, 1, p, p^2, \dots, p^n\}$ as the library.
- We assume that the protein regulates transcription rate:

$$\frac{d}{dt}p(t) = p(t) - \frac{m(t)}{t_D}$$

- We assume the transcription rate k_t is not constant but is regulated by the protein:

$$\frac{d}{dt}m(t) = k_t(p) - k_d m(t)$$

- $k_t(p)$ can take various forms :
  - Activation

$$k_t(p) = \alpha_0 + \alpha\frac{p/K}{1+p/K} \sim \alpha_0 + \alpha(\frac{p}{K} - (\frac{p}{K})^2)$$

  - Repression:

$$k_t(p) = \alpha_0 + \alpha\frac{1}{1+p/K} \sim \alpha_0 + \alpha(1 - \frac{p}{K})$$

  - Or other more complicated mechanism with multiple proteins
- $k_t(p)$ can always be represented by a polynomial library through Taylor expansion.

V. Rescaling of mRNA data

- Motivation: SINDy can only infer models with continuous parameters, whereas in the fundamental model, the transcription rate is a step function.
- Step1: Divide the mRNA number by 2 after the gene replicates.
- Step2: Set the original gene replication time as the starting time.

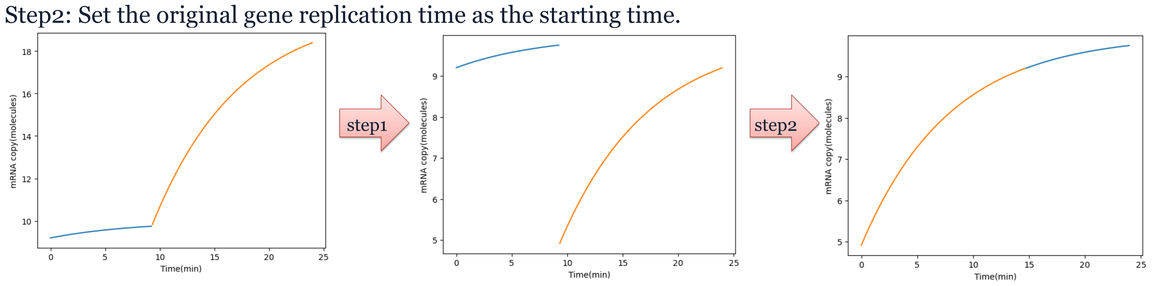

Fig 3. Illustration of mRNA data rescaling

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

VI. Results on simulated data

Example 1: mRNA activated model

- mRNA Activated Model for Simulation:

$$\frac{dm}{dt} = 1.26 + 2.52e^{-0.3t} - 0.126m$$

- Protein Model for Simulation:

$$\frac{dp}{dt} = m - \frac{p}{24}$$

- Inferred Model for mRNA

$$\frac{dm}{dt} = 5.1214 - 0.0076p - 0.3779m$$

- Reconstructed Protein-Repression Model:

$$\frac{dm}{dt} = 1.26 + 3.86\frac{1}{1+\frac{p}{505.74}} - 0.378m$$

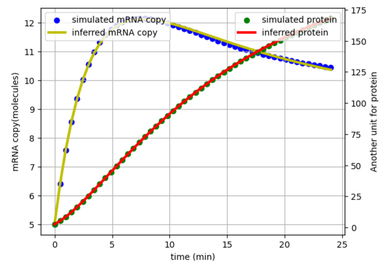

Fig 4. Comparison of simulated mRNA activated model and inferred model

Example 2: mRNA repressed model

- mRNA Repressed Model for Simulation:

$$\frac{dm}{dt} = 1.26 - 1.134e^{-0.3t} - 0.126m$$

- Protein Model for Simulation:

$$\frac{dp}{dt} = m - \frac{p}{24}$$

- Inferred Model for mRNA

$$\frac{dm}{dt} = 2.111 + 0.0570p - 0.0003p^2 - 0.5642m$$

- Reconstructed Protein-Activation Model:

$$\frac{dm}{dt} = 2.11 + 12.80\frac{\frac{p}{224.43}}{1+\frac{p}{224.43}} - 0.564m$$

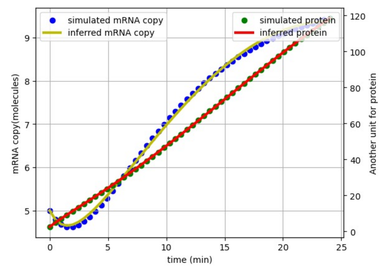

Fig 5. Comparison of simulated mRNA repressed model and inferred model

VII. Future research

- Apply SINDy to experimental data on *Escherichia coli* (*E. coli*)
- Compare the inferred models with databases such as EcoCyc to see if they are consistent with recorded regulatory mechanisms and whether they can predict unknown mechanisms.

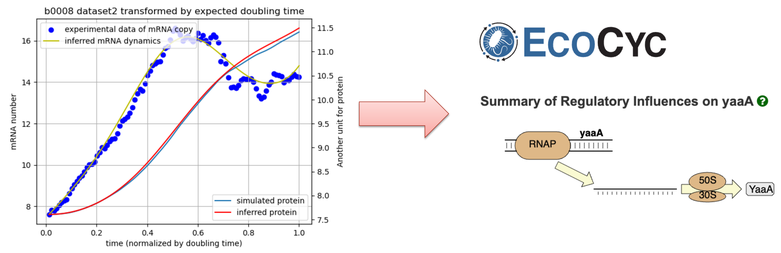

Fig. 6. Schematic comparing the inferred models with the EcoCyc database. The figure on the right is from ref. [4].

VIII. Conclusions

Using the SINDy algorithm, we have demonstrated that SINDy can correctly model activation and repression from time-series data of mRNA number.

IX. Acknowledgements

We are grateful to all members of the Golding lab, especially Prof. Ido Golding, Tianyou Yao, Kevin McDonald and Yuncong Geng for their valuable scientific discussions and guidance. We also gratefully acknowledge the computing resources provided by the National Center for Supercomputing Applications.

X. References

[1] Peterson, J. R., Cole, J. A., Fei, J., Ha, T., & Luthey-Schulten, Z. A. (2015). Effects of DNA replication on mRNA noise. *Proceedings of the National Academy of Sciences*, *112*(52), 15886-15891.

[2] Yanai, I., Pountain, A., Jiang, P., Yao, T., Homaee, E., Guan, Y., Podkowik, M., Shopsin, B., Torres, V., & Golding, I. (2023). Transcription-replication interactions reveal principles of bacterial genome regulation. *Research square*, rs.3.rs-2724389. https://doi.org/10.21203/rs.3.rs-2724389/v1

[3] Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, *113*(15), 3932-3937.

[4] Keseler, I. M., Gama-Castro, S., Mackie, A., Billington, R., Bonavides-Martínez, C., Caspi, R., Kothari, A., Krummenacker, M., Midford, P. E., Muñiz-Rascado, L., Ong, W. K., Paley, S., Santos-Zavaleta, A., Subhraveti, P., Tierrafría, V. H., Wolfe, A. J., Collado-Vides, J., Paulsen, I. T., & Karp, P. D. (2021). The EcoCyc Database in 2021. *Frontiers in microbiology*, *12*, 711077. https://doi.org/10.3389/fmicb.2021.711077

ILLINOIS