Universidade do Minho
Escola de Engenharia

José Pedro Lopes Faria

Modeling Microbes: New methods for integrated metabolic and regulatory network reconstruction

June 2015

Modeling Microbes: New methods for integrated metabolic and regulatory network reconstruction
José Pedro Lopes Faria
UMinho | 2015

Esta investigação foi financiada pela Fundação para a Ciência e Tecnologia através da concessão de uma bolsa de doutoramento (SFRH / BD / 70824 / 2010), co-financiada pelo POPH - QREN - Tipologia 4.1 -Formação Avançada - e comparticipados pelo Fundo Social Europeu (FSE) e por fundos nacionais do Ministério da Ciência, Tecnologia e Ensino Superior (MCTES).

Universidade do Minho
Escola de Engenharia

José Pedro Lopes Faria

Modeling Microbes: New methods for integrated metabolic and regulatory network reconstruction

PhD Thesis in Bioengineering

This work was executed under the supervision of:
Doctor Isabel Cristina de Almeida Pereira da Rocha

Co-supervisors:
Doctor Miguel Francisco de Almeida Pereira da Rocha
Doctor Christopher Scott Henry

June 2015

STATEMENT OF INTEGRITY

I hereby declare having conducted my thesis with integrity. I confirm that I have not used plagiarism or any form of falsification of results in the process of the thesis elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

University of Minho, June 22nd 2015

José Pedro Lopes Faria

Acknowledgements

So many people are responsible for me to be able to finish my PhD thesis that I have trouble to figure out how to start. More than anyone else, my advisors are main reason I am able to actually finish a PhD thesis. Thank you Miguel, Isabel and Chris for the guidance and good times outside the lab walls. I don't want to sound cheesy, but you guys are the best advisors a graduate student could ask for.

For my friend's at Argonne, thank you for a great work environment and good laughs with the Star Wars and Star Trek live size cut offs. I am looking at you guys, Ric, Janaka, Neal, Pam and Sam. Also a shout out to all the FIG people, with a special thanks for Veronika, Svetta and Ross for all your help. Thank you Ross for being a mentor more than a co-worker. Also a special thanks for everyone from the BisBii research group at University of Minho, you guys are the best, I know I can be a annoying sometimes, but you guys always put up with me. Thank you so much for that.

A very heartfelt thank you to my girlfriend Eileen. Thank you for listening to my never ending thesis rants, and for cheering me up when I was feeling down. Can't wait to see you!

It has been hard to live between the United States and Portugal, back and forward for 4 years during my PhD. A lot of good people on both sides made it easier though. There are so many of you for me to name only a few or to name all. I am so blessed to have so many good friends; I just want you all to know I could not have done it without you guys.

I also would also like to thank FCT for the financial support (SFRH / BD / 70824 / 2010) that allowed me to conduct the research necessary for this thesis.

Deixo para o final o obrigado mais importante para o meus pais a quem dedico esta tese. Pelo apoio que sempre me deram fosse qual fosse a minha decisão. E sobretudo por me terem dado o privilégio de poder estudar, muitas vezes com alguns sacrifícios, mas isso nunca os dissuadiu de fazerem de tudo para que eu pudesse ter sempre mais e melhor. Estou muito grato por tudo que me deram até hoje e continuarei a fazer tudo para que vocês se possam orgulhar de mim... já que eu tenho todo o orgulho de vocês.

ABSTRACT

The reconstruction of genome-scale metabolic models (GEMs) from genome functional annotations is, nowadays, a routine practice in Systems Biology (SB) research. The models have been successfully used to predict organisms' behavior, gene essentiality, growth phenotypes and to aid strain optimization via metabolic engineering strategies. As the community acknowledges the usefulness of GEMs, they also present limitations, most notably the inability to account for the impact of regulation on the metabolic activity. The overall objective of this thesis was to reconstruct and perform *in silico* phenotype simulations for integrated models of metabolism and regulation.

The number of genomes available in the public domain increased exponentially in the last decade. The overwhelming amount of data led to the introduction of automated pipelines for genome annotation, also facilitating the propagation of annotation inconsistencies from public repositories. In this work, we explore the use of GEMs as tools for annotation curation. A protocol for annotation curation with metabolic network reconstructions was designed and applied to the genus *Brucella*.

The high-throughput reconstruction and analysis of genome-scale transcriptional regulatory networks is a current challenge in SB research. In this work, the model organism *Bacillus subtilis* was chosen as a case study and a new manually curated network for its transcriptional regulation was introduced. We proposed a new methodology for the inference of regulatory interactions from gene expression data. The newly proposed methodology dubbed "atomic regulon inference" was shown to capture many sets of genes corresponding to regulatory units in the manually curated network.

Following this line of work, based on the proposed regulatory transcriptional regulatory network for *B. subtilis*, we introduced an integrated genome-scale model for the metabolism and transcriptional regulation in *B. subtilis.* Model validation was performed with *in silico* growth phenotype simulations for mutant strains described in the literature. The integrated model was able to predict transcription factor knockouts for growth in multiple environmental conditions, expanding the predictive capabilities of the metabolic model by itself.

RESUMO

A reconstrução de modelos metabólicos à escala genómica (MMEGs) a partir de anotações funcionais do genoma é, hoje em dia, uma prática comum na investigação em Biologia de Sistemas (BS). Estes modelos foram usados com sucesso para prever o comportamento de organismos, essencialidade de genes, fenótipos de crescimento e na optimização estirpes bacterianas com estratégias de engenharia metabólica. Com o reconhecimento pela comunidade da utilidade de MMEGs, várias limitações foram identificadas, principalmente a incapacidade destes modelos explicarem o impacto da regulação de genes na actividade metabólica. O objetivo global desta tese foi a reconstrução e execução de simulações de fenótipo *in silico* para modelos que integram metabolismo e regulação.

O número de genomas disponíveis no domínio público aumentou exponencialmente na última década. Com este aumento exponencial de dados, plataformas para anotação automática de genomas tornaram-se uma necessidade, o que facilita a propagação de inconsistências nas anotações em repositórios públicos. Neste trabalho exploramos o uso MMEGs como ferramentas para melhoramento de anotações. Um protocolo para o melhoramento de anotações com o uso de MMEGs foi desenvolvido neste trabalho e testado na melhoria de anotações do género *Brucella*.

A reconstrução e análise de redes regulatórias de fenómenos de transcrição à escala genómica é um desafio actual na investigação em BS. Neste trabalho, foi efectuada a reconstrução manual da rede regulatória da transcrição para o microrganismo *Bacillus subtilis*. Um novo método para inferência automática de interações de regulação, a partir de dados de expressão de genes, foi igualmente desenvolvido. Este novo método mostrou ser capaz de inferir interações regulatórias comparáveis às observadas na rede reconstruída manualmente.

Com base na rede regulatória proposta para a regulação da transcrição de *B. subtilis*, desenvolvemos um modelo à escala genómica que integra o metabolismo e regulação da transcrição em *B. subtilis*. O modelo foi validado com simulações de fenótipo de crescimento *in silico* para estirpes mutantes descritas na literatura. O modelo integrado foi capaz de prever o efeito da deleção de factores de transcrição no crescimento em várias condições ambientais, ampliando as capacidades de previsão do modelo metabólico por si só.

x

List of contents

LIST OF FIGURES

List of tables

CHAPTER 1

INTRODUCTION

1.1 CONTEXT AND MOTIVATION

In the mid 1970s, the pioneer work performed in DNA sequencing unveiled the "power" of DNA and triggered a paradigm shift in biomedical research, launching the genomic era [1]. This era was defined by major advances in Molecular Biology and culminated with complete sequencing of the human genome [2]. In the turn of the century, Bioinformatics and Computational Biology had already emerged as multidisciplinary fields to handle the wealth of data from sequencing projects, propelling the beginning of the post-genomic era [3].

As the first bacterial genome sequences became available, genome functional annotation quickly became one of the biggest challenges in the post-genomic era [4]. The lack of standards led to the introduction of inconsistencies and errors in the annotations [5]. As the number of sequenced genomes increased exponentially over the last decade, a rise of automatic pipelines for genome annotation in detriment of manual curation became common practice [6]. The heavy reliance on sequence homology in these pipelines propagates inconsistencies across multiple organisms and databases, as new genomes can be annotated with old and out of date references [7].

Simultaneously, the increase in the abundance of available experimental data in the beginning of the new millennia boosted the emergence of a new perspective on Biology, which was named as Systems Biology [8]. This approach aims to understand the dynamics and behavior of an organism by exploring the relationships between genes, the proteins they encode and their organization into pathways [9]. Genome-scale metabolic models (GEMs) have become major tools in Systems Biology, and reconstructions are already in place for a growing number of organisms, including prokaryotic, archaeal, and eukaryotic species [10]. Advances were made towards the use of these models and computational tools for the *in silico* design and optimization of enhanced microbial strains [11]. Indeed, Metabolic Engineering (ME) has made use of such models to successfully establish new metabolic enzyme functions and pathways or altering existing pathways for the optimization of the production of chemicals of interest [12].

As the use of GEMs in ME and biological discovery became a common practice, tools have been introduced to automate the reconstruction process [13]. Automation substantially decreased the effort for the reconstruction of new GEMs, but these metabolic reconstructions have been based on automated annotation pipelines and still require manual refinements and validation against experimental datasets. Also, in spite of their achievements, these models are not able to capture the impact of gene regulation or signaling networks. When we entered the post-genomic era, obtaining a whole cell model that can be used in simulation was proposed as a major goal for the field in the 21st century [14]. Today, nearly 15 years later, only one whole-cell model has been proposed for *Mycoplasma genitalium*, one of the smallest known genomes of any free-living organism [15], still presenting many limitations.

Early studies with GEMs pointed to the importance of the integration of gene expression data into metabolic models to account for regulatory effects [16]. The first endeavors for the integration of regulatory networks with metabolic models revealed increased accuracy in predicting *in silico* phenotypes for *Escherichia coli* [17] and unveiled novel regulatory mechanisms in *Saccharomyces cerevisiae* [18]. The main limitation of these modeling efforts is the requirement of extracting the knowledge from gene expression data and its conversion into Boolean gene regulatory rules.

New methods have been developed to tap into the wealth of transcriptomics data and its integration with GEMs. Some of these methods enabled the prediction of tissue-specific metabolic models of human tissues [19], the prediction of drug targets [20] and mycolic acid production in *Mycobacterium tuberculosis* [21]. Despite these cases of success, a recent evaluation of the most widely used methods that integrate metabolic models and gene expression data shows that the latter underperform when compared with predictions from methods that account just for metabolism [22]. This fact casts some doubts about the way transcriptional regulation affects metabolic fluxes [23] and motivates the development of novel models and methods to exploit this interaction.

1.2 RESEARCH OBJECTIVES

The overall objective of this thesis was to reconstruct and perform *in silico* phenotype simulations for integrated models of metabolism and regulation. To achieve this goal, studies were performed covering all individual components that should be integrated to fulfill the overall purpose, including analyses of genome annotations, reconstructions of GEMs, analyses of gene expression data and reconstruction of regulatory networks.

According to the challenges presented in the previous section, several aims were proposed to attain the global objective stated above:

- Assess the quality and caveats of current automated reconstructed GEMs by performing metabolic reconstructions for a large set of microbial genomes available in public databases.
- Examine the consistency of genome annotations across multiple databases and explore the development of methodologies that make use of metabolic reconstructions as tools for annotation curation.
- Review repositories and databases with notable bacterial transcriptional data, as well as methods for transcriptional regulatory network reconstruction.
- Perform the reconstruction and analysis of the regulatory network for *Bacillus subtilis* using manual curation and high throughput gene expression datasets.
- Build an integrated genome-scale model for the metabolism and transcriptional regulatory network of *B. subtilis*.
- Perform *in silico* growth phenotype simulations to validate the integrated model with experimental datasets and knowledge of regulatory interactions described in the literature.

1.3 THESIS OUTLINE

The thesis is structured in 5 chapters, as shown on Figure 1.1. This first chapter, "Introduction", details the motivations and the objectives of the thesis work, providing the outline of the texts structure.

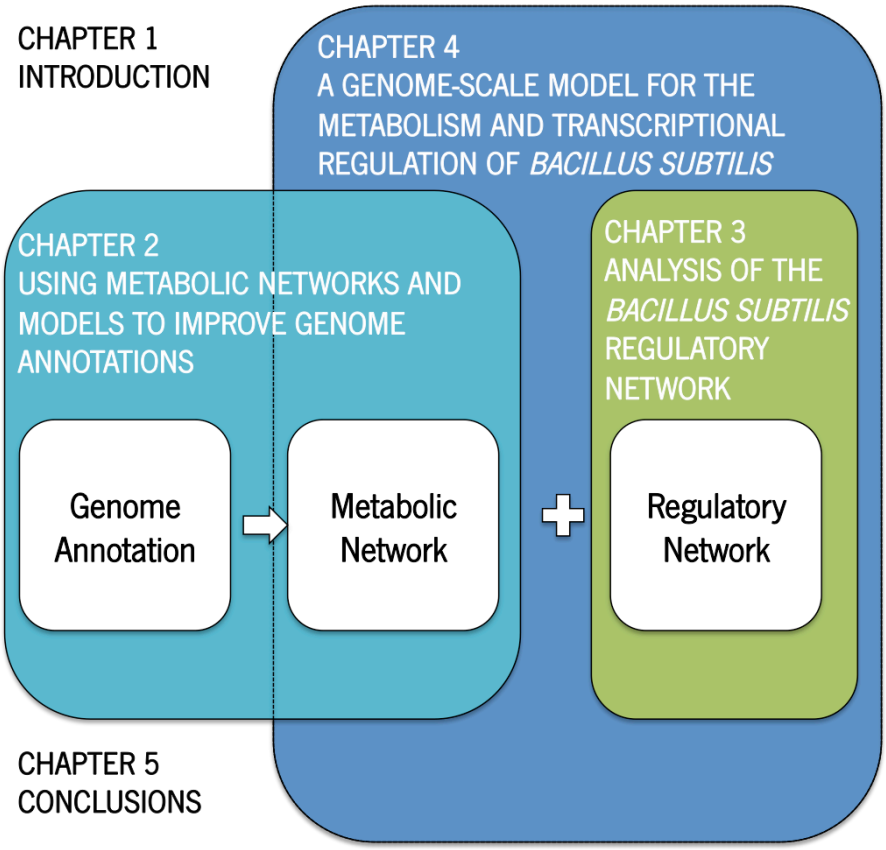


Figure 1.1 Thesis outline.

In order to develop integrated models of metabolism and regulation, three main elements have been identified as the "core" of the research conducted in this thesis: genome annotations, metabolic networks and regulatory networks. The three main chapters are organized according to each of those core elements. It is important to note that each chapter contains its own state of the art relevant to the specific work developed.

In chapter 2, "Using metabolic networks and models to improve genome annotations", we investigate the current state of automated reconstructed metabolic models. We attempt to achieve this objective by reconstructing GEMs for all genomes available in the PubSEED repository [24]. This process led to the identification of annotation errors and caveats of automated reconstruction processes. A study of the *Brucella* genus was conducted to assess the quality of genome annotations at the pan-genome level, and provide a protocol for genome annotation curation using metabolic models [25].

In chapter 3, "Analysis of the regulatory network of *Bacillus subtilis*" we explore the reconstruction of regulatory networks. A comprehensive review of databases for gene regulation data and methods for regulatory network reconstruction is presented [26]. Afterwards, we introduce a new manually curated regulatory network for *B. subtilis.* A new methodology we named "atomic regulon inference" is proposed for the inference of regulatory network elements from gene expression data. Finally, the conducted efforts to reconcile the manually curated and the automatically inferred regulatory networks are presented and discussed.

In chapter 4, "A genome-scale model for the metabolism and transcriptional regulation of *Bacillus subtilis*", we integrate the regulatory network presented in chapter 3 with the latest genome-scale metabolic model for *B. subtilis* proposed in the literature. We validate our model with growth phenotypes for knockout strains described in literature for multiple environmental constraints.

In chapter 5, we present the final conclusions. Additionally, we reflect on limitations and future work of the research developed for this thesis.

1.4 SCIENTIFIC OUTPUT

The results presented in this thesis have been partially presented elsewhere.

Publications

Faria, J.P., et al., "*Genome-scale bacterial transcriptional regulatory networks: reconstruction and integrated analysis with metabolic models*. Briefings in Bioinformatics", 2014. 15(4): p. 592-611.

Faria, J.P., *et al.*, "*Enabling comparative modeling of closely related genomes: example genus Brucella".* 3 Biotech, 2014: p. 1-5.

Faria, J. P., Overbeek, R., Taylor, R. C., Goelzer, A.,Fromion, V., Rocha, M., Rocha, I., &,Henry, C.

"*Reconciling gene expression data with regulatory network models–a stimulon-based approach for regulatory modeling of Bacillus subtilis*."

Abstract accepted for full manuscript submission, Frontiers in Systems Microbiology, 2015.

Faria, J. P., Rocha, M., Rocha, I., Henry, C. S.

"A genome-scale model for the metabolism and transcriptional regulation of Bacillus subtilis"

Manuscript in preparation, 2015

Poster presentations

Faria, J. P., Xia, F., Devoid, S., DeJongh, M., Best, A., Henry, C. S. & Stevens, R.

"*Modeling Microbial Life - Reconstruction of 3000 Metabolic Models*". E3 Forum: Education, Employment and Entrepreneurship. Lisboa, Portugal. June 29-30, 2011.

Oral presentations

Faria, J. P., Overbeek, R., Taylor, R. C., Goelzer, A., Fromion, V., Rocha, M., Rocha, I., & Henry, C. S. "*Reconciling gene expression data with regulatory network models–a stimulon-based approach for integrated metabolic and regulatory modeling of Bacillus subtilis*". American Institute of Chemical Engineers Annual Meeting 2013. San Francisco, USA. November 3-8, 2013.

1.5 REFERENCES

1. Sanger, F. and A.R. Coulson, *A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.* J Mol Biol, 1975. **94**(3): p. 441-8.

2. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

3. Lengauer, T. *Computational biology at the beginning of the post-genomic era*. in *Informatics*. 2001. Springer.

4. Eisenberg, D., et al., *Protein function in the post-genomic era.* Nature, 2000. **405**(6788): p. 823-6.

5. Devos, D. and A. Valencia, *Intrinsic errors in genome annotation.* Trends Genet, 2001. **17**(8): p. 429-31.

6. Richardson, E.J. and M. Watson, *The automatic annotation of bacterial genomes.* Brief Bioinform, 2013. **14**(1): p. 1-12.

7. Stothard, P. and D.S. Wishart, *Automated bacterial genome analysis and annotation.* Curr Opin Microbiol, 2006. **9**(5): p. 505-10.

8. Ideker, T., T. Galitski, and L. Hood, *A new approach to decoding life: systems biology.* Annu Rev Genomics Hum Genet, 2001. **2**: p. 343-72.

9. Kitano, H., *Systems biology: a brief overview.* Science, 2002. **295**(5560): p. 1662-4.

10. Feist, A.M., et al., *Reconstruction of biochemical networks in microorganisms.* Nat Rev Microbiol, 2009. **7**(2): p. 129-43.

11. Burgard, A.P., P. Pharkya, and C.D. Maranas, *Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.* Biotechnol Bioeng, 2003. **84**(6): p. 647-57.

12. Lee, J.W., et al., *Systems metabolic engineering of microorganisms for natural and non-natural chemicals.* Nat Chem Biol, 2012. **8**(6): p. 536-46.

13. Henry, C.S., et al., *High-throughput generation, optimization and analysis of genome-scale metabolic models.* Nat Biotechnol, 2010. **28**(9): p. 977-82.

14. Tomita, M., *Whole-cell simulation: a grand challenge of the 21st century.* Trends Biotechnol, 2001. **19**(6): p. 205-10.

15. Karr, J.R., et al., *A whole-cell computational model predicts phenotype from genotype.* Cell, 2012. **150**(2): p. 389-401.

16. Covert, M.W., C.H. Schilling, and B. Palsson, *Regulation of gene expression in flux balance models of metabolism.* J Theor Biol, 2001. **213**(1): p. 73-88.

17. Covert, M.W., et al., *Integrating high-throughput and computational data elucidates bacterial networks.* Nature, 2004. **429**(6987): p. 92-6.

18. Herrgard, M.J., et al., *Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae.* Genome Res, 2006. **16**(5): p. 627-35.

19. Shlomi, T., et al., *Network-based prediction of human tissue-specific metabolism.* Nat Biotechnol, 2008. **26**(9): p. 1003-10.

20. Chandrasekaran, S. and N.D. Price, *Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis.* Proc Natl Acad Sci U S A, 2010. **107**(41): p. 17845-50.

21. Colijn, C., et al., *Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production.* PLoS Comput Biol, 2009. **5**(8): p. e1000489.

22. Machado, D. and M. Herrgard, *Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism.* PLoS Comput Biol, 2014. **10**(4): p. e1003580.

23. Kochanowski, K., U. Sauer, and V. Chubukov, *Somewhat in control–the role of transcription in regulating microbial metabolic fluxes.* Curr Opin Biotechnol, 2013. **24**(6): p. 987-93.

24. Overbeek, R., et al., *The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).* Nucleic Acids Res, 2014. **42**(Database issue): p. D206-14.

25. Faria, J.P., et al., *Enabling comparative modeling of closely related genomes: example genus Brucella.* 3 Biotech, 2014: p. 1-5.

26. Faria, J.P., et al., *Genome-scale bacterial transcriptional regulatory networks: reconstruction and integrated analysis with metabolic models.* Brief Bioinform, 2014. **15**(4): p. 592-611.

CHAPTER 2

USING METABOLIC NETWORKS AND MODELS TO IMPROVE GENOME ANNOTATIONS

Work presented in this chapter includes the following article:

Faria, J. P., Edirisinghe, J. N., Davis, J. J., Disz, T., Hausmann, A., Henry, C. S., . . . Shukla, M.

Enabling comparative modeling of closely related genomes: example genus Brucella. 3 Biotech, p. 1-5.

2014.

ABSTRACT

Genome-scale metabolic models have emerged as a valuable resource for generating predictions of global organism behavior based on the sequence of nucleotides in the genome. These models can accurately predict essential genes, organism phenotypes, organism response to mutations, and metabolic engineering strategies. One of the host groups for this work has developed the Model SEED resource (http://seed-viewer.theseed.org/models/) for the high-throughput reconstruction of new genome-scale metabolic models for microbial genomes. We applied the Model SEED to produce draft metabolic models for over 3000 microbial genomes. We applied these models to study the diversity of microbial genomes, the completeness of our knowledge of these genomes, and the areas of our knowledge where more annotation gaps presently exist.

We also explored the application of draft models to exposing and reconciling inconsistencies in the genome annotations among a family of closely related genomes. Using fifteen strains of the genus *Brucella*, which contain pathogens of both humans and livestock, we developed a protocol for the comparative analysis of models of closely related genomes. This study resulted in the identification and subsequent correction of inconsistent annotations in the SEED database, as well as the identification of 31 biochemical reactions that are common to all *Brucella* genomes studied, which were not originally identified by automated metabolic reconstructions. We are implementing this protocol for improving automated annotations across the entire SEED database to facilitate the future creation of consistent annotation systems and high quality model reconstructions to support accurate phenotype predictions, including pathogenicity, media requirements, or respiration type.

2.1 INTRODUCTION

Genome-scale metabolic models (GEMs) are major tools used in the field of Systems Biology [1], and reconstructions are already in place for a growing number of organisms, including prokaryotic, archaeal, and eukaryotic species [2]. At the same time, many new systems engineering approaches are emerging for production and utilization of metabolic pathway reconstructions. As a result, pathways are being refined and the quality of metabolic models is improving [3]. Cases of success for the use of GEMs can be found across several fields.

Indeed, in industry, GEMs were used for the improved production of bioethanol [4, 5], one of Biotechnology's most notorious products [6] and for environmental remediation [7]. In research, GEMs have been used in studies for identifying new drug targets [8, 9], to better understand bacterial evolution [10] and to improve genome functional annotations [3].

The number of complete genome sequences have been growing exponentially, but metabolic models have been growing at a much slower pace. Figure 2.1 shows the shift towards the rapid increase in the availability of GEMs with the development of automated reconstruction tools, such as the Model SEED [11]. As the pipelines for automated GEM reconstruction become more robust, we are rapidly progressing towards having models for every available sequenced genome. However, the creation of accurate, high quality models requires a substantial investment in mining phenotypic data and an iterative reconciliation with experimental data [12]. These high quality models have a demonstrated capability to accurately predict gene essentiality, phenotypes and metabolic engineering strategies [13].

In this chapter, we focus on assessing the quality of automatically reconstructed models, the application of models to identify and fill gaps in functional annotations, the comparison of models to explore microbial diversity, and the use of models to rigorously study variation within a single family of microbial genomes. We begin by applying the Model SEED [11] to produce draft metabolic models for over 3000 microbial genomes, representing nearly all complete microbial genomes available in the SEED genomics database [14] (as of Winter 2011).

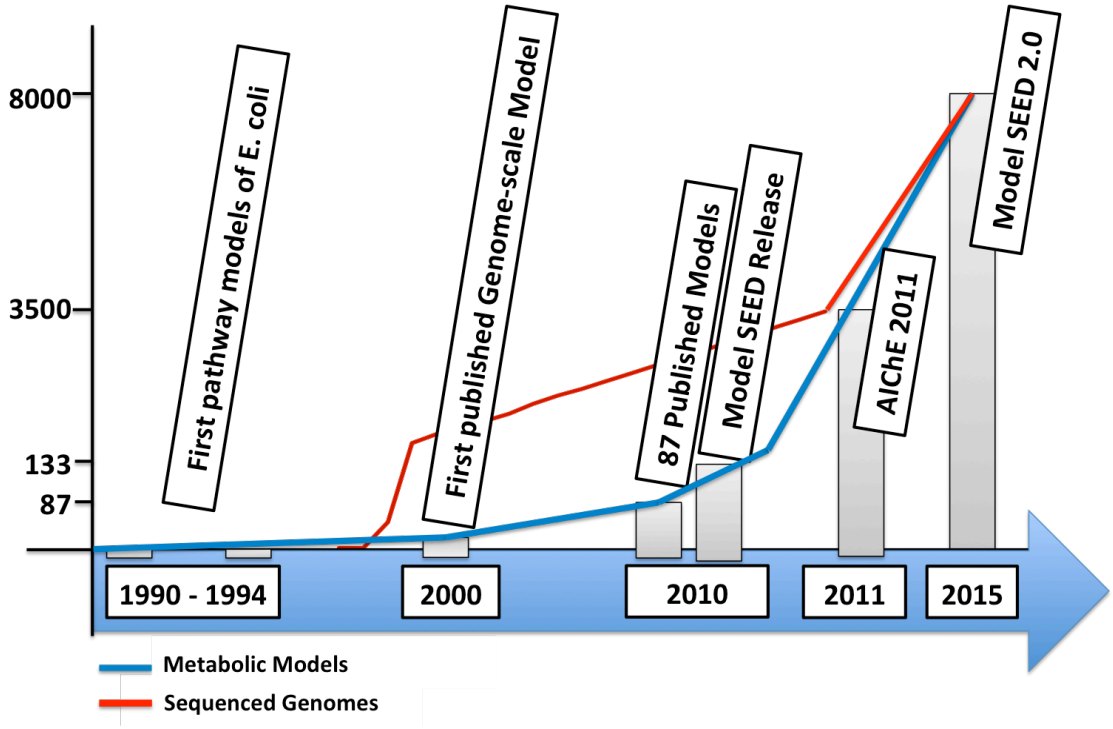


Figure 2.1 Development of metabolic models *versus* availability of genome sequences. Described in the picture is the timeline for the last 20 years of metabolic model development. Pictured are relevant events, such as the first pathway models for *E. coli* in the early 90s, the release of the first genome scale metabolic mode (GEM), the release of the Model SEED automated framework for GEM reconstruction and the 3000 models developed for this thesis work and presented at 2011 AIChE (American Institute for Chemical Engineers) Annual Meeting.

This was only possible due to the development of new algorithms for various steps of the model reconstruction process: iterative gap filling to enable the activation of all reactions in models; the generation of biomass reactions based on completeness of annotated pathways for biomass precursors; and, finally, new algorithms were applied for using the SEED tools to identify gene candidates that may be associated with the gap filled reactions.

This work reveals insights into the diversity of microbial genomes, the completeness of our knowledge of these genomes, and the areas of our knowledge where more gaps presently exist. This application of our analysis across all available prokaryotic genomes unveils systematic errors in annotation or model

reconstruction that may be subsequently corrected. This study also provides a rigorous assessment of the current state of the art for genome annotation and metabolic model reconstruction in the SEED.

In addition to our large-scale reconstruction of models for all available prokaryotic genome sequences, we also conducted a smaller scale study focused on performing a rigorous comparative analysis of models and annotations for a single family of closely related genomes. We conducted this study to evaluate the consistency of the annotations that form the foundation of our GEM reconstructions. In order to achieve this objective we developed a protocol [15] for improving the annotations and metabolic reconstructions for an entire genus. We demonstrate how this protocol has improved the annotations and metabolic reconstructions for the genus *Brucella*, a group of intracellular facultative bacterial pathogens of humans and livestock. While *Brucella* is an important pathogen for study, there is limited wet lab research on this family of genomes, and there is no curated metabolic model published in the literature. Thus, the development of a set of predictive metabolic models for this family of organisms is highly desirable.

2.2 STATE OF THE ART

2.2.1 Genome sequencing and annotation

It has been almost 20 years since the first whole genome sequence of any organism was released for *Haemophilus influenzae* Rd [16]. Sequencing costs have declined exponentially over the past two decades, leading to an exponential increase in available genome sequences [17].

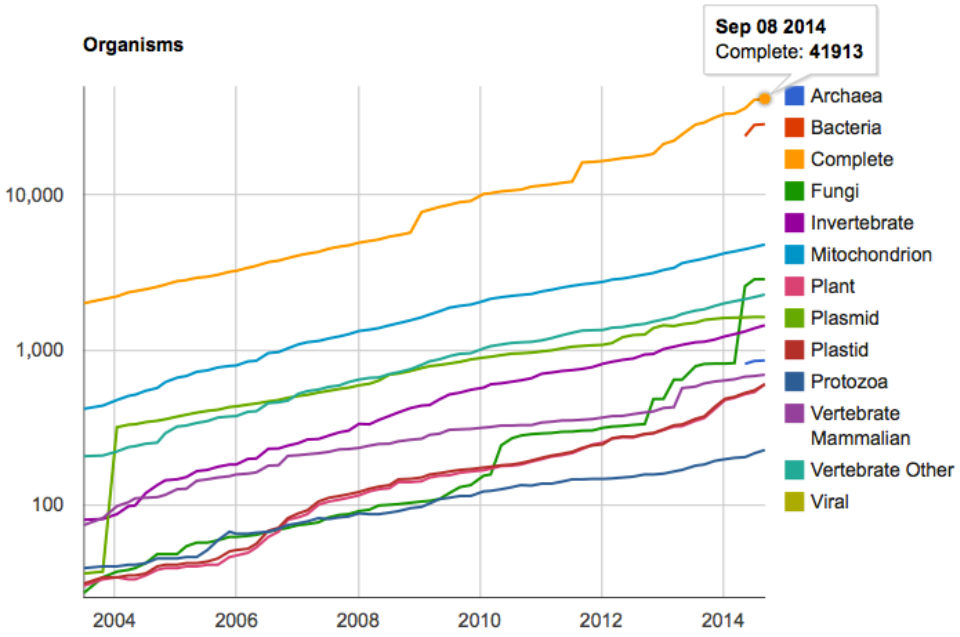
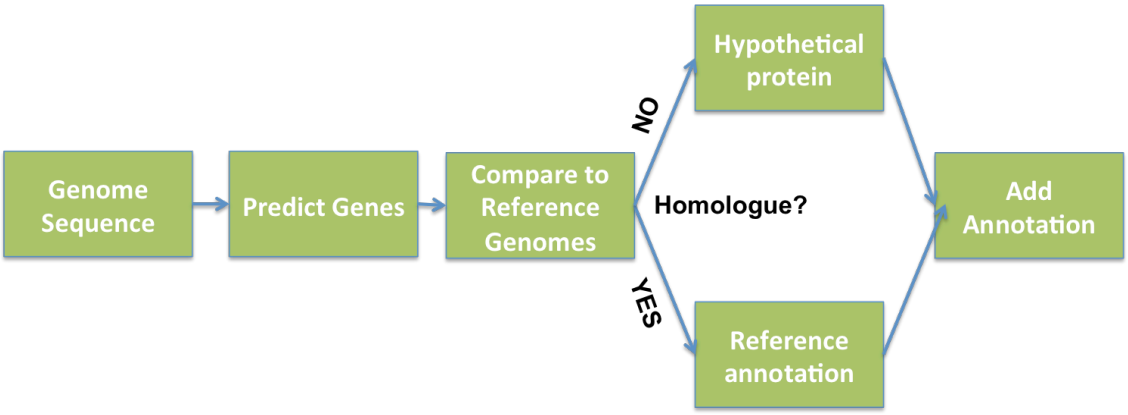


Figure 2.2 Number of complete genome sequences in the NCBI Reference Sequence (RefSeq) database [18].

The latest release of NCBI Reference Sequence (RefSeq) database has now over 40.000 complete sequenced genomes submitted [19]. The growth of the RefSeq database across different domains of life is shown in Figure 2.2 (http://www.ncbi.nlm.nih.gov/refseq/statistics/).

The rapid increase in genome sequence availability leads to the need of automated annotation pipelines [20]. Most automated genome annotation tools follow a generic pipeline structure as seen in Figure 2.3 (adapted from Richardson and Watson 2012 [21]).



.

Figure 2.3 Generic automated annotation pipeline. A gene prediction algorithm is applied for identification of coding genes. A homology search is conducted against reference genomes previously annotated. Annotation of reference genome is added to the new genome sequence when a homologue is found. Otherwise is annotated as hypothetical protein.

After sequencing, a gene prediction algorithm (e.g. GLIMMER [22] or GeneMark [23]) is used to identify the coding genes in the target genome. Each gene in the target genome is compared against reference genomes, which are usually curated, and a search for homologues is conducted (e.g. using the BLAST tool [24]). If no homologues are found in the reference genomes, the annotation is marked as "hypothetical protein". If a homologue is found, the reference annotation is adopted for the gene. This process comes at a price, as annotation errors in the reference genomes are easily propagated to new genome annotations. Data sharing protocols further propagate errors across multiple databases. Common errors introduced by automated pipelines have been identified [21]. One of the most common causes of inconsistent annotation is the fact that each research group uses different protocols for its annotations. Spelling mistakes in the annotation name are another common source of inconsistencies. The reliance on homology is still another issue, as genes are arbitrarily assigned a "hypothetical protein" annotation when a homologue cannot be found in the reference genomes.

The UniProt Consortium responsible for the UniProtKB [25] tried to address the issues above by creating TrEMBL [26], a dedicated database for automated annotations separated from its curated

database Swiss-Prot. Keeping the two databases separate allows UnitProt to use Swiss-Prot as the database of reference for the automated annotation conduced in TrEMBL.

On the other hand, to improve the quality of its annotations, RefSeq establishes collaborations with research groups responsible for highly curated organism-specific databases. Most notably, annotations are directly contributed to RefSeq from the Saccharomyces Genome Database (SGD) (*Saccharomyces cerevisiae)*, FlyBase (*Drosophila melanogaster)* and The Institute for Genomic Research (TIGR) (*Arabidopsis thaliana)*.

The RAST (Rapid Annotation using Subsystem Technology) applies a different approach to genome annotation based on the creation and curation of subsystems [27]. A subsystem is a set of functional roles that are part of a specific biological process and can be defined as a generalization of the term pathway [27]. It is often common that researchers have an expert knowledge of a specific set of a particular cellular machinery/pathway/subsystem. In contrast, it is significantly harder to have expert knowledge for the biology of a whole organism. The subsystem approach for annotation takes advantage of this fact as experts build/curate subsystems and annotate genes for subsystems across multiple genomes instead of focusing on annotating a single organism. Automated annotation becomes less error prone when the annotations are projected from a curated subsystem across multiple genomes in comparison to project single gene annotations based on homology. Pathway Tools uses a similar approach to RAST in its annotation framework [28].

2.2.2 Genome-scale metabolic model reconstruction

Metabolic models have evolved for almost 25 years now. Looking at the evolution of the model for *E. coli* [29] we see the growth in complexity of the models, as more information became available: the shift to genome scale modeling [30], the addition of Gene-Protein-Reaction associations (GPRs), the inclusion of thermodynamic information for reaction reversibility, and assignment of cellular compartments to all metabolites in the model. On Figure 2.4, we describe the main components of GEMs.

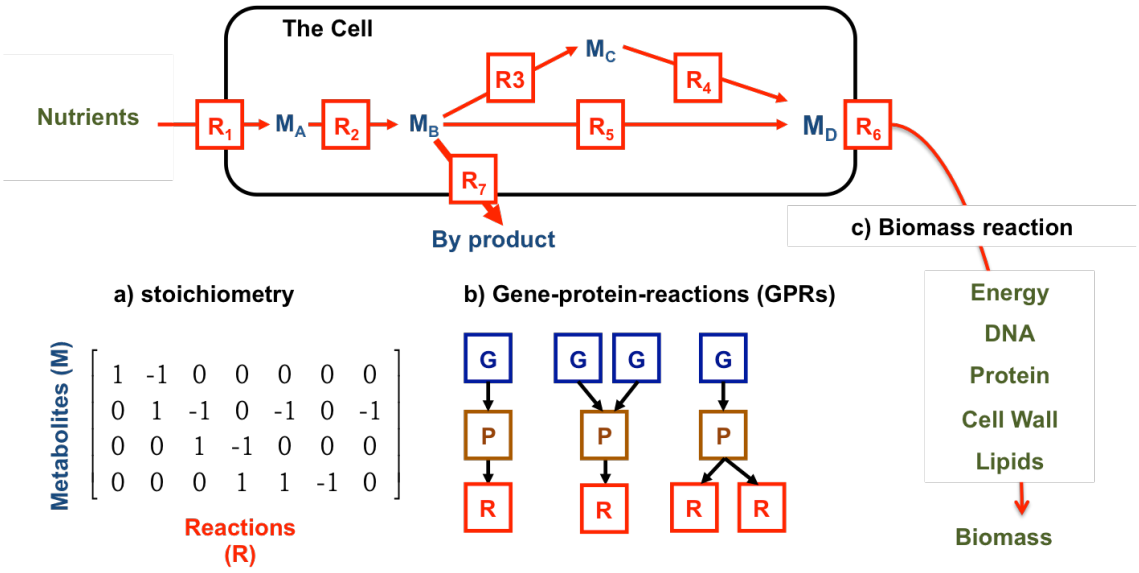


Figure 2.4 Main components of genome-scale metabolic models. a) Stoichiometry; b) Gene-Protein-Reaction associations (GPRs); c) Biomass reaction.

The first main component featured in Figure 2.4 is reaction stoichiometry. The stoichiometric matrix (Figure 2.4 a)) encodes the biological knowledge from the metabolic network (Metabolites (M) and reactions (R)) into mathematical terms [31, 32]. For more details on stoichiometric modeling from the constraint-based perspective [33] see section 4.2.1. GPRs are another main component of GEMs. These associations establish the relationship and dependence from genes to proteins and ultimately from proteins to the reactions they catalyze [34]. Figure 2.4 b) displays multiple scenarios for these associations, such as two (or more) genes (G) encoding one protein (P) and one subsequent reaction (R). The inclusion of GPRs in GEMs allows the development of Metabolic Engineering studies [35]. Gene deletion [36] and gene over/under expression [37] are among the studies enabled by the inclusion of GPRs. The third main component featured in Figure 2.4 c) is the biomass reaction. The biomass reaction defines the list of metabolic resources that a cell must produce in order to grow [38]. It allows the use of computational growth phenotype simulation methods, such as Flux Balance Analysis (FBA), [39] (for more details on FBA formulation see chapter 4 section 4.3.1.).

Detailed protocols have been established for GEM reconstruction and validation [12, 40]. Reconstruction of GEMs can be extremely time consuming, with the protocol proposed by Thiele and Palsson [12] being comprised of 96 individual steps. Figure 2.5 (adapted from Thiele and Palsson) shows the main stages involved in a GEM reconstruction and main tasks performed in each stage. Three main stages were defined as: Draft Reconstruction; Refinement and Curation; and Model Evaluation. The first stage, draft reconstruction, provides mainly the assignment of metabolic reactions to the metabolic functions predicted in the input genome annotation. GPRs, one of GEMs' main components discussed previously are generated during this stage.

