Structure-Based Inverse Reinforcement Learning for Quantification of Biological Knowledge

Amirhossein Ravari, Seyede Fatemeh Ghoreishi, and Mahdi Imani

2023 IEEE Conference on Artificial Intelligence (CAI) | 979-8-3503-3984-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/CAI54212.2023.00126

***Abstract*— Gene regulatory networks (GRNs) play crucial roles in various cellular processes, including stress response, DNA repair, and the mechanisms involved in complex diseases such as cancer. Biologists are involved in most biological analyses. Thus, quantifying their policies reflected in available biological data can significantly help us to better understand these complex systems. The primary challenges preventing the utilization of existing machine learning, particularly inverse reinforcement learning techniques, to quantify biologists' knowledge are the limitations and huge amount of uncertainty in biological data. This paper leverages the network-like structure of GRNs to define expert reward functions that contain exponentially fewer parameters than regular reward models. Numerical experiments using mammalian cell cycle and synthetic gene-expression data demonstrate the superior performance of the proposed method in quantifying biologists' policies.**

I. Introduction

Gene regulatory networks (GRNs) are crucial for a range of cellular functions, including stress response, DNA repair, and mechanisms that contribute to complex diseases like cancer [1]–[11]. Despite the rapid advances in statistical and computational techniques in systems biology, biologists remain integral components of most genomics analyses. Biologists are skilled at extracting knowledge from experience, which is evident in the acquired data following interventions such as drug prescriptions or perturbations. The biologists' decisions represent their understanding of the mechanisms of complex systems. Therefore, effectively quantifying that knowledge can significantly aid in various biological analyses.

Various techniques have been developed in the field of inverse reinforcement learning (IRL) to quantify expert policies in the presence of large amounts of expert-acquired data [12]–[21]. These techniques have found success in areas such as computer games by leveraging large amounts of expert data and employing methods such as maximum entropy IRL and other nonlinear reward function IRL techniques, such as neural networks. However, biological data is often extremely limited and carries a huge amount of uncertainty, which limit the applicability of existing techniques.

This paper introduces the network-based structure of GRNs to develop reward models that require significantly fewer parameters than regular reward models for non-structured systems. Specifically, our proposed models have a linear parameter growth rate based on the size of the GRN, which captures biological objectives, and enables the quantification of biologist policies in large GRNs with limited biological data.

A. Ravari and M. Imani are with the Department of Electrical and Computer Engineering, and S.F. Ghoreishi is with the Department of Civil and Environmental Engineering and Khoury College of Computer Sciences at Northeastern University. Emails: ravari.a@northeastern.edu, f.ghoreishi@northeastern.edu, m.imani@northeastern.edu

II. Mathematical Preliminaries

The GRN model can be represented by a Markov decision process (MDP) [22]–[28]. This MDP model of GRNs with d genes can be formally defined by a 5-tuple $\langle \mathcal{X}, \mathcal{A}, T, R, \gamma \rangle$, where $\mathcal{X} = \{0,1\}^d$ is the state space, $\mathcal{A}$ is the action space, $T : \mathcal{X} \times \mathcal{A} \times \mathcal{X}$ is the state transition probability function such that $T(\mathbf{x}, \mathbf{a}, \mathbf{x}') = p(\mathbf{x}' \mid \mathbf{x}, \mathbf{a})$ represents the probability of moving to state $\mathbf{x}'$ after taking action $\mathbf{a}$ in state $\mathbf{x}$, $R : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is a bounded reward function such that $R(\mathbf{x}, \mathbf{a})$ encodes the reward earned when action $\mathbf{a}$ is taken in state $\mathbf{x}$, and $0 < \gamma < 1$ is a discount factor.

