An integration framework for multiple high-throughput measurement types

Christian L. Barrett[1], Byung-Kwan Cho[1], and Bernhard O. Palsson[1]

Short Abstract — **A principle purpose of cellular regulatory systems is to effect different activity states of the genome. To a first order, activity states are manifested in the gene expression of different combinations of genes. High-throughput technologies are instrumental in enumerating these different gene combinations and associated promoters, but high-throughput data is inherently noisy and error prone. By integrating multiple high-throughput data types, though, errors can be significantly reduced and the functional units in the genome can be identified. Integrating these multiple types of data is an unsolved problem. In this work, we describe a computational framework for integrating multiple high-throughput data types for delineating the functional components of a genome.**

Keywords — **high-throughput data integration, systems biology, Markov models, Markov Random Fields.**

I. Purpose

It is now possible to generate large quantities of comprehensive data regarding different aspects of the activity state of a genome. These data include measurements of gene expression level, transcription start sites, protein content, and RNA polymerase binding locations. These measurement types are each subject to different failure modes, but taken together are complimentary and so can be used together to delineate the active functional units of a genome. As these types of data become easier to generate for the increasing number of organisms, it will be critical to have a computational framework for their integration.

The functional units of the genome, in the form of combinations co-transcribed genes and associated promoters, are a major component of the information processing systems in cells. Thus, integration of high-throughput data types is of central importance for understanding the complete path of information flows within cells.

II. Results

We have developed a framework for integrating multiple high-throughput measurements of a genome's expression and regulatory state. This framework is based on Markov models and Markov Random Fields and requires that each datum from each data type have an associated level of significance. These significance measurements are used as a common metric by which all different types of measurements can be compared to discern the activity state of all base pairs across a genome. This is accomplished by incorporating significance measurements via their "local FDR" values into a Markov model of gene structure. As genes often overlap in bacterial genomes, we utilize Markov Random Fields to allow for base pair locations in the genome to simultaneously participate in different states. This combination of techniques allows us to identify the possibly overlapping functional units of the genome.

In this work we demonstrate how this framework is utilized to integrate gene expression, transcription start site, RNA polymerase binding, and proteomic measurements to reveal the functional units utilized in different cellular states in *E. coli*.

III. Conclusion

Integration of different types of high-throughput measurements is an outstanding problem in systems biology. We have developed a framework to accomplish this goal and in this poster describe its underlying mathematical basis and application in a genome-wide setting.

Acknowledgements: This work was funded by NIH grant GM062791 and DOE DE-FC02-02ER63446.

[1]Department of Bioengineering, University of California, San Diego. E-mail: {cbarrett, bcho, bpalsson}@bioeng.ucsd.edu