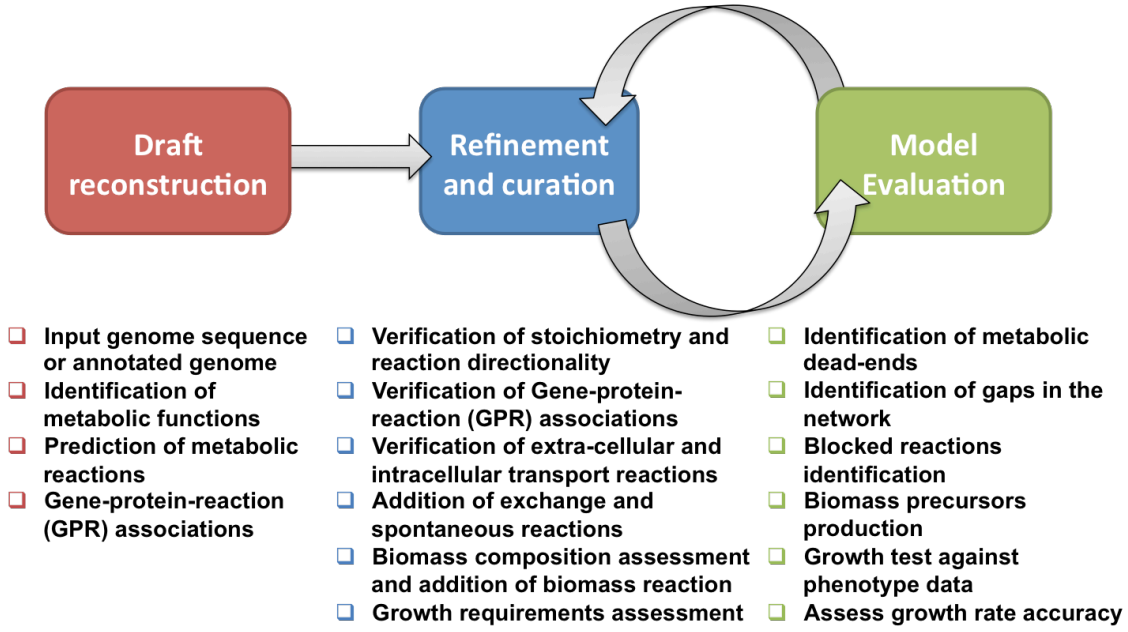


Figure 2.5 Genome-scale metabolic model reconstruction. Main tasks performed in the reconstruction process organized in three stages: Draft Reconstruction; Refinement and Curation; and Model Evaluation.

The second stage aims to refine and curate the original draft reconstruction. Reactions, stoichiometry and directionally are usually verified along with GPRs and extra and intra-cellular transport reactions. Spontaneous and exchange reactions are also added if necessary. Biomass composition is assessed, as biomass precursors vary depending on the organism's physiology. BioMog [41] is a recently introduced framework to address this issue, which uses high-throughput growth phenotype and fitness

datasets to generate *de novo* biomass components. Usually, at this stage, the biomass reaction is added to the model and specific media growth requirements are assessed to enable Model Evaluation in the next stage.

Identification of metabolic dead-ends aims to locate metabolites that cannot be consumed or generated by the network, being one of the first steps to evaluate the model. Verification of biomass precursor production is also performed at this stage, to assess if all molecules that enable growth can be produced from the medium components. The results from these two steps usually reveal candidates for gap filling in the network. Unbalanced metabolites and gaps in the network are a common cause of blocked reactions (reactions that are unable to carry flux through the network). Flux Variability Analysis (FVA) [42] is able to minimize and maximize flux through all reactions in the network and can be used to identify these blocked reactions. Growth tests can also be conducted using a simulation method, such as FBA with maximization of biomass to determine the model capabilities to represent known physiological properties of the organism. Single gene knockouts and growth data for different medium conditions are ideal for this task. The refinement and evaluation stages are handled as an iterative process. Inconsistencies identified during Model Evaluation will be curated, and this process is repeated until a refined model is ready for release.

2.2.3 Methods for automated GEM reconstruction

In this thesis, we focused our efforts on automated reconstruction of GEMs. Many tools have been developed for this task in the last decade [43]. Here, we chose to review some of the early methods along with more recently developed ones that produce higher quality automated reconstructions. A comparison of those methods is shown on Table 2.1.

GEM System [44] and AUTOGRAPH [45] were released in 2006 and were among the first methods attempting to automate GEM reconstruction. AUTOGRAPH uses published models as a template and performs an ortholog search from the target genome to the reference genomes to map genes and their gene-protein reactions. GEM System allows for the use of both annotated and un-annotated genomes. Homology and orthology searches are conducted against the SWISS-PROT and TrEMBL [26] databases

to match metabolic genes to the EC number of their protein products. Subsequently, the EC numbers are matched to the KEGG Orthology and Pathway database [46]. The downside of early methods was the limited capability to only produce draft reconstructions. A draft reconstruction cannot be used as a functional model, as it requires significant additional curation to properly simulate growth behavior. GEM System allows export draft reconstructions in the Systems Biology Markup Language (SBML) format [47]. A more complete list of GEM System features is shown on Table 2.1.

Most recent methods start with the draft reconstruction and provide tools to refine and evaluate the network reconstructions. Merlin (metabolic models reconstruction using genome-scale information) [48] provides several tools for annotation curation along with the tools for network reconstruction. The Model SEED [11] was the first platform to integrate the capability to generate draft models and perform network refinement and curation, automated gap filling [49] and network evaluation with FBA and phenotype datasets. Pathway Tools [28] stemmed from the development of EcoCyc [50] as a tool to create organism specific pathway databases. Since then, it has evolved and its latest release [51] features a full suite of tools for GEM reconstruction.

The RAVEN (Reconstruction, Analysis and Visualization of Metabolic Networks) Toolbox [52] is a software suite that focuses on providing tools for network visualization and analysis in addition to the network reconstruction tools. The SuBliMinal Toolbox [53] takes a modular approach to the reconstruction process. Different modules can be used independently to perform the tasks necessary for the reconstruction process. The modular infrastructure allows users to plugin multiple tools, such as the popular Cheminformatics software Marvin Beans (www.chemaxon.com) to determine metabolite charges.

The first feature shown on Table 2.1 is the input data. Much like the GEM System, merlin allows users to upload a genome sequence file. Merlin provides an array of tools that allow users to curate/re-annotate the functional annotations in the genome submitted. The RAVEN Toolbox [52] and Pathway Tools require annotated genomes as input. The Model SEED uses annotations from the RAST platform [14, 54]. Annotation tools are available in RAST, where users can use publicly available genomes or

submit their own. SuBliMinal Toolbox [53] uses KEGG and MetaCyc [55] limiting the reconstruction process to organisms available on those databases.

Table 2.1 Comparison between different resources for automated GEM reconstruction (Adapted from Hamilton, J.J. and J.L. Reed, 2014 [56])

	GEM System	Merlin	Model SEED	Pathway Tools	Raven Toolbox	SuBliMinal Toolbox
Input data	Annotated or un-annotated sequence	Annotated or un-annotated sequence	RAST annotation	Annotated sequence	Annotated sequence	Organisms in KEGG and MetaCyc
Reference Databases	KEGG, BioCyc	KEGG, TCDB	Model SEED database	MetaCyc	KEGG, Published models	KEGG, MetaCyc
Interface	Standalone (GUI)	Standalone (GUI)	Web	Standalone (GUI), Web	MATLAB	Standalone (cmd line)
Output	SBML	SBML	SBML, Excel	SBML, BioPax	SBML, Excel	SBML
Network Visualization	YES	YES	YES	YES	YES	NO
Simulation Support	NO	NO	YES	YES	YES	NO
Integrates Gap filling	NO	NO	YES	YES	YES	NO

Regarding reference databases, all the tools described in Table 2.1 use KEGG and additional databases. The downside of KEGG is that it does not feature charged and mass balanced metabolites and reactions (although all reactions for which the metabolites have formulae are balanced). The Model SEED internal database absorbs KEGG and published models. It is curated to address the lack of proper metabolite/reaction charges and balancing in KEGG. Pathway Tools follows the same "philosophy" as the Model SEED, using the curated database MetaCyc. Both the Model SEED and the RAVEN Toolbox use reactions from published models, as these are usually curated during their respective reconstruction processes. Merlin makes additional use of a transporters database TCDB [57] in the effort to better annotate and identify transport reactions. The SuBliMinal Toolbox, as mentioned before, absorbs the KEGG and MetaCyc databases.

All tools presented on Table 2.1 have different user interfaces. Merlin, Pathway Tools and the SuBliMinal Toolbox are distributed as standalone applications. The SuBliMinal Toolbox presents a command-line interface. On the other hand, both Merlin and Pathway Tools have a user-friendly graphical user interface (GUI). The Model SEED is a web application presenting also with a GUI and requiring users to set up an account to get started. A disadvantage of a web interface is that some researchers are still reluctant to upload their private genome sequences/annotations into the web. Pathway Tools also has a web interface distribution. RAVEN Toolbox is distributed as a MATLAB plugin, having the drawback of the need of a license for MATLAB.

SBML is the standard output for all the tools. The Model SEED and RAVEN Toolbox also provide output for the Microsoft Excel software. Pathway Tools provides additional output in the BioPax community standard for pathway sharing [58]. Network Visualization can be an important feature for both curation and model evaluation. RAVEN Toolbox uses manually curated CellDesigner [59] maps and Pathway Tools has their own tools to draw metabolic pathways. Merlin and the Model SEED provide a more basic visualization functionality, drawing on top of KEGG maps.

Support for simulations is also essential for model evaluation. FBA is the basic simulation method provided by Model SEED, Pathway Tools and the RAVEN Toolbox. RAVEN Toolbox and Pathway Tools provide additional flux balance simulation algorithms to support model evolution. Tools lacking simulation support need the use an additional platform, such as the COBRA Toolbox [60] and OptFlux [61].

Integrated gap filling is one of the most important features in a reconstruction, as it can be exceptionally time consuming. When performed manually, one has to identify the gaps and candidate reactions to complete a pathway/network. Model SEED and Pathway Tools provide algorithms that allow for completely automated gap filling. The RAVEN Toolbox approach suggests candidate reactions, but requires the users to assign the proper gene/GPR for reactions to be added. Merlin provides tools to find gaps, but no support is provided for automated gap filling.

Automated gap filling solutions still require inspection for further refinement, as reactions can be arbitrary added to restore model connectivity and pathway completeness. This fact calls for the continuous development of improved gap filling algorithms. To address that issue, all platforms that integrate gap filling usually provide their own algorithms. Those algorithms are variations of the GapFill algorithm, originally purposed by Kumar *et al*. [49]. The gap filled formulation, as it is implemented in the Model SEED, is detailed in section 2.3.2.

One of the drawbacks of GapFill and its variations is the use of Mixed-Integer Linear Programing (MILP) to determine the minimum set of reactions to be added to the model. GapFill was found to take over 14 hours to gap fill a single model of a prokaryote [62]. Databases such as MetaCyc and the Model SEED database are now comprised of over 10.000 reactions, which further extends the computation time for gap filling. Models of multiple cellular compartments add additional complexity to the gap-filling problem, further extending the computational time. In the development of the heavily compartmentalized human metabolic reconstruction (Recon 1) [63], the authors chose to de-compartmentalize the network to facilitate the gap filling process. This approach has the disadvantage of coupling reactions that do not co-occur in the same cellular compartment [64].

Recently, methods have been introduced that use linear programing (LP) to substantially reduce the computational time of gap filling. FASTCORE [65] adopts an LP formulation that was shown capable of obtaining good approximations to optimal solutions when compared with a MILP formulation. FastGapFill [64] presents itself as an expansion of FASTCORE that optimizes its LP formulation for heavily compartmentalized organisms. FastGapFill is available as an extension to the COBRA Toolbox [60]. A similar method, FastGapFilling [66] that also utilizes an LP formulation was used to effectively gap fill *E. coli* and yeast networks, performing up to three times faster when compared with a MILP formulation. FastGapFilling was integrated in the simulation framework of MetaFlux [67] that is distributed with Pathway Tools.

2.3 METHODS

2.3.1 The Model SEED

The Model SEED pipeline was used to build the models for different studies in this chapter. Figure 2.6 shows all the steps in the pipeline for automated reconstruction of GEMs in the Model SEED.

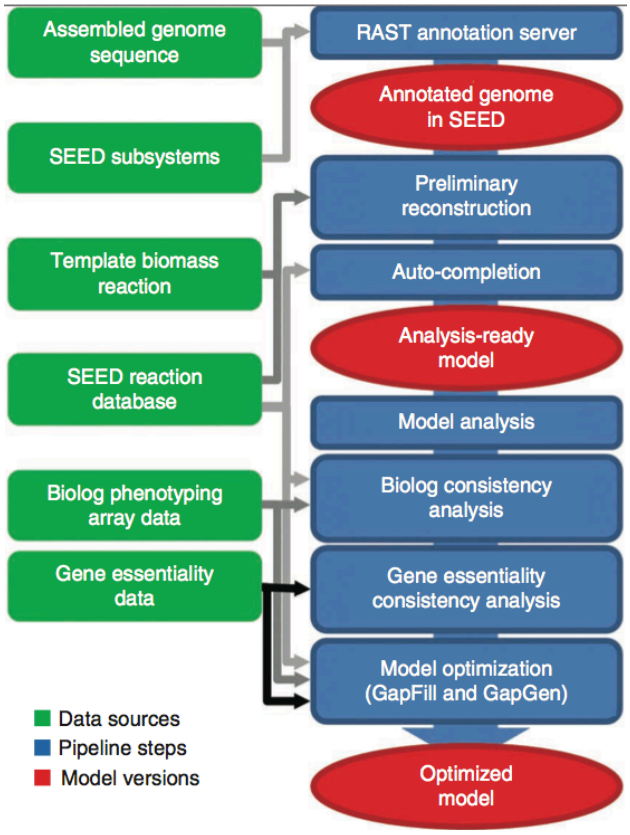


Figure 2.6 Model SEED genome scale reconstruction pipeline [11].

On Figure 2.6 we can see that the Model SEED proposes 7 different pipeline steps, tapping into multiple sources of data for the process of GEM reconstruction. Here, we detail the steps 1-4 as they were used to build the models necessary for this study. Steps 5-7 involve additional model curation and optimization of the models using experimental data when available. The Model SEED can use Biolog phenotyping arrays and gene essentiality data for this task. As we conducted reconstructions for

over 3000 organisms, this type of experimental data is only available for very few organisms, thus these steps were not performed in this research work.

1st Step - RAST annotation server

Users can upload genome sequences using the RAST server (http://rast.nmpdr.org/). The genomes will be annotated with the SEED Subsystem approach described in section 2.2.1. A search can also be conducted in the Model SEED for genomes already available in the SEED and previously annotated by RAST.

2nd Step - Preliminary reconstruction

The second step in the pipeline performs a preliminary reconstruction, where the RAST annotations are used to generate draft models. Draft models comprise a reaction network complete with GPR associations, predicted Gibbs free energy of reaction values and the biomass reaction. The biomass reaction includes non-universal cofactors, lipids and cell wall components. The biomass reaction is organism-specific, based on a biomass reaction template. The template makes use of the SEED subsystems and RAST functional annotations to assign non-universal (e.g., cofactors, cell wall components) biomass components that represent unique biological functions exhibited by an organism.

In order for an organism-specific biomass component to be added to the biomass reaction, its genome must contain the proper subsystems and annotations specified in the template. The GPR associations represent the mapping between the biochemical reactions and the standardized functional roles assigned to genes during the RAST annotation. This mapping allows to differentiate cases where protein products from multiple genes form a complex to catalyze a reaction, and cases where protein products from multiple genes can independently catalyze the same reaction. The draft model includes all reactions associated with one or more enzymes encoded in the genome. Additionally, spontaneous reactions are also added on this step.

3rd Step – Auto-completion

Draft models quality depends on the quality of the annotations used in the preliminary reconstruction. Due to this fact, these models usually contain gaps preventing the production of some biomass components. In this step, the Model SEED applies an optimization algorithm that identifies the minimal set of reactions that must be added to each model to fill these gaps [49, 68]. The gap filling algorithm is described in detail in section 2.3.2. Reactions to be used by gap filling are selected from the Model SEED internal database. This curated database contains mass and charge balanced reactions, standardized to aqueous conditions at neutral pH. The Model SEED reaction database integrates all the biochemistry contained in KEGG and 13 published genome-scale metabolic models. This step is conducted to ensure that every model is capable of simulating cell growth.

4th Step - Model analysis

Model analysis is performed to assess the capacity of reactions to carry flux and reaction essentiality. The Model SEED pipeline uses Flux Variability Analysis (FVA) to classify the reactions in the SEED models as essential, active or blocked. The detailed formulation of FVA is described on Section 4.3.2. Reactions that must carry flux for growth to occur are classified as essential; reactions that only optionally carry flux were classified as active; and reactions that are unable to carry flux were classified as blocked. Genes encoding reactions that were classified as essential were subsequently classified as essential, as long as alternative isozymes did not exist for these genes. Additionally, FBA is used to iteratively assess which compounds in the *in silico* media formulation are essential for the model to be able to produce biomass. These results provide clues for additional manual curation efforts.

2.3.2 Gap filling in the Model SEED

The gap filling as implemented in the Model SEED is detailed below as described by Henry *et al*. [69]. This method has been originally proposed by Kumar *et al*. [49] attempting to correct false negative predictions from the simulations provided by the original model. This is achieved by two alternative ways: (i) relaxing reversibility constraints on the model's reactions; (ii) adding new reactions to the existing model. For each condition, where the model simulation led to a false negative prediction, the

formulation detailed below was used based on a database of reactions consisting of every balanced reaction in the KEGG or in the published genome-scale models available in Model SEED.

$$minimize \sum_{i=1}^{r_{gapfilling}} (\lambda_{gapfill,i} Z_i) \quad (2.1)$$

subject to:

$$N_{reactionDB} \cdot v = 0 \quad (2.2)$$

$$0 \leq v_i \leq v_{max,i} Z_i \qquad i = 1, \dots, r \quad (2.3)$$

$$v_{bio} > v_{min} \quad (2.4)$$

The objective function (2.1) minimizes the number of reactions, which are not present in the model, but should be added for biomass to be produced in those conditions. Since, in this case, there is a false negative prediction, at least one reaction will need to be added.

In the formulation, all reactions are treated as reversible, being every reversible reaction decomposed into two reactions in the two directions. This allows for the independent addition of each direction in the algorithm. As a result of this, reactions represented in the formulation are the forward and backward components of the reactions in the database (from KEGG/models in SEED). In the objective function, $r_{gapfilling}$ represents the total number of reactions in the database; Z_i is a binary variable equal to zero if the flux through reaction i is zero and one otherwise; and, $\lambda_{gapfill,i}$ is a constant value stating the cost associated of adding reaction i to the model. If reaction i is already present in the model, $\lambda_{gapfill,i}$ is zero. Otherwise, $\lambda_{gapfill,i}$ is calculated using equation (2.5):

$$\lambda_{gapfill,i} = 1 + P_{KEGG,i} + P_{structure,i} + P_{known-\Delta G,i} + P_{unfavorable,i}\left(3 + \frac{\Delta_r G^{\circ m}_{i,est}}{10}\right) \quad (2.5)$$

Each of the *P* variables in equation (2.5) is binary, representing a penalty applied when adding different types of reactions to the model: they are equal to one if the penalty applies to the type of the particular reaction and equal to zero otherwise. $P_{KEGG,i}$ is related to reactions not in KEGG; $P_{structure,i}$ to the

addition of reactions involving metabolites with unknown structure. $P_{known-\Delta G,i}$ to reactions for which$\Delta_r G^{\circ}$cannot be calculated; $P_{unfavorable,i}$ to reactions operating in an unfavorable direction.

Equation (2.2) implements the mass balance constraints related to the steady-state assumption of FBA. Here, $N_{reactionDB}$ is the stoichiometric matrix, and v flux vector through reaction database.

Equation (2.3) enforces the bounds on reaction fluxes (v_i), and the values of the reaction use variables (Z_i). This equation ensures that each reaction flux, v_i, is zero unless Z_i is one. The $v_{max,i}$ term in equation (2.3) is the core to the simulation using FBA. If $v_{max,i}$ corresponds to a reaction associated with a knocked-out gene, $v_{max,i}$ is set to zero. If $v_{max,i}$ corresponds to the uptake of a nutrient not in the medium, $v_{max,i}$ is also set to zero.

Equation (2.4) constrains the biomass flux, v_{bio}, to a nonzero value, to ensure growth.

The result of the gap filling optimization includes a list of irreversible reactions from the model that should be made reversible, and a set of reactions not in the model that should be added to fix a false negative prediction. Recursive mixed integer linear programming (MILP) [70] is used to perform multiple gap filling to correct each false negative prediction.

2.3.3 Iterative gap filling of GEMs for reaction activation

We developed a new gap filling method with the aim of activating all blocked/inactive reactions in a metabolic model. A single "iteration" of our approach is identical to the fundamental gap filling algorithm described in the previous section.

The difference between the two formulations is the objective of the gap filling. As implemented in the Model SEED, gap filling is performed when the model is unable to produce biomass. In our new formulation, the gap filling is performed to activate all inactive reactions in the model, regardless of their impact in the biomass production. As the objective of the gap filling is the activation of blocked reactions in the network, a sizable number of reactions can be added depending on the number of inactive reactions in a given model. In each iteration of our algorithm, we force one reaction in our

model to be active, while minimizing the addition of new reactions. If a reaction is already active in the model, no new reactions will be added. If a reaction is involved in a pathway with a single gap, then reactions might be added. We apply this process iteratively to all annotated reactions in our original model, while integrating the new reactions identified in each iteration into the model, such that we accumulate an increasing number of gap filled reactions, while activating an increasing number of annotated reactions. We iterate through reactions in a specific order, starting with central carbon reactions, progressing to primary biosynthesis pathways, then catabolic and degradation pathways, and finally secondary metabolic pathways.

Often, reactions that are gap filled later in this iterative process may eliminate the need for reactions gap filled earlier in the process. For example, filling a gap in the biosynthesis pathway for a precursor metabolite may eliminate the need to add a transporter for the same precursor compound. For this reason, the final step of our iterative gap filling approach is a sensitivity analysis, where we remove gap filled reactions from our model one at a time and evaluate the impact of this removal using FVA. If the removal of a gap filled reaction has no impact (e.g. it causes no annotated reactions to become inactive), then we leave the reaction out of the model. Otherwise, we restore the reaction. A second benefit of this analysis is that it permits to associate gap filled reactions with the annotated reactions they correct, which enables us to use this analysis to quantify the quality of reaction annotations based on how much gap filling must be done to activate the reaction.

2.3.4 Assessing confidence in genome annotation

As described in the previous section 2.3.3, our iterative gap filling algorithm provides us with data that may be used to evaluate the confidence in each annotated reaction included in our model. Specifically, the gap filling process indicates how many un-annotated reactions must be gap filled in a model for each annotated reaction to function. It follows that a reaction that requires more gap filling to be activated in a particular model would have reduced confidence in that model.

To quantify this confidence, we developed a metric to compute the “cost” of an annotated reaction as the number of reactions that must be gap filled to activate the annotated reaction:

$$C_j = \sum_{i=0}^{N_{gaps,j}} \frac{1}{N_{act,i}} \tag{2.6}$$

In Equation 2.6, C_j is the cost of annotated reaction j; $N_{gaps,j}$ is the number of gap filled reactions whose activity is coupled to reaction j (e.g. meaning knockout of the gap filled reactions results in inactivation of reaction j); $N_{act,i}$ is the total number of annotated reactions that have been coupled to gap filled reaction i. We include $N_{act,i}$ in equation 2.6 because it is important to account for the fact that a single gap filled reaction will often correct many annotated reactions at once. Consider a linear pathway with ten steps, nine of which are annotated. Our gap filling algorithm will fill in the missing step and associate the gap filled reaction with all nine annotated reactions. Without the $N_{act,i}$ scaling factor, each annotated reaction would have a cost of 1, and we are failing to capture the fact that our annotated pathway is actually 90% complete. When we include $N_{act,i}$, every reaction has a cost of $1/9$, and we are successfully capturing the fact that our pathway is 90% complete, and the annotated reactions, therefore, have a low cost and a high confidence.

This same approach can also be applied to associate a value with each gap filled reaction, enabling us to identify the gap filled reactions that are most likely to be correct, and prioritizing the search for genes to associate with these gap filled reactions. In this case, we quantify the value of a gap filled reaction as the number of annotated reactions that were activated by the gap filled reaction:

$$V_i = \sum_{j}^{N_{act,i}} \frac{1}{N_{gaps,j}} \tag{2.7}$$

In Equation 2.7, V_i is the value of gap filled reaction i; $N_{act,i}$ is the number of annotated reactions whose activity has been coupled to reaction i; and $N_{gaps,j}$ is the number of other gap filled reactions that are also required to permit reaction j to function. As before, it is important to scale by $N_{gaps,j}$ when computing the gap filled reaction value because a single activated reaction may often require many gap filled reactions to correct it.

2.4 RESULTS AND DISCUSSION

We applied the Model SEED algorithm to construct draft models for over 3000 genomes, with the goal of exploiting our models as a tool to evaluate and improve genome annotations. We accomplished this goal with three studies: a global gene knockout and flux variability analysis to quantify essential genes and blocked reactions (section 2.4.1); applying our iterative gap filling formulation to identify important gaps and assess annotation confidence (section 2.4.2); and, a detailed comparison of all models within a single family of closely related genomes to evaluate annotation consistency (section 2.4.3).

2.4.1 Global analysis of the automated reconstructed GEMs

The Model SEED pipeline (Section 2.3.1) was used to reconstruct approximately 3000 GEMs, representing all genomes available in the SEED as of Winter 2011. We utilized FVA with these models to classify all reactions as active, blocked, or essential, as detailed in the methods' section. The results of this analysis are shown in Figure 2.7, and can be used to evaluate model quality, and by extension, annotation quality in three ways: (i) a poorer model and annotation will have more incomplete pathways resulting in a larger fraction of blocked reactions; (ii) a poorer model and annotation will generally have fewer genes mapped to metabolism and a greater fraction of gap filled reactions; and, (iii) a poorer model and annotation will have fewer essential genes relative to the number of essential reactions (a sign of many essential reactions having no genes).

Generally, we find that the number of essential reactions in our models varies little, despite wide variations in model size (red points in left panel of Figure 2.7). This is to be expected, as there is little variation in the biomass composition across our models, and all models were analyzed in a common rich medium condition. We found the number of active reactions to be roughly proportional to the overall model size (blue points in left panel of Figure 2.7); however, we found significant variation in the number of inactive/blocked reactions (green points in left panel of Figure 2.7). As previously mentioned, the number of inactive reactions is one of our indicators of model and annotation quality. Thus, this result indicates that there is also a significant variation in annotation quality among our 3000 genomes.

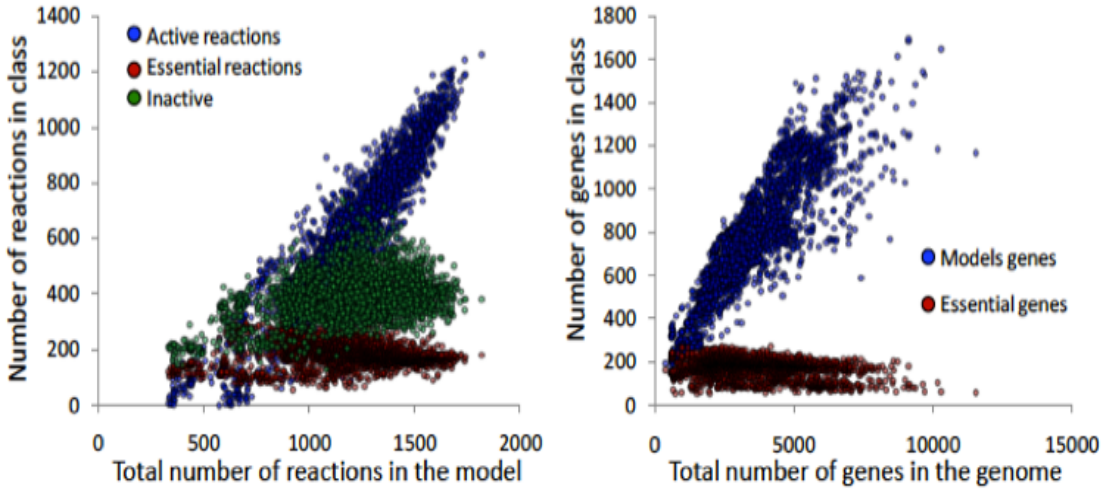


Figure 2.7 Characterization of reactions and genes across all models. Left: reactions were classified in 3 different classes; essential, active and inactive/blocked. Essential reactions are reactions that disrupt model growth when removed from a model. Reactions that carry flux are classified as active and reactions unable to carry flux were classified as inactive. Right; Genes encoding essential reactions were classified as essential. Model genes are all genes included in the metabolic model reconstruction.

There is one other possible explanation for the presence of inactive reactions, relating to the biomass composition of the models. In some cases, the biomass composition of the model is incomplete, failing to capture the biological distinctiveness of the organism being modeled. Thus, some inactive pathways may be responsible for producing a biomass component that has been left out of the biomass composition for the model, creating an erroneous dead-end in the model. We ultimately hope to apply our algorithms to recognize and correct these types of errors. Overall, the large fraction of inactive reactions in our models indicates that there is a significant opportunity to greatly improve models if gap filling algorithms could be adapted to correct these reactions.

We also compared the number of essential metabolic genes with the total number of genes in each model and genome (red versus blue points in right panel of Figure 2.7). As with essential reactions, we observe little variation across the 3000 models for essential genes.

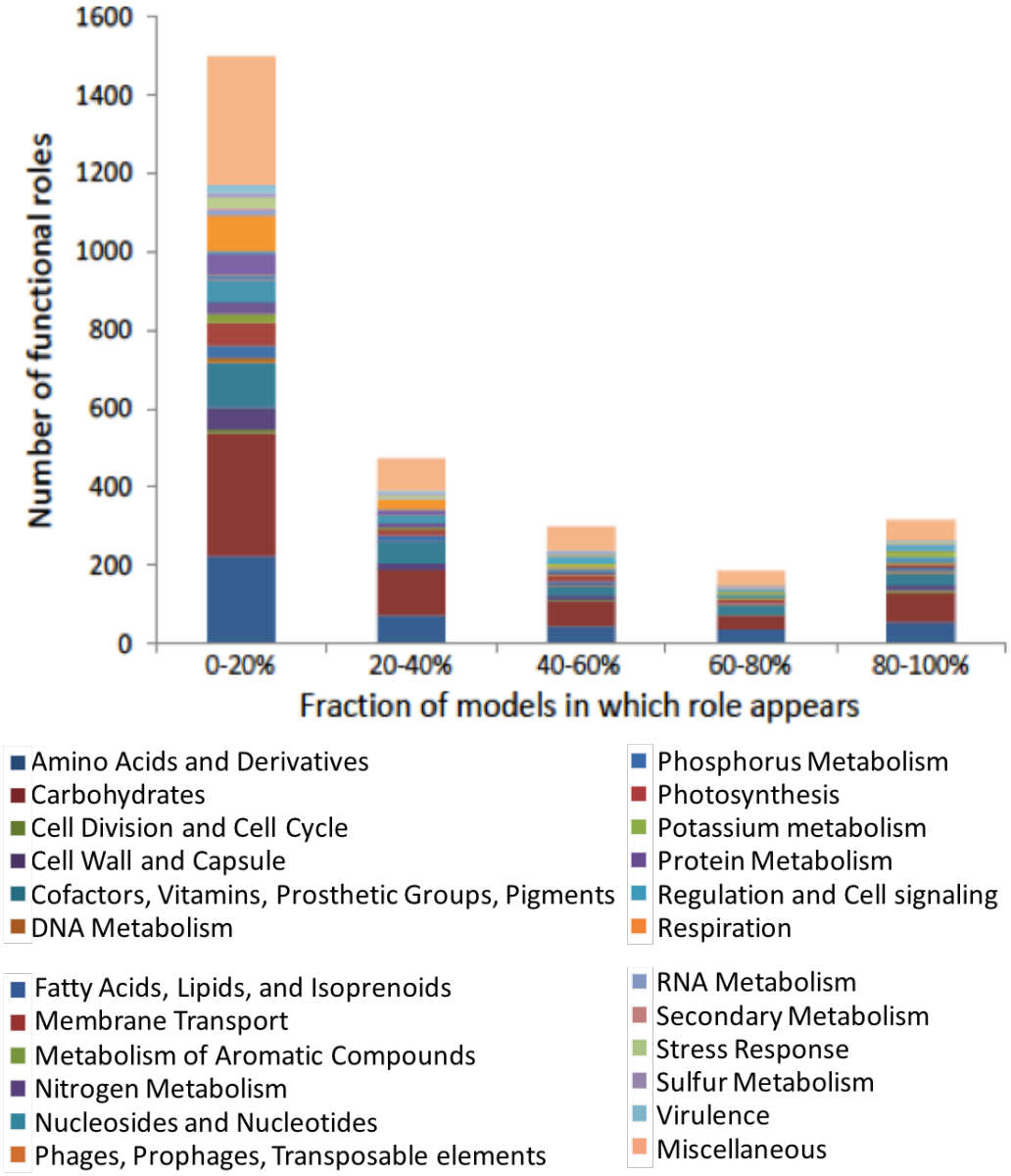


Figure 2.8 Distribution of functional roles across major cellular processes for 3000 genome-scale metabolic models. The functional roles and their respective subsystems were identified for all reactions in the models. The number of functional roles (y-axis) is shown for 5 different fractions (in 20% intervals) of the total of models used in this analysis (x-axis). The legend shows the 24 major cellular processes chosen to categorize the functional roles.

This fact shows that essentiality, at least as predicted by the genome-scale models, is not related to the genome size. This result is corroborated by previous studies, which found that larger genomes mostly contain additional non-essential functions (e.g. secondary metabolism) to improve their capacity to survive in environments where resources are scarce but diverse [71]. A recent study of the phenotypic evolution of bacteria using gene essentiality data also shows significant conservation of gene essentiality [72].

We also analyzed the functional roles associated with reactions included in our models. There were approximately 2800 functional roles assigned to reactions across all of our 3000 models. To give an overview of the roles present in the models, we grouped the functional roles by subsystem and subsystem categories, which consist of 24 major cellular processes (Figure 2.8). All data used in this analysis is available in Supplementary Material S2.1. Our results show that most functional roles are present in only 0-20% of the models (far left side of Figure 2.8). Within this group of rare functions, we see representatives from a wide range of subsystems, highlighting the diversity of our models.

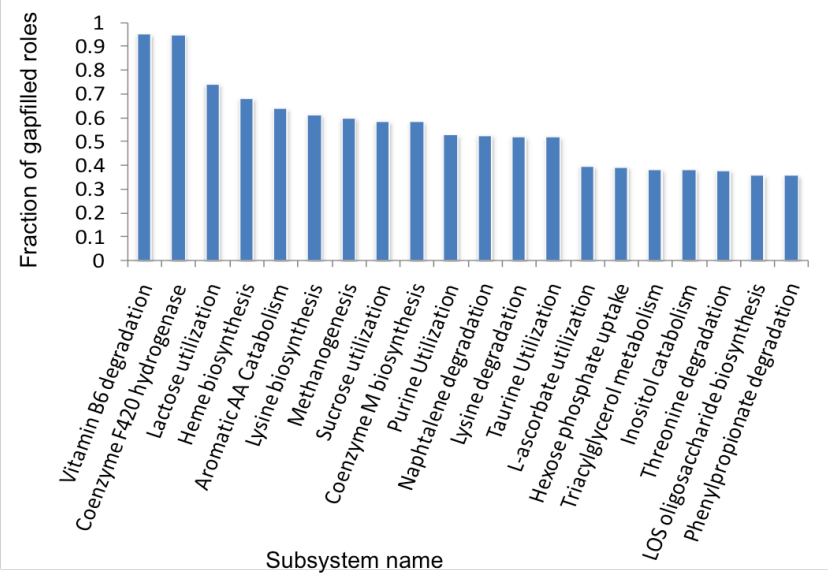


Figure 2.9 The 20 most gap filled subsystems in the 3000 genome-scale metabolic models. The fraction of gap filled functional roles represents, for each subsystem, the quotient of the number of gap filled roles by the total number of roles.

On the other side of the spectrum (far right side of Figure 2.8), we have ~300 functional roles that appear in 80-100% of the models. We observe that almost half of these ubiquitous

Table 2.2 Reactions associated with most gap filled subsystems

Subsystem	Name
Pyridoxin(Vitamin B6) Degradation Pathway	2-(acetamidomethylene)succinate hydrolase (EC 3.5.1.29)
Pyridoxin(Vitamin B6) Degradation Pathway	Pyridoxine 4-oxidase (EC 1.1.3.12)
Coenzyme F420 hydrogenase	Coenzyme F420 hydrogenase beta subunit (FrcB) (EC 1.12.98.1)
Coenzyme F420 hydrogenase	Coenzyme F420 hydrogenase gamma subunit (FruG) (EC 1.12.98.1)
Coenzyme F420 hydrogenase	Coenzyme F420 hydrogenase beta subunit (FruB) (EC 1.12.98.1)
Coenzyme F420 hydrogenase	Coenzyme F420 hydrogenase gamma subunit (FrcG) (EC 1.12.98.1)
Coenzyme F420 hydrogenase	Coenzyme F420 hydrogenase alpha subunit (FruA) (EC 1.12.98.1)
Coenzyme F420 hydrogenase	Coenzyme F420 hydrogenase alpha subunit (FrcA) (EC 1.12.98.1)
Lactose utilization	Beta-galactosidase (EC 3.2.1.23)
Lactose utilization	Lactose permease
Heme biosynthesis	Cytochrome cd1 nitrite reductase (EC:1.7.2.1)
Aromatic Amin Catabolism	Monoamine oxidase (1.4.3.4)
Aromatic Amin Catabolism	Amiloride-sensitive amine oxidase [copper-containing] precursor (EC 1.4.3.21)
Aromatic Amin Catabolism	Monoamine oxidase (1.4.3.4)
Aromatic Amin Catabolism	Phenylacetaldehyde dehydrogenase (EC 1.2.1.39)
Aromatic Amin Catabolism	4-hydroxyphenylacetate 3-monooxygenase (EC 1.14.13.3)
Aromatic Amin Catabolism	4-hydroxyphenylacetate 3-monooxygenase (EC 1.14.13.3)

functions were assigned to two subsystem categories: *amino acids and derivatives* and *carbohydrates*. The *amino acids and derivatives* category contains functions relating to amino acid biosynthesis and degradation. The *carbohydrates* category includes functions related to carbohydrate biosynthesis, central carbon metabolism, fermentation, etc. Thus, together these subsystems cover most the core

biochemistry and essential metabolite biosynthesis, as well as energy production. We would expect to find these functions in nearly all prokaryotic genomes.

Our analyses of reaction essentiality and functional role assignment both considered all of the reactions included within our models. However, some of these reactions were gap filled during the auto-completion step of the Model SEED reconstruction pipeline. We applied our models to identify which functional roles were most commonly associated with gap filled reactions (Figures 2.9-11 and Supplementary Material S2.2). In our first analysis, we identified the twenty subsystems for which the largest fraction of associated reactions in our models were gap filled (Figure 2.9). This fraction was computed by dividing, for each sub-system, the number of gap filled roles by the total number of roles. Vitamin B6 degradation was the most gap filled subsystem, followed by Coenzyme f420 hydrogenase and lactose utilization. The reactions associated with the most gap filled subsystems are shown on Table 2.2. These results provide valuable guidance to our future annotation efforts, emphasizing the need for additional work curating these subsystems. As these curation efforts proceed, annotations in these subsystems will be corrected in all genomes, and ultimately in all models and the Model SEED itself.

Our analysis of gap filled reactions in subsystems revealed that for a majority of cellular processes at least one subsystem required gap filling (Figure 2.10). The only category that required no gap filling was *phages, prophages and transposable elements* (detailed results available in the Supplementary material S2.2). This is a reasonable result, as that cellular process does not include essential metabolic functions that will typically be gap filled.