A deterministic stationary policy π for an MDP is a mapping $\pi : \mathcal{X} \to \mathcal{A}$ from states to actions. The expected discounted reward function at state $\mathbf{x} \in \mathcal{X}$ after taking action $\mathbf{a} \in \mathcal{A}$ and following policy π afterward is defined as:

$$Q^{\pi}(\mathbf{x}, \mathbf{a}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(\mathbf{x}_t, \mathbf{a}_t) \mid \mathbf{x}_0 = \mathbf{x}, \mathbf{a}_0 = \mathbf{a}, \mathbf{a}_{1:\infty} \sim \pi\right]. \tag{1}$$

According to (1), the expected return under the optimal policy $\pi^*(\mathbf{x}) = \operatorname{argmax}_{\pi \in \mathbf{\Pi}} Q^{\pi}(\mathbf{x}, \mathbf{a})$, for all $\mathbf{x} \in \mathcal{X}$ and $\mathbf{a} \in \mathcal{A}$. An optimal stationary policy π^* attains the maximum expected return for all states as: $\pi^*(\mathbf{x}) = \max_{\mathbf{a} \in \mathcal{A}} Q^*(\mathbf{x}, \mathbf{a})$.

III. The proposed Framework

Let D_T contain all available realizations of an expert/biologist, denoted by:

$$D_T = \{(\tilde{\mathbf{x}}_1, \tilde{\mathbf{a}}_1), (\tilde{\mathbf{x}}_2, \tilde{\mathbf{a}}_2), ..., (\tilde{\mathbf{x}}_T, \tilde{\mathbf{a}}_T)\}, \tag{2}$$

where $\tilde{\mathbf{a}}_r$ is the taken action/intervention by a biologist at the state $\tilde{\mathbf{x}}_r$ at time step r. These actions could be interventions in genomics, that are often drugs that flip the value of single or multiple genes to alter the dynamics of these systems.

Consider the mammalian cell cycle with 10 genes shown in Fig. 1(a) [29]. This gene expression contains $2^{10} = 1,024$ possible states: $\mathbf{x}^1 = [00\cdots0]^T, \mathbf{x}^2 = [00\cdots1]^T, ..., \mathbf{x}^{1,024} = [11\cdots1]^T$. Let also $\mathbf{a}^1 = [00\cdots0]^T$ and $\mathbf{a}^2 = [00\cdots01]^T$ be the actions/interventions available to biologists during the interventional process, where $\mathbf{a}^1$ denotes no intervention, and $\mathbf{a}^2$ alters the state value of CycB gene. A typical linear model for the expert reward function contains elements with the

979-8-3503-3984-0/23/$31.00 ©2023 IEEE
DOI 10.1109/CAI54212.2023.00126

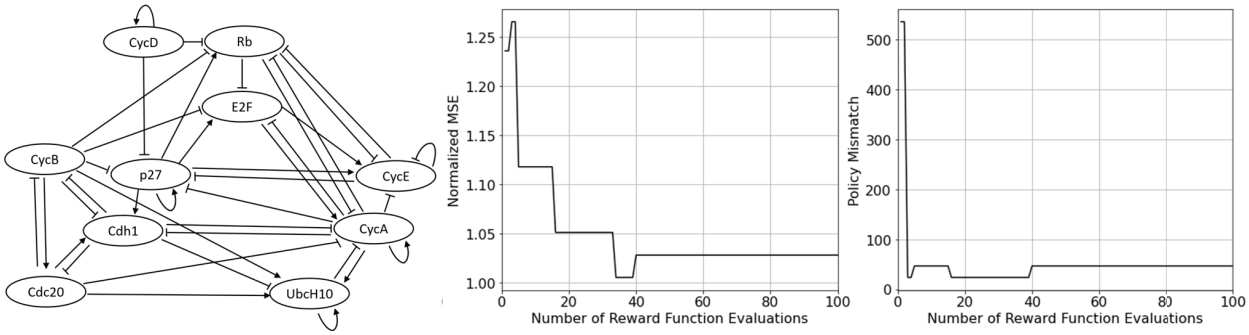

Fig. 1: (a) Pathway diagram for the cell-cycle mammalian network; (b) the normalized MSE of the inferred and the true model; (c) the mismatch between the inferred biologist policy and the true policy.

