Predictive genome-scale models arising from multi-relational networks

Minseung Kim[1,2], Athanasios Tsoukalas[1,2], Ian Davidson[2], and Ilias Tagkopoulos[1,2]

Accurate phenotypic predictions in novel environments is a challenging endeavor. Here, we present an integrative modeling methodology that unifies under a common framework the various biological processes and their interactions across multiple layers. For its training, we created a normalized compendium for the bacterium *Escherichia coli*, which incorporates gene expression, regulatory interactions, signal transduction and metabolic pathways, as well as growth measurements. Comparison with measured data demonstrates the enhanced ability of the integrative model to predict phenotypic outcomes in various environmental and genetic conditions. This work paves the way towards integrative techniques for the prediction and redesign of biological systems.

***Keywords* — genome-scale model, predictive modeling, systems and synthetic biology, genome engineering**

I. Introduction

Biological networks pose a formidable challenge to build, understand and control due to the complexity of cellular organization, partial knowledge of the underlying mechanisms, noisy data collection and scarcity of high-throughput techniques for targeted investigation of key functional layers. Over the years, several advances in regulatory network inference and in network analysis techniques that specifically target molecular networks have helped to elucidate biological processes and organization. Still, when it comes to predictive modeling, the impact of network biology is very limited as the former is mostly focused on local, well-studied pathways as an integrative genome-scale model would require a wealth of genome-scale data and the methods to link them under one unifying framework.

II. Approach

Towards this goal, we have developed a comprehensive normalized gene expression compendium for the model bacterium Escherichia coli that consists of 4,291 genes over 2,262 microarray and RNA-Seq datasets, covering various genetic (knock-out, overexpression, rewiring) and environmental perturbations [1]. We then created an integrative model that contains three sub-models that bridge the transcriptional, signal transduction and metabolic layers. This model covers 3704 regulatory interactions, 151 instances of signal transduction systems and 2251 metabolic reactions. Parameters in the transcriptional sub-model were determined by fitting the gene expression level of 328 transcription factors over four sets of constraints (phenomenological, capacity, environmental and genetic constraints). To reconstruct the regulatory network, we used a consensus method that resulted in better predictive performance than previous attempts [2], mainly because of the larger, integrated dataset. Application of guided learning algorithms for role discovery to the resulting multi-relational network, resulted in a data-driven gene ontology (GO) and discrepancies to the manually curated E. coli GO can guide experimentation and future model enrichment [3].

The integrated model was evaluated by performing cross-validation on the various datasets for growth and gene expression prediction, as well as predicting de novo experimentally measured data on the growth rate of 10 single-gene knock-outs for E. coli strains over different environments (28 genotype-phenotype combinations in total). Results show that our model can predict growth rates with 0.6 to 0.8 Pearson correlation coefficient between the experimentally measured and computationally-derived predictions, which is significantly higher than M models and on par to other ME models so far [4]. Furthermore, the constructed model can sense environmental changes and translated them to changes in gene expression and growth, which is a significant step forward for synthetic biology tools [5] and the targeted bioengineering of microbial strains.

Acknowledgements: The genome-scale model is based on work with J. Carrera, R. Estrella, J. Luo and N. Rai at UC Davis. The Gene Ontology work is in collaboration with M. Kramer and T. Ideker at UCSD.

[1]Genome Center, UC Davis, 451 Health Sciences Dr., Davis, CA, 95616
[2]Department of Computer Science, UC Davis, One Shields Ave, Davis, CA, 95616

III. References

[1] J. Carrera, R. Estella, J. Luo, N. Rai, A. Tsoukalas, I. Tagkopoulos (in press) An integrative, multi-layer, genome-scale model reveals the phenotypic landscape of *Escherichia coli*, *Molecular Systems Biology*.
[2] Marbach D, et al (2012) Wisdom of crowds for robust gene network inference. *Nature Methods* 9(8): 796-804.
[3] Dutkowski J. et al. (2013) A gene ontology inferred from molecular networks. *Nat Biotechnology*, 31(1):38-45.
[4] E. J. O'Brien, J. A. Lerman, R. L. Chang, D. R. Hyduke, and B. A. Palsson (2013) Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular Systems Biology*, vol. 9, no. 1.
[5] L. Huynh, A. Tsoukalas, M. Köppe, and I. Tagkopoulos (2013) SBROME: A scalable optimization and module matching framework for automated biosystems design. *ACS Synthetic Biology*, vol. 2, no. 5, pp. 263-273.