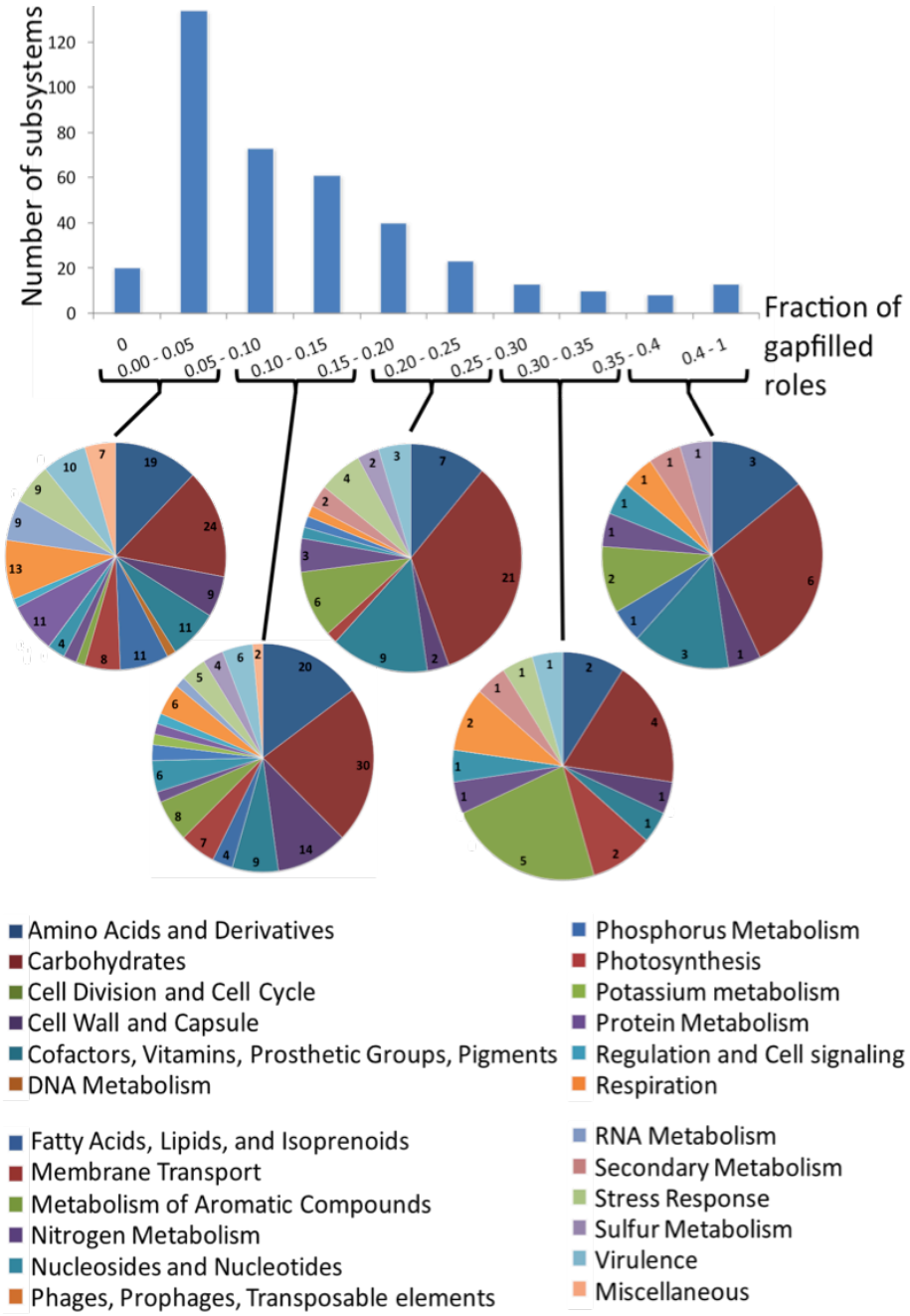


Figure 2.10 Fraction of functional roles that were gap filled in the 3000 genome-scale models. The fraction of gap filled functional roles represents for each subsystem, the quotient of the number of gap filled roles by the total number of roles. The number of subsystems that required gap filling is shown for 10 intervals (x-axis) representing different fractions of gap filled roles. The legend shows the 24 major cellular processes chosen to categorize the subsystems.

We also observe that the largest number of subsystems (~150) fell in the category 0 – 0.05 (0 – 5%), indicating many subsystems requiring only a small amount of gap filling; in contrast, only 20 subsystems required extensive gap filling of 40% to 100% of their reactions (corresponding to the interval of fraction of gap filled roles 0.4 – 1 in Figure 2.10).

The large number of subsystems requiring little gap filling and the small number of subsystems requiring extensive gap filling points to the quality of the manually curated SEED subsystems, as little effort would be required to curate a majority of the subsystems. It also demonstrates how these data can greatly improve SEED annotations by identifying key subsystems where more curation is needed.

Additionally, we recognize that the size of the subsystem can cause noise in the previous analysis, as subsystems with few functional roles can more easily have a high fraction/percentage of gap filled reactions.

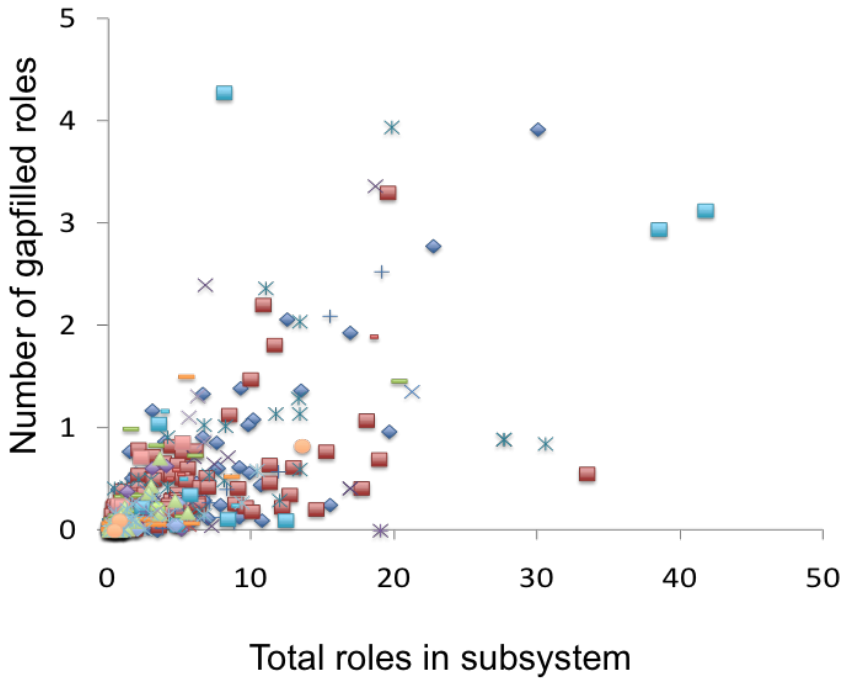


Figure 2.11 Distribution of gap filled functional roles in the 3000 genome-scale metabolic models. The number of gap filled roles per subsystem (y-axis) is shown for the total number of roles in the subsystem (x-axis)

To verify this hypothesis, we examined the total roles in a subsystem vs. the number of gap filled roles

(Figure 2.11 and details in Supplementary material S2.2). We found that many subsystems indeed have few functional roles. We also found that subsystems with a larger number of roles also tend to have a low percentage of gap filled roles. Overall, the roles that are most often gap filled are associated with subsystems with a large number of roles, like amino acids biosynthesis, regulation and cell signaling or carbohydrates. This result was not totally unexpected as many roles associated with those cellular processes can be essential and will be prioritized for gap filling when missing from genome annotation.

2.4.2 Assessing confidence in genome annotations

In the previous study, we determined that many of the reactions in our 3000 GEMs are inactive (Figure 2.7). We suggested these inactive reactions were due to a combination of missing annotations and limitations in our biomass composition reaction. Gap filling provides a means of correcting inactive reactions that are a result of missing annotations by adding additional reactions to the models. The issue is that the gap filling applied as part of the Model SEED auto completion step (as described in Section 2.3.2) adds only the minimal set of reactions needed to produce all biomass precursors, and this gap filling is performed on extremely rich media. This leaves many possible gaps in the network related with pathways that are not utilized by the model for biomass production, with reactions involved in synthesis of nutrients present in the rich gap filling media, or with reactions involved in degradation of nutrients not present in the rich media. Our previous studies also demonstrated, however, how even this very limited form of gap filling provided valuable guidance to our annotation curation efforts.

Motivated by these results, we developed a new gap filling formulation that favors the activation of inactive reactions in the network. This new methodology was described in section 2.3.2. We can then quantify how many reactions were required to be gap filled to activate each inactive reaction. We can also quantify which biomass precursors required the most gap filled reactions. This solution analysis allows us to use the data as a metric to quantify the impact of gap filling on models and assess the quality of genome annotations.

To assess quality of the genome annotations, we used the data from this analysis to compute two different scores: cost of annotated reaction and value of gap filled reaction. The formulation of these scores can be found in Section 2.3.3. The *cost of annotated reaction* score aims to quantify, across all models, the number of gap filled reactions required to enable each inactive gene associated reaction to carry flux in the network. The higher the cost of an annotated reaction, the lower the confidence in that genome annotation. We use these data to compute the distribution of annotated reaction costs across all 3000 of our GEMs (Figure 2.12 and Supplementary Material S2.3).

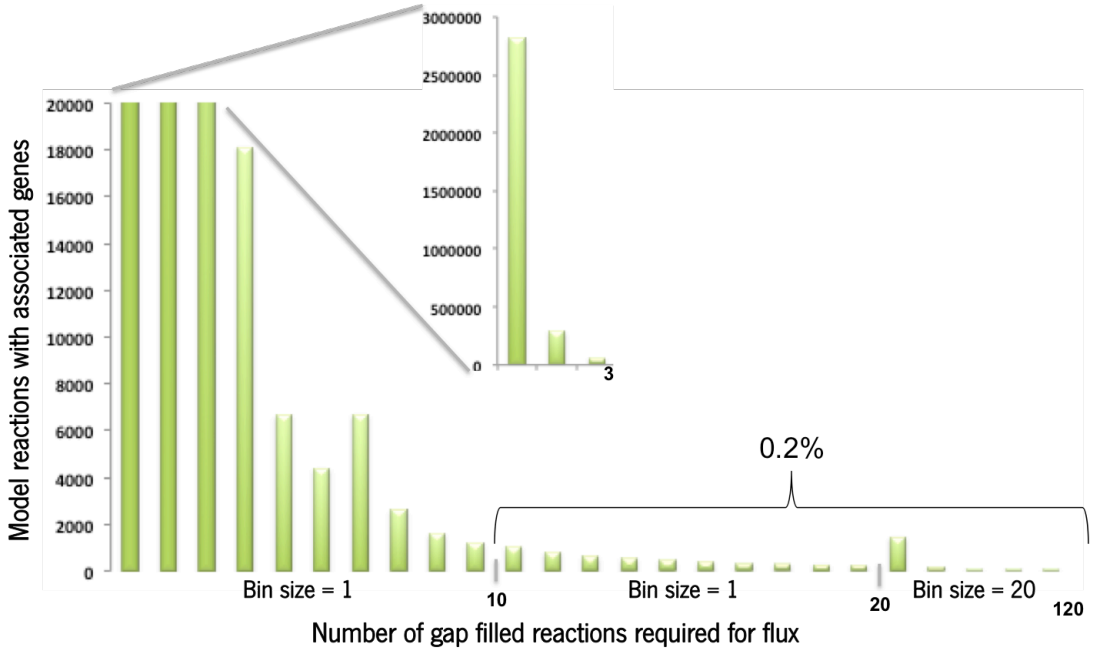


Figure 2.12 Distribution of values for the cost of annotated reaction for the 3000 genome-scale metabolic models. The cost of annotated reaction is shown as the number of gap filled reactions required for an inactive reaction to carry flux (x-axis). To facilitate comprehension, the x-axis is not linear, being values organized in bins of different sizes to better show cases where a large number of gap-filled reactions is necessary. The distribution of the cost is shown for 25 intervals. The number of genes associated with model reactions associated is shown in the y-axis.

One of the most clear results that we can observe by looking at the first interval in Figure 2.12 (that required a change of scale) is that the vast majority of gene associated reactions required one gap filled reaction to be able to carry flux through the network. This is a very positive result for the quality of

annotations used to build the models. Another good indicator of the current state of the annotations is that only 0.2% of reactions require 10 or more gap filled reactions to become active. We can use these results to drive the removal of over annotations from the genomes and models. The 0.2% of reactions that require 10 or more gap filled reactions can probably be safely removed, as there seems to be very little contextual evidence for their presence in the network.

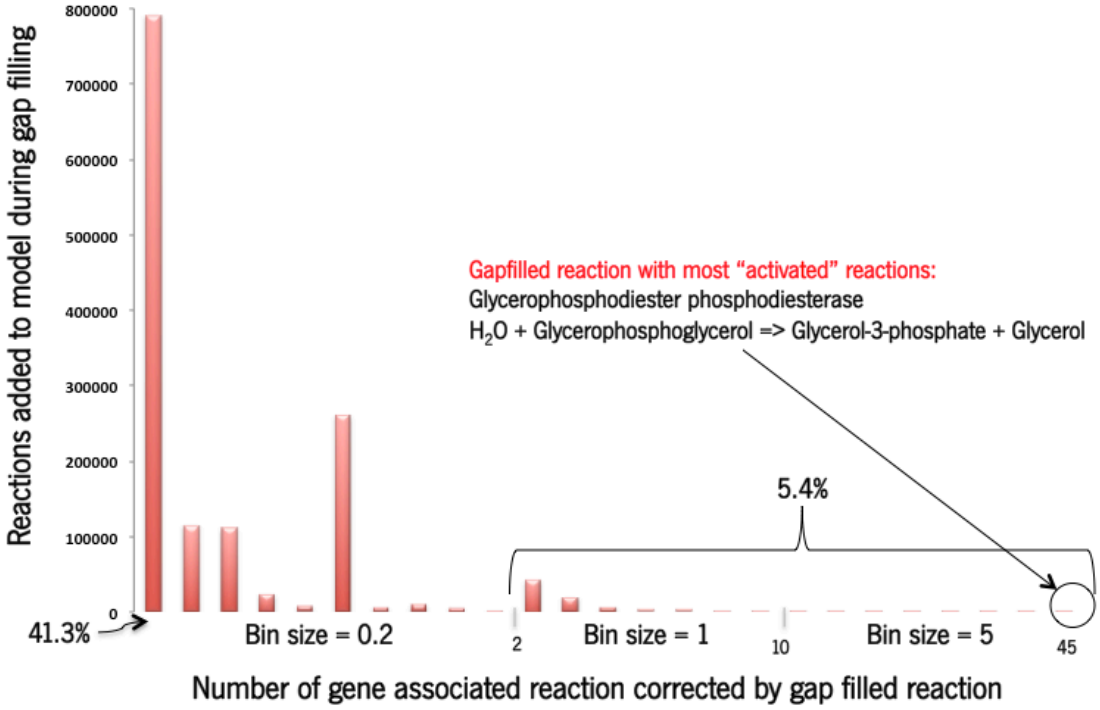


Figure 2.13 Distribution of the value of gap filled reaction for 3000 genome-scale metabolic models. The value of a gap filling reaction is shown as the number of gene associated reactions in the network that are corrected by gap filling (x-axis). The distribution is shown for 25 intervals. The number genes added to the model by gap filling is shown in the y-axis.

The second score we considered computes the value of a gap filling reaction by calculating the number of gene associated reactions in the network that are repaired by this gap filling. Figure 2.13 (Supplementary material S2.4) displays a histogram of the number of gene-associated reactions that were corrected by each gap filled reaction. The first column in the distribution is quite large, comprising of 41.3% of gap filled reactions correcting less than 0.2 gene associated reactions in the network. This indicates that many gap filled reactions are of low quality, but the large size of this

column is the result of diminishing returns when gap filling many reactions to fix only a single gene associated reaction.

Additionally, we observed that a large set of reactions were activated/corrected by the addition of only 1 gap filled reaction. Only 5% of reactions were activated/corrected by the addition of 2 or more gap filled reactions. The reactions with higher gap filling value are prime candidates for inspection to fill in missing gene annotations.

We inspected the most extreme case found on this analysis with a gap fill value of 43.65 (highlighted in Figure 2.13). Glycerophosphodiester phosphodiesterase (Glycerophosphoglycerol) was the reaction associated with the highest score (Model SEED reaction *rxn08699*): H_2O +Glycerophosphoglycerol=> Glycerol-3-phosphate + Glycerol

This reaction is present in 2943 models, but it was gap filled in only 278 models. The highest value was found for the model of *Thermobaculum terrenum* ATCC BAA-798 (Model SEED model ID *Seed525904.4.796*). Since our 3000 models were reconstructed for all genomes publicly available on the SEED, we used the SEED tools to investigate annotations associated with this reaction.

Table 2.3 Curated genome annotations

Genome	Old annotation	Gene ID	New annotation
Thermobaculum terrenum ATCC BAA-798	Glycerophosphoryl diester phosphodiesterase	fig|525904.4.peg.2926	Glycerophosphoryl diester phosphodiesterase (EC 3.1.4.46)
Desmospora sp. 8437	Glycerophosphoryl diester phosphodiesterase	fig|997346.3.peg.2633	Glycerophosphoryl diester phosphodiesterase (EC 3.1.4.46)
Desmospora sp. 8437	Glycerophosphoryl diester phosphodiesterase	fig|997346.3.peg.2632	
Streptomyces griseoaurantiacus M045	secreted hydrolase	fig|996637.3.peg.5832	Glycerophosphoryl diester phosphodiesterase (EC 3.1.4.46)

We searched the SEED for the gene(s) associated with gap filled glycerophosphodiester phosphodiesterase (EC 3.1.4.46) functional role for that organism. The gene (SEED id *fig|525904.4.peg.2926*) associated with the gap filled role had an incomplete annotation name, lacking the proper EC number. Using the compare regions feature tools in the SEED, we were able to assess that the same annotation inconsistency was occurring in additional organisms (Figure 2.14). In the genome annotation of *Streptomyces griseoaurantiacus* M045 the gene associated with the gap filled functional role was annotated as "secreted hydrolase". We were able to use the SEED annotation tools to easily fix this error across all occurrences in the cluster. The fixes introduced by this work are shown on Table 2.3.

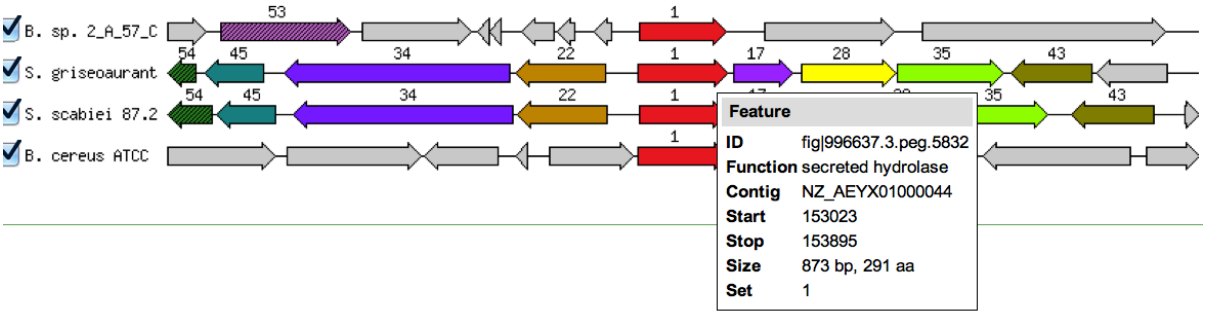


Figure 2.14 Comparative analysis of genes (represented in red) associated with the gap filled function glycerophosphodiester phosphodiesterase (EC 3.1.4.46).

This same process was repeated with other gap filled reactions leading to the correction of multiple errors in the SEED database of genome annotations.

After assessing the cost of genome annotations and value of gap filled reactions, we performed an analysis to assess the gap filling of missing biomass precursors. Table 2.4 shows the most commonly gap filled biomass compounds. The average score in Table 2.4 represents the average number of reactions that had to be gap filled to enable the production of the specific biomass compound.

Table 2.4 Most commonly gap filled biomass components

Compound	Models	Gap filled	Avg. score
Cardiolipin	2134 (62.20%)	1618 (75.82%)	4.49
Phosphatidylglycerol	1743 (50.80%)	1176 (67.47%)	4.30
Calcium	3431 (100%)	2301 (67.06%)	0.71
Ubiquinone-8	1919 (55.93%)	1242 (64.72%)	4.98
Heme B	69 (2.91%)	43 (62.32%)	0.73

Cardiopilin was found to be the most gap filled compound, present in 2134 models (62% of total models) and being gap filled in 75% of those. This is due to the fact that the Model SEED uses a template biomass reaction in its network reconstruction process. Some compounds are marked as universal (e.g. amino acids) being present in the biomass reaction for all models. Other compounds are associated with a specific subsystem or function role or class (e.g. gram negative vs gram positive) and are only included in models that have fit those criteria. Cardiopilin is a non-essential phospholipid for growth in prokaryotes [73-75] and was wrongly associated with a high number of organisms that do not require Cardiopilin as a biomass precursor, causing it to be gap filled in the vast majority (75%) of the models in which it is present. Phosphatidylglycerol is another phospholipid that, like Cardiopilin, can be used for optimal growth but is not essential [75].

Calcium is an essential co-factor [76, 77] being present in all models and being gap filled in 67% of the models. To further investigate why an essential compound such as calcium was being highly gap filled, we looked at the most gap filled reactions (Table 2.5). We found that 3/4 of the most gap filled reactions were associated with calcium transport. Particularly, the calcium transport via ABC system is gap filled in ~60% of the models. This revealed an error in the Model SEED reaction database regarding the calcium transporters not being properly mapped to reactions in the model.

Table 2.5 Most commonly gap filled reactions

Reaction	Number of times gap filled	Functional role
Calcium transport via ABC system (H_2O + ATP + Ca^{2+} [e] => ADP + Phosphate + Ca^{2+} + H^+)	2024	Calcium-transporting ATPase (EC 3.6.3.8)
Citrate-Ca^{2+} :H+ symporter (Ca^{2+} [e] + H^+[e] + Citrate[e] <=> Ca^{2+} + H^+ + Citrate)	217	Ca^{2+}/citrate complex secondary transporter
Citrate-Mg^{2+} :H^+ symporter (H^+[e] + Citrate[e] + Mg[e] <=> H^+ + Citrate + Mg)	72	
Calcium / sodium antiporter (Ca^{2+} + Na^+[e] <=> Ca^{2+} [e] + Na^+)	42	Ca^{2+}/Na+ antiporter

The correction of these errors impacted a substantial percentage of the models generated for this study, as well as future reconstructions generated using the Model SEED. In the first study we conducted, we analyzed gap filling at a subsystem level, pointing to subsystems that potentially require additional curation of genome annotations. With the analysis conducted in this section, we are able to find problems in genome annotations in specific genomes. This large-scale study also provided perspective on the quality of automated GEMs generated by the Model SEED. Finally, we were able to use this analysis to find and fix multiple errors in genome annotations and in the Model SEED reaction database.

2.4.3 Analysis of strains from the genus *Brucella*

The previous studies were done for all genomes available in the SEED. As part of on-going collaborations with the SEED research group, it was possible to conduct those studies at large scale to infer about the quality of automated GEMs and genome annotations. We were also able to fix gene annotations for multiple organisms across all bacteria phyla. However, most researchers focus their efforts on a specific organism or a small set of organisms. Inspired by that fact, we developed a protocol that can be used to assess annotation consistency with the aid of metabolic models for a

small set of closely related organisms. The results of that protocol applied to 15 species of *Brucella* are shown below.

Description of the Protocol

Step 1. Choice of genomes for analysis. We have chosen fifteen genomes representing the major species, biovars and clades of the genus *Brucella* [78] (Table 2.6).

Table 2.6 *Brucella* genomes used in this study with their SEED [14, 27] and PATRIC [79, 80] identifiers, sizes, number of contigs, and number of protein coding sequences (CDSs).

Genome Name	PubSEED ID	PATRIC Genome ID	Genome Size (bp)	Number of Contigs	Number of CDSs
Brucella abortus bv. 1 str. 9-941	262698.4	15061	3286445	2	3413
Brucella canis ATCC 23365	483179.4	25663	3312769	2	3394
Brucella ceti str. Cudo	595497.3	28239	3389269	7	3578
Brucella ceti M13/05/1	520460.3	83544	3337230	22	3367
Brucella melitensis bv. 1 str. 16M	224914.11	92729	3294931	2	3446
Brucella microti CCM 4915	568815.3	92249	3294931	2	3374
Brucella neotomae 5K33	520456.3	114381	3329623	11	3383
Brucella ovis ATCC 25840	444178.3	136990	3275590	2	3499
Brucella pinnipedialis M292/94/1	520462.3	74143	3373519	15	3356
Brucella sp. 83/13	520449.3	75385	3153851	20	3152
Brucella inopinata BO1	470735.4	109945	3366774	55	3361
Brucella inopinata-like BO2	693750.4	146994	3305941	174	3276
Brucella sp. NVSL 07-0026	520448.3	103899	3297137	17	3442
Brucella suis 1330	204722.5	107850	3315175	2	3402
Brucella suis bv. 5 str. 513	520489.3	73489	3323676	19	3316

Step 2. Potential mobile element proteins are identified and removed from consideration. In order to find potential mobile protein elements, we first identified repeat regions in each chromosome. BLASTN [81] was used to compare each of the fifteen genomes against itself. Any DNA region (other than rRNA

operons) occurring more than once in the genome with a nucleotide identity ≥ 90% and a length ≥ 200 nucleotides was considered to be a repeat. Although there are many ways to identify mobile element proteins that could be substituted within this framework [82], for the purposes of this study, we define a potential mobile element protein as one that overlaps a repeat region by at least 10 bp. All of the 15 *Brucella* genomes were then compared to the list of potential mobile element proteins using BLASTP, and matching proteins with identity larger or equal to 50% and coverage larger or equal to 80% were also considered to be potential mobile element proteins regardless of proximity to a repeat region. This resulted in the creation of 50 mobile element protein families, containing a total of 410 proteins. These proteins were excluded from subsequent steps due to their variability and because they are not currently used for metabolic model reconstructions.

Step 3. Families of core proteins are generated. To find the core proteins, the remaining genes from each of the *Brucella* genomes were compared. Two proteins were placed in the same protein family if they were bi-directional best hits between a pair of genomes with greater than 50% identity and 80% coverage, and the genes occurred within a conserved genomic context [83, 84]. We considered the context of the matched pairs to be conserved if there were at least 3 pairs of bi-directional best hits co-occurring within a 10 Kb region. This resulted in 5,038 families (with two or more proteins) containing a total of 52,626 proteins. From these initial families we generated *core protein families*, which are defined as families containing at most one protein from each genome, where 80% of the genomes are represented in the family. Similar to Step 2, it would be possible to use other methods for finding orthologous genes at this step as well [85].

Step 4. Annotation inconsistencies are removed. The core protein families of the RAST-annotated *Brucella* genomes were compared and inconsistencies (defined as two or more family members having different annotations) were evaluated. We manually curated a total of 398 families containing 4,848 proteins. We defined two metrics to measure progress.

The first:

Given a protein family (i.e., from one of the 5,038 families we constructed), at what frequency has any given pair of proteins within the family been assigned precisely the same annotation by RAST [14]?

We report this property before and after manual cleanup, and compare our annotations to other public annotation resources (Table 2.7).

Table 2.7 The consistency of annotations across different resources. For each protein in a *Brucella* protein family used in this study, all of the proteins with identical sequences were found in various databases and the percentage of pairs that were inconsistently annotated was computed. Annotations were collected from RefSeq [18], UniProt Knowledgebase (UniProtKB) [25], the Translated EMBL Nucleotide Sequence Data Library (TrEMBL) [26], the Integrated Microbial Genomes (IMG) system [86] and the SEED [14, 27].

Source	Number of Pairs	Number of Pairs Inconsistently Annotated	% of Pairs Inconsistently Annotated
RefSeq	562597217	383808122	68.2
IMG	101525838	52434525	51.6
UniProtKB/TrEMBL	112735194	46284849	41.1
UniProtKB/SwissProt	803819	42429	5.3
SEED	271622566	9056551	3.3
Original RAST Output	16349603	102097	0.6
RAST After Manual Curation	16349603	47504	0.3

The second:

How many Brucella-universal-reactions have been assigned to each genome?

By universal reactions we mean the reactions that are present in all *Brucella* genomes used in this study. We chose this second metric to demonstrate that improvements in annotations lead to improvements in the metabolic reconstructions.

Step 5. Annotation and reaction database improvements are made based on metabolic network reconstructions. Metabolic reconstructions were built for the fifteen *Brucella* genomes (Tables S1 and

S2 on the online supplementary material), using the Model SEED automated reconstruction pipeline. Starting with the manually improved genomes, we focused on the reactions that were non-universal among the 15 *Brucella* strains. The annotations relating to these reactions were manually evaluated and corrected, if needed.

The initial set of metabolic reconstructions from the original RAST annotations contained 1011 *Brucella*-universal-reactions. The second set of reconstructions from the manually curated annotations (step 4) contained 1016, of which 20 were found to be new core-reactions and 15 were removed from the set due to annotation errors. Finally, the third set, after using the metabolic reconstructions to guide the annotation cleanup, contained 1047 *Brucella*-universal-reactions, of which 31 previously unrecognized core reactions were found.

Annotation Improvements

As a way to eliminate sequencing, annotation and modeling errors from true strain-specific differences, we manually examined the 86 non-universal reactions from the second set of metabolic reconstructions. This revealed problems with the automated assertion or omission of reactions in certain genomes (Table S3 of the online Supplementary material). We verified the absence of 39 reactions from the set of genomes and identified 31 cases of *Brucella*-universal-reactions that had not been identified in the first round of metabolic reconstruction. The leading cause for the omission of reactions was insufficient sequencing quality (e.g., frame shifts, incomplete ORFs at the end of contigs or stretches of low quality sequence) that resulted in gene calling errors. We also found 16 annotation errors (outdated functional roles), errors in the reaction database (labeled as "functional role ambiguities" in Table S3) and one gene fusion.

More importantly, this process resulted in the identification of five unique non-universal reactions in the *Brucella inopinata* BO1 and *Brucella inopinata*-like BO2 strains. Those reactions are involved in rhamnose-containing glycan synthesis and confirm the findings for those strains reported in [78]. Additionally we proposed candidate proteins in all *Brucella* for the N-acetyl-L,L-diaminopimelate

deacetylase, the missing step in the diaminopimelate pathway (DAP) of leucine biosynthesis. All *Brucella* non-universal reactions for each genome are provided in Tables S4 and S5.

2.5 CONCLUSIONS

The studies performed within this chapter resulted in the analysis of huge amounts of data. It was the first time that metabolic models were reconstructed and analyzed for such a large number of prokaryotic genomes. The analysis of reactions revealed that even with the improved gap filling methods, there are still a large number of reactions that remain inactive. This fact is probably due to lack of components on the biomass reactions or to errors in annotations. Another interesting fact was how the number of essential genes did not vary even with genomes with various sizes. This reveals that gene essentiality is likely not dependent on the genome size. The study of the functional roles across the 3000 models also showed interesting results.

One of the most relevant was to see how diverse the models are, since the majority of functional roles are only present in about 20% of the models. The closer look at the gap filled reactions in subsystems unveiled the most gap filled subsystems. This can help to correct errors and to improve the genome annotations. When looking at the fraction of subsystems that required gap filling, we reached two conclusions: many subsystems may have at least one gap filled role, but at the same time fewer than 5% of roles were gap filled. We showed that this issue was due to the fact that several subsystems have a small number of roles. Gap filling is a computational attempt to fill gaps in the biological knowledge, and we showed the importance of properly analyzing the gap filling solution to obtain biological meaningful results.

This study aimed to show the importance of the development of high throughput tools for model reconstruction and how models can be tools to refine and curate genome annotation. To aid this effort, we developed measures of confidence in genome annotations. These measures allowed us to identify and fix multiple errors in SEED annotations. Similarly, we assessed the confidence in biomass compounds leading to the discovery of errors associated with Calcium transport in the Model SEED reaction database. We were also able to verify the use of a template biomass reaction as one of the caveats of automated model reconstruction, as we observed compounds being wrongfully assigned as universal biomass compounds.

In this chapter, we also described a workflow for improving the annotations of a genus utilizing metabolic reconstructions as a measure of annotation consistency. This has resulted in the production of an accurate and consistent collection of annotations and initial estimates of the metabolic network for the genus *Brucella*. By manual curation of 398 protein families (used in metabolic models) whose members had inconsistent annotations for isofunctional homologs, we have lowered the percentage of inconsistently annotated pairs of genes from 0.6% to 0.3%. Those improvements have led to changes in the metabolic reconstructions, generating a larger set of *Brucella*-universal reactions and highlighting the real metabolic differences between organisms. We believe that knowledge of the real differences will be of importance when deciding on sets of "representative models" to portrait the entire genus. The "representative models" will aid in the research of less studied or newly-sequenced strains.

With this work, we have demonstrated that the use of a controlled vocabulary for the annotation of genomes is key for the construction of reaction networks and future predictive comparative models. The automated annotations provided by the RAST system and the SEED's controlled vocabulary provide a good start, but annotation inconsistencies caused by sequencing and propagation errors have to be manually processed. The methods devised in different studies of this chapter reduce the workload of researchers who are trying to build models, but also clearly exposed bottlenecks where future computational tools must be built that can meet and exceed the skill level of an expert human annotator. This work has improved the annotations in the SEED and RAST and the reaction databases in Model SEED by flagging ambiguities in current functional roles. It has also improved the *Brucella*-specific collections of protein families that are propagated to RAST and PATRIC, the PathoSystems Resource Integration Center, which is dedicated to enabling bioinformatics research for bacterial pathogens and has particularly strong ties to the *Brucella* research community.

With this proof of concept, we plan to use these methodologies to improve annotations of other conserved genera, as well the entire PubSEED.

2.6 REFERENCES

1. Terzer, M., et al., *Genome-scale metabolic networks.* Wiley Interdiscip Rev Syst Biol Med, 2009. **1**(3): p. 285-97.

2. Reed, J.L., et al., *Towards multidimensional genome annotation.* Nat Rev Genet, 2006. **7**(2): p. 130-41.

3. Reed, J.L., et al., *Systems approach to refining genome annotation.* Proc Natl Acad Sci U S A, 2006. **103**(46): p. 17480-4.

4. Bro, C., et al., *In silico aided metabolic engineering of Saccharomyces cerevisiae for improved bioethanol production.* Metab Eng, 2006. **8**(2): p. 102-11.

5. Ng, C.Y., et al., *Production of 2,3-butanediol in Saccharomyces cerevisiae by in silico aided metabolic engineering.* Microb Cell Fact, 2012. **11**: p. 68.

6. Otero, J.M., G. Panagiotou, and L. Olsson, *Fueling industrial biotechnology growth with bioethanol.* Adv Biochem Eng Biotechnol, 2007. **108**: p. 1-40.

7. Singh, J.S., et al., *Genetically engineered bacteria: an emerging tool for environmental remediation and future research perspectives.* Gene, 2011. **480**(1-2): p. 1-9.

8. Chavali, A.K., et al., *A metabolic network approach for the identification and prioritization of antimicrobial drug targets.* Trends Microbiol, 2012. **20**(3): p. 113-23.

9. Kim, H.U., et al., *Integrative genome-scale metabolic analysis of Vibrio vulnificus for drug targeting and discovery.* Mol Syst Biol, 2011. **7**: p. 460.

10. Pal, C., et al., *Chance and necessity in the evolution of minimal metabolic networks.* Nature, 2006. **440**(7084): p. 667-70.

11. Henry, C.S., et al., *High-throughput generation, optimization and analysis of genome-scale metabolic models.* Nat Biotechnol, 2010. **28**(9): p. 977-82.

12. Thiele, I. and B.O. Palsson, *A protocol for generating a high-quality genome-scale metabolic reconstruction.* Nat Protoc, 2010. **5**(1): p. 93-121.

13. Stephanopoulos, G., *Metabolic engineering.* Biotechnol Bioeng, 1998. **58**(2-3): p. 119-20.

14. Overbeek, R., et al., *The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).* Nucleic Acids Res, 2014. **42**(Database issue): p. D206-14.

15. Faria, J.P., et al., *Enabling comparative modeling of closely related genomes: example genus Brucella.* 3 Biotech, 2014: p. 1-5.

16. Fleischmann, R.D., et al., *Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.* Science, 1995. **269**(5223): p. 496-512.

17. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.

18. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Res, 2007. **35**(Database issue): p. D61-5.

19. Tatusova, T., et al., *RefSeq microbial genomes database: new representation and annotation strategy.* Nucleic Acids Res, 2014. **42**(Database issue): p. D553-9.

20. Angiuoli, S.V., et al., *Toward an online repository of Standard Operating Procedures (SOPs) for (meta)genomic annotation.* OMICS, 2008. **12**(2): p. 137-41.

21. Richardson, E.J. and M. Watson, *The automatic annotation of bacterial genomes.* Brief Bioinform, 2013. **14**(1): p. 1-12.

22. Delcher, A.L., et al., *Improved microbial gene identification with GLIMMER.* Nucleic Acids Res, 1999. **27**(23): p. 4636-41.

23. Besemer, J. and M. Borodovsky, *GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W451-4.

24. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

25. UniProt, C., *The Universal Protein Resource (UniProt) in 2010.* Nucleic Acids Res, 2010. **38**(Database issue): p. D142-8.

26. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.* Nucleic Acids Res, 2003. **31**(1): p. 365-70.

27. Overbeek, R., et al., *The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.* Nucleic Acids Res, 2005. **33**(17): p. 5691-702.

28. Karp, P.D., S. Paley, and P. Romero, *The Pathway Tools software.* Bioinformatics, 2002. **18 Suppl 1**: p. S225-32.

29. Feist, A.M. and B.O. Palsson, *The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli.* Nat Biotechnol, 2008. **26**(6): p. 659-67.

30. Edwards, J.S. and B.O. Palsson, *The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities.* Proc Natl Acad Sci U S A, 2000. **97**(10): p. 5528-33.

31. Llaneras, F. and J. Pico, *Stoichiometric modelling of cell metabolism.* J Biosci Bioeng, 2008. **105**(1): p. 1-11.

32. Kuepfer, L., *Stoichiometric modelling of microbial metabolism.* Methods Mol Biol, 2014. **1191**: p. 3-18.

33. Price, N.D., et al., *Genome-scale microbial in silico models: the constraints-based approach.* Trends Biotechnol, 2003. **21**(4): p. 162-9.

34. Reed, J.L., et al., *An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR).* Genome Biol, 2003. **4**(9): p. R54.

35. Patil, K.R., M. Akesson, and J. Nielsen, *Use of genome-scale microbial models for metabolic engineering.* Curr Opin Biotechnol, 2004. **15**(1): p. 64-9.

36. Burgard, A.P., P. Pharkya, and C.D. Maranas, *Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.* Biotechnol Bioeng, 2003. **84**(6): p. 647-57.

37. Pharkya, P., A.P. Burgard, and C.D. Maranas, *Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock.* Biotechnol Bioeng, 2003. **84**(7): p. 887-99.

38. Feist, A.M. and B.O. Palsson, *The biomass objective function.* Curr Opin Microbiol, 2010. **13**(3): p. 344-9.

39. Varma, A. and B.O. Palsson, *Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use.* Bio/technology, 1994. **12**.

40. Feist, A.M., et al., *Reconstruction of biochemical networks in microorganisms.* Nat Rev Microbiol, 2009. **7**(2): p. 129-43.

41. Tervo, C.J. and J.L. Reed, *BioMog: a computational framework for the de novo generation or modification of essential biomass components.* PLoS One, 2013. **8**(12): p. e81322.

42. Mahadevan, R. and C.H. Schilling, *The effects of alternate optimal solutions in constraint-based genome-scale metabolic models.* Metab Eng, 2003. **5**(4): p. 264-76.

43. Garcia-Albornoz, M.A. and J. Nielsen, *Application of genome-scale metabolic models in metabolic engineering.* Industrial Biotechnology, 2013. **9**(4): p. 203-214.

44. Arakawa, K., et al., *GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes.* BMC Bioinformatics, 2006. **7**: p. 168.

45. Notebaart, R.A., et al., *Accelerating the reconstruction of genome-scale metabolic networks.* BMC Bioinformatics, 2006. **7**: p. 296.

46. Ogata, H., et al., *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Res, 1999. **27**(1): p. 29-34.

47. Hucka, M., et al., *The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.* Bioinformatics, 2003. **19**(4): p. 524-31.

48. Dias, O., et al. *Merlin: Metabolic Models Reconstruction Using Genome-Scale Information*. in *Computer Applications in Biotechnology*. 2010.

49. Satish Kumar, V., M.S. Dasika, and C.D. Maranas, *Optimization based automated curation of metabolic reconstructions.* BMC Bioinformatics, 2007. **8**: p. 212.

50. Karp, P.D., et al., *The EcoCyc Database.* Nucleic Acids Res, 2002. **30**(1): p. 56-8.

51. Karp, P.D., et al., *Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology.* Brief Bioinform, 2010. **11**(1): p. 40-79.

52. Agren, R., et al., *The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum.* PLoS Comput Biol, 2013. **9**(3): p. e1002980.

53. Swainston, N., et al., *The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks.* J Integr Bioinform, 2011. **8**(2): p. 186.

54. Aziz, R.K., et al., *The RAST Server: rapid annotations using subsystems technology.* BMC Genomics, 2008. **9**: p. 75.

55. Caspi, R., et al., *The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.* Nucleic Acids Res, 2008. **36**(Database issue): p. D623-31.

56. Hamilton, J.J. and J.L. Reed, *Software platforms to facilitate reconstructing genome-scale metabolic networks.* Environ Microbiol, 2014. **16**(1): p. 49-59.

57. Saier, M.H., Jr., C.V. Tran, and R.D. Barabote, *TCDB: the Transporter Classification Database for membrane transport protein analyses and information.* Nucleic Acids Res, 2006. **34**(Database issue): p. D181-6.