size of state space, such as $R_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{a}) = \theta^1 1_{\mathbf{x}=\mathbf{x}^1} + \theta^2 1_{\mathbf{x}=\mathbf{x}^2} + \cdots + \theta^{1023} 1_{\mathbf{x}=\mathbf{x}^{2023}} + \theta^{1024} 1_{\mathbf{x}=\mathbf{x}^{1024}} + \theta^{1025} 1_{\mathbf{a}=\mathbf{a}^2}$, where θ^i denotes the reward value if the system is at state $\mathbf{x}^i$, for $i = 1, ..., 1,024$. The positive values of θ^i correspond to the desirability of being at $\mathbf{x}^i$, whereas negative values are associated with the undesirability of being at $\mathbf{x}^i$ (e.g., $\mathbf{x}^i$ is associated with cancer). Meanwhile, θ^{1025} denotes the cost of taking action $\mathbf{a}^2$, which could come from real expenses associated with action $\mathbf{a}^2$ (i.e., drug cost) or possible side effects of taking this action. The parameters of typical reward models for non-structured systems often grow with the number of states (e.g., exponentially with the number of genes). For instance, the simplest linear reward model for a GRN with d genes and m actions contains $2^d + m$ parameters. This large number of parameters prevents reliable quantification of biologists' knowledge, given often limited biological data.

In genomics, the desirable or undesirable activities of one or some specific genes often correspond to undesirable conditions. Taking advantage of this biological knowledge and the network-like structure of GRNs, the expert reward function can be represented component-wise as: $R_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{a}) = \theta^1 \mathbf{x}(1) + \theta^2 \mathbf{x}(2) + \cdots + \theta^{10} \mathbf{x}(10) + \theta^{11} 1_{\mathbf{a}=\mathbf{a}^2}$, where $\theta^i \in [-1, 1]$ corresponds to desirability or undesirability of the activation of the ith gene (for $i \leq 10$) and θ^{11} denotes the action cost. This reward model represented the mammalian cell cycle network with 10 genes and 2 actions consisting of 11 parameters (instead of 1025 parameters in non-structured cases). It should be noted that the biologically-inspired reward model can be more complex (e.g., nonlinear and stochastic), depending on the available prior biological knowledge and the application.

We model the biologists as semi-optimal decision-makers, where the imperfections of their decisions are taken into account for quantifying the relevant biological objective. Given the available biologist-acquired data, D_T in (2), the optimal quantification of biological knowledge can be expressed as:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} \log P(D_T \mid \boldsymbol{\theta}), \tag{3}$$

where $P(.)$ is a probability mass function, and

$$\log P(D_T \mid \boldsymbol{\theta}) = \sum_{k=1}^{T} \log P(\tilde{\mathbf{a}}_k \mid \tilde{\mathbf{x}}_k, \boldsymbol{\theta})$$

Boltzmann:

$$= \sum_{k=1}^{T} \log \left[\frac{\exp\left(\eta\, Q^*_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_k, \tilde{\mathbf{a}}_k)\right)}{\sum_{\mathbf{a}\in\mathcal{A}} \exp\left(\eta\, Q^*_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_k, \mathbf{a})\right)} \right] \tag{4}$$

ϵ-greedy:

$$= \sum_{k=1}^{T} \log \left[\left(1 - \epsilon + \frac{\epsilon}{|\mathcal{A}|}\right) 1_{\tilde{\mathbf{a}}_k = \pi^*_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_k)} + \frac{\epsilon}{|\mathcal{A}|} 1_{\tilde{\mathbf{a}}_k \neq \pi^*_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_k)} \right]$$

