A Data-Driven Optimization Method for Coarse-Graining Gene Regulatory Networks

Cristian Caranica[1], Mingyang Lu[1]

***Short Abstract* — We developed a coarse-graining method for large gene regulatory networks (GRNs) based on gene expression data and network topology. The purpose is to take advantage of network redundancy to build small circuits that can capture network's behavior. Our method is based on clustering the genes and the models describing the network's behavior, followed by an optimization method that samples the best coarse-grained circuits (CGCs) producing the same expression pattern as the GRN. We benchmark and demonstrate the usage of the method with different synthetic, literature-based, and bioinformatics-derived GRNs.**

***Keywords* — Coarse-graining, gene regulatory networks, hierarchical clustering, Markov Chain Monte Carlo methods.**

I. PURPOSE

ONE of the biggest challenges in systems biology is to understand how various genes in a regulatory network collectively perform their functions and control network dynamics. This task becomes extremely hard to tackle in the case of big networks with hundreds of genes/edges, many of which have redundant regulatory roles and functions.

While a variety of methods have been developed to model reduction of biochemical networks [1], they usually require detailed mathematical description of a dynamical system and the corresponding kinetic parameters, which in many cases are not available.

Here, we present a data-driven method for coarse-graining large GRNs, using ensemble-based mathematical modeling, dimensionality reduction and gene circuit optimization. Our approach requires GRN's topology as the only input for a robust network coarse graining.

II. METHODS

The GRN coarse-graining method consists of three major steps. First, simulations of a large GRN are performed for an ensemble of models with randomly selected kinetic parameters using RACIPE [2]. Second, hierarchical clustering analysis is applied to the simulated data to identify the coarse-grained nodes, the gene grouping scheme, the putative regulatory interactions, and the nature of the interactions (activation or inhibition) between the coarse-grained nodes. Third, an enhanced sampling is performed using either Metropolis-Hastings (MH), simulated annealing (SA) or parallel tempering (PT) algorithm [3] to identify optimized coarse-grained circuits.

III. RESULTS

We benchmarked the robustness of our method using two synthetic circuits. Each circuit is expanded to large networks with different levels of redundancy and edge perturbation. In both cases our method can successfully recover the original circuits even when edge perturbation level is as high as 60%. As expected, the PT algorithm overall produced the best CGCs, but it was computationally less efficient when sampling space was large. SA performed almost as well as PT, and it was usually much faster. MH was fast but was outperformed by the other sampling algorithms when edge perturbation level was above 40%.

We also applied our method to several literature-based and bioinformatics-derived GRNs, and we observed the following strengths of our approach. First, in all cases we could reproduce GRN's expression patterns using CGCs of small sizes. Second, the method worked well to handle large networks, e.g., those containing as many as 33 genes and 357 edges. Third, the identified CGCs preserved rare populations of cellular states. Fourth, the performance of the method was insensitive to the choice of number of coarse-grained nodes. In our test cases, 3 or 4 nodes were usually sufficient to capture network state distributions.

IV. CONCLUSION

We developed a robust coarse-graining method for gene regulatory networks. Our data-driven method can be used to bridge the gap between a top-down and a bottom-up modeling approach.

Acknowledgements: This work was funded by grant NIH- R35GM128717.

[1]Department of Bioengineering, Northeastern University
Contact E-mail: c.caranica@northeastern.edu

REFERENCES

[1] Snowden TJ, et al., Methods of Model Reduction for Large-Scale Biological Systems: A Survey of Current Methods and Trends. Bull Math Biol. 2017;79: 1449–1486. doi: 10.1007/s11538-017-0277-2

[2] B Huang, et al., Interrogating the topological robustness of gene regulatory circuits by randomization. *PLOS Comput Biol* **13**, e1005456 (2017)

[3] Liu, J. S. (Ed.) (2001). *Monte Carlo strategies in scientific computing*. Berlin: Springer-Verlag.