58. Demir, E., et al., *The BioPAX community standard for pathway data sharing.* Nat Biotechnol, 2010. **28**(9): p. 935-42.

59. Funahashi, A., et al., *CellDesigner 3.5: a versatile modeling tool for biochemical networks.* Proceedings of the IEEE, 2008. **96**(8): p. 1254-1265.

60. Schellenberger, J., et al., *Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0.* Nat Protoc, 2011. **6**(9): p. 1290-307.

61. Rocha, I., et al., *OptFlux: an open-source software platform for in silico metabolic engineering.* BMC Syst Biol, 2010. **4**: p. 45.

62. Brooks, J.P., et al., *Gap detection for genome-scale constraint-based models.* Adv Bioinformatics, 2012. **2012**: p. 323472.

63. Rolfsson, O., B.O. Palsson, and I. Thiele, *The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions.* BMC Syst Biol, 2011. **5**: p. 155.

64. Thiele, I., N. Vlassis, and R.M. Fleming, *fastGapFill: efficient gap filling in metabolic networks.* Bioinformatics, 2014. **30**(17): p. 2529-31.

65. Vlassis, N., M.P. Pacheco, and T. Sauter, *Fast reconstruction of compact context-specific metabolic network models.* PLoS Comput Biol, 2014. **10**(1): p. e1003424.

66. Latendresse, M., *Efficiently gap-filling reaction networks.* BMC Bioinformatics, 2014. **15**: p. 225.

67. Latendresse, M., et al., *Construction and completion of flux balance models from pathway databases.* Bioinformatics, 2012. **28**(3): p. 388-96.

68. Kumar, V.S. and C.D. Maranas, *GrowMatch: an automated method for reconciling in silico/in vivo growth predictions.* PLoS Comput Biol, 2009. **5**(3): p. e1000308.

69. Henry, C.S., et al., *iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations.* Genome Biol, 2009. **10**(6): p. R69.

70. Lee, S., et al., *Recursive MILP model for finding all the alternate optima in LP models for metabolic networks.* Computers & Chemical Engineering, 2000. **24**(2): p. 711-716.

71. Konstantinidis, K.T. and J.M. Tiedje, *Trends between gene content and genome size in prokaryotic species with larger genomes.* Proc Natl Acad Sci U S A, 2004. **101**(9): p. 3160-5.

72. Plata, G., C.S. Henry, and D. Vitkup, *Long-term phenotypic evolution of bacteria.* Nature, 2014. **advance online publication**.

73. Miyazaki, C., et al., *Genetic manipulation of membrane phospholipid composition in Escherichia coli: pgsA mutants defective in phosphatidylglycerol synthesis.* Proc Natl Acad Sci U S A, 1985. **82**(22): p. 7530-4.

74. Nishijima, S., et al., *Disruption of the Escherichia coli cls gene responsible for cardiolipin synthesis.* J Bacteriol, 1988. **170**(2): p. 775-80.

75. Arias-Cartin, R., et al., *Cardiolipin binding in bacterial respiratory complexes: structural and functional implications.* Biochim Biophys Acta, 2012. **1817**(10): p. 1937-49.

76. Shemarova, I.V. and V.P. Nesterov, *[Evolution of mechanisms of Calcium signaling: the role of Calcium ions in signal transduction in prokaryotes].* Zh Evol Biokhim Fiziol, 2005. **41**(1): p. 12-7.

77. Williams, R.J., *The evolution of calcium biochemistry.* Biochim Biophys Acta, 2006. **1763**(11): p. 1139-46.

78. Wattam, A.R., et al., *Comparative genomics of early-diverging Brucella strains reveals a novel lipopolysaccharide biosynthesis pathway.* MBio, 2012. **3**(5): p. e00246-11.

79. Gillespie, J.J., et al., *PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species.* Infect Immun, 2011. **79**(11): p. 4286-98.

80. Wattam, A.R., et al., *PATRIC, the bacterial bioinformatics database and analysis resource.* Nucleic Acids Res, 2014. **42**(Database issue): p. D581-91.

81. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.

82. Davis, J.J. and G.J. Olsen, *Characterizing the native codon usages of a genome: an axis projection approach.* Mol Biol Evol, 2011. **28**(1): p. 211-21.

83. Overbeek, R., et al., *Use of contiguity on the chromosome to predict functional coupling.* In Silico Biol, 1999. **1**(2): p. 93-108.

84. Overbeek, R., et al., *The use of gene clusters to infer functional coupling.* Proc Natl Acad Sci U S A, 1999. **96**(6): p. 2896-901.

85. Li, L., C.J. Stoeckert, Jr., and D.S. Roos, *OrthoMCL: identification of ortholog groups for eukaryotic genomes.* Genome Res, 2003. **13**(9): p. 2178-89.

86. Markowitz, V.M., et al., *IMG: the Integrated Microbial Genomes database and comparative analysis system.* Nucleic Acids Res, 2012. **40**(Database issue): p. D115-22.

2.7 SUPPLEMENTARY MATERIAL

The supplementary material is available online at http://darwin.di.uminho.pt/jplfaria/phdthesis/Chapter_2_SupplMaterial.zip.

S2.1 Analysis of functional roles in the 3000 models (Figure 2.8).

S2.2 Analysis of gap filled function roles in the 3000 models (Figures 2.9-11).

S2.3 Distribution of values for the cost of annotated reaction for the 3000 genome-scale metabolic models (Figure 2.12).

S2.4 Distribution of the value of gap filled reaction for 3000 genome-scale metabolic models (Figure 2.13).

Additional supplementary material is available online via the PATRIC website at: http://enews.patricbrc.org/annotation_protocol_brucella/

Table S1 -Initial Set of *Brucella*-universal-reactions

Table S2 - Improved set of *Brucella*-universal-reactions

Table S3 - Non-universal reactions analysis

Table S4 - Non-universal reactions per organism

Table S5 - Non-universal reactions functional roles

CHAPTER 3

ANALYSIS OF THE *BACILLUS SUBTILIS* REGULATORY NETWORK

Work presented in this chapter comprises the following articles:

Faria, J. P., Overbeek, R., Xia, F., Rocha, M., Rocha, I., & Henry, C. S.

Genome-scale bacterial transcriptional regulatory networks: reconstruction and integrated analysis with metabolic models. Briefings in Bioinformatics. doi: 10.1093/bib/bbs071

2014

Faria, J. P., Overbeek, R., Taylor, R. C., Goelzer, A.,Fromion, V.,Rocha, M.,Rocha, I., &,Henry, C. S.

Reconciling gene expression data with regulatory network models –a stimulon-based approach for regulatory modeling of *Bacillus subtilis*.

Abstract accepted for full manuscript submission, Frontiers in Systems Microbiology.

2015

ABSTRACT

Advances in sequencing technology are resulting in the rapid emergence of large numbers of complete genome sequences. High-throughput annotation of these genomes and metabolic modeling of the corresponding organisms is now a reality. The high-throughput reconstruction and analysis of genome-scale transcriptional regulatory networks represents the next frontier in microbial bioinformatics. The fruition of this next frontier will depend on the integration of numerous data sources relating to mechanisms, components, and behavior of the transcriptional regulatory machinery, as well as the integration of the regulatory machinery into genome-scale cellular models. In this chapter, we review existing repositories for different types of transcriptional regulatory data, including expression data, transcription factor data, and binding site locations; and we explore how these data are being used for the reconstruction of new regulatory networks. From template network-based methods to *de novo* reverse engineering from expression data, we discuss how regulatory networks can be reconstructed and integrated with metabolic models to improve model predictions and performance.

We then introduce a manually curated regulatory network for *Bacillus subtilis*, tapping into the notable resources for *B. subtilis* regulation. We propose the concept of *Atomic Regulon*, as a set of genes that share the same "ON" and "OFF" gene expression profile across multiple samples of experimental data. Atomic regulon inference uses prior knowledge from curated SEED subsystems, in addition to expression data to infer regulatory interactions. We show how atomic regulons for *B. subtilis* are able to capture many sets of genes corresponding to regulated operons in our manually curated network. Additionally, we demonstrate how atomic regulons can be used to help expand/validate the knowledge of the regulatory networks and gain insights into novel biology.

3.1 INTRODUCTION

As a model organism, literature for the bacterium *Bacillus subtilis* regulation is extensive and several resources/databases are available. A regulatory network model for the central carbon metabolism was made available by Goelzer *et al*. in 2008 [1]. Multiple inferred networks based on expression data have also been proposed in the literature [2, 3]. RegPrecise [4], a resource for transcription factor binding site (TFBS) based network inference also provides a network for *B. subtilis* [5]. Subtiwiki [6, 7] is a community collaborative resource for *B. subtilis* that includes a vast compendium of regulatory information. DBTBS [8] is another *B. subtilis* comprehensive resource of regulatory data with promoters, transcription factors (TFs), TFBS, motifs and regulated operons. Our novel genome-scale reconstruction of the *B. subtilis* regulatory network integrates the previous work from the Goelzer *et al*, literature and the other notable resources for regulation described above [4, 7-9].

We reconciled our new model against a large set of high-quality gene expression data [10, 11]. For the process of reconciliation with expression data, we introduce the concept of *Atomic Regulons*. Atomic regulons are sets of co-regulated genes that share the same "ON" and "OFF" expression profile (meaning the genes in these sets are "ON" and "OFF" in the same conditions). While predicting traditional operon structures can be a difficult task [12], our approach for computing atomic regulons from expression data is fairly easy. Our approach begins by constructing draft regulons using a combination of crude operon predictions and SEED subsystem technology [13-15]. We then decompose and expand these draft regulons based on consistency with expression data. This process results in a set of co-regulated gene clusters, now called Atomic Regulons. We show how atomic regulons can be used to help expand/validate the knowledge of the regulatory network and gain insights into novel biology.

3.2 STATE OF THE ART

3.2.1 Introduction

Systems biology has provided numerous tools for modeling biological systems [16], many of which depend on the reconstruction of genome-scale metabolic models (GEM). These models now exist for a growing number of organisms, including prokaryotic, archaeal, and eukaryotic species [17]. With the advent of next-generation sequencing, the development of GEMs has become routine [17, 18], and many steps involved in the reconstruction and optimization of draft GEMs have been automated [19]. Algorithms and methods for GEM reconstruction have been reviewed in detail elsewhere [20-22], and in the previous chapter.

However, nearly all-existing GEMs fail to account for the impact of gene expression regulation on metabolic activity. In order to capture the impact of regulation on the behavior of an organism, a GEM must integrate some abstraction of regulatory mechanisms, which include the activity of RNA polymerase, transcription factors (TFs), promoters, transcription factor binding sites (TFBS), and sigma factors. Sigma factors allow the recognition of the enzyme by the promoter region, enabling transcription to begin. TFs bind to specific TFBSs in the promoter region and can act as activators, repressors, or both (dual regulators). In eukaryotes, TFs are able to perform other tasks affecting regulation, such as chromatin-modifying activities [23]. Other elements have been identified as taking part in the control of transcription regulation in bacteria, such as riboswitches [24], RNA switches [25], antisense RNA [26], or microRNAs [27]. Here, we focus on regulation by transcription factors, a mechanism illustrated in Figure 3.1. Also displayed are some of the technologies, tools, and resources necessary for reconstructing transcriptional regulatory networks.

The integration of these regulatory mechanisms in GEMs requires methods for the reconstruction and analysis of transcriptional regulatory networks (TRNs). Once a regulatory model has been constructed for an organism, it can be integrated with GEMs to improve predictive accuracy and reveal new biological insights. For example, some cellular processes exhibit a dominance of regulatory mechanisms, affecting their behavior and leading to incorrect predictions when only metabolism is

accounted for [28]. The first genome-scale integrated metabolic and regulatory model for *E. coli* [29] revealed that regulation significantly affects growth phenotype predictions, and these predictions improved with the addition of regulatory constraints. Simultaneously, the study of TRNs has unveiled novel interactions; in *Salmonella enterica,* 14 regulators were identified that affect the same genes leading to a systemic infection [30]. Similar studies led to the discovery of novel regulatory mechanisms in *Saccharomyces cerevisiae* [31].

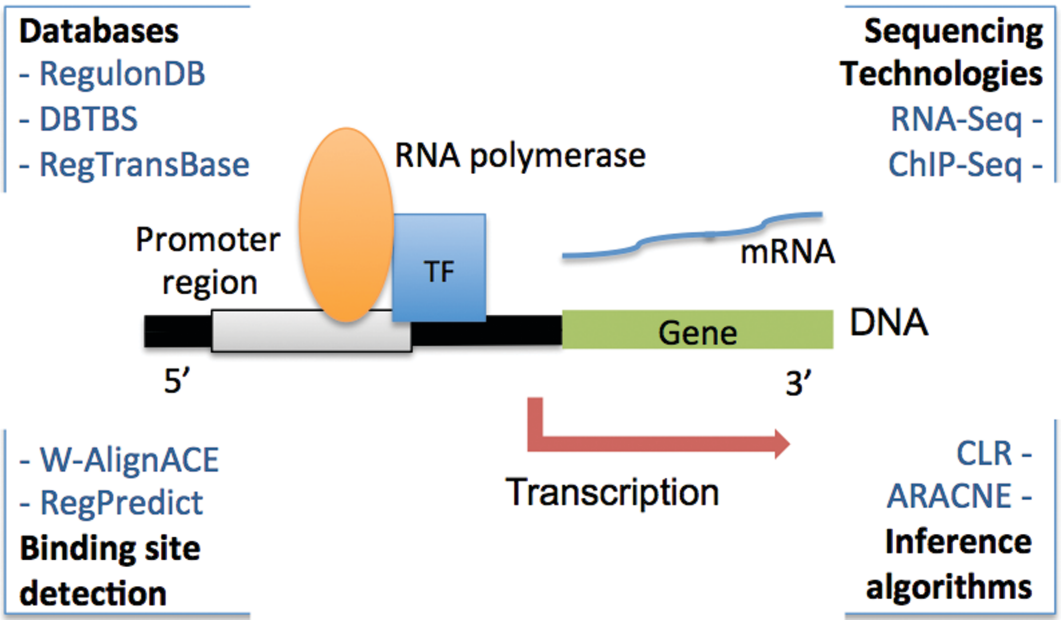


Figure 3.1 Technologies, tools, and resources for transcriptional regulatory network modeling and reconstruction.

In this chapter, we review the reconstruction of TRNs, by firstly exploring the data available for TRN reconstruction, covering the most prominent databases of expression data and repositories of TF/TFBS data. Next, we examine how data availability triggered the development of a variety of TRN inference methods, including reverse engineering from expression datasets [32-36], network inference from TFBS site data [37-39], and knowledge-based template methods [40].

3.2.2 Regulation data for TRN reconstruction – From standards and technologies to databases

The development of microarray technologies gave rise to a revolution in biomedical research [41], also bringing new problems such as quality control of experiments [42] and selection of an appropriate level of detail [43]. To address these issues, the Functional Genomics Data Society (FGED) launched a proposal to standardize the publishing and sharing of microarray data (MIAME) [44]. The majority of the community adopted the proposal, requiring authors to follow the MIAME guidelines. Publishers also required authors to store data [45] in either NCBI's Gene Expression Omnibus (GEO) [46] or EBI's ArrayExpress [47], the major public gene expression data repositories, both MIAME compliant.

These databases integrate data from a variety of technologies that can help determine regulatory interactions, although expression profiling and genome binding and occupancy studies have become the most prevalent. Expression profiling techniques vary from the traditional array oligonucleotide hybridization technology for measuring gene expression level to mRNA quantification methodologies, such as serial analysis of gene expression (SAGE) [48, 49] or reverse transcriptase PCR (RT-PCR). Genome binding and occupancy experiments have the advantage of identifying the spots corresponding to DNA-protein binding targets. Chromatin immunoprecipitation with array hybridization (ChIP-chip) [50, 51] is used to overcome limitations of common expression profiling. Other ChIP technologies have also been developed in combination with different expression techniques, such as SAGE (ChIP-SAGE [52]), to achieve a particular level of detail, depending on the organism and tissue studied [53]. With the development of next-generation sequencing technologies, ChIP-Seq [54] and RNA-Seq emerged [55, 56]. ChIP-Seq enables whole-genome ChIP assays, while RNA-Seq provides a capacity for direct measurement of mRNA, small RNA, and noncoding RNA abundances [57]. ChIP methods have been widely used to collect expression data from *E. coli* [58-60]; and, more recently, RNA-Seq methods have been adjusted for studying bacterial transcriptomes [61, 62]. RNA-Seq has been also successfully used to detect transcription start sites [63] that can be used for regulon inference.

Data available for TRN inference can be categorized into two major groups: (i) databases of gene expression data (including genome binding experimental data), and (ii) databases of TF and TFBS. Table 3.1 shows the most notable databases of the former group.

Table 3.1 Gene expression repositories with bacterial transcriptional data.

Database	Main Features
GEO [46]	NCBI's database for expression data. Supports multiple expression studies platforms for all organisms. Browsing tools available.
ArrayExpress [47]	EBI's database for expression data. Data submitted by users and imported from GEO. Advanced queries and ontology-driven searches.
M3D [64]	Data uniformly normalized from Affymetrix microarrays for *Escherichia coli, Saccharomyces cerevisiae* and *Shewanella oneidensis*.
SMD [65]	Partially public database with data from around 60 organisms. *Escherichia coli, Mycobacterium tuberculosis* and *Streptomyces coelicor* are among the most represented microbes. Data analysis framework embedded.
COLOMBOS [66]	Cross-platform expression compendia for *E. coli*, *B. subtilis*, and *S. enterica* subspecies serovar Typhimurium. Provides tools for expression analysis and extraction of relevant information.

We surveyed GEO as the major expression database, gathering statistics on the type of studies conducted, availability of data, quantification of bacterial data, and the most represented microbes (Figure 3.2). These statistics clearly indicate that most of the current data are from expression profiling, with 18,498 experimental series (85%). Although next-generation sequencing technologies were introduced recently [67], we can already see a change in the types of experiments being performed (Figure 3.2 b). Examining the organisms for which expression data are available, we find that only 7% of datasets are from bacteria (Figure 3.2 c), with *Escherichia coli* being the most represented prokaryote (Figure 3.2 d).

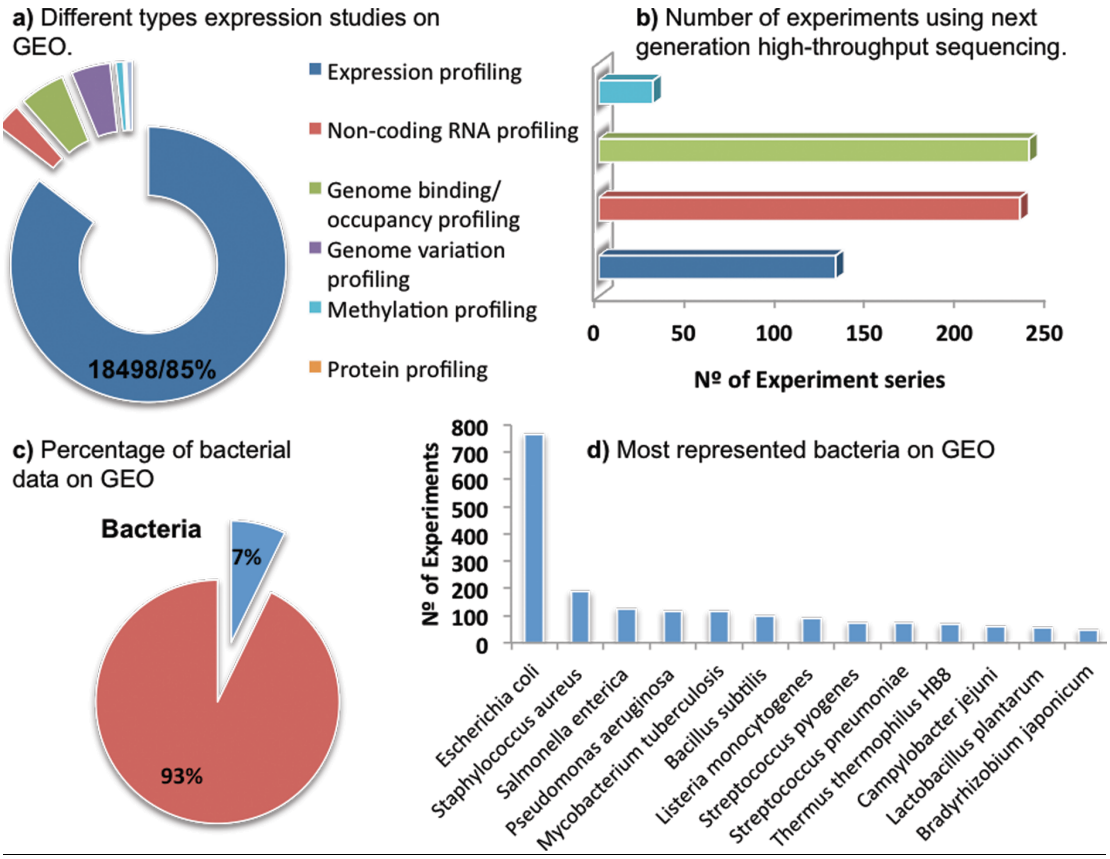


Figure 3.2 Survey of the GEO database. (a) Types of expression studies on the database [68]. (b) Number of series of experiments available from next-generation sequencing technologies [68]. (c) Percentage of data from bacteria in the entire database: from a total of 28,150 series of experiments only 2,196 represent bacterial organisms. (d) Most represented bacteria on GEO. The organisms presented have at least a minimum of 43 series of experiments. Data for (c) and (d) were obtained with GEO tools [69] in April 2012.

Table 3.1 also includes other notable databases, from which we highlight the Many Microbe Microarrays Database (M3D) [64] currently holding around 2,000 microarrays for *Escherichia coli*, *Saccharomyces cerevisiae,* and *Shewanella oneidensis*. The data available are all from Affymetrix single channel microarrays, allowing a uniform normalization procedure and higher-quality data. The *E. coli* data have already been applied for TRN inference [70].

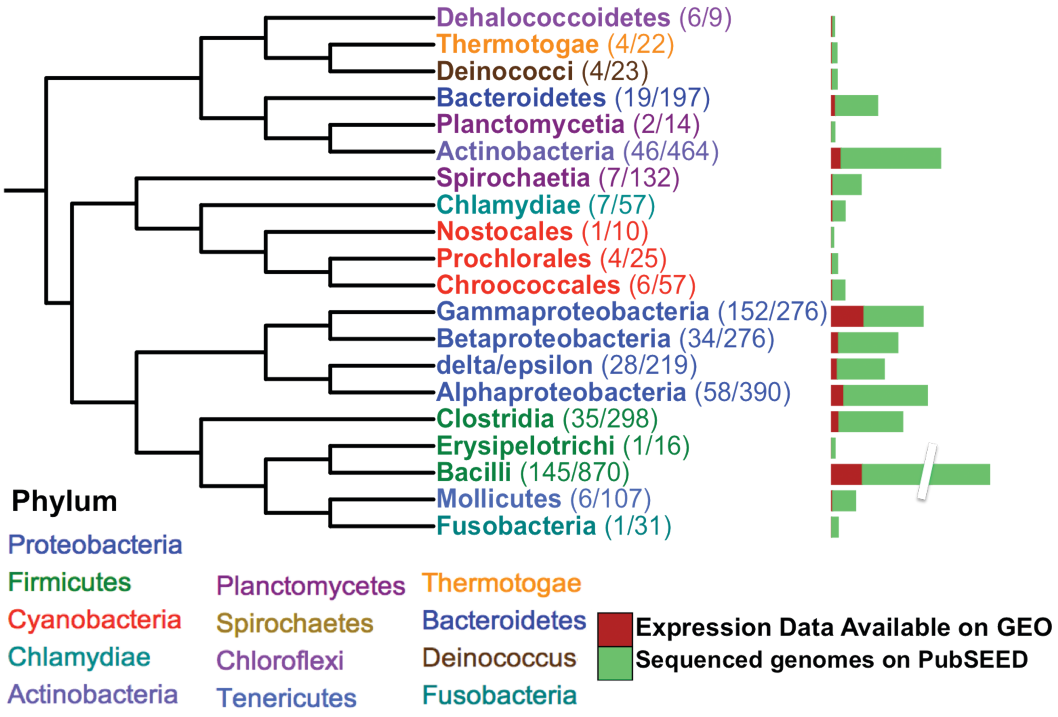


Figure 3.3 Comparison of bacterial genomes with expression data in GEO versus genomes with complete DNA sequences in the PubSEED [13]. The 20 bacterial families that contain genomes with expression data in GEO are arranged in a topological tree. For each family, the most abundantly sampled species in the PubSEED was picked to represent that family, and the alignment of their 16S sequences was used to reconstruct the bacterial family tree. The color-coding of the tree nodes denotes the phyla they belong to. Most phyla contain only one family, with the exception of Cyanobacteria (3 families), Bacteroidetes (4 families), and Firmicutes (3 families). The last two phyla are especially overrepresented in terms of both sequenced genomes and expression data. The numbers on the right of each tree node denote the number of genomes with GEO expression data (566 in total) and the number of genomes present in the PubSEED (3,493 in total). Archaea organisms were removed from this study since we aim to survey only bacterial genomes. In the horizontal bar plot, we show the fraction of each bacterial family for which expression data is available (in dark red). The tree was designed with the Interactive Tree of Life Tool [71, 72].

Figure 3.3 shows the discrepancy between the number of sequenced genomes and the number of genomes for which any type of expression data exists. In this study, we cluster bacterial genomes available in the PubSEED [13] (a large repository of genomes and annotations) at the taxonomical level of family. The set of 20 bacterial families associated with expression data in GEO are shown in the

phylogenetic tree. On average, 16.2% of the 3,493 PubSEED genomes that fall into these families have expression data linked to them. Expression data are available for 55% of the genomes in the Gammaproteobacteria family, demonstrating the extensive amount of data available for this taxonomic clade. In contrast, more than half of the bacterial phyla have expression data for less than 10% of their species, revealing that numerous phylogenetically distinct clusters of microbes have little gene expression experimentally characterized.

Repositories with regulatory interactions also hold valuable information. Table 3.2 shows the most comprehensive resources available for prokaryotes. Organism-specific databases are available for well-known organisms such as *E. coli*, *B. subtilis,* and *M. tuberculosis*, including a comprehensive collection of regulatory information. Among those, RegulonDB is the most comprehensive resource for regulatory interactions data of any single organism (*E. coli*). In its latest release, genetic sensory response units are introduced to better represent the biology of gene regulation [73], trying to capture all the phenomena involved in regulation, from the initial signal to gene response. Another major resource for *E. coli* data is EcoCyc [74], integrating RegulonDB and curated data from over 21,000 publications and TRN descriptions that include genes, ligands, and regulators with their targets. DBTBS [8] is the major resource for *B. subtilis* regulatory data.

Less comprehensive databases present fewer types of different regulatory information (sometimes only TFBS predictions or TF information) but cover a wide range of bacteria (Table 3.2). Notable examples are ODB [75], which stores known operon data for about 10,000 operons in 56 organisms and putative operons for over 1000 genomes; RegTransBase [76], which collects regulatory data from the literature; and RegPrecise [4], a repository of manually curated regulons that provides tools for regulon propagation.

Reconstruction of TRNs can use different types of data, and the accurate selection of data/database(s) for the method of choice is paramount in the reconstruction process. Organism-specific databases are particularly useful for reverse engineering methodologies as training datasets and essential for validation. Methodologies based on comparative genomics approaches make good use of less comprehensive databases but cover a wider range of organisms.

Table 3.2 Databases with notable bacterial transcriptional data.

Database	Organism(s)	Main Features
Organism specific		
DBTBS [8]	*B. subtilis*	Compendium of regulatory data with promoters, TFs, TFBS, motifs and regulated operons
RegulonDB [73]	*E. coli*	Compendium of regulatory data, promoters, TFs, TFBS, transcription units, operons and regulatory network interactions.
EcoCyc [74]	*E. coli*	Comprehensive database with gene products, transcriptional, post-transcriptional data and operon organization
DPInteract [77]	*E. coli*	DNA binding proteins and binding site data.
MTBRegList [78]	*M. tuberculosis*.	TFBS and regulatory motifs
Organism class/family		
CoryneRegNet [79]	Corynebacteria	TF and regulatory networks
cTFbase [80]	Cyanobacteria	Putative TFs
TractorDB [81]	Gamma-proteobacteria	TFBS predictions
MycoRegNet [82]	Mycobacteria	TF and regulatory networks
Non-organism specific		
ExtraTrain [83]	Bacteria and Archaea	Transcriptional data and extragenic regions
DBD [84]		TF predictions
RegTransBase [76]		Regulatory interactions from literature and TFBS
PRODORIC [85]	Bacteria	TFs, TFBSs, regulon lists, promoters, expression profiles
sRNAMap [86]		Small noncoding RNAs and regulators
ODB [75]		Known and putative operons
RegPrecise [4]		Regulon database

3.2.3 TRN Reconstruction – From template networks and inference algorithms to integration with GEMs

TRN reconstruction aims to make sense of gene expression and binding site data by revealing the interactions between the different elements of the cell's regulatory machinery. Different methodologies have been proposed for TRN inference. However, there is no consensus for classification in the literature. Some reviews classify methods as bottom-up and top-down [87], others focus on inference from a specific type of data such as gene expression [88], while others present methods and computational tools [89].

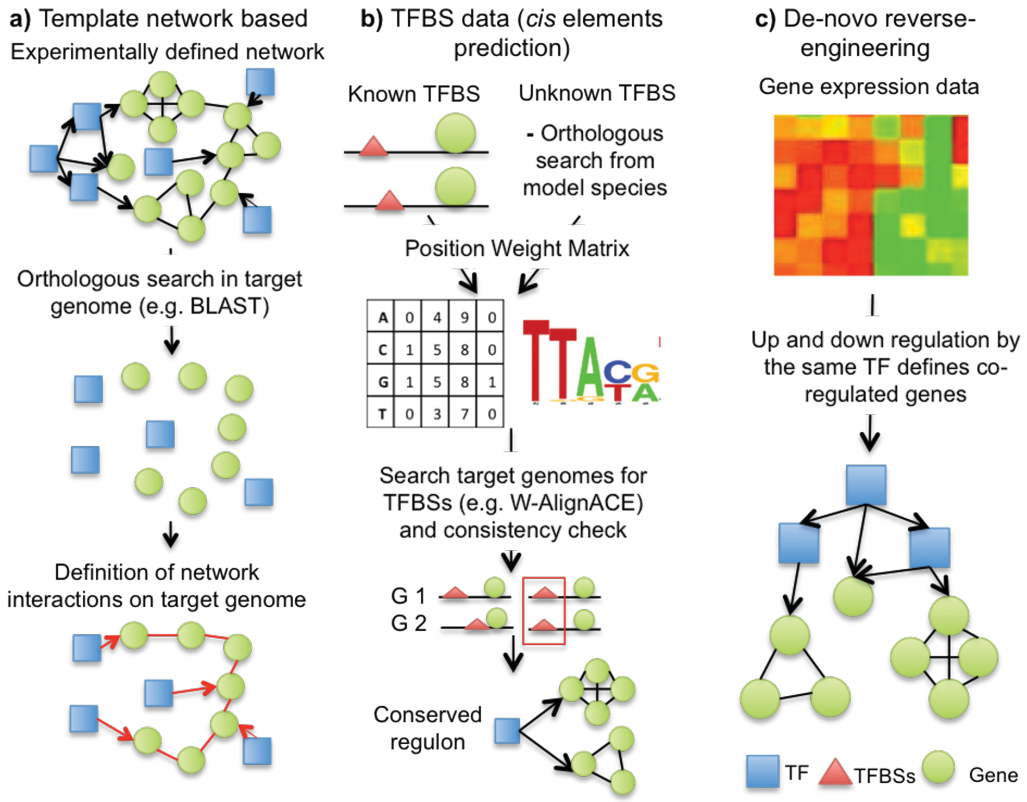


Figure 3.4 TRN reconstruction methodologies. (a) Template network based methods. b) TFBS data based via regulatory cis elements. (c) *De novo* reverse engineering.

Here, we review and categorize different methodologies within two major types: genomics-driven and data-driven. The first uses comparative genomics approaches, while the second refers to *de novo*

reverse engineering from expression data. Within the genomics-driven approaches, we describe two methodologies: template network-based methods and TFBS data-based methods via prediction of *cis*–regulatory elements, including propagation from known regulons and *ab initio* regulon inference. The comparative genomics approaches are described in Figure 3.4 a) and b); Figure 3.4 c) describes data-driven methods from expression data.

Template network-based methods

Template-based methods [90] rely on one or more well-characterized networks to serve as a starting point for the reconstruction. These methods exploit the conservation of prokaryotic gene networks [91-94] to reconstruct TRNs (Figure 3.4 a). Starting with a well-characterized network, a search for orthologous genes (e.g. using bidirectional best hits [95]) is conducted on the genome of interest. With the orthologous TFs and their targets noted on the target genome, random networks are generated from the template network to confer statistical strength to the new reconstructed interactions in the target genome, since this shows the significant trends. After this analysis, the new interactions on the target genome are reconstructed. This approach can be useful for propagation of TRNs to other strains of a model organism or to closely related organisms.

This methodology presents some limitations, however. The first is intrinsic: the need for a high-quality template network derived for an organism that is phylogenetically close to the organism being studied. A long phylogenetic distance between the template and the target organisms can generate meaningless interactions; hence, the choice of the template network is of paramount importance for the reconstruction. Another limitation is the scale of the network to be reconstructed; here, our focus is genome-scale network reconstruction, and reconstructions on this scale depend on the availability of a template network that also exists at the genome scale.

TFBSs data-based methods via prediction of *cis* – regulatory elements

TRN reconstruction from binding site data can also be defined as a comparative genomics approach. Prior to the development of the first binding-site approaches, most methods relied almost entirely on functional information from expression data [34, 96]. The GRAM (Genetic Regulatory models) algorithm

[97] was the first to combine the use of expression data and binding site data in a genome wide inference process, enabling the inclusion of information about physical interactions between regulatory genes and their targets. Other work focused on the conservation of the regulatory machinery across different organisms.

Regulogger [98] was introduced to generate *regulogs*, or sets of genes that are co-regulated and have their regulation processes conserved across several organisms. Using *Staphylococcus aureus*, regulogs were produced for well-known sets of genes and provide clues about the functions of unannotated genes. Studies of δ-proteobacteria [38] revealed that very diverse species of proteobacteria have similar regulatory mechanisms.

The principles behind this methodology were reviewed by Rodionov [99]. Figure 3.4 b) describes one of the two strategies proposed. The first step is to gather all available information related to TFs and TFBSs in a selected model organism. These data are then used as a training set for the TFBS model. The accuracy of the methodology is closely connected to the quality and quantity of sequences used for training. *E. coli* is usually used as a model species for gram-negative bacteria, and *B. subtilis* for gram-positive bacteria. If the TFBSs corresponding to a particular TF are unknown, all genes regulated by the TF in the model species are identified, and then orthologues for these genes in closely related genomes are found. With a TFBS training set built by this process or experimentally determined (see Table 3.2), positional weight matrices (PWMs) are constructed for the collection of binding sites. Several algorithms are available that perform motif pattern recognition [100] to construct PWMs. One of the first algorithms developed for this task was AlignACE [101]. This algorithm was recently upgraded to W-AlignACE [102] incorporating a new learning approach [103] and showing increased accuracy in obtaining PWMs for gene sequences, gene expression data, and ChIP-chip data [102]. Using the PWMs, one can perform a genome wide search for putative TFBSs on the target genomes.

This comparative-genomics-based approach requires a high-quality training set; using genomes that are not closely related can lead to generation of false positive TFBS predictions. Even for a set of closely related genomes, selecting a threshold for binding site detection can be difficult. The final step of the TFBS prediction involves the verification of site consistency. Early studies on *E. coli* and *H.*

influenzae regulon predictions showed conservation of co-regulated genes by orthologous TFs [104]. Based on this principle, a search is conducted for binding sites upstream from the operons regulated by each TF. If the site is conserved, the TFBS prediction is assumed to be correct. On the other hand, if matches to the predicted TFBS motif are found dispersed across the genome, the prediction is assumed to be a false positive. By accounting for changes in the operon structure, further consistency checks are possible. This method showed improved results in binding site detection in several studies such as nitrate and nitrite respiration in γ-Proteobacteria [105] and nitrogen metabolism in gram-positive bacteria [106].

These methodologies have been implemented in the RegPredict web resource [107], a state-of-the-art tool for TRN reconstruction with TFBS data. The webserver comprises a large set of comparative genomics tools available in two reconstruction frameworks; the first reconstructs regulons for known PWMs, and the second performs *de novo* regulon inference for unknown binding sites using analysis of regulon orthologues across closely related genomes. One of the novelties of RegPredict is the concept of CRONs (Clusters of co-Regulated Orthologous Operons) to facilitate and improve consistency check. This semi-automated approach provides the community with a more swift reconstruction, curation and storage of regulons. RegPredict was used for TRN reconstruction of the central metabolism of the *Shewanella genus* [108], for the analysis of the regulation of the hexunorate metabolism in Gammaproteobatceria [109], and for the elucidation of control mechanisms for proteobacterial central carbon metabolism by the HexR regulator [110]. FITBAR [111] is another web tool for prokaryotic regulon prediction that aims to fill the gap of the lack of statistical comparison for calculating the significance of the predictions.

Techniques also exist for predicting TFBSs when the available regulatory information is not sufficient for regulon-based approaches. Phylogenetic footprinting [112] identifies highly conserved untranslated regions (UTRs) upstream from the genes of interest, since these are prime regulatory site candidates. An orthologous search for these regions is performed across closely related genomes; candidate binding sites are identified; and these sites are used to perform a regulatory motif search across all analyzed genomes. This technique successfully identified the FabR regulon in *E. coli* and regulon

members in several cyanobacteria genomes [113]. Another approach has been described as subsystem oriented [99] based on the hypothesis that one TF regulates the genes on the same metabolic pathway. A search for orthologous genes on the same metabolic pathway of closely related genomes is conducted. Using the orthologous operons from the same subsystem, one can perform a motif search to build the PWM and search for TFBS. Concepts of this approach were also implemented in RegPredict with the introduction of the SEED subsystems [13] for regulon reconstruction and curation.

De novo reverse engineering

As gene expression data became available through microarray technologies, development began on methods for inference of regulatory networks from expression data [114]. Early reviews describe several mathematical formalisms such as Bayesian networks, Boolean networks, and differential equations to represent regulatory networks [115], together with appropriate algorithms to support network inference.

The development of these methodologies led to the creation of the DREAM (Dialogue for Reverse Engineering Assessments and Methods) project in 2007 [116], bringing together experts from different areas and aiming to provide tools to enable the unbiased evaluation of various methods [117], hosting annual challenges. The lessons gained from the results obtained in those challenges have provided improved methods for network inference [118]. Each year different methods are ranked as top performers on specific sub challenges that differ in either the type of data or network size.

Past reviews have categorized reverse engineering network inference methods according to (i) mathematical modeling approach [88, 119], (ii) module-based or direct inference methods [87, 120], and (iii) unsupervised and (semi)-supervised methodologies [87, 121, 122].

In the first category [88, 123], the differential equation (ODEs)-based [124, 125], mutual information-based [126, 127], and Bayesian network-based methods [128, 129] are the most popular approaches. Other notable approaches are based on Boolean networks [130], neural networks [131, 132], correlation analysis [133], and relevance networks [134].

The second category divides methods into those based on a modular view of regulatory networks that infer regulatory programs for sets of co-expressed genes and those able to infer the regulatory behavior of individual genes (direct inference) [79]. Module-based inference is inspired by evidence that regulatory networks exhibit a modular structure of co-expressed genes [135, 136], using a separate algorithm for the module inference step, typically based on clustering or biclustering algorithms, such as cMonkey [137]. Direct inference methods search for single interactions between targets and their regulators [70, 138] (Figure 3.5 a)).

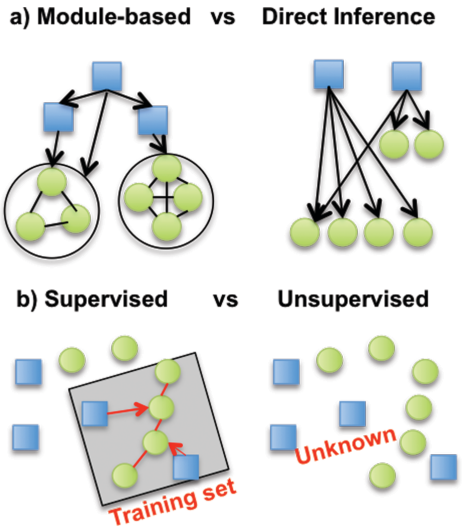


Figure 3.5 Network inference methods classification. (a) Network node Module Based vs Direct Inference. (b) Supervised vs unsupervised. Supervised methods require a training set of previous known interactions.

