Reconciling Gene Expression Data with Regulatory Network Models – A Stimulon-Based Approach for Integrated Metabolic and Regulatory modeling of *Bacillus subtilis*

José P. Faria[1,2], Ross Overbeek[3], Ronald C. Taylor[6], Anne Goelzer[5], Vincent Fromion[5], Miguel Rocha[4], Isabel Rocha[4], and Christopher S. Henry[1]

[1]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA
[2]IBB-Institute for Biotechnology and Bioengineering/Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal
[3]Fellowship for the Interpretation of Genomes, Burr Ridge, IL 60527, USA
[4]CCTC, University of Minho, Campus de Gualtar, Braga, Portugal
[5]INRA, UR1077, Mathématique Informatique et Génome, F-78350 Jouy-en-Josas, France
[6]Computational Biology & Bioinformatics Group, Pacific Northwest National Laboratory (U.S. Dept of Energy), Richland, Washington, USA

Abstract

The reconstruction of genome-scale metabolic models from genome annotations has become a routine practice in Systems Biology research. The potential of metabolic models for predictive biology is widely accepted by the scientific community, but these same models still lack the capability to account for the effect of gene regulation on metabolic activity. Our focus organism, *Bacillus subtilis* is most commonly found in soil, being subject to a wide variety of external environmental conditions. This reinforces the importance of the regulatory mechanisms that allow the bacteria to survive and adapt to such conditions. Existing integrated metabolic regulatory models are currently available for only a small number of well-known organisms (e.g *E. coli* and *B. subtilis*). The *E. coli* integrated model was proposed by Covert *et al* in 2004 and has slowly improved over the years. Goelzer *et al*. introduced the *B. subtilis* integrated model in 2008, covering only the central metabolic pathways. Different strategies were used in the two modeling efforts. The *E. coli* model is defined by a set of Boolean rules (turning genes ON and OFF) accounting mostly for transcription factors, gene interactions, involved metabolites, and some external conditions such as heat shock. The *B. subtilis* model introduces a set of more complex rules and also incorporates sigma factor activity into the modeling abstraction.

Here we propose a genome-scale model for the regulatory network of *B. subtilis*, using a new stimulon-based approach. A stimulon is defined as the set of genes (that can be a part of the same operon(s) and regulon(s)) that respond in the same set of stimuli. The proposed stimulon-based approach allows for the inclusion of more types of regulation in the model. This methodology also abstracts away much of the complexity of regulatory mechanisms by directly connecting the activity of genes to the presence or absence of associated stimuli, a necessity in the many cases where details of regulatory mechanisms are poorly understood.

Our model integrates regulatory network data from the Goelzer *et al* model, in addition to other available literature data. We then reconciled our model against a large set of high-quality gene expression data (tiled microarrays for 104 different conditions). The stimulons in our model were split or extended to improve consistency with our expression data, and the stimuli in our model were adjusted to improve consistency with the conditions of our expression experiments. The reconciliation with gene expression data revealed a significant number of exact or nearly exact matches between the manually curated regulons/stimulons and pure correlation-based regulons. Our reconciliation analysis of the 2011 SubtiWiki regulon release suggested many gene candidates for regulon extension that were subsequently included in the 2013 SubtiWiki update. Our enhanced model also includes an improved coverage of a wide range of different stress conditions.

We then integrated our regulatory model with the latest metabolic reconstruction for *B. subtilis*, the iBsu1103V2 model (Tanaka *et al*. 2012). We applied this integrated metabolic regulatory model to the simulation of all growth phenotype data currently available for *B. subtilis*, demonstrating how the addition of regulatory constraints improved consistency of model predictions with experimentally observed phenotype data. This analysis of growth phenotype data unveiled phenotypes that could only be characterized with the addition of regulatory network constraints.

All tools applied in the reconstruction, simulation, and curation of our new regulatory model are now publicly available as a part of the KBase framework. These tools permit the direct simulation of gene expression data using the regulon model alone, as well as the simulation of phenotypes and growth conditions using an integrated metabolic and regulatory model. We will highlight these new tools in the context of our reconstruction and analysis of the *B. subtilis* regulatory model.

Keywords: Regulatory Model; Stimulon; Integrated Metabolic Regulatory Model