The $\pi^*_{\boldsymbol{\theta}}$ and $Q^*_{\boldsymbol{\theta}}$ in (4) are the optimal policy and Q-value associated with the reward function parameterized by $\boldsymbol{\theta}$ (i.e., $R_{\boldsymbol{\theta}}$). The Boltzmann policy models the expert decisions as $p(\mathbf{a} \mid \mathbf{x}, \boldsymbol{\theta}) \propto \exp\left(\eta\, Q^*_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{a})\right)$, where η models the exert imperfectness, whereas the ϵ-greedy model considers taking optimal action (i.e., $\pi^*_{\boldsymbol{\theta}}$) with probability $1-\epsilon$ and random actions with probability ϵ. These two well-known models allow for the quantification of non-optimal biologists. Meanwhile, the parameters representing the expert confidence (η and ϵ) can also be included as part of parameter $\boldsymbol{\theta}$ and be inferred during the optimization in (3).

The performance of the proposed framework is examined using the mammalian cell-cycle network using 40 gene-expression data acquired by a biologist during intervention with the goal of suppressing the activation of CycD and Rb genes. i.e., $\boldsymbol{\theta}^* = [-1, -1, 0, \cdots, 0, 0]$. Fig. 1(b)-(c) represents the normalized mean square error (MSE) of inferred and true reward function and policy mismatch between the inferred biologist's policy and the expert's true policy. One can see that the policy mismatch and normalized MSE values decrease with each iteration of the proposed method. This reduction is significant for policy mismatch, indicating that despite a relatively large normalized MSE in the early steps, the policy under those inferred biologist reward functions becomes similar to the policy under the true reward function.

IV. Conclusion

In conclusion, this paper developed a framework for quantifying biologists' knowledge using limited and uncertain biological data. By leveraging the network-like structure

of gene regulatory networks, this paper introduces expert reward functions with exponentially fewer parameters than traditional models. This alleviates the challenges of utilizing machine learning techniques, specifically inverse reinforcement learning, and enables the quantification of biologists' knowledge. Through numerical experiments on mammalian cell cycle and synthetic gene-expression data, the proposed method has shown to be highly effective in quantifying biologists' policies, thus paving the way for a better understanding of complex biological systems.

ACKNOWLEDGMENT

The authors acknowledge the support of the National Institute of Health award 1R21EB032480-01, National Science Foundation award IIS-2202395, ARMY Research Office award W911NF2110299, and Oracle for Research program.

REFERENCES

[1] V. Tiwari and D. M. Wilson III, "DNA damage and associated DNA repair defects in disease and premature aging," *The American Journal of Human Genetics*, vol. 105, no. 2, pp. 237–257, 2019.

[2] E. H. Davidson, *The regulatory genome: gene regulatory networks in development and evolution*. Elsevier, 2010.

[3] I. Shmulevich, E. R. Dougherty, and W. Zhang, "From Boolean to probabilistic Boolean networks as models of genetic regulatory networks," *Proceedings of the IEEE*, vol. 90, no. 11, pp. 1778–1792, 2002.

[4] Y. Li, J. Xiao, L. Chen, X. Huang, Z. Cheng, B. Han, Q. Zhang, and C. Wu, "Rice functional genomics research: past decade and future," *Molecular plant*, vol. 11, no. 3, pp. 359–380, 2018.

[5] A. Pratapa, A. P. Jalihal, J. N. Law, A. Bharadwaj, and T. Murali, "Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data," *Nature methods*, vol. 17, no. 2, pp. 147–154, 2020.

[6] X. Qian and E. R. Dougherty, "Intervention in gene regulatory networks via phenotypically constrained control policies based on long-run behavior," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 1, pp. 123–136, 2011.

[7] M. Imani and U. M. Braga-Neto, "Finite-horizon LQR controller for partially-observed Boolean dynamical systems," *Automatica*, vol. 95, pp. 172–179, 2018.

[8] M. Imani and U. M. Braga-Neto, "Point-based methodology to monitor and control gene regulatory networks via noisy measurements," *IEEE Transactions on Control Systems Technology*, 2018.