A comparison between representative methods of both approaches showed that none can be defined as the best solution [120]: the module-based method LeMoNe [139, 140] is able to retrieve more efficiently targets for regulators with a high number of targets, and the direct-inference method CLR [70] is preferable for detecting regulators with one or few targets. Thus, these methods can be seen as complementary when handling genome-scale regulatory model reconstruction.

The third category divides methods into supervised [141, 142] and unsupervised [143, 144] (Figure 3.5 b)). The former use a training set of known interactions creating classification problems (e.g., to

infer whether a given gene is regulated by a transcription factor) (Figure 3.5). Some supervised methods are known as semi-supervised [145, 146]. Supervised methods have shown to provide more accurate predictions than unsupervised methods [147], with successes in expanding the compendium of TF-gene interactions in *E. coli* [145]. At the same time, when inferring interactions for an organism that is not well known, the lack of a proper training set can lead to a better performance by unsupervised methods.

A detailed review of the mathematical formalisms and detailed inference algorithms is out of the scope of this review. From the overwhelming number of methods available, we chose to briefly describe 10 methods, including the most widely used, the most recent [87], and the best performing from the DREAM challenges [117, 118, 148-150]. We focus our review on methods that produce genome-scale regulatory network reconstructions in the form of regulatory models that may be integrated with GEMs. While no method currently exists that completely satisfies these criteria, several algorithms, given in Table 3.3, can provide important results in the route to achieve the goal of fully integrated genome-scale models.

ARACNE [138] is one of the most widely used methods, first applied to infer regulatory interactions on human B cells [151]. Also, it has shown capacity for genome-wide inference in bacterial species such as *Streptomyces coelicor* [152]. CLR (context likelihood of relatedness) introduced the use of data from different experimental conditions for the same organism to infer regulatory interactions and enabled the identification of over 700 novel interactions in *E. coli* [70]. Being one of the most cited methods with an ability to predict edges in the RegulonDB, CLR is the method of choice for regulatory interactions studies [153]. It was recently used to unveil virulence factors in *Salmonella* [154]. A newer algorithm based on CLR, called SA-CRL (synergy augmented-CLR) [155], was the best-performing method in the DREAM2 genome-scale inference challenge, exploiting the concept of synergy among multiple interacting genes [156], where a pair of genes is used to infer the expression of a third to increase prediction accuracy.

Table 3.3 Methods for reverse engineering of gene regulatory networks from expression data.

		Inference Approach		Semi / Supervised	
Algorithm	**Modeling Approach**	DI*	MB**	Yes	No
ARACNE [138]		X			X
CLR [70]	Mutual Information (MI)	X			X
SA-CRL [155]		X			X
tlCLR [157]	+ MI		X		X
Inferelator [158]	ODE Model		X		X
Yip *et al.* [159]	+ Noise Model	X			X
GENIE3 [142]	Regression tress	X		X	
SEREND [160]	Logistic regression	X		X	
GPS [161]	Fuzzy Clustering		X		X
DISTILLER [162]	Association rules (itemsets)		X		X

*DI – Direct Inference | **MB – Module-Based

The Inferelator [158] was applied for genome wide reconstruction of *Halobacterium*. A mixed approach combining this method with CLR was one of the top performers in the DREAM3 *in silico* network challenge [157], using a modified version of CLR to compute mutual information values that are subsequently used by Inferelator to produce an ODE model. This method, called tlCLR (time-lagged CLR), takes advantage of two types of data: steady-state data from knockout experiments and time series gene expression data. Another method using different types of data was introduced by Yip *et al.* [159] gathering steady-state data from a noise model and time series data from an ODE model; this method was the top performer of the DREAM3 *in silico* challenge. Most algorithms in Table 3.3 can use steady-state or time series data, thus showing the benefits of integrating both types of data.

DREAM5 featured a genome-scale network inference challenge with a large dataset from a compendium of microarray data for *E. coli* comprising 805 chips, 334 TFs, and 4,511 genes. Large datasets were also provided for network inference on *Saccharomyces cerevisiae* and *Staphylococcus aureus*. GENIE3 (GEne Network Inference with Ensemble of trees) [142] uses tree-based methods

[163] decomposing the inference problem of *p* size into *p* distinct regression models. This method was the best performer overall and the top performer in the *in silico* network. GENIE3 had already been the best performer in the DREAM4 *in silico* inference for the 100-gene-multifactorial subchallenge, where only multifactorial data were provided, and showed equal capacity in successfully inferring networks from real data when compared with widely used methods such as CLR and ARACNE [142].

Several methods integrate multiple data types (e.g., inference from expression, binding site data) to facilitate TRN reconstruction. SEREND (SEmi- supervised REgulatory Network Discoverer) [160] uses a semi-supervised and iterative approach to unveil regulatory interactions. SEREND depends on a curated set of TF-gene interactions and TF-gene motif scores as a training set to construct a logistic regression model. The known predictions are then expanded and the predictions validated with ChiP-chip and time-series expression data. This approach was used to better predict and to give new insights into the factors involved in activation and repression in the aerobic/anaerobic regulation mechanism in *E. coli* [160].

GPS (Gene promoter Scan) [161] is also able to integrate other types of data; but as a module-based method, it follows a different approach. GPS is a machine learning method that builds promoter models and their relationships computed from a dataset. In the next step, characterized profiles (groups of promoters) are generated. The best profiles are used as candidates for genome wide predictions. Studies with *E. coli* and *S. enterica* using GPS unveiled previously unknown interactions and novel members of the PhoP protein controlled regulon [161].

DISTILLER [162] is another method that exploits the concept of regulation modularity integrating other sources of data for network inference. This framework can be applied to any organism and incorporate motif and ChiP-Chip data. The integrated approach was used to study the *FNR* regulon in *E. coli* identifying novel predictions that were experimentally validated. These studies provided insights on modularity dynamics pointing to the existence of polycistronic transcription [164].

A search for the best inference method usually turns to benchmarking studies; but the choice of benchmark datasets presents a problem, with different studies showing very sparse results [165, 166].

Lessons from all the DREAM challenges show that there is no individual best method. Results from community predictions, a combination of several reverse engineering methods, are closer to a state-of-art/best method, outperforming results from individual algorithms. The determination of error profiles enables the advantages and limitations of each inference method to be assessed in order to determine which method is "the best" for a specific inference problem.

The methods described above show recent advances, providing a good summary of the huge number of approaches that have been put forward. However, the underlying problem is complex, given the large search spaces involved and the still restricted availability of data that leads to an undetermined problem where many solutions can explain the data equally well. Hence, most of the methods rely on heuristic methods using different strategies to simplify the problem. The most important simplification is to reduce the search for a network or model explaining the data, with a huge number of possible interactions between the different entities involved, to the search of individual interactions or to small clusters or modules. This allows in some cases for distinct methods to be integrated to better support the results and, in the most elaborate methods, being followed by steps of determining regulatory programs based on these individual interactions.

3.3 METHODS

In this chapter, we explore the reconstruction of regulatory networks using 2 different approaches. In a first approach, we combine the information available in databases with notable regulatory transcriptional data for *B. subtilis* [5, 8, 9] into a comprehensive manually curated regulatory network.

In the second approach, we developed a methodology, dubbed “atomic regulon inference”, to infer regulatory interactions from gene expression data. For this purpose, we chose a dataset comprised of 269 samples across 104 different experimental conditions [10, 11]. To take advantage of both approaches and expand our knowledge of the *B. subtilis* regulatory network, we propose a process to reconcile the output from both approaches.

3.3.1 Atomic regulon inference

We define an *Atomic Regulon* as a set of genes with identical binary (ON/OFF) expression profiles. That is, in any given state of the cell, all of the genes in an atomic regulon will either be expressed or "not expressed". This notion has meaning only in a simplified model of the cell in which genes are either ON or OFF in any condition. Thus, we must have the ability to accurately assign genes to these binary states based on their normalized expression values from a variety of experimental samples.

Atomic regulons differ subtly from existing abstractions for describing the co-regulation of genes: regulons (set of genes that respond to the same regulator), and stimulons (set of genes that respond to the same stimuli) (Figure 3.6). Figure 3.6 a) shows a set of six genes (G1-G6) being regulated by three regulators (R1-R3) and effected by two stimuli (S1 and S2). Figure 3.6 b) overlays the theoretical atomic regulons with the previous figure.

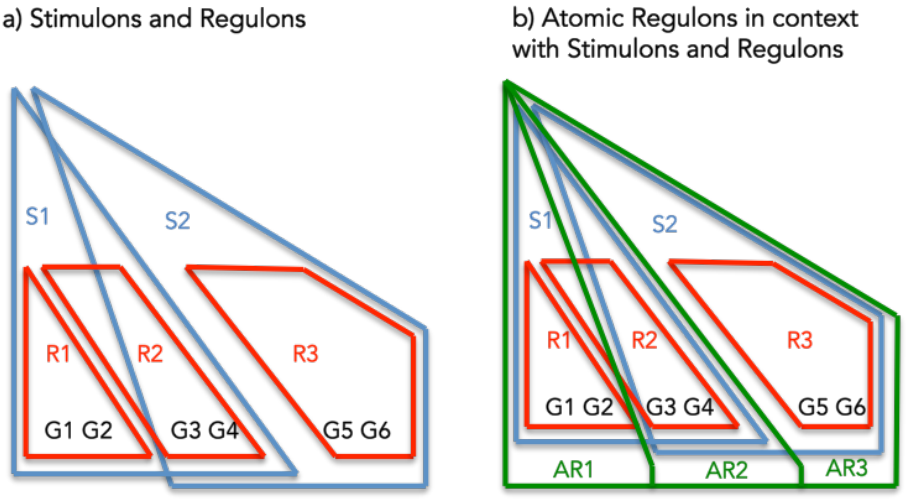


Figure 3.6 The interplay between Stimulons, Regulons and Atomic Regulons. a) The representation features six genes (G1-G6), three regulators (R1-R3) and two stimuli (S1-S2). The red lines define the regulons and the blue lines define the stimulons. b) Features in addition to a): the Atomic Regulons (AR1-AR3) in the representation as sets of genes that have identical binary profiles.

We compute atomic regulons using a three-step process (Figure 3.7) a) Inference of initial atomic regulons, b) Estimation of gene ON and OFF calls from expression data, and c) Merge regulons with similar expression profiles. Note that steps (a) and (b) are interrelated, and that step (c) is the final step merging the results of the previous steps.

In step (a) of our atomic regulon inference pipeline (Figure 3.7 a)), we perform 4 different computations.

(i) First we compute a set of hypotheses of the form:

Genes G1 and G2 should be in the same atomic regulon

These hypothesis are motivated largely by estimates of location in the chromosome via operon prediction and descriptions of SEED Subsystems [13]. The operon prediction gives us hypothetical atomic regulons based of position in strand. Sets of close genes in the same strand within 200 base pairs up and down stream are predicted to be in the same operon (Figure 3.7 a) i)).

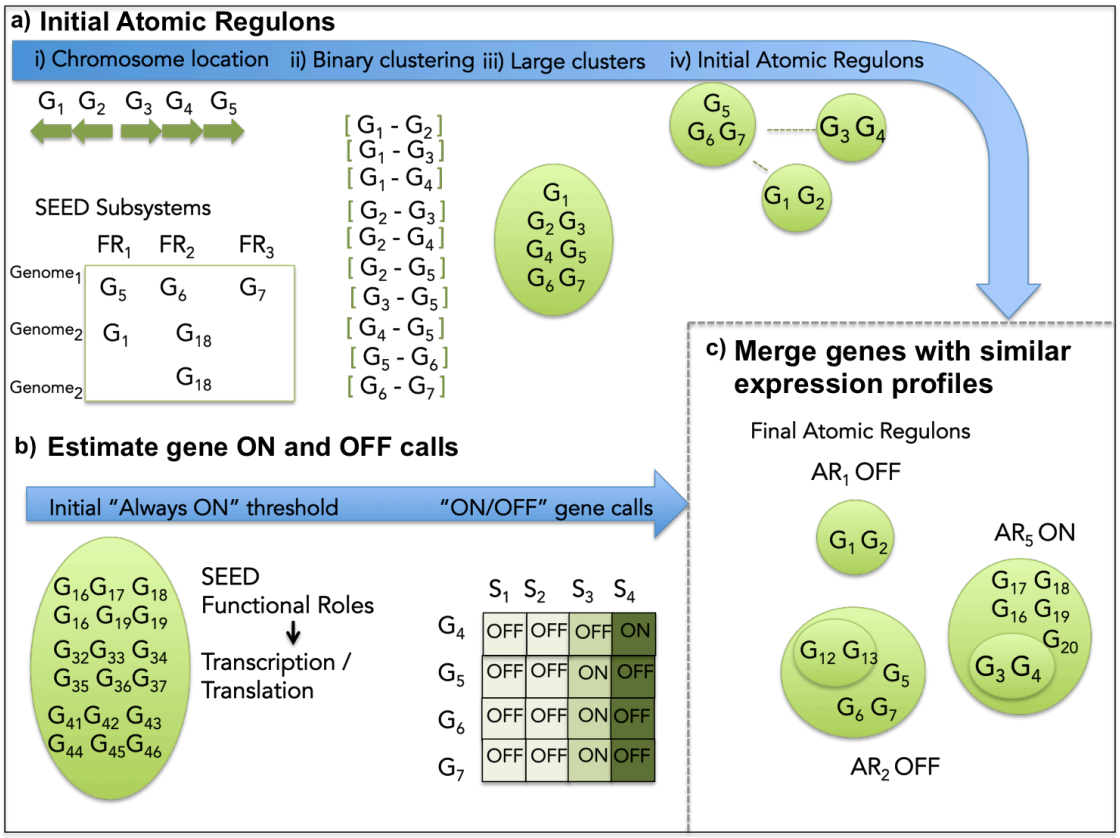


Figure 3.7 Atomic Regulon Inference. a) Inference of initial Atomic Regulons. Gene (G) location in the chromosome and genome information from SEED subsystem [13] are both used to generate the initial gene clusters. Single-linkage clustering generates large clusters from the original set. Initial Atomic Regulons are created using the Pearson correlation coefficient for the normalized expression values to break down the large clusters. b) The ON and OFF gene calls are calculated. The initial threshold for always ON profile is determined by a large set of genes with functional roles (FR) associated with transcription and translation that are assumed to be "always ON". The ON and OFF profile is calculated for each gene in every sample (S). c) Merging genes with similar expression profiles. Genes with the same expression profile are merged to create the final atomic regulons.

In order to understand how we use the descriptions of the SEED subsystem, it is important to understand the notion of "subsystem" and "populated subsystem". A subsystem is a set of functional roles (described as FR in Figure 3.7) that represent a biological process/pathway. A populated subsystem describes the exact genes that implement the functional roles of the subsystem across the specific genomes in which the FR is present. Each column in the subsystem corresponds to a FR with

each row representing a genome. Each individual cell identifies the genes within each specific genome that encode proteins, which implement the specific FRs (Figure 3.7 a) i)).

(ii) We generate binary connections between the genes predicted to be in the same operon. In the same manner, for a populated subsystem we generated binary connections between the genes in the row that corresponds to our genome of interest.

(iii) Then, we use the binary connections and form large clusters using transitive closure (if A is connected to B and B is connected to C, then A is connected to C). This leads to a situation in which any 2 genes that are connected are in the same cluster.

(iv) When necessary, we split large clusters based on Pearson correlation coefficients of the normalized gene expression values. We chose the simple approach of asserting a connection between two adjacent genes on the chromosome if they have a Pearson correlation coefficient greater than or equal to 0.4. For the split, a notion of "distance" (3.1) between genes X and Y is introduced:

$$Distance = \frac{\big(2 - (PCC + 1)\big)}{2} \tag{3.1}$$

where PCC is the Pearson correlation coefficient based on the normalized experimentally-derived expression values. Then, the genes from a single, perhaps too large, cluster from step iii) are used to construct sub-clusters.

Sub-clusters are formed by taking the two closest genes and methodically adding other genes to the growing sub-cluster. At each point, the gene with the minimum average distance to genes in the growing sub-cluster is added to the sub-cluster, until no such gene exists with an average distance less than or equal to 0.25. If this simple accretion algorithm produces a single sub-cluster, no splitting is required, If not, the sub-clusters become the set of tentative/initial atomic regulons.

In step (b) of our atomic regulon inference pipeline (Figure 3.7 b), we attempt to estimate ON and OFF profiles for each gene in our expression data. For each sample (S), we do the following computations:

(i) First, we determine the threshold for a gene to be considered ON based on the normalized expression of genes associated with functions that are believed to be universally expressed. For this task, we constructed a list of 175 functional roles from the SEED [167], largely from translation and transcription. Genes believed to implement these roles are thought of as "almost always ON".

(ii) We create an initial "ON threshold" as the 10^{th} percentile of observed expression values of the genes believed to be almost always on. The initial OFF threshold is computed as the 80^{th} percentile of the observed expression values below the initial ON threshold.

(iii) We then adjusted the ON/OFF boundaries for samples in which the difference in the ON and OFF thresholds was quite low. Specifically, we computed the differences between ON and OFF thresholds for all samples, and then looked at samples for which the difference was below the 25^{th} percentile. In those cases, we reset the OFF threshold to the value of the ON threshold minus the 25^{th} percentile of the difference scores.

(iv) All genes with values above the ON threshold were treated as "ON". All genes with expression values below the OFF threshold for a sample were classified as "OFF". Genes with expression values between the thresholds were labeled as "UNDECIDED".

In step (c) of our Atomic Regulon inference pipeline (Figure 3.7 c), we merge our initial atomic regulons together, if they have identical ON/OFF expression profiles, and we split them if the profiles of genes within the atomic regulon are not internally consistent. Finally, we estimate the ON/OFF status of each atomic regulon in any specific experimental sample by a simple voting algorithm using the ON/OFF estimates for the genes that make up the atomic regulon. After picking ON/OFF/UNDECIDED values for both the genes and atomic regulons, we make one final pass. For each sample, if the expression values for the gene and atomic regulon are incompatible, the value for the gene is altered to match that of the atomic regulon.

It is important to note that the resultant set of reconciled atomic regulons is not comprehensive (that is, not all genes are placed into an atomic regulon), but this set attempts to capture many of the operational groups of genes.

3.3.2 Atomic regulon curation

Genes contained within the same atomic regulon must share a common expression profile; so we assume they must respond to the same stimuli. Following that principle, three assertions were made for the reconciliation of our manually curated network with expression data:

1. Each *regulon* is a subset of at least one *stimulon*
2. Genes often take part in multiple *stimulons*, and they will vary in whether they are induced or suppressed in the stimulon.
3. A set of genes that all take part in identical sets of *stimuli* with identical induction/suppression profiles comprises an atomic regulon.

Figure 3.6 b) demonstrates these criteria: AR1 includes genes only affected by S1, AR2 includes genes affected by S1 and S2, and AR3 includes genes only affected by S2.

In order to curate the atomic regulons with these assertions, we organized the relevant data into Entities and Relationships. We begin by creating a basic Entity-Relationship model that will organize the data we propose to use for studying the notion of atomic regulons, as they might apply to *Bacillus subtilis*:

Entities

Our data is organized into the following entities:

- PEG: Protein-Encoding Gene (unique id).
- Peg function: RAST Annotation [167].
- Gene name: which is normally a 3-4-character string that is the most common name of the gene.
- Locus id: id assigned by the sequencing project.
- Atomic regulon: has an associated description (a description of the molecular mechanism), relating to the PEGs it contains.
- Stimulus: stimuli/effectors from the manually curated regulatory network.

- Study: the expression estimates were gathered as part of specific studies. Each study has two associated fields; a description of the study experimental conditions, and a general explanation of what was sought by doing the study.
- Sample set: each study includes sample sets, which are sets of associated samples from the expression data.
- Sample: includes estimates of activity for every PEG.

Relationships Between Entities

We support the following relationships:

- AtomicRegulon-PEG: connects an AR to the PEGs it includes.
- Sample-AtomicRegulon: connects a sample to the AR.
- AtomicRegulon-Stimulus: connects AR to stimuli that either turn the AR "ON" or "OFF".
- SampleSet-Sample: connects a sample set to the samples it includes.
- Sample-PEG: connects samples to the PEGs, showing the ON/OFF values as intersection data.
- PEG-Stimulus: connects PEGs to the stimuli that control their expression (when known). The relationship should contain a sign indicating activation or deactivation, but for now the connection just indicates relevance.
- Study-SampleSet: connects a study to the sample sets that were gathered.
- PEG-PEG: relationship showing the calculated correlation of expression values. There is a single field as intersection data – the Pearson correlation coefficient, which ranges from -1 to 1.

The Web Site

To display and analyze the data according to the entity relationship model described above, we developed a web site. The initial version of the web site can be seen at http://tinyurl.com/AtomicRegulons. An improved version is being prepared for the upcoming submission of the full manuscript comprising the work described in this chapter. This initial page will get the user to a list of the ARs, some of which have general descriptions.

The website can be used for other analysis of the expression data, under the notion of AR for other studies out of the scope the work presented in this chapter.

3.4 RESULTS AND DISCUSSION

3.4.1 Draft regulatory network of *Bacillus subtilis* from manual curation

The model that we have manually constructed and curated describes the current state of knowledge of the transcriptional network of *B. subtilis.* Our model corresponds to an updated and enlarged version of the regulation network in the central metabolism originally proposed in 2008 [1]. We have extended that original network to the whole genome by including the information from the DBTBS database [8]. The DBTBS compendium of regulatory data includes promoters, TFs, TFBS, motifs and regulated operons. The addition of the DBTBS led to a significant increase in the size of the regulatory network (Table 3.4). Additionally, we consolidated our network with all the information on regulation included in the Subtiwiki [7, 9] as of March 2013. Subtiwiki is the reference community-curated resource for *B. subtilis*. This consolidation with Subtiwiki resulted in some revision of regulatory data included in the original network by Goelzer *et al* [1]. Also, it significantly enlarged the network with respect to other microbial processes. All the above data reflect experimentally-validated regulatory interactions. Additionally, RegPrecise [4], a database that provides tools [107] for prediction and curation of regulons, recently released their reconstruction of the regulatory network for *B. subtilis* [5]. Reconciliation with the RegPrecise inferred network resulted in the addition of a total of 39 regulators to our experimentally validated network.

Table 3.4 Comparison between notable resources for *Bacillus subtilis* regulatory network modeling

Resource	TFs	Sigma Factors	RNA Regulators	Effectors	Regulated Genes
Goelzer *et al*. 2008	65	9	21	95	434
Leyn *el at*. 2013	129	-	33	130	1065
This work	177	19	60	169	1993

We compared our reconstruction with previously described reconstructions in the literature (Table 3.4). This comparison exposes a substantial increase in network coverage from the original Goelzer et al..

This increase is due in large part to an expansion of the scope of our model from the central carbon metabolism to genome-scale, as well as our effort to include most of the regulation mechanisms for *B. subtilis* that have been described in the literature to date. Our model includes 177 regulators, representing a wide variety of regulatory mechanisms: TFs conditioned by metabolites, accessory proteins, phosphorylated proteins and stress factors. Sigma factor regulation was included as it plays a role in governing many major cell functions such as sporulation (sigE, sigF, sigG and sigH), regulation of flagella, motility and chemotaxis (sigD), cell wall surface properties and stress (sigX, sigW and sigV). Elements relating to anti-sense RNA, riboswitches, RNA switches, RNA antiterminators and small regulatory RNAs compose the 60 RNA regulators described in our network. The increase in the number of regulators in our model led to a corresponding increase in effectors. We distinguish our effectors into two categories; biochemical (involving metabolites) and environmental effectors (e.g. DNA damage and heat shock). The 177 regulators in our model are linked to a set of regulons comprised of a total of 2000 genes. However, notably, the detailed regulatory mechanisms associated with some of the regulons on our model, particularly in cases of sigma factor and RNA regulation, remain unclear or unknown. All details related to our new regulatory model are provided on Supplementary material S3.1 and S3.2.

3.4.2 Atomic regulon computation

Once the initial reconstruction of our new regulatory model of *B. subtilis* was complete, we validated and reconciled our model with available gene expression datasets for *B. subtilis*. We began this process by surveying the data present in current expression databases. Currently (as of January 2014) there are approximately 1750 datasets in GEO [46] related to *B. subtilis* strains. To reconcile expression data with our manually curated network, we used the Atomic Regulon Inference methodology described in the Methods section.

To infer atomic regulons for *B. subtilis*, we utilized a subset of the existing expression data for *B. subtilis*. Numerous expression experiments have been performed for *B. subtilis*, spanning a wide range of academic labs and gene expression measurement platforms. Utilization of all of these data poses a challenge, as protocols vary from lab to lab, and various expression platforms can produce different

results. Thus, we selected a dataset emerging from a single publication, comprised of 269 samples across 104 different conditions [10, 11]. Figure 3.8 shows an overview of the inferred atomic regulons for *B. subtilis*. The entire set of atomic regulons is available in Supplementary material S3.3

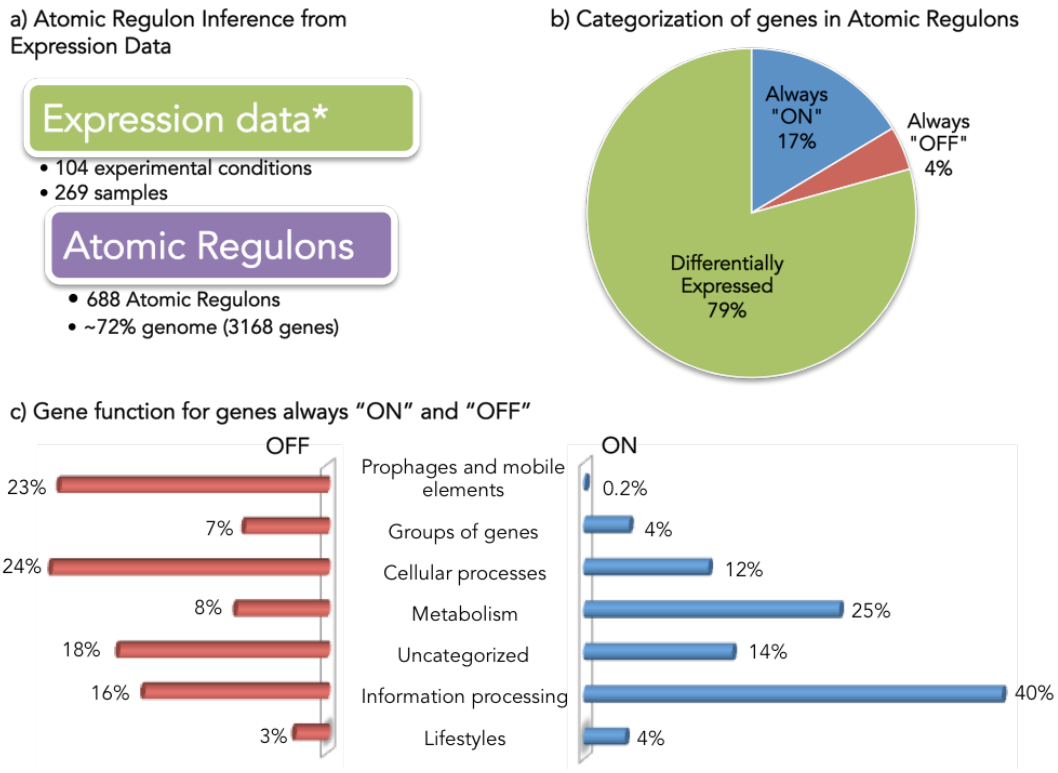


Figure 3.8 Overview of *B. subtilis* atomic regulons. a) Expression data used for atomic regulon inference in *B. subtilis*. b) Categorization of genes in atomic regulons. Genes have been categorized based on the expression profile as always "ON", always "OFF" and differentially expressed. C) Gene function for genes always "ON" and "OFF". The genes were classified among 6 different major groups of cellular functions defined in the SubtiWiki [7].

A total of 688 atomic regulons were computed comprising 3168 genes (approximately 72% of the genome) (Figure 3.8 a). We categorized these ARs according to their expression profile (Figure 3.8 b): only 4% (137) of the genes were always OFF in all conditions, while 17% (523) of the genes were ON in all conditions. This result was consistent with the claim by the authors of the study that 95% of the

genes in *B. subtilis* had been expressed in at least one condition. We explored the functions associated with the genes that we found to be always ON or always OFF (Figure 3.8 c).

40% of the always-ON genes (211) are categorized as information processing, which encompasses: RNA synthesis and degradation (transcription); protein folding, modification and degradation and (translation); and, DNA replication. 25% (129) of the genes always-ON were metabolic, including Central carbon, nucleotide, and lipid metabolism. Finally, 12% (63) of the always-ON genes were associated with cellular process, including cell wall biosynthesis, cell division, transporters and homeostasis.

The small set of genes (137) found to be OFF in all conditions is comprised of genes across a diverse set of functions. To verify that no gene found to be OFF in all conditions was an essential gene, we compared the set with a list of *B. subtilis* essential genes [168, 169]. No essential *B. subtilis* gene was found to be OFF in all conditions.

3.4.3 Reconciling expression data with the draft regulatory network for *Bacillus subtilis*

Our definition of AR states that genes contained within the same AR must respond to the same set of stimuli (Figure 3.6). We can use this principle to identify and reconcile inconsistencies that exist between the stimuli mapped to the genes in our *B. subtilis* model and the set of genes comprising each AR. Considering sucrose as an example (Figure 3.9), we can explore the set of ARs computed for the genes comprising the Sucrose stimulon. We have 8 genes in the Sucrose stimulon; *ywdA*, *sacA* and *sacP* [170] are all effected by fructose-biphosphate and glucose-6-phosphate; sacX and sacY are effected by an uncharacterized stimulus [171]; sacB and levB are effected by two uncharacterized stimuli [171]; yveA shares the same uncharacterized stimulus as the previous genes plus another uncharacterized stimulus [172].

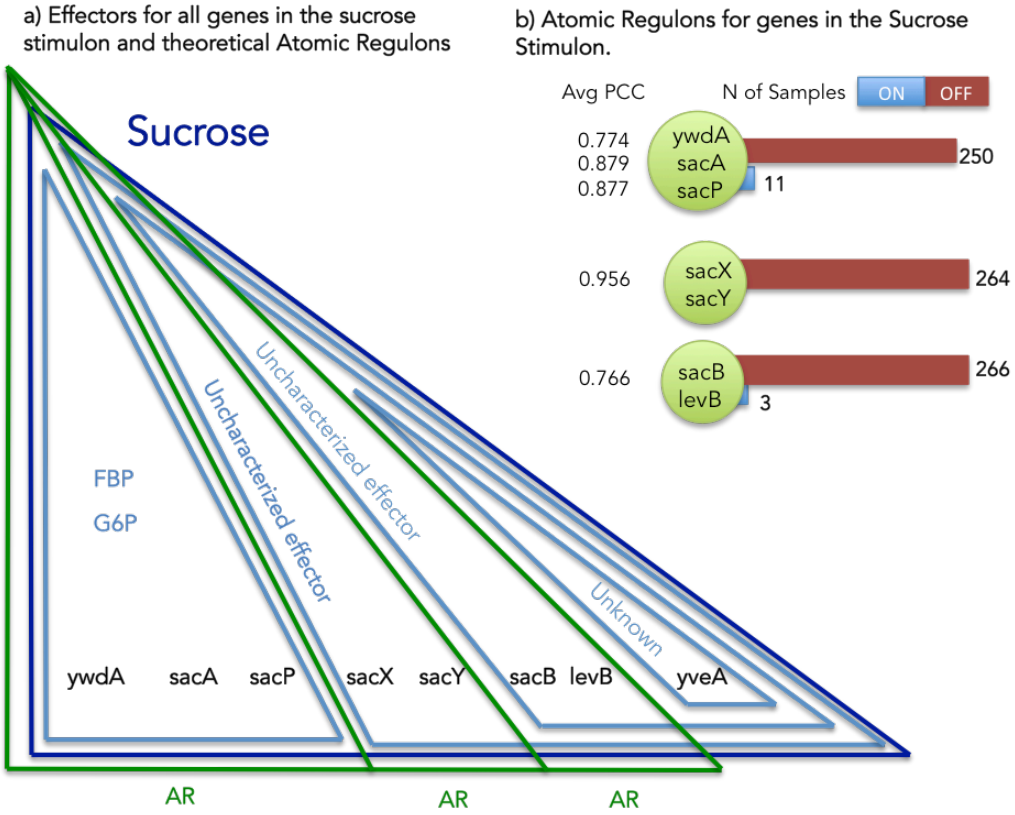


Figure 3.9 Atomic regulons for the sucrose stimulon. a) Effectors for all genes in the sucrose stimulon and theoretical atomic regulons. Eight genes compose the sucrose stimulon (dark blue triangle). Fructose-biphospshate (FBP), Glucose-6-phosphate and uncharacterized effectors are also effectors (light blue triangles). The theoretical atomic regulons (AR) are represented in green triangles. b) Atomic regulons inferred for the sucrose stimulon. The atomic regulons that were inferred are shown with the average Pearson correlation coefficient (PCC) with other members of the AR. Number of samples ON an OFF for each AR is also shown.

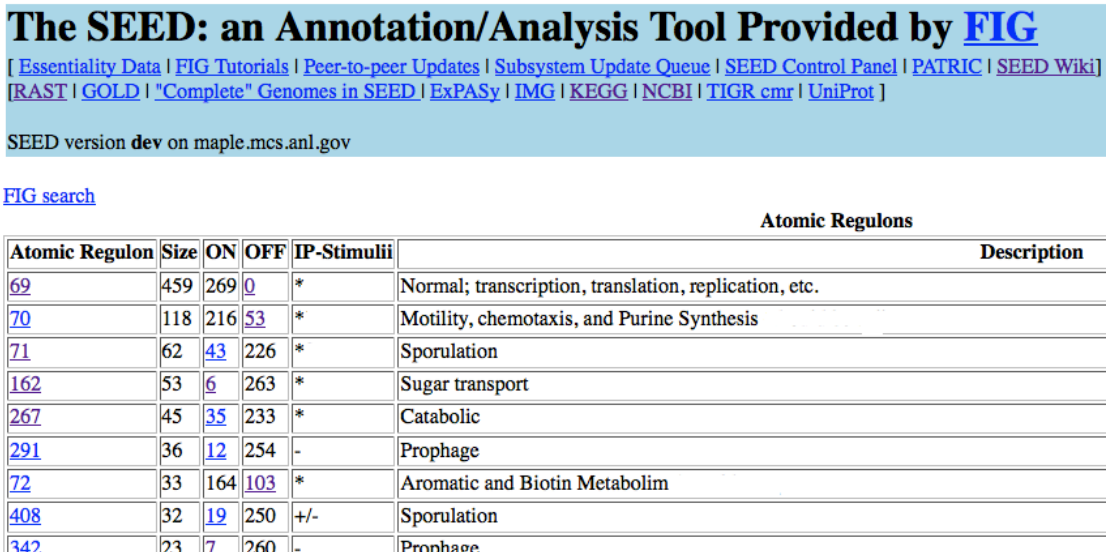
Based on the available expression data, we initially divided the sucrose stimulon into three ARs (green triangles in Figure 3.9 a)) and lists in Figure 3.9 b): (*ywdA* , *sacA* and *sacP*), (sacX and sacY) and (sacB and levB). yveA was not placed into an ARs due to the assertions previously discussed on the inference methodology. Figure 3.9 b) shows the average Pearson correlation coefficient (PCC) for each gene; the average PCC is computed for each gene relative to the other members of the AR to which the gene was assigned. Figure 3.9 b) also shows the number of studies in our expression data set in which each AR was considered to be ON and OFF. The capacity of our AR inference methodology to divide the

genes that respond different stimuli into different ARs demonstrates the robustness of the approach. To further investigate the Sucrose stimulon, we also developed a set of web resources to display and analyze the ARs. Table 3.5 shows the Sucrose stimulon as featured in the analysis web page. The web resource allows a user to explore the AR data, the effector/stimuli data from the network reconstruction, and the metadata from the genome expression experiments. The web resource is available at: http://tinyurl.com/AtomicRegulons.

Table 3.5 Sucrose stimulon represented in the AR web analysis resource

Atomic Regulon	PEG	Stimuli
254	fig|224308.113.peg.3951, sacX, BSU38410	Uncharacterized, Sucrose
254	fig|224308.113.peg.3952, sacY, BSU38420	Uncharacterized, Sucrose
376	fig|224308.113.peg.3912, ywdA, BSU38030	D-fructose-1,6-bisphosphate,Glucose-6-Phosphate, Sucrose
376	fig|224308.113.peg.3913, sacA, BSU38040	D-fructose-1,6-bisphosphate,Glucose-6-Phosphate, Sucrose
376	fig|224308.113.peg.3914, sacP, BSU38050	D-fructose-1,6-bisphosphate, Glucose-6-Phosphate, Sucrose
625	fig|224308.113.peg.3544, sacB, BSU34450	Uncharacterized, uncharacterized, Sucrose
625	fig|224308.113.peg.3545, levB, BSU34460	Uncharacterized, uncharacterized ,Sucrose

In our AR analysis web resource (link above), we display the size of each atomic regulon and the samples in which each AR has been called as being ON or OFF (Figure 3.10). The first column shows the AR number arbitrarily assigned by the inference algorithm (the ARs described in Figure 3.9 were assigned AR numbers 254, 376 and 625). From this page, it is possible to check all genes in each AR, associated stimuli from the regulatory network (if available), and genome annotation. All gene identifiers are derived from and linked to the PubSEED (http://pubseed.the-seed.org), which also provides a series of comparative genomics tools that allow for further analysis. It is also possible to retrieve a list of the experiments in which each gene is characterized as being either ON or OFF.



The SEED: an Annotation/Analysis Tool Provided by FIG

[Essentiality Data | FIG Tutorials | Peer-to-peer Updates | Subsystem Update Queue | SEED Control Panel | PATRIC | SEED Wiki]
[RAST | GOLD | "Complete" Genomes in SEED | ExPASy | IMG | KEGG | NCBI | TIGR cmr | UniProt]

SEED version **dev** on maple.mcs.anl.gov

FIG search

Atomic Regulons

Atomic Regulon	Size	ON	OFF	IP-Stimulii	Description
69	459	269	0	*	Normal; transcription, translation, replication, etc.
70	118	216	53	*	Motility, chemotaxis, and Purine Synthesis
71	62	43	226	*	Sporulation
162	53	6	263	*	Sugar transport
267	45	35	233	*	Catabolic
291	36	12	254	-	Prophage
72	33	164	103	*	Aromatic and Biotin Metabolim
408	32	19	250	+/-	Sporulation
342	23	7	260	-	Prophage

Figure 3.10 Atomic regulon analysis web resource.

Figure 3.9 shows that the genes in 2 ARs (254 and 625) respond to uncharacterized effectors. It also shows that those ARs are only ON in very few experiments (Figura 3.9 b). For AR 625 we checked the experiments in which the genes were ON (Table 3.6).

Table 3.6 Experiments in which AR 625 was found to be "ON"

Study	Sample	Study explanation
study0003	S6/t_2_hyb42359702	tested gene expression at regular intervals after sporulation was induced
study0003	S6_2_hyb29634602	tested gene expression at regular intervals after sporulation was induced
study0003	S8_5_hyb43271102	tested gene expression at regular intervals after sporulation was induced

Genes in AR 625 were found to be "ON" in experiments that tested gene expression at regular intervals after sporulation was induced. A detailed description of the study can be found by checking the associated study number (in this case "study0003"). Sporulation was induced with the use of CH medium [173], and cells were harvested at hourly intervals, with the genes in our ARs being ON in the late intervals of the study. This fact tells us that our uncharacterized effector can be related with

sporulation and that a compound in the CH medium can be a candidate for the uncharacterized effector.

To assess the consistency of all ARs when compared with the stimuli in the regulatory network, we organized all computed ARs into four categories: consistent, consistent with missing stimuli, inconsistent, and empty (counts in Table 3.7).

Table 3.7 Consistency of the Atomic regulons with the regulatory network. Reflects the consistency of the original ARs (V1) and the curated ARs (V2)

Classification	V1	V2
Consistent (+)	151	174
Consistent with missing stimuli (+/-)	74	45
Inconsistent (*)	48	32
Empty (-)	415	425
Total	688	676

(+) All ARs members have the same stimuli/effectors in the regulatory network.

(+/-) Some members of the AR have the same stimuli/effectors while other members have no described stimuli in the regulatory network.

(*) ARs members have different stimuli associated in the regulatory network.

(-) No stimuli described in the regulatory network

The category consistent comprises ARs members that have the same stimuli/effectors in the regulatory network. Consistent with missing stimuli comprises cases in which a member of the AR has the same stimuli/effectors while other members have no described stimuli in the regulatory network. The inconsistent category displays ARs with members that have different stimuli associated in the regulatory network. ARs with no stimuli/effectors described in the regulatory network were categorized as “empty”. Additionally, we used the web resource described above to curate and improve the ARs (results in Table 3.7).

In Table 3.7 (V1) we can see that 151 ARs were found to be consistent, as all genes shared the same set of stimuli from the regulatory network. 74 were found to be consistent with missing stimuli, meaning some members of the AR have no stimuli assigned in the regulatory network. 48 ARs showed

inconsistencies; and 415 ARs had no stimuli associated. Here, we highlight some of the considerations we made during the process of manual curation.

