Cross-evaluation of curated *E. coli* transcription unit structures against a whole-cell model

Gwanggyu Sun*[1], Mialy M. DeFelice*[1], Taryn E. Gillies[1], Travis A. Ahn-Horst[1], Cecelia J. Andrews[1], Jerry H. Morrison[1], and Markus W. Covert[1]

***Abstract*— Computational models can be used to evaluate the cross-consistency of heterogeneous data collected from the same biological system. Here, we attempted to integrate the curated structures of *E. coli*'s transcription units into a whole-cell model of *E. coli*. We identified discrepancies between the proposed transcription unit structures and the remaining data used by the model, which were resolved through iterative analyses of the model's outputs. The resulting conflict-resolved model was able to simulate the growth of *E. coli* while expressing polycistronic transcripts, and provide quantitative insights on the utility of operon structures in the bacterial genome.**

***Index Terms*— deep curation, whole-cell modeling, transcription unit structure, operons, transcriptional regulation**

I. Introduction

Whole-cell models are mechanistic simulations of living cells that take into account the known functions of every molecule within the cell. A decade ago, such a model was first constructed for the simplest culturable bacterium, *Mycoplasma genitalium* [1]. Since then, this approach has been extended to *Escherichia coli*, the most well-characterized model organism in biology [2].

The original version of the *E. coli* whole-cell model integrated more than 19,000 parameters that were collected from hundreds of heterogeneous data sources into a unified model of an *E. coli* cell [3]. We have demonstrated that while most of these parameters were consistent with each other, there were some that were not compatible. This method of cross-evaluating heterogeneous datasets – a process that we call "deep curation" – lead us to discoveries of key discrepancies between the datasets that would not be apparent if these parameters are analyzed independently of each other.

In this work, we have taken the deep curation approach to a dataset that was not considered in the original whole-cell model – the collection of *E. coli*'s polycistronic transcription units curated by the EcoCyc database [4]. By integrating this data into the whole-cell model, we were able to identify conflicts between this data and the remaining datasets that were used to parameterize the model, and make model-guided suggestions on how these conflicts could be resolved.

This work was funded by NIH grant 1R01GM140008-01A1.

[1]Department of Bioengineering, Stanford University, Stanford, CA, United States. Email: ggsun@stanford.edu

*These authors contributed equally to this work.

II. Results

A naive integration of the transcription unit dataset revealed that the proposed transcription unit structures for certain operons were at odds with other parameters that characterize the same operon. The proposed transcripts for the rplKAJL-rpoBC operon, for example, restricted the stoichiometry of the transcribed mRNAs in a way that led to the whole-cell model having doubling times that are approximately 1.5x longer than their expected values. Through iterative analyses of model outputs, we were able to suggest alternative transcription units for these operons that better align with the rest of the model, whose structures were further validated by R-end sequencing [5].

The resulting conflict-resolved model successfully replicated the key phenotypes of the original whole-cell model while transcribing polycistronic transcripts. By comparing the outputs of the new model against the original, we were able to demonstrate that the addition of multi-gene operons significantly improves the efficiency by which the *E. coli* model allocates its resources. For example, the production of subunits for protein complexes was more stoichiometrically balanced in the operon version of the model. These results suggest that one of the reasons operons were evolved in bacteria was to enable a tighter control over the inherent stochasticity of the transcription process.

III. Conclusion

Deep curation with the *E. coli* whole-cell model allowed us to test the cross-consistency of the proposed transcription unit structures of *E. coli* against other heterogeneous data. The resulting, updated whole-cell model provided quantitative insights on how operon structures can improve the efficiency of transcription in *E. coli*.

References

[1] J. Karr *et al.*, "A Whole-Cell Computational Model Predicts Phenotype from Genotype," *Cell*, vol. 150, no. 2, pp. 389–401, 2012.

[2] G. Sun, T. A. Ahn-Horst, and M. W. Covert, "The E. coli Whole-Cell Modeling Project," *EcoSal Plus*, vol. 9, no. 2, pp. eESP–0001–2020, 2021.

[3] D. N. Macklin *et al.*, "Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation," *Science*, vol. 369, no. 6502, 2020.

[4] I. M. Keseler *et al.*, "The EcoCyc Database in 2021," *Frontiers in Microbiology*, vol. 12, p. 711077, 2021.

[5] J.-B. Lalanne *et al.*, "Evolutionary Convergence of Pathway-Specific Enzyme Expression Stoichiometry," *Cell*, vol. 173, no. 3, pp. 749–761.e38, 2018.