[9] M. Imani and U. M. Braga-Neto, "Control of gene regulatory networks with noisy measurements and uncertain inputs," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 2, pp. 760–769, 2018.

[10] M. Imani and U. Braga-Neto, "Control of gene regulatory networks using Bayesian inverse reinforcement learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 4, pp. 1250–1261, 2019.

[11] M. Imani, R. Dehghannasiri, U. M. Braga-Neto, and E. R. Dougherty, "Sequential experimental design for optimal structural intervention in gene regulatory networks based on the mean objective cost of uncertainty," *Cancer informatics*, vol. 17, p. 1176935118790247, 2018.

[12] J. Rivera-Villicana, F. Zambetta, J. Harland, and M. Berry, "Exploring apprenticeship learning for player modelling in interactive narratives," in *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, pp. 645–652, 2019.

[13] G. Neu and C. Szepesvári, "Apprenticeship learning using inverse reinforcement learning and gradient methods," pp. 295–302, 2007.

[14] R. Takanobu, H. Zhu, and M. Huang, "Guided dialog policy learning: Reward estimation for multi-domain task-oriented dialog," *arXiv preprint arXiv:1908.10719*, 2019.

[15] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the twenty-first international conference on Machine learning*, p. 1, ACM, 2004.

[16] A. Y. Ng and S. J. Russell, "Algorithms for inverse reinforcement learning.," in *Icml*, vol. 1, p. 2, 2000.

[17] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning.," in *AAAI*, vol. 8, pp. 1433–1438, Chicago, IL, USA, 2008.

[18] M. Imani and S. F. Ghoreishi, "Scalable inverse reinforcement learning through multi-fidelity Bayesian optimization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 8, pp. 4125–4132, 2022.

[19] M. Imani and U. Braga-Neto, "Optimal control of gene regulatory networks with unknown cost function," in *Proceedings of the 2018 American Control Conference (ACC 2018)*, pp. 3939–3944, IEEE, 2018.

[20] S. Arora and P. Doshi, "A survey of inverse reinforcement learning: Challenges, methods and progress," *Artificial Intelligence*, vol. 297, p. 103500, 2021.

[21] C. You, J. Lu, D. Filev, and P. Tsiotras, "Advanced planning for autonomous vehicles using reinforcement learning and deep inverse reinforcement learning," *Robotics and Autonomous Systems*, vol. 114, pp. 1–18, 2019.

[22] I. Shmulevich and E. R. Dougherty, *Probabilistic Boolean networks: the modeling and control of gene regulatory networks*. SIAM, 2010.

[23] M. Alali and M. Imani, "Reinforcement learning data-acquiring for causal inference of regulatory networks," in *American Control Conference (ACC)*, IEEE, 2023.

[24] M. Alali and M. Imani, "Inference of regulatory networks through temporally sparse data," *Frontiers in control engineering*, vol. 3, 2022.

[25] A. Ravari, S. F. Ghoreishi, and M. Imani, "Optimal recursive expert-enabled inference in regulatory networks," *IEEE Control Systems Letters*, vol. 7, pp. 1027–1032, 2022.

[26] M. Imani and S. F. Ghoreishi, "Optimal finite-horizon perturbation policy for inference of gene regulatory networks," *IEEE Intelligent Systems*, 2020.

[27] M. Imani and U. Braga-Neto, "Gene regulatory network state estimation from arbitrary correlated measurements," *EURASIP Journal on Advances in Signal Processing*, vol. 2018, no. 1, pp. 1–10, 2018.

[28] L. D. McClenny, M. Imani, and U. Braga-Neto, *BoolFilter package vignette*, 2017.

[29] A. Fauré, A. Naldi, C. Chaouiya, and D. Thieffry, "Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle," *Bioinformatics*, vol. 22, no. 14, pp. e124–e131, 2006.