Table 3.8 Atomic Regulon 56

PEG	Stimuli	Avg. PCC
fig|224308.113.peg.2590, zur, BSU25100		0.664
fig|224308.113.peg.2903, hemL, BSU28120	Hydrogen_peroxide	0.885
fig|224308.113.peg.2904, hemB, BSU28130	Hydrogen_peroxide	0.900
fig|224308.113.peg.2905, hemD, BSU28140	Hydrogen_peroxide	0.905
fig|224308.113.peg.2906, hemC, BSU28150	Hydrogen_peroxide	0.906
fig|224308.113.peg.2907, hemX, BSU28160	Hydrogen_peroxide	0.893
fig|224308.113.peg.2908, hemA, BSU28170	Hydrogen_peroxide	0.804

We found multiple occurrences in which members of an AR included regulatory proteins. These genes are the genes responsible for imposing the regulatory mechanism. This contradicts our definition of ARs in which we are trying to represent as sets of regulated genes. Table 3.8 shows one of such cases in AR 56. All members of AR 56 except *zur* have hydrogen peroxide as stimuli. Upon further inspection we noted that AR 56 is capturing the *hemAXCDBL* operon [174], which has been found to be regulated by PerR and is induced by hydrogen peroxide [175]. Zur is a PerR paralogous protein [176], involved in regulation of the zinc homeostasis as the zinc uptake repressor [177]. *zur* was removed from AR 56 making this AR consistent according to our previously described categorization. We subsequently used the information from our manually curated network to remove all known regulatory proteins from the ARs.

AR 612 is comprised of 10 genes having functions associated with heme/iron transport (Table 3.9). From our regulatory network we have "Iron" associated with 8 out of the 10 AR members. A survey for *yetG* (now *hmoA*) revealed that the gene has been recently characterized to encode a heme monooxygenase [178]. *hmoA* has also been shown to be regulated by *Fur*, the same regulator as the

other members of AR 612. We suggest the expansion of regulatory information for *hmoA* to be consistent with AR 612 (the latest release of Subtiwiki has already implemented this change). *yetH* was removed from AR 612 as it was found not to be related with other members of AR 612 and it had the lowest average PCC among all members of AR 612. We applied this same logic to suggest multiple additions to the regulatory network.

Table 3.9 Atomic Regulon 612

PEG	Stimuli	Avg. PCC
fig|224308.113.peg.385, yclN, BSU03800	Iron	0.804
fig|224308.113.peg.386, yclO, BSU03810	Iron	0.807
fig|224308.113.peg.387, yclP, BSU03820	Iron	0.815
fig|224308.113.peg.388, yclQ, BSU03830	Iron	0.806
fig|224308.113.peg.747, yetG, BSU07150		0.716
fig|224308.113.peg.748, yetH, BSU07160		0.605
fig|224308.113.peg.780, yfmF, BSU07490	Iron	0.621
fig|224308.113.peg.781, yfmE, BSU07500	Iron	0.702
fig|224308.113.peg.782, yfmD, BSU07510	Iron	0.769
fig|224308.113.peg.783, yfmC, BSU07520	Iron	0.764
fig|224308.113.peg.1063, yhfQ, BSU10330	Iron	0.779

We also analyzed the ARs that were flagged as inconsistent. Some inconsistencies were caused by AR members with low average PCC that were found to be unrelated to the other AR genes and subsequently removed. Another inconsistency involved genes in ARs where all members of the AR did not share the same set of stimuli. An example of this case is AR 332 (Figure 3.10).

Table 3.10 Atomic regulon 332

PEG	Stimuli	Avg. PCC	Function
fig|224308.113.peg.810,treP,BSU07800	D-fructose-1,6-bisphosphate,Glucose-6-Phosphate,phosphate,D-trehalose-6-phosphate	0.807	PTS system, trehalose-specific IIB component (EC 2.7.1.69)
fig|224308.113.peg.811,treA,BSU07810	D-fructose-1,6-bisphosphate,Glucose-6-Phosphate,phosphate,D-trehalose-6-phosphate	0.814	Trehalose-6-phosphate hydrolase (EC 3.2.1.93)
fig|224308.113.peg.812,treR,BSU07820	D-fructose-1,6-bisphosphate,Glucose-6-Phosphate,phosphate,D-trehalose-6-phosphate	0.734	Trehalose operon transcriptional repressor
fig|224308.113.peg.813,yfkO,BSU07830	Disulfide_stress_conditions	0.706	Oxygen-insensitive NAD(P)H nitroreductase (EC 1.-.-.-)

On a first analysis we noted that 3 out of 4 members of the AR 332 share the same effectors. These 3 members (*treP*, *treA* and *treR*) were found to comprise the tre operon [179]. TreR is a transcriptional repressor, involved in the regulation of trehalose utilization and it is inhibited by trehalose-6-phosphate [180]. The additional stimuli, D-fructose-1,6-bisphosphate and Glucose-6-Phosphate, relate to the activity of the carbon catabolite repression global regulator CcpA [181]. The fourth member of the AR, *yfkO*, has been described in the literature as a nitroreductase [182]. Upon inspection of this region of the chromosome we found *yfkO* up-stream of the transcriptional regulator TreR, and not a member of the tre operon/TreR regulon. Due to this analysis we removed *yfkO* from AR 332. As noted before we also removed TreR as the protein imposing the regulatory activity. This curated AR was classified as "Trehalose Utilization".

Table 3.7 also shows that 415 ARs were found to have no associated stimuli in the regulatory network. Previously, we attempted to use details in the gene expression experiments to aid in the characterization of unknown effectors (Figure 3.9). We applied the same approach to genes for which there are no effectors in the regulatory network.

Table 3.11 Atomic Regulon 651.

PEG	Stimuli	Avg. PCC	Function	OFF	ON
fig|224308.113.peg.2068,yosW,BSU19980		0.906	unknown	261	4
fig|224308.113.peg.2069,yosV,BSU19990		0.948	unknown	261	4
fig|224308.113.peg.2070,yojW,BSU19999		0.941	unknown	261	4

As an example we looked at AR 651 (Table 3.11). In addition to having no regulatory information in the network, all genes in AR 651 also have unknown functions. The members of AR 651 show a high average PCC and are only "ON" in a very small number of samples. In the experiment that activated AR 651, cells were grown in LB medium at 37°C with vigorous shaking. During exponential growth (O.D.600 approx. 0.25), the cell culture was divided into two subcultures: one subculture acted as control [no mitomycin C, M0], while mitomycin was added to the second subculture at a concentration of 40 ng/mL [mitomycin, M40]. Samples were harvested at 0, 45 and 90 minutes after mitomycin addition [t0, t45 and t90]. Addition of mitomycin C promoted prophage induction. To verify that the prophage induction occurred, we developed the capability in our web tools to compare the difference between "ON" and "OFF" profiles among experimental conditions. This web resource can be found at: http://tinyurl.com/ARStudies. We are able to do pairwise comparisons for all samples in the expression data used to compute the atomic regulons. For this study, we compared the control sample (grown in LB media only) against the sample grown with mitomycin C. The results can be found at http://tinyurl.com/LBvsMitomycin. In the results, we see several AR, associated with prophages being ON in the experiment where mitomycin C was added. Mitomycin C serves to stimulate the expression of these specific genes, leading to its addition as a stimulus for these ARs.

The "case studies" presented before were part a larger manual curation effort. All changes made to ARs during the curation process can be found on supplementary material S3.4. In Table 3.7 we can see the impact of the curation process (V2) across our four different categories. We see a decrease in ARs categorized as inconsistent and consistent with missing stimuli. This led to a subsequent increase in the number of consistent ARs. The improved set (Supplementary materials S3.5 and S3.6) contain new versions of ARs and the Regulatory network that reflect all the changes made during the manual curation process.

3.4.4 Atomic Regulons in the SEED

In the previous sections, we used the atomic regulons in combination with comparative genomics tools developed in the PubSEED environment to expand our knowledge of the *B. subtilis* transcriptional regulatory network. Atomic regulons also show potential in elucidating unknown gene functions. In order to exploit this functionality, we collected expression datasets from GEO for an additional set of 21 organisms and computed ARs.

Unfortunately, these organisms for which we computed AR (and most organisms in general) have far fewer experimental data points in their expression data than *E. coli* or *B. subtilis* (Figure 3.3). To assess the impact of this data sparseness on our AR inference algorithm, we analyzed all the ARs that were computed for this study.

We can see by the results of Table 3.12 that, when compared with the model organisms' *B. subtilis* and *E. coli*, other organisms have significantly fewer atomic regulons. Additionally, we see that for most organisms, a significant percentage of their total genes are being placed in a small number of ARs. This results in large ARs that fail to account for the diversity of cellular machinery.

This indicates that most expression series lack data from the wide variety of conditions that are necessary for the inference algorithm to capture diverse and unique ARs.

To make use of the ARs to aid in genome annotation efforts, we integrated the computed ARs for this study in the SEED database. They can be accessed in the interface for all genes that are members of a

computed AR. Figure 3.11 shows an example of the integration of the atomic regulon information in the gene features page in the SEED website.

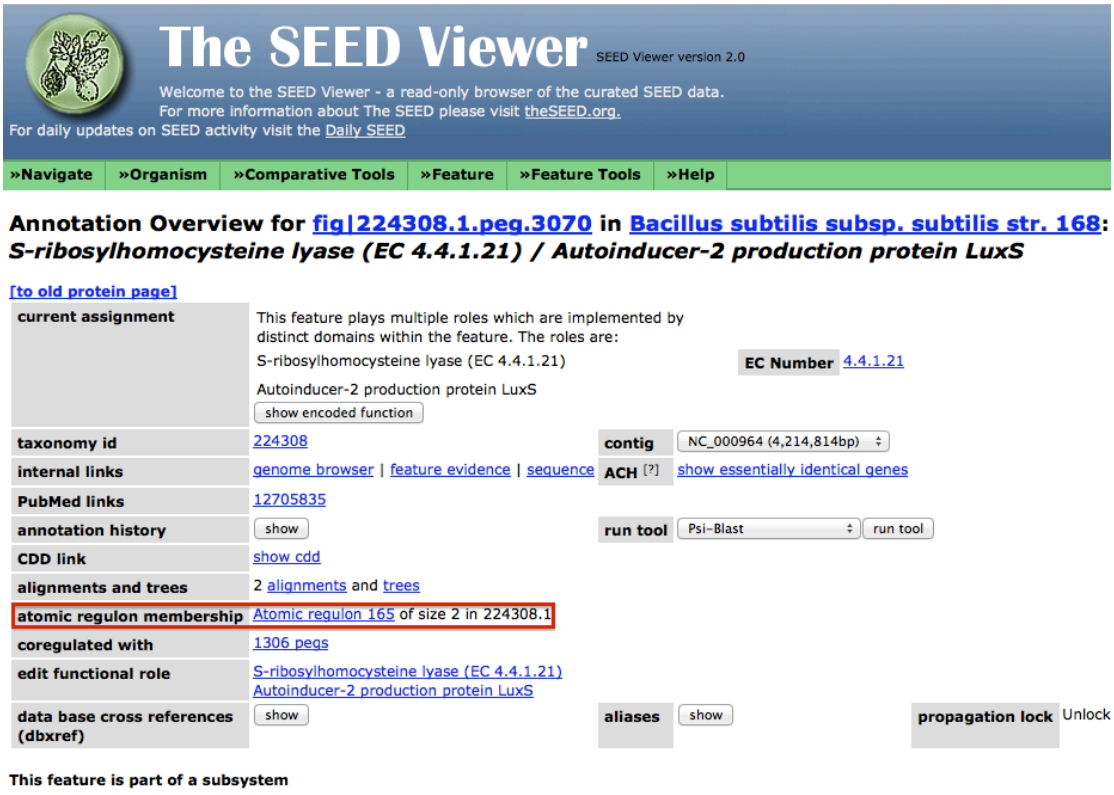


Figure 3.11. Integration of Atomic Regulons in the SEED website.

Table 3.12 Organisms with computed Atomic Regulons (ARs) available in the PubSEED.

Organism	PubSEED ID	% of genes in ARs	Number of ARs	Genome size
Shewanella oneidensis MR-1	211586.9	38%	343	4167
Thermus thermophilus HB8	300852.3	63%	168	2239
Vibrio parahaemolyticus RIMD 2210633	223926.6	81%	194	4664
Salmonella enterica subsp. enterica serovar Typhimurium str. LT2	99287.12	38%	334	4969
Bacillus anthracis str. Ames	198094.1	91%	129	5665
Vibrio fischeri ES114	312309.3	92%	116	3798
Bradyrhizobium japonicum USDA 110	224911.1	58%	642	8594
Pasteurella multocida subsp. multocida str. Pm70	272843.1	77%	111	2026
Rhodopseudomonas palustris CGA009	258594.1	85%	161	4891
Staphylococcus aureus subsp. aureus Mu50	158878.1	60%	431	2770
Rhodobacter sphaeroides 2.4.1	272943.3	67%	227	4127
Helicobacter pylori HPAG1	357544.13	55%	73	1596
Streptomyces coelicolor A3(2)	100226.1	57%	275	8154
Escherichia coli K12	83333.1	60%	626	4309
Eubacterium rectale ATCC 33656	515619.6	72%	144	3194
Bacillus subtilis subsp. subtilis str. 168	224308.113	74%	676	4292
Bacteroides thetaiotaomicron VPI-5482	226186.1	84%	256	4832
Mycoplasma pneumoniae M129	272634.1	83%	36	689
Streptococcus pyogenes MGAS5005	293653.3	72%	128	1865
Synechococcus elongatus PCC 7942	1140.3	52%	106	2729
Rickettsia rickettsii str. Iowa	452659.3	57%	43	1599
Pseudomonas aeruginosa PAO1	208964.1	43%	432	5682

3.5 CONCLUSIONS

In this work, we started by conducting a survey of the available resources for gene regulatory network data. This survey revealed the current status of expression data available in major public repositories. As part of this analysis, we compared the availability of expression data sets in GEO versus the number of genomes available in the NCBI. The results of this comparison show how only a small portion of sequenced organisms have available expression data. This is due to the recent exponentially rise in genome sequencing (as discussed in Chapter 2) vs the price of gene expression studies. We also surveyed databases with notable bacterial transcriptional regulatory data. This survey showed that detailed information of regulatory networks is only available for a small number of organisms. In addition to our survey, we extensively reviewed methods for regulatory network inference. The results of this data survey and inference methods review were published in the February 2013 issue of the journal *Briefings in Bioinformatics*.

Taking in to account the data survey conducted, we introduced a more comprehensive regulatory network for *B. subtilis,* compiling information from multiple notable sources of gene regulatory data. We show that our reconstruction is more comprehensive than other previous versions found in the literature. We also introduced a new methodology called *Atomic Regulon Inference* to reconcile our proposed network with available gene expression data. We show how this methodology is able to elucidate details of the regulatory network. The reconciliation process allowed us to extend our knowledge of the regulatory network. We were also able to provide clues for putative gene function assignments for genes with unknown functions. ARs were integrated into the PubSEED, and they can be used as part of the annotation curation tools available in that framework. A web resource was also created showing the relationship between the ARs and the expression data used for their computation. This resource is available for the public and can be used to conduct analysis of the ARs and expression data sets beyond the objectives of the work described in this chapter.

During the reconciliation process, we were able to see how many ARs represent the same regulatory mechanisms we found in our manually curated network. This fact highlights the convenience of using

ARs to study the regulatory network of an organism without a huge effort from initial manual curation. As part of the manuscript in preparation for the work discussed in this chapter, we plan to provide AR inference as a pipeline, in which users can submit their own data for inference.

As new algorithms are proposed for the task of regulatory gene network inference, no algorithm can be defined as the best algorithm for this task. In the state of the art section we discuss this issue, as algorithms were best performers in different DREAM network inference challenges. Due to this fact, the community has been advocating to the wisdom of crowds, as integration of multiple methods shows better results that any individual method [183]. We believe AR inference can be valuable for this type wisdom of crowds approach, as it leverages the prior knowledge from SEED Subsystems instead of relying purely in inference from expression data. This can be extremely useful especially for organisms lacking high quality expression data.

We expect that, with the growth of next generation high throughput sequencing data, we are able to use the wealth of data to better characterize regulatory networks.

3.6 REFERENCES

1. Goelzer, A., et al., *Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of Bacillus subtilis.* BMC Syst Biol, 2008. **2**: p. 20.

2. de Hoon, M.J., et al., *Inferring gene regulatory networks from time-ordered gene expression data of Bacillus subtilis using differential equations.* Pac Symp Biocomput, 2003: p. 17-28.

3. Fadda, A., et al., *Inferring the transcriptional network of Bacillus subtilis.* Mol Biosyst, 2009. **5**(12): p. 1840-52.

4. Novichkov, P.S., et al., *RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes.* Nucleic Acids Res, 2010. **38**(Database issue): p. D111-8.

5. Leyn, S.A., et al., *Genomic reconstruction of the transcriptional regulatory network in Bacillus subtilis.* J Bacteriol, 2013. **195**(11): p. 2463-73.

6. Michna, R.H., et al., *SubtiWiki-a database for the model organism Bacillus subtilis that links pathway, interaction and expression information.* Nucleic Acids Res, 2014. **42**(1): p. D692-8.

7. Florez, L.A., et al., *A community-curated consensual annotation that is continuously updated: the Bacillus subtilis centred wiki SubtiWiki.* Database (Oxford), 2009. **2009**: p. bap012.

8. Sierro, N., et al., *DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information.* Nucleic Acids Res, 2008. **36**(Database issue): p. D93-6.

9. Mader, U., et al., *SubtiWiki–a comprehensive community resource for the model organism Bacillus subtilis.* Nucleic Acids Res, 2012. **40**(Database issue): p. D1278-87.

10. Nicolas, P., et al., *Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis.* Science, 2012. **335**(6072): p. 1103-6.

11. Buescher, J.M., et al., *Global network reorganization during dynamic adaptations of Bacillus subtilis metabolism.* Science, 2012. **335**(6072): p. 1099-103.

12. Salgado, H., et al., *Operons in Escherichia coli: genomic analyses and predictions.* Proc Natl Acad Sci U S A, 2000. **97**(12): p. 6652-7.

13. Overbeek, R., et al., *The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.* Nucleic Acids Res, 2005. **33**(17): p. 5691-702.

14. Ermolaeva, M.D., O. White, and S.L. Salzberg, *Prediction of operons in microbial genomes.* Nucleic Acids Res, 2001. **29**(5): p. 1216-21.

15. Price, M.N., et al., *A novel method for accurate operon predictions in all sequenced prokaryotes.* Nucleic Acids Res, 2005. **33**(3): p. 880-92.

16. Chuang, H.Y., M. Hofree, and T. Ideker, *A decade of systems biology.* Annu Rev Cell Dev Biol, 2010. **26**: p. 721-44.

17. Reed, J.L., et al., *Towards multidimensional genome annotation.* Nat Rev Genet, 2006. **7**(2): p. 130-41.

18. Covert, M.W., et al., *Metabolic modeling of microbial strains in silico.* Trends Biochem Sci, 2001. **26**(3): p. 179-86.

19. Henry, C.S., et al., *High-throughput generation, optimization and analysis of genome-scale metabolic models.* Nat Biotechnol, 2010. **28**(9): p. 977-82.

20. Terzer, M., et al., *Genome-scale metabolic networks.* Wiley Interdiscip Rev Syst Biol Med, 2009. **1**(3): p. 285-97.

21. Ruppin, E., et al., *Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks.* Curr Opin Biotechnol, 2010. **21**(4): p. 502-10.

22. Feist, A.M., et al., *Reconstruction of biochemical networks in microorganisms.* Nat Rev Microbiol, 2009. **7**(2): p. 129-43.

23. Struhl, K., *Fundamentally different logic of gene regulation in eukaryotes and prokaryotes.* Cell, 1999. **98**(1): p. 1-4.

24. Nudler, E. and A.S. Mironov, *The riboswitch control of bacterial metabolism.* Trends Biochem Sci, 2004. **29**(1): p. 11-7.

25. Mironov, A.S., et al., *Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria.* Cell, 2002. **111**(5): p. 747-56.

26. Wagner, E.G. and R.W. Simons, *Antisense RNA control in bacteria, phages, and plasmids.* Annu Rev Microbiol, 1994. **48**: p. 713-42.

27. Chen, K. and N. Rajewsky, *The evolution of gene regulation by transcription factors and microRNAs.* Nat Rev Genet, 2007. **8**(2): p. 93-103.

28. Covert, M.W., C.H. Schilling, and B. Palsson, *Regulation of gene expression in flux balance models of metabolism.* J Theor Biol, 2001. **213**(1): p. 73-88.

29. Covert, M.W., et al., *Integrating high-throughput and computational data elucidates bacterial networks.* Nature, 2004. **429**(6987): p. 92-6.

30. Yoon, H., et al., *Coordinated regulation of virulence during systemic infection of Salmonella enterica serovar Typhimurium.* PLoS Pathog, 2009. **5**(2): p. e1000306.

31. Herrgard, M.J., et al., *Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae.* Genome Res, 2006. **16**(5): p. 627-35.

32. Friedman, N., *Inferring cellular networks using probabilistic graphical models.* Science, 2004. **303**(5659): p. 799-805.

33. Gardner, T.S., et al., *Inferring genetic networks and identifying compound mode of action via expression profiling.* Science, 2003. **301**(5629): p. 102-5.

34. Segal, E., et al., *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.* Nat Genet, 2003. **34**(2): p. 166-76.

35. Tegner, J., et al., *Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling.* Proc Natl Acad Sci U S A, 2003. **100**(10): p. 5944-9.

36. Yeung, M.K., J. Tegner, and J.J. Collins, *Reverse engineering gene networks using singular value decomposition and robust regression.* Proc Natl Acad Sci U S A, 2002. **99**(9): p. 6163-8.

37. Mwangi, M.M. and E.D. Siggia, *Genome wide identification of regulatory motifs in Bacillus subtilis.* BMC Bioinformatics, 2003. **4**: p. 18.

38. Rodionov, D.A., et al., *Reconstruction of regulatory and metabolic pathways in metal-reducing delta-proteobacteria.* Genome Biol, 2004. **5**(11): p. R90.

39. Rodionov, D.A., et al., *Dissimilatory metabolism of nitrogen oxides in bacteria: comparative reconstruction of transcriptional networks.* PLoS Comput Biol, 2005. **1**(5): p. e55.

40. Babu, M.M., B. Lang, and L. Aravind, *Methods to reconstruct and compare transcriptional regulatory networks.* Methods Mol Biol, 2009. **541**: p. 163-80.

41. Young, R.A., *Biomedical discovery with DNA arrays.* Cell, 2000. **102**(1): p. 9-15.

42. Eisenstein, M., *Microarrays: quality control.* Nature, 2006. **442**(7106): p. 1067-70.

43. Edgar, R., *Challenge of choosing right level of microarray detail.* Nature, 2006. **443**(7110): p. 394.

44. Brazma, A., et al., *Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.* Nature genetics, 2001. **29**(4): p. 365.

45. *Microarray standards at last.* Nature, 2002. **419**(6905): p. 323.

46. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Res, 2002. **30**(1): p. 207-10.

47. Brazma, A., et al., *ArrayExpress–a public repository for microarray gene expression data at the EBI.* Nucleic Acids Res, 2003. **31**(1): p. 68-71.

48. Velculescu, V.E., et al., *Serial analysis of gene expression.* Science, 1995. **270**(5235): p. 484-7.

49. Velculescu, V.E., et al., *Characterization of the yeast transcriptome.* Cell, 1997. **88**(2): p. 243-51.

50. Ren, B., et al., *Genome-wide location and function of DNA binding proteins.* Science, 2000. **290**(5500): p. 2306-9.

51. Iyer, V.R., et al., *Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.* Nature, 2001. **409**(6819): p. 533-8.

52. Roh, T.Y., et al., *High-resolution genome-wide mapping of histone modifications.* Nat Biotechnol, 2004. **22**(8): p. 1013-6.

53. Kim, T.H. and B. Ren, *Genome-wide analysis of protein-DNA interactions.* Annu Rev Genomics Hum Genet, 2006. **7**: p. 81-102.

54. Johnson, D.S., et al., *Genome-wide mapping of in vivo protein-DNA interactions.* Science, 2007. **316**(5830): p. 1497-502.

55. Nagalakshmi, U., et al., *The transcriptional landscape of the yeast genome defined by RNA sequencing.* Science, 2008. **320**(5881): p. 1344-9.

56. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Methods, 2008. **5**(7): p. 621-8.

57. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.

58. Herring, C.D., et al., *Immobilization of Escherichia coli RNA polymerase and location of binding sites by use of chromatin immunoprecipitation and microarrays.* J Bacteriol, 2005. **187**(17): p. 6166-74.

59. Wade, J.T., et al., *Extensive functional overlap between sigma factors in Escherichia coli.* Nat Struct Mol Biol, 2006. **13**(9): p. 806-14.

60. Grainger, D.C., et al., *Association of nucleoid proteins with coding and non-coding segments of the Escherichia coli genome.* Nucleic Acids Res, 2006. **34**(16): p. 4642-52.

61. Perkins, T.T., et al., *A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi.* PLoS Genet, 2009. **5**(7): p. e1000569.

62. Croucher, N.J. and N.R. Thomson, *Studying bacterial transcriptomes using RNA-seq*. Current opinion in microbiology, 2010.

63. Price, M.N., et al., *Evidence-based annotation of transcripts and proteins in the sulfate-reducing bacterium Desulfovibrio vulgaris Hildenborough.* J Bacteriol, 2011. **193**(20): p. 5716-27.

64. Faith, J.J., et al., *Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata.* Nucleic Acids Res, 2008. **36**(Database issue): p. D866-70.

65. Sherlock, G., et al., *The Stanford Microarray Database.* Nucleic Acids Res, 2001. **29**(1): p. 152-5.

66. Engelen, K., et al., *COLOMBOS: access port for cross-platform bacterial expression compendia.* PLoS One, 2011. **6**(7): p. e20938.

67. Barrett, T., et al., *NCBI GEO: archive for high-throughput functional genomic data.* Nucleic Acids Res, 2009. **37**(Database issue): p. D885-90.

68. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets–10 years on.* Nucleic Acids Res, 2011. **39**(Database issue): p. D1005-10.

69. Barrett, T., et al., *NCBI GEO: mining tens of millions of expression profiles–database and tools update.* Nucleic Acids Res, 2007. **35**(Database issue): p. D760-5.

70. Faith, J.J., et al., *Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.* PLoS Biol, 2007. **5**(1): p. e8.

71. Letunic, I. and P. Bork, *Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation.* Bioinformatics, 2007. **23**(1): p. 127-8.

72. Letunic, I. and P. Bork, *Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W475-8.

73. Gama-Castro, S., et al., *RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units).* Nucleic Acids Res, 2011. **39**(Database issue): p. D98-105.

74. Keseler, I.M., et al., *EcoCyc: a comprehensive database of Escherichia coli biology.* Nucleic Acids Res, 2011. **39**(Database issue): p. D583-90.

75. Okuda, S. and A.C. Yoshizawa, *ODB: a database for operon organizations, 2011 update.* Nucleic Acids Res, 2011. **39**(Database issue): p. D552-5.

76. Kazakov, A.E., et al., *RegTransBase–a database of regulatory sequences and interactions in a wide range of prokaryotic genomes.* Nucleic Acids Res, 2007. **35**(Database issue): p. D407-12.

77. Robison, K., A.M. McGuire, and G.M. Church, *A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome.* J Mol Biol, 1998. **284**(2): p. 241-54.

78. Jacques, P.E., et al., *MtbRegList, a database dedicated to the analysis of transcriptional regulation in Mycobacterium tuberculosis.* Bioinformatics, 2005. **21**(10): p. 2563-5.

79. Pauling, J., et al., *CoryneRegNet 6.0–Updated database content, new analysis methods and novel features focusing on community demands.* Nucleic Acids Res, 2012. **40**(Database issue): p. D610-4.

80. Wu, J., et al., *cTFbase: a database for comparative genomics of transcription factors in cyanobacteria.* BMC Genomics, 2007. **8**: p. 104.

81. Perez, A.G., et al., *Tractor_DB (version 2.0): a database of regulatory interactions in gamma-proteobacterial genomes.* Nucleic Acids Res, 2007. **35**(Database issue): p. D132-6.

82. Krawczyk, J., et al., *From Corynebacterium glutamicum to Mycobacterium tuberculosis–towards transfers of gene regulatory networks and integrated data analyses with MycoRegNet.* Nucleic Acids Res, 2009. **37**(14): p. e97.

83. Pareja, E., et al., *ExtraTrain: a database of Extragenic regions and Transcriptional information in prokaryotic organisms.* BMC Microbiol, 2006. **6**: p. 29.

84. Wilson, D., et al., *DBD–taxonomically broad transcription factor predictions: new content and functionality.* Nucleic Acids Res, 2008. **36**(Database issue): p. D88-92.

85. Grote, A., et al., *PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes.* Nucleic Acids Res, 2009. **37**(Database issue): p. D61-5.

86. Huang, H.Y., et al., *sRNAMap: genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes.* Nucleic Acids Res, 2009. **37**(Database issue): p. D150-4.

87. De Smet, R. and K. Marchal, *Advantages and limitations of current network inference methods.* Nat Rev Microbiol, 2010. **8**(10): p. 717-29.

88. Bansal, M., et al., *How to infer gene networks from expression profiles.* Mol Syst Biol, 2007. **3**: p. 78.

89. Karlebach, G. and R. Shamir, *Modelling and analysis of gene regulatory networks.* Nat Rev Mol Cell Biol, 2008. **9**(10): p. 770-80.

90. Madan Babu, M., S.A. Teichmann, and L. Aravind, *Evolutionary dynamics of prokaryotic transcriptional regulatory networks.* J Mol Biol, 2006. **358**(2): p. 614-33.

91. Madan Babu, M. and S.A. Teichmann, *Evolution of transcription factors and the gene regulatory network in Escherichia coli.* Nucleic Acids Res, 2003. **31**(4): p. 1234-44.

92. Teichmann, S.A. and M.M. Babu, *Gene regulatory network growth by duplication.* Nat Genet, 2004. **36**(5): p. 492-6.

93. Gelfand, M.S., *Evolution of transcriptional regulatory networks in microbial genomes.* Curr Opin Struct Biol, 2006. **16**(3): p. 420-9.

94. Lozada-Chavez, I., S.C. Janga, and J. Collado-Vides, *Bacterial regulatory networks are extremely flexible in evolution.* Nucleic Acids Res, 2006. **34**(12): p. 3434-45.

95. Overbeek, R., et al., *The use of gene clusters to infer functional coupling.* Proc Natl Acad Sci U S A, 1999. **96**(6): p. 2896-901.

96. Pilpel, Y., P. Sudarsanam, and G.M. Church, *Identifying regulatory networks by combinatorial analysis of promoter elements.* Nat Genet, 2001. **29**(2): p. 153-9.

97. Bar-Joseph, Z., et al., *Computational discovery of gene modules and regulatory networks.* Nat Biotechnol, 2003. **21**(11): p. 1337-42.

98. Alkema, W.B., B. Lenhard, and W.W. Wasserman, *Regulog analysis: detection of conserved regulatory networks across bacteria: application to Staphylococcus aureus.* Genome Res, 2004. **14**(7): p. 1362-73.

99. Rodionov, D.A., *Comparative genomic reconstruction of transcriptional regulatory networks in bacteria.* Chem Rev, 2007. **107**(8): p. 3467-97.

100. Tompa, M., et al., *Assessing computational tools for the discovery of transcription factor binding sites.* Nature Biotechnology, 2005. **23**(1): p. 137-144.

101. Roth, F.P., et al., *Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.* Nature Biotechnology, 1998. **16**(10): p. 939-945.

102. Chen, X., et al., *W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data.* Bioinformatics, 2008. **24**(9): p. 1121-1128.

103. Chen, X., et al., *Learning position weight matrices from sequence and expression data.* Comput Syst Bioinformatics Conf, 2007. **6**: p. 249-60.

104. Tan, K., et al., *A comparative genomics approach to prediction of new members of regulons.* Genome Research, 2001. **11**(4): p. 566-584.

105. Ravcheev, D.A., et al., *Comparative genomics analysis of nitrate and nitrite respiration in gamma proteobacteria.* Molecular Biology, 2005. **39**(5): p. 832-846.

106. Doroshchuk, N.A., M.S. Gelfand, and D.A. Rodionov, *Regulation of nitrogen metabolism in gram-positive bacteria.* Molecular Biology, 2006. **40**(5): p. 829-836.

107. Novichkov, P.S., et al., *RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W299-307.

108. Rodionov, D.A., et al., *Comparative genomic reconstruction of transcriptional networks controlling central metabolism in the Shewanella genus.* BMC Genomics, 2011. **12 Suppl 1**: p. S3.

109. Suvorova, I.A., et al., *Comparative genomic analysis of the hexuronate metabolism genes and their regulation in gammaproteobacteria.* J Bacteriol, 2011. **193**(15): p. 3956-63.

110. Leyn, S.A., et al., *Control of proteobacterial central carbon metabolism by the HexR transcriptional regulator: a case study in Shewanella oneidensis.* J Biol Chem, 2011. **286**(41): p. 35782-94.

111. Oberto, J., *FITBAR: a web tool for the robust prediction of prokaryotic regulons.* BMC Bioinformatics, 2010. **11**: p. 554.

112. McCue, L., et al., *Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes.* Nucleic Acids Res, 2001. **29**(3): p. 774-82.

113. Su, Z., et al., *Comparative genomics analysis of NtcA regulons in cyanobacteria: regulation of nitrogen assimilation and its coupling to photosynthesis.* Nucleic Acids Res, 2005. **33**(16): p. 5156-71.

114. Brazhnik, P., A. de la Fuente, and P. Mendes, *Gene networks: how to put the function in genomics.* TRENDS in Biotechnology, 2002. **20**(11): p. 467-472.

115. De Jong, H., *Modeling and simulation of genetic regulatory systems: a literature review.* Journal of computational biology, 2002. **9**(1): p. 67-103.

116. Stolovitzky, G., D. Monroe, and A. Califano, *Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference.* Ann N Y Acad Sci, 2007. **1115**: p. 1-22.

117. Marbach, D., et al., *Revealing strengths and weaknesses of methods for gene network inference.* Proc Natl Acad Sci U S A, 2010. **107**(14): p. 6286-91.

118. Stolovitzky, G., R.J. Prill, and A. Califano, *Lessons from the DREAM2 Challenges.* Ann N Y Acad Sci, 2009. **1158**: p. 159-95.

119. Hache, H., H. Lehrach, and R. Herwig, *Reverse engineering of gene regulatory networks: a comparative study.* EURASIP J Bioinform Syst Biol, 2009: p. 617281.

120. Michoel, T., et al., *Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks.* BMC Syst Biol, 2009. **3**: p. 49.

121. Elati, M. and C. Rouveirol, *Unsupervised Learning for Gene Regulation Network Inference from Expression Data: A Review.* Algorithms in Computational Molecular Biology, 2011: p. 955-978.

122. Cloots, L. and K. Marchal, *Network-based functional modeling of genomics, transcriptomics and metabolism in bacteria.* Curr Opin Microbiol, 2011. **14**(5): p. 599-607.

123. Cantone, I., et al., *A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches.* Cell, 2009. **137**(1): p. 172-181.

124. Gustafsson, M., et al., *Reverse engineering of gene networks with LASSO and nonlinear basis functions.* Ann N Y Acad Sci, 2009. **1158**: p. 265-75.

125. di Bernardo, D., et al., *Chemogenomic profiling on a genomewide scale using reverse-engineered gene networks.* Nature Biotechnology, 2005. **23**(3): p. 377-383.

126. Butte, A.J. and I.S. Kohane. *Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements*. 2000.

127. Meyer, P.E., et al., *Information-theoretic inference of large transcriptional regulatory networks.* EURASIP J Bioinform Syst Biol, 2007: p. 79879.

128. Yu, J., et al., *Advances to Bayesian network inference for generating causal networks from observational biological data.* Bioinformatics, 2004. **20**(18): p. 3594-3603.

129. Friedman, N., et al., *Using Bayesian networks to analyze expression data.* Journal of computational biology, 2000. **7**(3-4): p. 601-620.

130. Liang, S., S. Fuhrman, and R. Somogyi, *Reveal, a general reverse engineering algorithm for inference of genetic network architectures.* Pac Symp Biocomput, 1998: p. 18-29.

131. Hache, H., et al. *Reconstruction and validation of gene regulatory networks with neural networks*. 2007.

132. Grimaldi, M., G. Jurman, and R. Visintainer, *Reverse Engineering Gene Networks with ANN: Variability in Network Inference Algorithms.* Arxiv preprint arXiv:1009.4824, 2010.

133. Rice, J.J., Y. Tu, and G. Stolovitzky, *Reconstructing biological networks using conditional correlation analysis.* Bioinformatics, 2005. **21**(6): p. 765-773.

134. Butte, A.J. and I.S. Kohane, *Unsupervised knowledge discovery in medical databases using relevance networks.* Journal of the American Medical Informatics Association, 1999: p. 711-715.

135. Ihmels, J., et al., *Revealing modular organization in the yeast transcriptional network.* Nature genetics, 2002. **31**(4): p. 370-377.

136. Bonneau, R., *Learning biological networks: from modules to dynamics.* Nat Chem Biol, 2008. **4**(11): p. 658-664.

137. Reiss, D.J., N.S. Baliga, and R. Bonneau, *Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks.* BMC Bioinformatics, 2006. **7**.

138. Margolin, A.A., et al., *ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.* BMC Bioinformatics, 2006. **7**.

139. Segal, E., et al., *Learning module networks.* Journal of Machine Learning Research, 2005. **6**: p. 557-588.

140. Joshi, A., et al., *Module networks revisited: computational assessment and prioritization of model predictions.* Bioinformatics, 2009. **25**(4): p. 490-496.

141. Mordelet, F. and J.P. Vert, *SIRENE: supervised inference of regulatory networks.* Bioinformatics, 2008. **24**(16): p. i76-82.

142. Huynh-Thu, V.A., et al., *Inferring Regulatory Networks from Expression Data Using Tree-Based Methods.* PLoS One, 2010. **5**(9).

143. Bonneau, R., et al., *The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo.* Genome Biol, 2006. **7**(5): p. R36.

144. Lemmens, K., et al., *DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli.* Genome Biol, 2009. **10**(3): p. R27.

145. Ernst, J., et al., *A semi-supervised method for predicting transcription factor-gene interactions in Escherichia coli.* PLoS Comput Biol, 2008. **4**(3): p. e1000044.

146. You, Z.H., et al., *A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network.* BMC Bioinformatics, 2010. **11**: p. 343.

147. Cerulo, L., C. Elkan, and M. Ceccarelli, *Learning gene regulatory networks from only positive and unlabeled data.* BMC Bioinformatics, 2010. **11**: p. 228.

148. Marbach, D., et al., *Generating realistic in silico gene networks for performance assessment of reverse engineering methods.* J Comput Biol, 2009. **16**(2): p. 229-39.

149. Prill, R.J., et al., *Towards a rigorous assessment of systems biology models: the DREAM3 challenges.* PLoS One, 2010. **5**(2): p. e9202.

150. Greenfield, A., et al., *DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models.* PLoS One, 2010. **5**(10): p. e13397.

151. Basso, K., et al., *Reverse engineering of regulatory networks in human B cells.* Nat Genet, 2005. **37**(4): p. 382-90.

152. Castro-Melchor, M., et al., *Genome-wide inference of regulatory networks in Streptomyces coelicolor.* BMC Genomics, 2010. **11**: p. 578.

153. Glass, K., et al., *Implications of functional similarity for gene regulatory interactions.* J R Soc Interface, 2012.

154. Yoon, H., et al., *Systems analysis of multiple regulator perturbations allows discovery of virulence factors in Salmonella.* Bmc Systems Biology, 2011. **5**.

155. Watkinson, J., et al., *Inference of Regulatory Gene Interactions from Expression Data Using Three-Way Mutual Information.* Challenges of Systems Biology: Community Efforts to Harness Biological Complexity, 2009. **1158**: p. 302-313.

156. Anastassiou, D., *Computational analysis of the synergy among multiple interacting genes.* Molecular Systems Biology, 2007. **3**.

157. Madar, A., et al., *DREAM3: Network Inference Using Dynamic Context Likelihood of Relatedness and the Inferelator.* PLoS One, 2010. **5**(3).

158. Bonneau, R., et al., *The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo.* Genome Biology, 2006. **7**(5).

159. Yip, K.Y., et al., *Improved Reconstruction of In Silico Gene Regulatory Networks by Integrating Knockout and Perturbation Data.* PLoS One, 2010. **5**(1).

160. Ernst, J., et al., *A semi-supervised method for predicting transcription factor-gene interactions in Escherichia coli.* Plos Computational Biology, 2008. **4**(3).

161. Zwir, I., H. Huang, and E.A. Groisman, *Analysis of differentially-regulated genes within a regulatory network by GPS genome navigation.* Bioinformatics, 2005. **21**(22): p. 4073-4083.

162. Lemmens, K., et al., *DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli.* Genome Biology, 2009. **10**(3).

163. Geurts, P., A. Irrthum, and L. Wehenkel, *Supervised learning with decision tree-based methods in computational and systems biology.* Molecular Biosystems, 2009. **5**(12): p. 1593-1605.

164. Balaji, S., M.M. Babu, and L. Aravind, *Interplay between network structures, regulatory modes and sensing mechanisms of transcription factors in the transcriptional regulatory network of E-coil.* Journal of Molecular Biology, 2007. **372**(4): p. 1108-1122.

165. Schaffter, T., D. Marbach, and D. Floreano, *GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods.* Bioinformatics, 2011. **27**(16): p. 2263-2270.

166. Narendra, V., et al., *A comprehensive assessment of methods for de-novo reverse-engineering of genome-scale regulatory networks.* Genomics, 2011. **97**(1): p. 7-18.

167. Aziz, R.K., et al., *The RAST Server: rapid annotations using subsystems technology.* BMC Genomics, 2008. **9**: p. 75.

168. Kobayashi, K., et al., *Essential Bacillus subtilis genes.* Proc Natl Acad Sci U S A, 2003. **100**(8): p. 4678-83.

169. Commichau, F.M., N. Pietack, and J. Stulke, *Essential genes in Bacillus subtilis: a re-evaluation after ten years.* Mol Biosyst, 2013. **9**(6): p. 1068-75.

170. Debarbouille, M., et al., *The sacT gene regulating the sacPA operon in Bacillus subtilis shares strong homology with transcriptional antiterminators.* J Bacteriol, 1990. **172**(7): p. 3966-73.

171. Tortosa, P. and D. Le Coq, *A ribonucleic antiterminator sequence (RAT) and a distant palindrome are both involved in sucrose induction of the Bacillus subtilis sacXY regulatory operon.* Microbiology, 1995. **141 (Pt 11)**: p. 2921-7.

172. Pereira, Y., M.F. Petit-Glatron, and R. Chambert, *yveB, Encoding endolevanase LevB, is part of the sacB-yveB-yveA levansucrase tricistronic operon in Bacillus subtilis.* Microbiology, 2001. **147**(Pt 12): p. 3413-9.

173. Sterlini, J.M. and J. Mandelstam, *Commitment to sporulation in Bacillus subtilis and its relationship to development of actinomycin resistance.* Biochem J, 1969. **113**(1): p. 29-37.

174. Hansson, M., et al., *The Bacillus subtilis hemAXCDBL gene cluster, which encodes enzymes of the biosynthetic pathway from glutamate to uroporphyrinogen III.* J Bacteriol, 1991. **173**(8): p. 2590-9.

175. Fuangthong, M., et al., *Regulation of the Bacillus subtilis fur and perR genes by PerR: not all members of the PerR regulon are peroxide inducible.* J Bacteriol, 2002. **184**(12): p. 3276-86.

176. Fuangthong, M. and J.D. Helmann, *Recognition of DNA by three ferric uptake regulator (Fur) homologs in Bacillus subtilis.* J Bacteriol, 2003. **185**(21): p. 6348-57.

177. Gaballa, A., et al., *Functional analysis of the Bacillus subtilis Zur regulon.* J Bacteriol, 2002. **184**(23): p. 6508-14.

178. Gaballa, A. and J.D. Helmann, *Bacillus subtilis Fur represses one of two paralogous haem-degrading monooxygenases.* Microbiology, 2011. **157**(Pt 11): p. 3221-31.

179. Schock, F. and M.K. Dahl, *Expression of the tre operon of Bacillus subtilis 168 is regulated by the repressor TreR.* J Bacteriol, 1996. **178**(15): p. 4576-81.

180. Burklen, L., F. Schock, and M.K. Dahl, *Molecular analysis of the interaction between the Bacillus subtilis trehalose repressor TreR and the tre operator.* Mol Gen Genet, 1998. **260**(1): p. 48-55.

181. Henkin, T.M., *The role of CcpA transcriptional regulator in carbon metabolism in Bacillus subtilis.* FEMS Microbiol Lett, 1996. **135**(1): p. 9-15.

182. Prosser, G.A., A.V. Patterson, and D.F. Ackerley, *uvrB gene deletion enhances SOS chromotest sensitivity for nitroreductases that preferentially generate the 4-hydroxylamine metabolite of the anti-cancer prodrug CB1954.* J Biotechnol, 2010. **150**(1): p. 190-4.

183. Marbach, D., et al., *Wisdom of crowds for robust gene network inference.* Nat Methods, 2012. **9**(8): p. 796-804.

3.7 SUPPLEMENTAL MATERIAL

The supplementary material is available online at http://darwin.di.uminho.pt/jplfaria/phdthesis/Chapter_3_SupplMaterial.xlsx.

The following tables comprise the supplementary material:

Table S3.1 Manually curated regulatory network for *B. subtilis.*

Table S3.2 Regulators described in the regulatory network.

Table S3.3 Atomic regulons for *B. subtilis.*

Table S3.4. Curation of atomic regulons.

Table S3.5 Atomic regulons or *B. subtilis* (version 2).

Table S3.6 Manually curated regulatory network for *B. subtilis* (version 2)

CHAPTER 4

A GENOME-SCALE MODEL FOR THE METABOLISM AND TRANSCRIPTIONAL REGULATION OF *BACILLUS SUBTILIS*

Work presented in this chapter comprises the following articles:

Faria, J. P., Overbeek, R., Xia, F., Rocha, M., Rocha, I., & Henry, C. S.

Genome-scale bacterial transcriptional regulatory networks: reconstruction and integrated analysis with metabolic models. Brief Bioinform. doi: 10.1093/bib/bbs071

2014.

Faria, J. P., Rocha, M., Rocha, I., Henry, C. S.

A genome-scale model for the metabolism and transcriptional regulation of *Bacillus subtilis*

2015

(Manuscript in preparation)

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, where it is subject to a wide variety of external environmental conditions. This reinforces the importance of the regulatory mechanisms that allow bacteria to survive and adapt to such conditions.

In this chapter, we present the first attempt to simulate the metabolism and regulation of *B. subtilis* at genome-scale. Both expression data and a regulatory network were used to generate and impose regulatory constraints in the genome-scale metabolic model for *B.* subtilis. We validated our integrated model with mutant phenotypes described in the literature, considering the knockout of transcription factors represented in the regulatory network. The impact in our model of different environmental constraints was assessed across a sizable variety of growth media. The integrated regulatory and metabolic model was able to replicate the regulatory behavior described in the literature for different environmental constraints.

4.1 INTRODUCTION

Phenotype simulation using reconstructed biochemical networks has been one of the major goals and challenges of Systems Biology since the reconstruction of the first metabolic models [1-3]. At the same time, early works on the integration of metabolic networks with gene expression data revealed cellular phenotypes that cannot be described by the metabolic flux distribution itself [4]. The ultimate goal of whole cell modeling and simulation has been described as one of the great challenges of the century [5]. Integration of regulatory networks was identified as one of the key factors in achieving this goal [6]. Significant advances have been made in the reconstruction of metabolic, regulatory and signaling networks [7, 8], and in the integrated simulation of these three network types [9, 10]. However, we are still far from a whole-cell model. Here, we focus on the potential for the simulation of integrated metabolic and regulatory networks and the challenges which will arise as we attempt to achieve this objective [11].

The integration of regulatory and metabolic networks for predictive modeling is possible only with the development of integrated phenotype simulation methods. The most widely used approach for simulating genome-scale metabolic models (GEMs) is flux balance analysis (FBA) [12]. To account for regulatory information, FBA was expanded with new methodologies, including rFBA [13] and SR-FBA [14].

In this chapter, we introduce a genome-scale model that integrates the metabolic and regulatory networks of *B. subtilis*. This was achieved by making use of the manually curated regulatory network and gene expression data sets [15, 16], introduced in the previous chapter. We apply the probabilistic regulation of metabolism (PROM) [17] formulation to integrate the regulatory network and the gene expression data with the latest published genome-scale metabolic model of *B. subtilis* [18]. A previously existing model for *B. subtilis* metabolism and regulation only covered the central carbon metabolism [19].

To validate our model we ran *in silico* growth phenotype simulations for knockout strains to attempt to replicate mutant growth phenotypes described in the literature. This validation was performed across

multiple medium conditions to assess the ability of our model to represent regulatory effects imposed by different environmental constraints. The results showed both the ability of the model to represent the regulatory interactions described in the literature and limitations of the model and simulation framework. All the methods necessary for the work on this chapter were implemented in the DOE KnowledgeBase of Systems Biology (www.kbase.us) facilitating their use and analysis.

4.2 STATE OF THE ART

4.2.1 Constraint-based modeling

Several mathematical formalisms such as Boolean and Bayesian networks and constraint-based models (among others), have been applied to model different types of biochemical networks. Using these formalisms, the modeling community has developed different types of models, such as stoichiometric and kinetic models. Modeling approaches and mathematical formalisms for integrated metabolic and regulatory network reconstruction and analysis have been reviewed recently [20-23]. Here, we will focus on the methods described in the literature that can be applied in genome-scale, mainly stoichiometric and regulatory models using the constraint-based approach [24, 25]. Several efforts have been made recently to produce a genome-scale kinetic model of yeast metabolism [26], but the lack of data and complexity of these models has driven the community to primarily using constraint-based modeling [27].

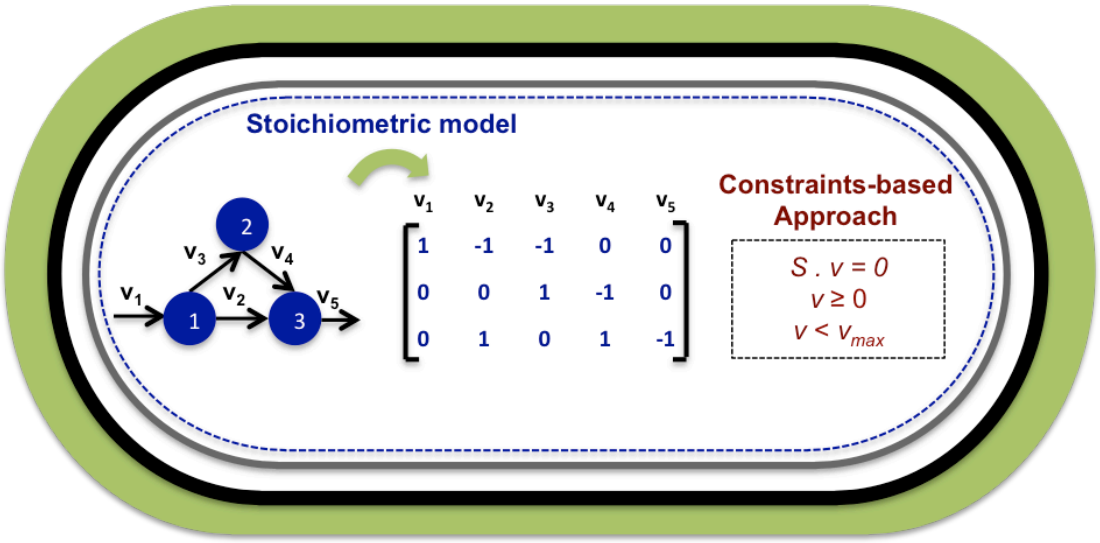


Figure 4.1 Stoichiometric modeling. The metabolic network is used to construct the stoichiometric matrix using mass balances of the metabolites. The constraint-based approach is used to impose constraints to the stoichiometric model. S.v = 0 – pseudo steady-state assumption; v > 0 – reversibility constraint; v < v_{max} – capacity constraint.

Constraint-based stoichiometric models do not account for intercellular dynamics, as they assume a pseudo steady state of the cell in which metabolite accumulation does not occur. That pseudo steady state is defined mathematically by a set of linear constraints over the fluxes through each metabolic reaction, defined by the mass balance around each internal metabolite (Figure 4.1):

$$S \cdot v = 0$$

where S represents the stoichiometric matrix and v represents the vector of fluxes through all metabolic reactions. The set of fluxes that satisfy the steady-state constraints define the feasible space for all reaction fluxes in the cell's metabolism. The constraint-based approach relies on the assumption that biological phenomena are coordinated by a set of constraints that limit and control their behavior. Constraints can be imposed on reaction reversibility and directionality ($v > 0$), on enzyme capacity ($v < vmax$), and on nutrient availability and uptake.

Several methods were developed to analyze and simulate phenotypes using constraint-based models. Most of those methods have initially been developed for stoichiometric metabolic models, but extensions have been made to accommodate constraints derived from regulatory interactions. Figure 4.2 shows existing methods for the analysis and simulation of integrated metabolic and regulatory networks. Global network analysis methods, such as Extreme Pathway Analysis [28], a pathway-based method [29], were developed to analyze specific pathway properties, such as length and redundancy, and were used successfully to characterize changes in the solution space with the addition of regulatory constraints [30].

Before methods like regulatory FBA (rFBA) [13], Steady-State Regulatory FBA (srFBA) [31], integrated FBA (iFBA) [9] or integrated dynamic FBA (idFBA) [10] can be applied, transcriptional regulatory networks (TRNs) must be translated into Boolean network models that connect external stimuli to internal metabolic reactions activity. The methodologies that make use of omics data use gene expression to impact reaction fluxes, without the need to develop Boolean gene regulatory network rules.

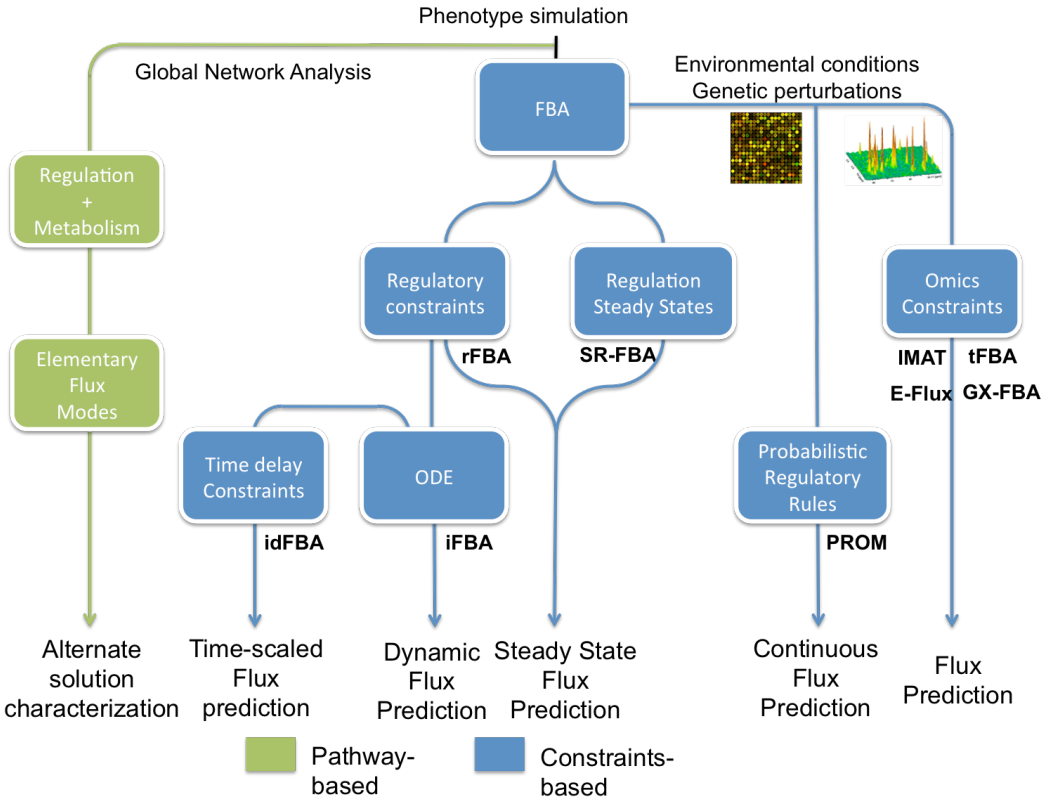


Figure 4.2 Pathway-based and Constraint-based methods for the analysis and simulation of integrated metabolic and regulatory networks. FBA (Flux Balance Analysis); rFBA (regulatory FBA); SR-FBA (Steady-State Regulatory FBA); idFBA (integrated dynamic FBA); iFBA (integrated FBA); PROM (Probabilistic Regulation of Metabolism); tFBA (transcriptional controlled FBA); iMAT (The integrative metabolic analysis tool); gene-expression data FBA (GX-FBA)

4.2.2 Simulation of integrated models

The FBA approach utilizes linear programming to identify the specific flux distributions that satisfy all problem constraints and best reflect the state of the cell [32, 33] (detailed FBA formulation is available in Section 4.3.3). FBA was expanded to account for regulatory information with the development of rFBA, which uses Boolean logic formalisms, as additional constraints that specify which genes in the network are ON or OFF based on specified stimuli (e.g. stress). This approach was successfully applied with the first genome-scale integrated model of metabolism and regulation in *E. coli*, resulting in the correction of several phenotype predictions compared with the use of mass balance and flux boundary

constraints alone [34]. However, this approach requires the integrated model to be initialized at a relevant state for the regulatory components of the system. The Boolean regulatory constraints are then applied to determine how the state of the regulatory components will change over time in response to stimuli. Selection of a relevant initial condition for the model remains a challenge for this methodology, since many equally consistent states exist for a set of stimuli, with equally valid associated flux distributions.

To address some of the limitations of rFBA, SR-FBA was introduced, differing from rFBA in that it accounts for metabolic and regulatory constraints in a single step and quantifies the impact of these constraints on the flux distribution. This methodology enables the rapid exploration of feasible combined regulatory and metabolic states and it rapidly identifies constraints that are internally inconsistent, preventing their simultaneous enforcement in a single steady-state. Yet, therein lies the substantial limitation of this approach, since inconsistent regulatory constraints often arise, because regulatory mechanisms exist to manage transitions between states of the cell in response to stimuli. Some of these transitions involve a cascade of intermediate unstable states that cannot be captured by the SR-FBA formalism. The constraints that manage these cascade transitions are not designed to be simultaneously enforced with all other constraints in the cell, meaning they appear to be internally inconsistent.

As more information became available, the quest for the whole-cell model moved the community efforts towards the development of methods that can also integrate signaling networks. Two methods have been proposed; iFBA [9] and idFBA [35]. iFBA is an expansion of the rFBA approach that aims to further integrate ordinary differential equations (ODEs) that might be associated with metabolic, regulatory or signaling networks. To perform their studies, an rFBA model for the central metabolism of *E. coli* [36] was combined with an ODE kinetic model for the phosphotransferase system (central metabolism). An algorithmic approach is suggested to integrate both models at different time steps. The first step involves the computation of regulatory constraints and numeric integration of the ODE model. Following this step, the FBA optimization is performed with specific boundaries to match/combine fluxes from the ODE integration. The last step comprises the update of external

metabolite concentrations and biomass. An important consideration is the length of the time-steps that have to be tuned properly to maintain the FBA pseudo steady-state assumption. The comparison with both individual rFBA and ODE models showed improved predictions. The authors suggested that prediction improvements arose from the improved accounting of internal metabolite concentrations enabled by iFBA.

idFBA [35] is an FBA-based approach for integrated analysis with a focus on the integration of signaling networks. The novelty of idFBA is the incorporation of slow and fast reactions into the stoichiometric framework on the three types of networks integrated. Slow reactions are incorporated directly into the stoichiometric matrix with a time-delay, while fast reactions rely on the pseudo steady-state assumption of the FBA approach. As with the iFBA approach, an algorithm is used to deal with the different time-steps of the integrated approach. idFBA was applied to the analysis of yeast metabolism, with a comparison performed on an integrated kinetic pathway model of *S. cerevisae* [37]. This analysis demonstrated an approximation for the time-course prediction of time-delayed reactions, with the advantage of requiring fewer measured parameters then full kinetic modeling.

4.2.3 Metabolic and regulatory modeling with omics data

Multiple methods have been proposed in the literature to integrate omics data with genome-scale metabolic models [17, 38-41]. A recent review evaluated the performance of methods that integrate transcriptomics data into constraint-based metabolic models [23]. In this review, Machado and Herrgård propose a categorization of this type of methods according to the way they use expression data. The distinction is made between methods that integrate discrete or continuous levels of expression, and between the use of absolute values for a single condition, or relative expression levels between different conditions. Due to the multitude of methods available in the literature, here we choose to review 5 widely used methods spanning across those proposed strategies for use of expression data.

PROM [42] was introduced as a constraint-based method for the generation of integrated models directly from high-throughput expression data. The first step of the framework introduced by PROM is the definition of the stoichiometric matrix for the metabolic network. The next is the evaluation of the regulatory structure from microarray data and regulatory interactions from databases such as RegulonDB (for *E. coli*). The regulatory interactions (gene-TF) and gene states are represented as probabilities inferred from the expression data. This concept of probabilities aims to circumvent the Boolean approaches that would consider a gene as either ON or OFF. When all the constraints are properly set, optimal cellular growth is calculated via linear programming using FBA. A more detailed description of the PROM methodology is available in section 4.3.1. PROM aimed to overcome limitations of the rFBA and SR-FBA models, and the authors showed a comparison study with PROM outperforming rFBA on phenotype prediction in *E. coli* for a validation set of 1875 growth phenotypes. The differences in the predictions are attributed to the Boolean formalism of rFBA that sets up more "rigid" flux restrictions, while PROM presents a more continuous flux restriction. The approach was also used for a reconstruction of *M. tuberculosis* and can be extended to other organisms when data are available. The reconstruction of the integrated model for *M. tuberculosis* also showed a potential use of PROM for drug target prediction. PROM can be seen as the closest methodology for semi-automated reconstruction of integrated metabolic and regulatory networks.

Transcriptional controlled FBA (tFBA) [43] is another method that uses experimental expression data for assessment of the regulatory state. Like PROM, tFBA authors aim to surpass the rigid ON/OFF gene states imposed by the Boolean formulation. They introduce the concept of up/down constraints, as being more relaxed than ON/OFF constraints. The concept of the new type of constraints lies in the fact that, as more experimental data are available, the level of expression of a gene can be observed to change under specific conditions. The authors refer to this assessment of expression levels as "relative" gene expression in an effort to effectively predict "relative" intercellular fluxes for all the pairs of conditions in the expression compendia. The regulatory constraints are defined between all pairs of conditions, generating an FBA model for each condition. This method shows how the addition of large

quantities of expression data can provide a way to improve FBA-based methods in the absence of kinetic parameters for metabolites and reactions.

The integrative metabolic analysis tool (iMAT) [40] is a web implementation of the method for integration of expression data with metabolic models originally proposed by Sholomi *et al*. [44]. The method was originally developed to find tissue-specific activities using the human metabolic genome-scale model, and was more recently applied to identify regulators of virulence in *Listeria monocytogenes* [45]. To account for the impact of non-transcriptional regulatory effects that are not represented by the expression data in the metabolic flux activity, expression levels do not determine directly an enzyme activity. In the iMAT framework reactions are classified based on associated gene-expression data as either highly expressed or lowly expressed. Flux maximization is performed to identify a possible steady-state flux distribution among those that maximize the number of reactions with predicted flux, consistent with the gene-expression data and the model stoichiometric constraints. As changes in gene expression levels seem to be key to control tissue specific metabolic functions [46], this method laid the foundations for rapid development of tissue-specific models.

E–Flux was introduced to make use of continuous expression levels to model the maximum flux possible for all reactions in the metabolic network. This method was originally used to predict mycolic acid production in *Mycobacterium tuberculosis* [38]. As a way to define maximum fluxes/bounds for all reactions, it normalizes the expression of each gene by the maximum gene expression level across all genes. An analogy can be made as "setting the width of pipes" in the network, as loose constraints (allowing for a higher possible flux) are applied to the flux of reactions encoded by highly expressed genes. In the same manner, tight constraints (limiting the maxing flux possible) are applied for reactions encoded by lowly expressed genes. FBA is then applied subject to the reaction bounds set by the expression levels of the genes associated with each reaction.

The gene-expression data FBA (GX-FBA) [39] is yet another method that incorporates expression data into the FBA formulation. It is similar to E-Flux as it makes use of continuous expression levels, but it makes use of relative expression between a reference/control condition and a perturbed condition, like tFBA. The method was proposed to enhance FBA with the ability to better predict responses of the cell

to the changes in the environment. It accomplishes this by first generating a wild type flux distribution for a reference/control condition. In a second step, the maximum and minimum fluxes possible for each reaction in the perturbed condition are determined and expression levels are used to constrain the reactions. A new objective function for the perturbed condition is then formulated with these constraints and the reference wild type flux distribution.

4.3 METHODS

After the review of methods available in literature, we chose PROM to perform our simulations of the metabolism and regulation. This choice was motivated by the results of the previous chapter in which we introduced a curated regulatory network for *B. subtilis*. Additionally, we also commented on the quality of a regulatory dataset that we used to infer regulatory interactions. The PROM framework uses gene expression data and the regulatory network interactions between transcription factor and target genes. The inference of the regulatory interactions from expression data under the PROM formalism allows us to simulate our regulatory network without having to build a Boolean network model that connects external stimuli to internal metabolic reactions activity. The reconstruction of Boolean network model simplifies regulatory interactions to a binary process, where genes can either be "ON" or "OFF" and logical functions, including the use of operators AND, OR and NOT, are used to represent the relationships between genes (e.g. between regulated genes and transcriptional factors) and between genes and stimuli. The generation of Boolean gene regulatory rules is an extremely time consuming task, since an algorithm to automate the inference of these rules is yet to be proposed in the literature. Manual reconstruction of these rules also limits the amount of regulatory interactions that can be modeled. This limitation has been described as the main reason for the existence of very few models that integrate metabolism and regulation in the literature [17]. In this context, PROM was considered the most suitable method to be applied to the data we have available. We also proposed modifications to the original PROM formulation, which can be found below in section 4.3.3.

4.3.1 Flux Balance Analysis (FBA) and Parsimonious Enzyme Usage FBA (pFBA)

FBA is a constraint-based simulation method used to define the limits on the metabolic capabilities of a microorganism as it calculates the flow of metabolites through the metabolic network. FBA is formulated as a linear programming problem that maximizes or minimizes a configured objective function. $v_{objective}$ specifies the flux being optimized, and the maximization of biomass production/growth rate is usually the choice [47].

$$maximize\ /\ minimize \quad v_{objective} \tag{4.1}$$

subject to:

$$\sum_j S_{ij} \cdot v_j = 0\,, \qquad \forall i \tag{4.2}$$

$$v_j^L \leq v_j \leq v_j^U\,, \qquad \forall j \tag{4.3}$$

where $v_j$corresponds to the flux of reaction j, and S_{ij}stands for stoichiometric coefficient of metabolite i in reaction j. The objective function (4.1) allows to calculate the steady-state fluxes that satisfy the stoichiometric constraints (4.2). Constraint (4.3) sets the upper (v_j^U) and lower (v_j^L) bounds on the individual fluxes.

Maximization of biomass production with FBA aims to represent the assumption that growth selection pressure will select for the fastest growing strains. In addition to that assumption, Lewis *et al.* [48] proposed that there would be a growth advantage to the more efficient cells using the least amount of enzymes [48]. This method was named parsimonious enzyme usage FBA (pFBA).

This approach employs a two-step linear program formulation to minimize enzyme-associated fluxes, subject to optimal biomass. In the first step, the FBA simulation is performed as described in (4.1-3); the constraints for the second step are detailed bellow:

$$Minimize \sum_j |vj| \tag{4.4}$$

Subject to

$$v_{objective} = \alpha\ max/\min v_{objective} \quad \alpha \in \mathbb{R} \tag{4.5}$$

Constraints (4.2) and (4.3)

where $v_{objective}$ is the optimized objective flux, α is a relaxing coefficient applied to the optimized objective (4.5).

4.3.2 Flux Variability Analysis (FVA)

Flux variability analysis (FVA) [49] allows to determine the range of permissible fluxes in the optimal solutions of a constraint-based analysis problem. Using FVA, we can determine the minimum and maximum possible flux through a reaction for a given growth rate. FBA is used to calculate the growth rate v_g^*, followed by FVA to assess the variability of fluxes in the network:

$$minimize\ /maximize\ \ v_r \tag{4.6}$$

subject to

$$v_{growth} \geq v_g^* \tag{4.7}$$

Constraints (4.2) and (4.3)

This process is typically repeated for all reactions *r* in the model.

4.3.3 Probabilistic Regulation of Metabolism (PROM)

The PROM methodology (as previously described in the State of the Art) makes use of probabilities to assess gene states and interactions between genes and transcription factors to enable the integration of regulatory and metabolic networks.

In order to apply the PROM methodology [50] three elements are necessary:

1) Metabolic network.
2) Transcriptional regulatory network with transcription factor and target gene interactions.
3) High throughput gene expression data

To produce our integrated model with the PROM formulation, we chose the GEM for *B. subtilis* *i*Bsu1103V2 [18]. The iBsu1103V2 is the second iteration of the *i*Bsu1103 model [51] with additional curation using a growth dataset of 157 gene deletion intervals. For the regulatory network component, we used our own manually curated model that was described on chapter 3. The high throughput gene expression data of choice was the high quality dataset proposed by Nicolas *et al.* and Buescher *et al.* [15, 16]. This dataset comprehends a huge variety of conditions as recommended for PROM. We extensively analyzed this dataset, and more details can be found on chapter 3.

PROM uses conditional probabilities for modeling transcriptional regulation and uses FBA for modeling metabolic networks. This methodology introduces probabilities to represent interactions between a transcription factor (TF) and the subsequent gene states. The probability of target gene (TG) "A" being active when the TF factor "B" is not active is represented by:

$$P(A = 1|B = 0) \quad (8)$$

while the probability of TG being active if TF is also active is:

$$P(A = 1|B = 1) \quad (9)$$

The information from the high throughput gene expression data is then used to determine the relationship between TFs and TGs. To assess this relationship, a preprocessing of the data is performed to discretize the data to ON and OFF states. This allows for the representation of all gene states as either ON or OFF. The probability of TG "A" being ON, when the TF factor "B" is OFF, is given by the number of times (*N*) that this combination of states was observed, over the total of times the TF was OFF in the expression data. This description is represented by the following formula:

$$P(A = 1|B = 0) = \frac{N(A = 1|B = 0)}{N(B = 0)} \quad (10)$$

For example, when observed that in 90% of the samples the TG is found to be ON when the TF is OFF, then the probability $P(A = 1|B = 0) = 0.9$. For TFs that affect multiple genes, this relationship is calculated for all its TGs. This information is used to constrain the fluxes of the reactions encoded by each TG. The flux through the reaction regulated by gene "A" (v_A) when its corresponding regulator "B "is turned OFF, is constrained by:

$$P \cdot v_A^L \leq v_A \leq P \cdot v_A^U \quad (11)$$

where P is the probability of the gene being active under the specific phenotype observed in the expression data. For irreversible reactions, only upper bounds are defined. Estimates of the reaction lower and upper bounds are given by running FVA on the metabolic network without the regulatory constraints.

The PROM algorithm is able to violate the regulatory constraints and exceed reaction bounds to maximize growth. This capability was implemented to set regulatory constraints as soft constraints to account for the inherent uncertainty that comes from experimental techniques, lack of knowledge of the regulatory mechanisms and non-transcriptional regulation. A penalty is applied to prevent this from happening often. The final formulation for the PROM model, is given by:

$$maximize \quad v_{objective} - \sum_j \kappa \cdot (\alpha_j + \beta_j) \quad (12)$$

subject to

$$\sum_j S_{ij} \cdot v_j = 0\,, \quad \forall i \quad (13)$$

$$P \cdot v_j^L - \alpha_j \leq v_j \leq P \cdot v_j^U + \beta_j \ , \quad \forall j \tag{14}$$

$$\alpha_j, \beta_j \geq 0 \ , \quad \forall j \tag{15}$$

The PROM objective function adds a term to the FBA objective function that represents the penalty for exceeding an upper or lower bound $\kappa \cdot (\alpha_j + \beta_j)$. It is subject to the same steady state constraints, where S_{ij} represents the stoichiometric coefficient of the metabolite i and reaction j in the network.

$P \cdot v_j^U$ and $P \cdot v_j^L$ represent the transcriptional regulation bounds, α_j and β_j are positive variables that allow the described violation of the reactions bounds. κ represents the cost of reaction bounds violation. The higher the value of k, the greater the constraint on the system based on transcriptional regulation. For values of k significantly greater than 1, the regulatory constraints become hard, and for values less than 0.1, they become insignificant. A k value of 1 is typically used to balance this effect on the regulatory constraints.

PROM also allows the incorporation of interactions for which strong evidence from the literature or experimentation exists. Probabilities can be manually set to assign 0 or 1 for a specific interaction, setting the corresponding TG to either fully active or completely inactive.

4.3.4 Modifications to the original PROM formulation

In this chapter, we applied PROM with 3 changes to the original methodology. First we decided to remove the penalty term $\kappa \cdot (\alpha_j + \beta_j)$ for the relaxation constraints that allows reaction bounds to be violated. As we experimented with the method, we noticed that the inclusion of this term resulted in unpredictable behavior from the method that led to prediction artifacts, including: (i) regulatory constraints being entirely ignored at random; and, (ii) suboptimal solutions being selected to exploit

weaker penalties on alternative fluxes. Generally, removing this term makes the regulatory constraints hard, allowing us to make a better assessment of the impact of the regulatory constraints on the metabolic network.

The second change introduced to the PROM algorithm was the use of an FBA formulation with minimization of fluxes (pFBA). The FBA objective function can return alternate optimal solutions, achieving the same growth rate with different flux distributions. [33]. The flux variability of alternate optimal solutions was also shown to be dependent on environmental conditions [49]. As we performed *in silico* simulations across approximately 100 different environments, it would be important to minimize these flux variability effects in our solutions. The use of pFBA provides the minimal flux distribution that complies with the constraints imposed in the FBA simulation, eliminating most of the variability.

The third change was introduced to properly model isoenzymes. In the methodology, there was no description of how the methodology handles flux restrictions when a TF knockout affects the flux of an isoenzyme. As we experimented with the methodology, we noted flux being restricted for a reaction encoded for a given isoenzyme when a TF was knocked out, when the regulator did not affect the other isoenzyme. To correct this effect, we adjusted the methodology to properly handle the restriction of reaction fluxes of isoenzymes when different regulators affect them.

4.4 RESULTS AND DISCUSSION

To analyze our integrated model of metabolism and regulation for *B. subtilis*, we conducted 2 different studies. In the first study (section 4.4.1), we searched SubtiWiki [52] for mutant phenotypes for regulators in our model and performed *in silico* simulations to attempt the validate those phenotypes. In a second study, we wanted to assess if our model was representing the regulatory interactions imposed by different environmental constraints for knockout strains. To make this assessment, we performed *in silico* knockout growth simulations for over 80 media conditions and for all regulators in our model. We then searched the literature to validate observed lethal phenotypes.

All the necessary methodologies to run the studies in this chapter were performed with the tools implemented on the DOE Knowledge Base of Systems Biology (www.kbase.us). A "reviewer" account was created to provide access to all data and simulation results performed. Additional information on how to access these results is available in section 4.7.

4.4.1 Model validation with transcription factor mutant phenotypes

To validate GEMs, it is common practice for researchers to make use of growth phenotypes for knockout strains [53]. The *i*Bsu1103V2 model was validated with an extensive dataset of multiple gene deletions. We validated our integrated model against this dataset and obtained the same results observed for *i*Bsu11033 (these results are available in Supplementary material S4.1). To validate the specific addition of regulatory constraints into the model, we adopted the same strategy to validate our integrated model, simulating mutant phenotypes for TFs included in the regulatory network. For this purpose, we performed a search for all regulators in our model in SubtiWiki, as it provides in the gene entries a report of observed mutant phenotypes in the literature. We were able to find mutant phenotypes for 35 TFs. A majority of those we are not able to simulate with our model, as they report non-metabolic defects in colony or biofilm formation, delayed sporulation, etc. From that list we picked 6 mutant phenotypes to validate with our model, which are shown in Table 4.1. The full list of phenotypes found in SubtiWiki is available in the Supplementary material S4.2.

Table 4.1 Transcription factor mutant phenotypes reported in SubtiWiki

TF KO	Locus ID	Observed phenotype	*In silico* phenotype
AlsR	BSU36020	No acetoin production [54]	True
CitT	BSU07590	Unable to grow with citrate as sole carbon source [55]	False
CcpN	BSU25250	Impaired growth on glucose [56]	False
CysL	BSU37650	Unable to grow with sulfate or sulfite as the sole sulfur source [57]	True*
GltC	BSU18460	Auxotrophic for glutamate [58]	True
PutR	BSU03230	Unable to grow with proline as single source of carbon or nitrogen [59]	True*

*Model curation necessary to achieve phenotype

We were able to simulate *in silico* the proposed phenotypes *in silico* Δ*alsR*, Δ*c*ysL, Δ*gltC* and Δ*putR*. For the knockouts of the CysL and PutR, additional model curation was necessary to achieve the observed growth phenotype. We were unable to simulate the phenotypes observed for CitT and CcpN. Here, we discuss in detail the regulatory mechanisms associated with each mutant phenotype and their simulation with our integrated model.

ΔalsR

AlsR is the regulator of acetoin synthesis and the mutant of this transcription factor was found to disrupt the production of acetoin in *B. subtilis* [54]. Acetoin is a major compound of interest for the industry as a flavor agent [60] and is produced by multiple microorganisms as a glycolytic product. Recently, efforts have been conducted in metabolic engineering of *B. subtilis* to increase the production of acetoin [61, 62]. Acetolactate synthase and acetolactate decarboxylase are enzymes involved in acetoin formation. The genes *alsS* and *alsD* have been found to encode these enzymes and are

regulated by AlsR [54]. In addition to the *alsD* and *alsS*, *alsR* was found to be essential for anaerobic expression of *lctE* and *lctP* [63].

To simulate this AlsR mutant phenotype *in silico* with our integrated metabolic regulatory model, we first assessed if the iBsu1103V2 was able to produce acetoin. The iBsu1103V2 has 2 reactions capable of producing acetoin. The model reaction ID, reaction name and model GPRs are show in Table 4.2.

Table 4.2 *i*Bsu1103 model reactions that can produce acetoin

Reaction ID	Reaction Name	GPR
rxn02112	Acetoin reductase/2,3-butanediol dehydrogenase	*bdhA*
rxn02113	Acetolactate decarboxylase	*alsD*

Alpha-acetolactate decarboxylase (rxn02113) is encoded by *alsD* (Table 4.2). The additional reaction (rxn02112) is associated with the production of 2,3-butanediol dehydrogenase from acetoin by fermentation [64]. We ran a wild-type FBA simulation with a rich media (LB) formulation to verify if the model can produce acetoin. We observe 0 flux through reaction rxn02113, meaning no acetoin is not being produced by the model. This result was not totally unexpected as acetoin is a fermentation product and the FBA simulation conducted has biomass maximization as its objective function. When maximizing biomass in a nutrient rich media, it is expected for the model not to produce some byproducts of fermentation during cellular growth.

Due to this fact, we decided to complement our FBA simulation with a FVA analysis. FVA minimizes and maximizes the fluxes through all reactions in the model, allowing us to see if rxn02113 is capable of carrying flux. Reaction rxn02113 was found to be able to carry a maximum flux of 23.3 mmol. gDW^{-1} .h^{-1} under the defined simulation conditions. After verifying the ability of the model to produce acetoin, the next step involved inspecting our PROM regulatory constraints. Table 4.3 shows the PROM constraints generated for the genes regulated by AlsR.

Table 4.3 PROM Constraints for AlsR regulated genes

Gene Name (TG)	Locus ID	Prob TG ON | TF OFF	Prob TG ON| TF ON
lctE	BSU03050	0.318	0.375
lctP	BSU03060	0.212	0.208
alsD	BSU36000	0.0980	0.375
alsS	BSU36010	0.0980	0.375

On Table 4.3, we have the probabilities for a target gene (TG) to be ON if the TF is ON or OFF. For our analysis, we can see that the genes encoding acetolactate synthase (*alsS*), and acetolactate decarboxylase (*alsD*) have an approximate probability of zero to be ON, when AlsR is knocked out. A probability of 0 means the reactions encoded by those genes will be OFF in the metabolic model. To simulate the AlsR knockout *in silico* and verify if the model becomes unable to produce acetoin, we ran a PROM simulation with FVA. The results show that rxn02113 is unable to carry flux and the FVA results also show maximum flux of 0, validating the phenotype observed in the literature.

ΔcitT

The second phenotype listed on Table 4.1 reports no growth with citrate as sole carbon source for the CitT mutant. *citT* and *citS* genes compose a two-component system [65] that was shown to positively regulate the expression of *citM* [55]. The gene *yflN* was found to be polycistronically transcribed with *citM,* and these two genes represent the *citM-yflN* operon. As part of the two-component system, *citS* acts as the sensor kinase and *citT* as the response regulator. In *B. subtilis*, *citM* encodes the transport of the citrate-Mg complex [66]. The function of the *yflN* is currently unknown.

The first step for validation of this phenotype was to inspect both the metabolic and regulatory models to ensure that interactions described in the literature are being captured by our integrated model. The analysis of the GPRs for *yflN* and *citM* revealed that no reaction in the iBsu1103 model is associated with *yflN*. *citM* is associatd with Citrate-Mg^{2+} :H^{+} symporter reaction (rxn05214). The sensor kinase *citS* is also not represented in the metabolic model. Our regulatory model captured the regulation of the yflN-citM operon by CitT and the respective PROM constraints are shown on Table 4.4.

Table 4.4 PROM constraints for CitT regulated genes

Gene Name (TG)	Locus ID	Prob TG ON | TF OFF	Prob TG ON | TF ON
yflN	BSU07620	0.0792	0.333
citM	BSU07610	0	0.310

The PROM constraints shown on Table 4.4 reveal that the knockout of CitT is turning *citM* ON with a probability 0, and this will subsequently cause reaction rxn05214 to be unable to carry flux, stopping the activity of the Citrate-Mg symporter.

Table 4.5 Flux through citrate transport model reactions in the *ΔcitT* with citrate as sole carbon source

Reaction	Flux	Equation	GPR
rxn05211	0	Citrate[e] + H^+[e] <= Citrate[c] + H^+[c]	*cimH*
rxn05213	13.4	Citrate[e] + H^+[e] + Ca^{2+}[e] <=> Citrate[c] + H^+[c] + Ca^{2+}[c]	*citH*
rxn05214	0	Citrate[e] + H^+[e] + Mg[e] <=> Citrate[c] + H^+[c] + Mg[c]	*citM*
rxn05557	0.003	Fe^{3+}[e] + Citrate[e] + ATP[c] + H_2O[c] => Fe^{3+}[c] + H^+[c] + Citrate[c] + Phosphate[c] + ADP[c]	(*yfiY* AND *yfiZ* AND *yfhA*) OR *yusV*

All fluxes are in mmol. gDW^{-1} $.h^{-1}$

To simulate this phenotype *in silico* we created a minimal media formulation containing only citrate as carbon source and ran a PROM simulation with TF CitT knocked out. The results show that the model was able to obtain growth under these conditions. Since the PROM constraints are turning OFF the Citrate-Mg symporter, we inspected the PROM results for citrate transport in the model. We found 4 reactions (Table 4.5) capable of transporting citrate into the cell.

As expected, no flux is observed through reaction rxn05214 encoded by *citM.* Citrate is entering the cell via a reaction encoded by *citH*. The gene *citH* was found to be a homologous gene of *citM* sharing 60% of its identity [66]. The two transporters differ in cation specificity and preference. The cation specificity of *citM*, in order of preference is Mg^{2+}, Mn^{2+}, Ba^{2+}, Ni^{2+}, Co^{2+}, Ca^{2+} and Zn^{2+}[67]. For *citH*, the cation specificity in order of preference, is Ca^{2+}, Ba^{2+} and Sr^{2+} [68]. These transporters belong to the secondary transport of metal-citrate complexes, the CitMHS family [69]. As said, mutant strains of *citM* were found not to be able to grow with citrate as sole carbon source suggesting that *citH* is not able to uptake citrate under these conditions. [55]. Citrate uptake by *citH* was found to be inhibited in the presence of Mg^{2+}[66] To understand how the model is handling citrate transport, we performed mutant simulations for *ΔcitT, ΔcitM, ΔcitH*. Those results are shown on Table 4.6. Additional transport reactions of the cations involved in the metal-citrate transport complexes (Ca^{2+} and Mg^{2+}) are also shown on the table.

Table 4.6 Fluxes through the citrate, Ca^{2+} and Mg^{2+} model transport reactions for the wild-type, *ΔcitT, ΔcitM, ΔcitH* mutants with citrate as sole carbon source.

Reaction ID	WT*	*ΔcitT*	*ΔcitM*	*ΔcitH*	Equation
rxn05211	0	0	0	0	Citrate[e] + H^+[e] <= Citrate [c] + H^+[c]
rxn05213	13.3	13.4	13.4	0	Citrate[e] + H^+[e] + Ca^{2+}[e] <=> Citrate[c] + H^+[c] + Ca^{2+}[c]
rxn05214	0.099	0	0	0	Citrate[e] + H^+[e] + Mg[e] <=> Citrate[c] + H^+[c] + Mg[c]
rxn05557	0.003	0.003	0.003	0	Fe^{3+}[e] + Citrate[e] + ATP[c] + H_2O[c] => Fe^{3+}[c] + H^+[c] + Citrate[c] + Phosphate[c] + ADP[c]
rxn05513	0	0	0	0	ATP[c] + Ca^{2+}[c] + H_2O[c] => H^+[c] + Ca^{2+}[e] + Phosphate[c] + ADP[c]
rxn05514	13.3	13.4	13.4	0	H^+[e] + Ca^{2+}[c] <=> H^+[c] + Ca^{2+}[e]
rxn05616	0	-0.099	-0.099	0	Mg [c] <= Mg [e]

*Wild-type

All fluxes are in mmol. gDW^{-1} $.h^{-1}$

Analyzing the wild type flux distribution through all transport reactions, we can see that rxn05213, the Citrate-Ca symporter (previously described as encoded by *citH*) is the preferred reaction for citrate transport into the cell. This fact appears to be caused by a cycle of Ca^{2+}. Ca^{2+} is co-transported with citrate the cell via the Citrate-Ca symporter (rxn05213) and excreted back to the extracellular environment via an antiport transport mechanism (rxn05514). Flux through rxn05214 encoded by *citM* is very low. The flux distribution for *ΔcitT* and *ΔcitM* was found to be the same as expected. With the knockout of *citM* we observe the transport of Mg to be conducted via a uniport transporter (rxn05616). The knockout of *citH* was found to be lethal as we see no flux through all reactions in mutant *ΔcitH*. As these results contradict the experimental observations, we formulated 2 hypotheses. The first is that an additional transporter of Mg is not described in the model, to allow for a similar functionality as the calcium transporters. The second hypothesis is that an unknown regulatory effect, or the reported inhibition of *citH* in the presence of Mg, may block the ability of this citrate transporter to function.

ΔccpN

Mutants of the TF CcpN were reported to show impaired growth on glucose [56]. CcpN has been described as a regulator of carbon catabolite repression (CCR) in *B. subtilis* [70]. CCR repression via CccN was shown to be independent of CcpA, the major regulator for CCR in B. *subtilis* [71]. CcpN regulates the activity of the genes *gapB* and *pckA*. These encode the gluconeogenic enzymes glyceraldehyde-3-phosphate dehydrogenase and phosphoenolpyruvate carboxykinase, respectively. As a transcriptional repressor, CcpN prevents fluxes through these enzymes, in the presence of glucose (and other glycolytic substrates). Additionally, CcpN was found to regulate the activity of the small noncoding regulatory RNA sr1 [72]. This regulatory RNA is involved in arginine catabolism and was found to have a minor impact in the phenotype exhibited by the CcpN mutants [64]. The impaired growth on glucose shown by the CcpN mutant is due to a shift in the internal fluxes from glycolysis to the pentose-phosphate pathway. This shift of internal flux to the pentose-phosphate pathway is attributed to the derepression of *pckA,* activating the activity of phosphoenolpyruvate carboxykinase (EC

4.1.1.49). Analyses of the fluxomic data [56] revealed an extensive futile cycling through the pyruvate kinase (EC 2.7.1.50), pyruvate carboxylase (EC 6.4.1.1) and phosphoenolpyruvate carboxykinase (EC 4.1.1.49). Figure 4.3 illustrates the futile cycle through these reactions.

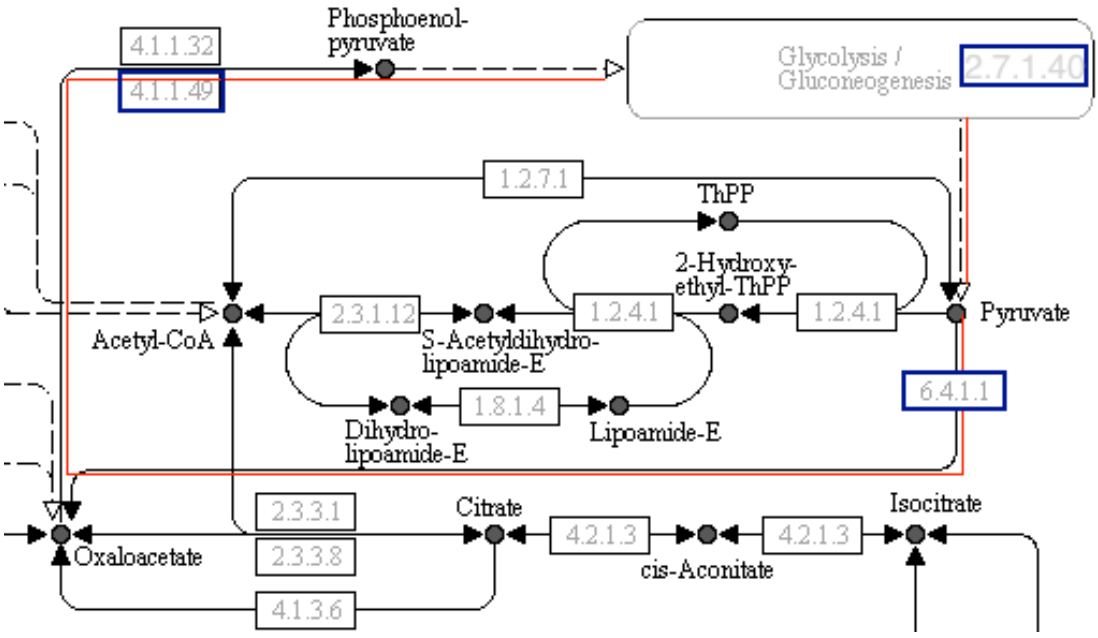


Figure 4.3 Partial KEGG metabolic map of the citric acid cycle (TCA). The futile cycle through pyruvate kinase (EC 2.7.1.50), pyruvate carboxylase (EC 6.4.1.1) and phosphoenolpyruvate carboxykinase (EC 4.1.1.49) is highlighted in red in the metabolic map.

The futile cycle causes dissipation of ATP and causes a drain of the citric acid cycle (TCA) intermediates leading to the reduce growth of a ccpN mutant. The high flux through the PP pathway in the CcpN mutant is modulated by the flux through the glyceraldehyde-3-phosphate dehydrogenases, *gapA* and *gapB*. The derepression of *gapB* was shown to increase the concentration of intermediates in upper glycolysis indicating that *gapB* overexpression leads to a metabolic jamming of this pathway and the observed increased flux through the pentose-phosphate pathway [56].

The analysis of the PROM constraints (Table 4.7) for the genes regulated by CcpN shows that the probability of all target genes being ON is 1 if the TF is OFF. These results seem to be capturing the derepression effect of the CcpN knockout to turn ON the gluconeogenic enzymes encoded by *gapB* and *pckA*.

Table 4.7 PROM constraints for CcpN regulated genes

Gene Name (TG)	Locus ID	Prob TG ON | TF OFF	Prob TG ON | TF ON
sr1	BSU14629	1	0.337
gapB	BSU29020	1	0.288
pckA	BSU30560	1	0.397

When we perform a PROM simulation in a medium with glucose, we do not observe any impaired growth rate when compared with the wild type. This fact shows us a limitation of the FBA (used by the PROM) methodology to simulate the activation of an enzyme, in this case phosphoenolpyruvate carboxykinase. With the objective function set to growth maximization, the FBA solution does not return an optimal that would be caused by a futile cycle. In addition to this limitation, we also cannot simulate the effect of accumulation of upper glycolysis intermediates, as the steady-state assumed in our simulation does not allow the accumulation of metabolites.

ΔcysL

The TF CysL was described as a regulator involved in cysteine biosynthesis in *B. subtilis* [57]. CysL regulates the activity of the *cisJI* operon, which was found to be part of the sulfate reduction pathway encoding sulfite reductase [73]. The sulfite reductase enzyme catalyses the reaction responsible for the reduction of sulfite to sulfide, one of the substrates of cysteine biosynthesis. The sulfate reduction pathway is represented on Figure 4.4. The mechanisms previously described are highlighted in the red dotted box.

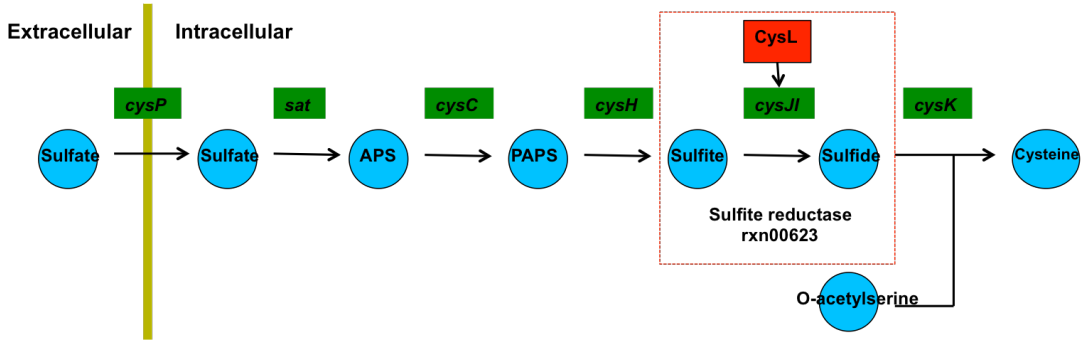


Figure 4.4 Sulfate reduction pathway. The sulfite reductase encoded by *cysJI* and regulated by the TF CysL is highlighted in the red dotted box.

CysL mutants are unable to grow with sulfate or sulfite as sole sulfur sources, as the cell is not capable of producing cysteine in these conditions [57]. Additionally, researchers reported that the CysL mutant was able to grow on other sources of sulfate, such as cysteine and methionine. To simulate the observed phenotypes *in silico*, we created four minimal media formulations varying only the sulfur source. The results of the PROM simulation under these conditions are shown on Table 4.8.

Table 4.8 Model growth on different sulfur sources for wild type and CysL mutant

	*i*Bsu1103V2 + PROM Constraints		*i*Bsu1103V3 + Adjusted PROM constraints	
Sulfur source	WT	*ΔcysL*	WT	*ΔcysL*
Sulfate	0.619	0.619	0.619	0
Sulfite	0	0	0.619	0
Cysteine	1.11	1.11	1.11	1.11
Methionine	1.19	1.19	1.19	1.19

Growth rate unites are given in h^{-1}

The *i*Bsu1103V2 was able to grow with sulfate, cysteine and methionine as unique sulfur sources, but not sulfite. We inspected the model and noted the absence of a mechanism of transport for sulfite, lacking a transporter to the intracellular environment. We added the following sulfite passive transport reaction to the model:

rxn12453: Sulfite[c] <=> Sulfite[e]

For *ΔcysL*, we observe the same behavior as in the wild type, not being in accordance with the experimental observations. To evaluate the cause for growth on sulfate in *ΔcysL* we inspected the PROM constraints in our model (Table 4.9). The PROM constraints show a high probability (~80%) for the *cysI and cysJ* to be ON if the TF CysL is knockout out, thus the false positive result.

Table 4.9 PROM constraints for CysL regulated genes

Gene Name (TG)	Locus ID	Prob TG ON | TF OFF	Prob TG ON | TF ON
cysI	BSU33430	0.816	1
cysJ	BSU33440	0.781	0.923

The PROM constraints are generated automatically from the expression data. The probabilities for a TF knockout are given from the estimation of the number of microarray samples in which the target gene was ON when the TF is OFF. The PROM constraints were inferred from a rich expression data set, with several different experimental conditions with multiple rich and minimal media. The high probability leads us to believe that in a majority of experiments (especially rich media) multiple sulfur sources were available rather than just sulfate or sulfite. To compensate for this fact, we manually adjusted the PROM constraints to 0 and ran the previous simulations again with the model including the sulfite transport reaction (*i*Bsu1103V3). The results are shown on Table 4.8. With the new manually curated model and regulatory constraints, we are able to observe the phenotypes described in the literature. The addition of the transport reaction allowed the model to grow on sulfite and the manual tune of the PROM constraints did not affect growth on other sulfur sources.

ΔgltC

GltC is a regulator of the *gltA-gltB* operon, which encodes the enzyme glutamate synthase in *B. subtilis* [58]. Glutamate synthase plays a major role as a link between the carbon and nitrogen metabolism [74]. The regulation of glutamate synthase was found to be nutrient dependent [75]. When grown in

complex media, alternative pathways such as amino acid degradation pathways can be used to produce glutamate [76]. In minimal media, when no source of glutamate is present in the media, it was found essential under the regulation of GltC [58, 77]. The different mechanisms for glutamate biosynthesis are described in Figure 4.5.

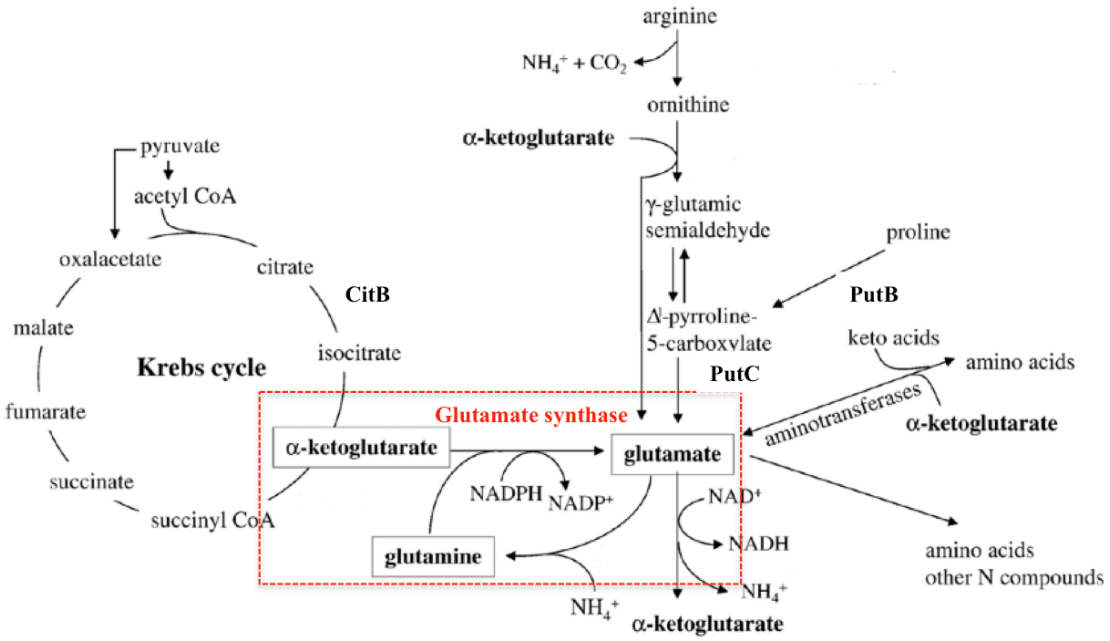


Figure 4.5 Glutamate biosynthesis. Glutamate synthase enzyme is highlighted in the red dotted box. Figure adapted from Picossi *et al.* [75].

To determine if our integrated model is able to replicate the phenotypes described in the literature, we performed wild type and *ΔgltC* simulations for 3 media formulations: a rich media (Luria-Bertani) and two minimal media with glucose and glutamine as sole carbon sources. The results are shown onTable 4.10.

Our integrated model was able to grow in the rich media for both for the wild type and *ΔgltC*. Regarding the minimal media formulations, we were able replicate the dependency of the *ΔgltC* on a source of glutamate, in this case glutamine.

Table 4.10 Model growth on LB, glucose minimal and glutamine minimal media for wild type and CysL mutant

Media	WT	*ΔgltC*
LB	30.5	30.5
Glucose minimal	0.629	0
Glutamine minimal	0.429	0.429

Growth rate units are given in h^{-1}

ΔputR

PutR is a regulator of the *putBCP* operon, responsible for proline utilization in *B. subtilis* [78]. The *putBCP* operon genes encode proline uptake and two-step oxidation of proline to glutamate. Its activity was found to be essential with proline as sole source of carbon and nitrogen [59]. Proline dehydrogenase (*putB*), 1-pyrroline-5-carboxylate dehydrogenase (*putC*), and a proline uptake protein (*putP*), are encoded by the genes in this operon. With proline as sole carbon and nitrogen source, the activity of this operon becomes essential, as it becomes the only pathway available for glutamate production. Mutants of the glutamate dehydrogenase were found to be able grow on these conditions [79]. To assist in the understanding of this mechanism, we represented *putB* and *putC* on Figure 4.5. To assess if our model is able to represent the behavior reported in literature, we performed wild-type and mutant simulations for a proline minimal medium (proline as sole source of carbon and nitrogen) and for a glucose minimal medium supplemented with proline. The results are shown on Table 4.11.

Table 4.11 Model growth on proline minimal and glucose minimal (supplemented with proline) media for wild type and PutR knockout mutant

Media	WT	*ΔputR*	*ΔputRΔrocA*
Proline minimal	0.429	0.429	0
Glucose minimal (supplemented with proline)	1.23	1.23	0.632

Growth rate unites are given in h^{-1}

We observed no change in the predicted growth of in the *ΔputR*, when compared with the wild type. As we did in previous mutant phenotypes studies, we inspected the PROM constraints to verify if the knockout effect was being captured by the automatically inferred regulatory constraints. The analysis of the regulatory constraints revealed that PROM assigned a probability of 0.1 to the genes encoding the reaction of the two-step oxidation of proline to glutamate.

Table 4.12 PROM constraints for PutR regulated genes

Gene name (TG)	Locus ID	Prob TG ON | TF OFF	Prob TG ON | TF ON
putB	BSU03200	0.122	0.778
putC	BSU03210	0.102	0.778
putP	BSU03220	0.594	0.833

With such low probabilities restricting the reaction fluxes encoded by *putB* and *putC,* a very reduce growth or knockout phenotype was expected. When we inspected the model for the reactions encoded by *putB* and *putC* we noted an extra GPR association for the reaction encoded by *putC*. The gene *rocA* also catalises an 1-pyrroline-5-carboxylate dehydrogenase, making *putC* and *rocA* isoenzymes. Induced by arginine, *rocA* encodes a 1-pyrroline-5-carboxylate dehydrogenase as the third step in the arginine degradation pathway, also leading to the production of glutamate [80]. The activity of *rocA* requires induction by arginine, thus its *in vivo* inactivity during growth on proline minimal media. On the other hand, the representation of the isoenzymes with the logic "OR" in the model allows *rocA* to encode the reaction associated with 1-pyrroline-5-carboxylate dehydrogenase, when we perform the knockout of *putR* that only inactivates *putC*. To attempt to simulate the phenotype described in the literature, we performed a double knockout *ΔputRΔrocA* simulation. The results on Table 4.11 show the expected lethal phenotype for the proline minimal medium and a reduced growth for the glucose minimal medium supplemented with proline. To properly simulate this phenotype with our model, an additional regulatory constraint, accounting for this information on the arginine dependency of *rocA*, would have to be added to the model. This is a limitation of PROM that infers regulatory interactions from gene expression data and does not account for the presence of specific substrates in the medium. Manual

addition of a Boolean gene regulatory constraint to connect external stimuli in the medium (in this case arginine) to internal metabolic reaction activity would allow to proper simulate this phenotype. This is a direct consequence of our decision to infer of regulatory interactions from expression data in detriment of manual design of Boolean gene regulatory rules.

4.4.2 Impact of different environmental conditions

In the previous study, we validated our model with growth phenotypes for 6 TF knockouts. For those 6 phenotype studies, we performed simulations across different media as the regulators knocked out responded to different environmental conditions. Inspired by those findings, we decided to perform *in silico* mutant simulations for all regulators in our model across different medium conditions. We chose 4 different medium formulations, 2 minimal and 2 rich media commonly used for bacterial cultures. The composition of the 4 media is available on Table 4.13.

Table 4.13 Bacterial culture growth medium composition

M9	GMM	NMS	LB
Citrate, Molybdate, Na^+, NH_3, Ni^{2+}, Sulfate	L-Tryptophan, Na^+, NH_3, Sulfate Glutamine	Biotin, Citrate, H_2S, Molybdate, Niacin, Pyridoxal, Riboflavin, Thiamin, Vitamin B12, Amino acids[1]	Adenosine, Arsenate, Cd^{2+}, Chromate, CMP, Deoxyadenosine, Deoxycytidine, Folate, GMP, Heme, Hg^{2+}, Hypoxanthine, Inosine, Na^+, Niacin, Pyridoxal, Riboflavin, Shikimate, Sulfate, Thiamine phosphate, Thymidine, Uracil, Uridine, Amino acids[2]
Common compounds			
Mn^{2+}, Cl^-, Fe^{2+}, Ca^{2+}, D-Glucose, Cu^{2+}, H_2O, Mg, K^+, Phosphate, Fe^{3+}, Zn^{2+}, O_2, Co^{2+}			

1 All amino acids but glycine

2 All amino acids but asparagine

We choose as minimal medium the M9 salts minimal medium supplemented with citrate, and glucose minimal medium (GMM). We also performed *in silico* knockouts for two variations of the GMM, one

variation without oxygen to simulate anaerobic environments, and another replacing ammonia with glutamine. As rich media, we opted for nitrate mineral salts (NMS) and Luria-Bertani (LB). The results of the knockout for all regulators in our model are available in the Supplementary material S4.3. On Figure 4.6, we show the results for the knockouts that have reduced growth of 10% or more, when compared with the growth achieved in the wild type.

Analyzing the results on Figure 4.6, we see that the first 5 gene knockouts are unable to grow in any medium. These five genes (*dnaA, birA, hbs*, *rplT* and *walR*) have been described in the literature as essential genes in rich LB medium [81] and that are thus also essential in more strict conditions. We also found mutants that show shifts on growth behavior on minimal media (*citB*, *tnrA* and *PhoP* mutants) that will be further analysed below.

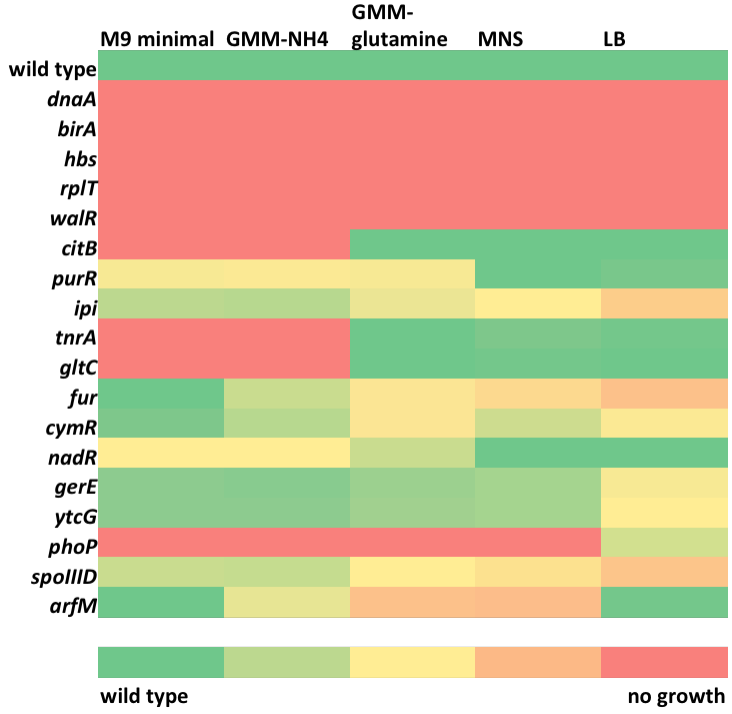


Figure 4.6 *In silico* gene knockouts for 5 different bacterial growth media. M9 salts minimal media, glucose minimal media (GMM), GMM with glutamine (instead of ammonia), nitrate mineral salts (NMS) and Luria-Bertani (LB). Red represents a lethal phenotype.

The mutant of CitB is only able to grow in GMM-Glu and rich media compositions. CitB has been described as a trigger enzyme [82], meaning it is able to act as regulator in addition to its metabolic capabilities. CitB encodes the TCA enzyme aconitase that convert citrate to isocitrate (see Figure 4.5) requiring an iron-sulfur cluster for its activity. CitB was also found to be able to bind iron responsive elements to genes controlling iron homeostasis, when iron is limited [83]. The lethal phenotype we are observing is due to its enzymatic activity as the interruption of the TCA cycle makes the cell auxotrophic for glutamate [84], thus the growth on media conditions that have glutamine precursors. The mutant of GltC as we discussed in the previous section is also auxotrophic for glutamate.

TnrA is the regulator responsible for the global regulation of nitrogen assimilation in *B. subtilis* [85]. In conditions of nitrogen limitation, TnrA is responsible for activating the expression of genes involved in the assimilation of various nitrogen sources. TnrA mutants were found not to be able to assimilate sources poor in nitrogen such as allantonin, nitrate or urea [85], but no effect of the mutation was observed for cultures grown on medium where ammonium or glutamine were added [86]. Our *in silico* simulation the TnrA mutant showed no growth defect with glutamine as sole nitrogen source, but was unable to grow in ammonium as sole nitrogen source. As the global regulator of the nitrogen source metabolism, TrnA is affecting the regulation of multiple operons, either as an activator or a suppressor, being involved in the regulation of over 100 genes. In our *in silico* knockout study, all these regulatory effects are simulated at once, not properly capturing the complexity of the regulation by TnrA. This fact makes the knockout of the TrnA impact the flux of over 100 reactions in our network, which may be causing a flux restriction that makes the model unable to grow with ammonium as sole nitrogen source.

This seems to be the case for another global regulator, PhoP, responsible for the regulation of the phosphate metabolism in conditions of phosphate limitation [87]. Studies on the regulation of PhoP described a very complex regulatory network affected by stresses other than phosphate starvation and post-exponential-phase processes that influence the expression of PhoP [88]. In our *in silico* simulation, we observe no growth in all media but LB, with no media conditions representing a scenario of phosphate starvation. Our inability to proper represent these growth phenotypes may be due to the

complexity of the regulatory mechanism, and also possibly due to heavy flux restrictions as we hypothesized for the TnrA mutant. Other global regulators and regulators affecting a large number of genes are also represented on Figure 4.6, showing small changes in growth predictions, when compared with the wild type. Some examples are the regulator of sporulation (SpoIIID) [89] and the regulator of iron homeostasis (Fur) [90].

This notion that heavy flux restrictions may be causing some of the observed *in silico* phenotypes is supported by our initial knockout studies performed with PROM using relaxation of constraints. Relaxed constraints permit reaction bounds to be exceeded to allow flux maximization (results available in Supplementary material S4.4). In these results, we observe no predicted growth rate changes when compared with the wild type for the knockout mutants of Phop, SpoIIID and Fur.

The relaxation of the constraints was introduced in the initial PROM formulation to account for lack of knowledge of the regulatory mechanisms and regulatory effects other than transcriptional regulation. The observation that, without the relaxation constraints a wider impact is felt, was not completely unexpected, as we purposely enforced hard regulatory constraints to better observe the impact of the regulatory network in our model, without the possibility of reaction bounds being exceeded.

As we saw very few variations in the growth predictions across our 5 media formulations, we decided to perform knockout growth simulations for additional 78 minimal media conditions, varying only in the carbon source. The full results of this study are available in Supplementary material S4.5. Due to scale of the study conducted with had to narrow down the predicted *in silico* phenotypes to analyze. We selected, for this analysis, regulators that only showed lethal phenotypes for growth in 1/78 medium conditions. The selected regulators and carbon sources, in which the lethal phenotype was predicted, are shown on Table 4.14. Literature evidence (when available) assessing the lethality of the regulator knockout was also added to Table 4.14. A growth phenotype was considered lethal when less than 5% of the wild type growth rate is observed.

Table 4.14 Regulator mutants predicted to have a lethal phenotype *in silico.*

Regulator Name	Locus ID	Carbon Source	Literature evidence
GlpP	BSU09270	Glycerol	Lethal [91]
IolR	BSU39770	*myo*-inositol	Non-lethal [92]
RbsR	BSU35910	Deoxyribose	Non-lethal [93]
TreR	BSU07820	Trehalose	Non-lethal [94]
RhaR	BSU31210	Rhamnose	Non-lethal* [95]
ManR	BSU12000	Mannose	Lethal [96]
HxlR	BSU03470	β–methyl-D-glucoside	Unknown
GabR	BSU03890	γ-Aminobutyric acid	Lethal [97]
XylR	BSU17590	Xylose	Non-lethal [98]
GutR	BSU06140	Glucitol	Lethal [99]

*Phenotype only predicted by bioinformatics analysis

B. subtilis can grow with glycerol as sole carbon source [91]. In our study, the GlpP mutant exhibited an *in silico* lethal phenotype during growth on glycerol. The *glpP* gene is a member of the *glp* regulon, which comprises the necessary genes for growth with glycerol or glycerol-3-phosphate as sole carbon sources [100]. Within the *glp regulon,* mutants of GlpK (glycerol kinase) and GlpD (glycerol-3-phosphate dehydrogenase) are not able to grow with glycerol as sole carbon source [91]. Additionally, mutants of GlpP were found to exhibit a pleiotropic phenotype causing the non-induction of GlpK and GlpD pointing to the activity of GlpP as a regulator. GlpP encodes a regulatory protein that regulates the expression of the GlpK and GlpD via a mechanism of transcriptional antitermination [101]. This mechanism prevents transcriptional termination of the enzymes catabolizing glycerol, making GlpP essential for growth on glycerol as sole carbon source. We also tested growth on glycerol-3-phosphate and achieved a significant growth reduction to approximately 15% growth rate when compared with the wild-type.

Inositol is a compound widely available in the environment, especially in the soil. Several soil bacteria, including *B. subtilis* are capable to grow with inositol as sole carbon source [92]. The *iol* operon comprises the genes responsible for catabolism of inositol in multiple steps converting it to acetyl-CoA. In *B. subtilis*, the *iol* operon and the gene *iolT* (encoding an inositol transporter [102]) are regulated by the repressor *iolR* [103]. In the absence of inositol in the medium, *iolR* represses expression of the *iol*

operon and *iotT*. In the presence of inositol, *iolR* is antagonized and the *iol* operon expressed. A knockout of *iolR* leads to the constitutive expression of its target genes [92].

Table 4.15 PROM constraints for IolR regulated genes

Gene Name (TG)	Locus ID	Prob TG ON | TF OFF	Prob TG ON | TF ON
iolA	BSU39760	0	0.422
iolB	BSU39750	0	0.441
iolD	BSU39730	0	0.451
iolI	BSU39680	0	0.471
iolT	BSU06230	0	0.167
IolG	BSU39700	0	0.471
iolH	BSU39690	0.0120	0.539
iolF	BSU39710	0	0.382
iolE	BSU39720	0	0.392
iolJ	BSU39670	0	0.520
iolC	BSU39740	0	0.441
iolS	BSU39780	0.964	1

Our *in silico* knockout study showed a lethal phenotype for the *iloR* mutant in medium containing inositol as the carbon source. We inspected the PROM regulatory constraints for the *iolR* regulation (Table 4.15) and observed that most genes have a 0 probability of being ON (with the exception of *iolS* that is co-transcribed with *iloR* [92]) when *iolR* is inactive. This leads us to believe that the dataset used to infer the regulatory interactions lacks conditions in which inositol was used a sole carbon source to allow the observation of the regulatory interaction described in the literature.

Additional transcriptional repressors on Table 4.14 showed similar behavior to IolR, with the PROM constraints revealing that the expression data did not capture the proper regulatory interactions between the regulators and their target genes. This is the case for TheR, repressor of the threalose utilization operon [104], which is induced by threalose-6-phosphate [94]. The transcriptional repressor XylR is another example, as the regulator of the xylose utilization operon being induced by xylose [98]. The last example of this behavior is RbsR, the transcriptional repressor of the ribose transport operon

[93]. For all these cases we manually adjusted the regulatory constraints to properly represent the knockout phenotypes.

The transcriptional regulator RhaR was predicted to control the utilization of rhamnose in multiple *Bacillales* species, including *B. subtilis* using the bioinformatics tool in the RegPredict [95] (RegPredict [105, 106] is described in section 3.2.3). We included this predicted regulator in our model as part of the manual curation effort for the reconstruction of a more comprehensive *B. subtilis* regulatory network. A description of the predicted rhamnose utilization operons, predicted by RegPredict, is shown in Figure 4.7.

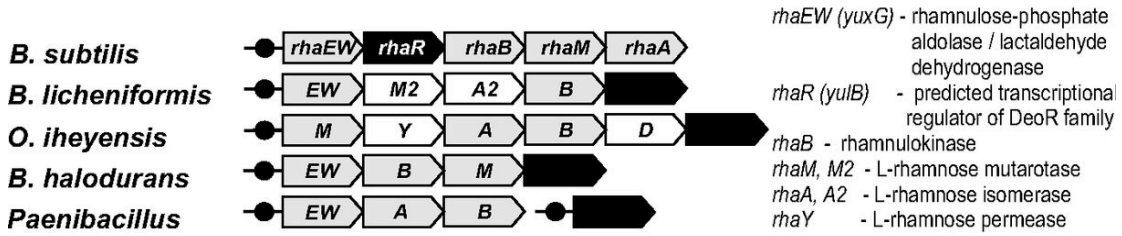


Figure 4.7 Predicted RhaR regulon for rhamnose utilization in *Bacillales* species. Arrows represent genes. TF genes are colored in black, catabolic enzymes in gray, and genes lacking orthologs in *B. subtilis* are in white. Adapted from Leyn *el al.* [95].

This predicted regulator was found to be a DeoR-like regulator, a family of transcriptional repressors [107]. Additional comparative genomics analysis, aimed to characterize the poorly studied rhamnose catabolic pathways in bacteria, suggested a novel enzyme for the rhamnose catabolism in *B. subtilis* that was verified *in vivo* [108]. This same study confirmed the repression activity by RhaR and its negative regulation when induced by rhamnose in *Chloroflexus aurantiacus*. Our *in silico* phenotype simulation predicted lethal phenotype for the mutant of RhaR with rhamnose as sole carbon source. Trusting the described mechanism in *Chloroflexus aurantiacus* functions likewise in *B. subtilis*, we are lead to believe that similarly to the activity previously described for other transcription repressors, the experimental data did not have growth conditions capable of capturing this regulatory interaction. We also manually fixed the regulatory constraints to correct the knockout phenotype.

Many organisms, including *B. subtilis,* are capable of using mannose as sole carbon source [109]. Our *in silico* knockout study predicted the *ΔmanR* mutant to be lethal. In *B. subtilis*, the mannose utilization operon (*manPA-yjdF*) is composed by *manP, manA* and *yjdF*. *manP* encodes the mannose specific phosphotransferase transporter system [110], *manA* encodes the mannose-6-phosphate isomerase which converts mannose-6-phostphate to fructose-1,6-bisphosphate, and the function of *yjdF* is still unknown [96]. ManR was found to be the transcriptional activator of the *manPA-yjdF* operon and the *ΔmanR* mutant was found to be unable to grow in minimal medium with mannose as sole carbon source [96].

A similar mechanism of transcriptional activation was described for the GutR, the regulator of the glucitol utilization operon [111]. GutR mutant strains show a complete loss of glucitol induction by this operon, showing the role of GutR regulator for induction of glucitol [99]. As *B. subtilis* can grow in glucitol as sole carbon source [112], the knockout of the regulator required for its induction causes a lethal phenotype. Our model also predicted the knockout of GutR as lethal in a medium condition with glucitol as sole carbon source.

B. subtilis was found able to grow with β-methyl-D-glucoside as substrate, but the genes coding for its transport and utilization remain unknown [113]. In our model, *B. subtilis* becomes unable to grow on β-methyl-D-glucoside in the mutant of gene the *hxlR*. This gene was described as a transcriptional activator of the ribulose monophosphate pathway in *B. subtilis* [114]. This pathway was originally described in methylotrophs [115], but was found to be present in many other bacteria involved in formaldehyde fixation and detoxification [116]. The 3-hexulose-6-phosphate synthase and 6-phospho-3-hexuloisomerase are the enzymes regulated by HxlR. These enzymes are responsible for the synthase of hexulose-6-phosphate from formaldehyde and ribulose-5-phosphate, and for the isomerization of fructose-6-phosphate from hexulose-6-phosphate.

Our model was able to achieve growth with β-methyl-D-glucoside. Since this is a process that is still unknown in the literature, we investigated the iBsu1103V2 uptake and catabolism of this substrate to assess its relation to the ribulose monophosphate pathway and HxlR. Regarding the transport into the

cell, β-methyl-D-glucoside is uptaken via a proton symport mechanism described by reaction rxn11397:

H^+ [e] + (1) beta-Methylglucoside[e] <=> H^+ [c] + (1) beta-Methylglucoside[c]

We found that this transport reaction was added to the *i*Bsu1103V2 model via gap filling (this process is described in detail in Section 2.2.2) and, therefore, has no GPR association in the model. After the transport into the cell, β-methyl-D-glucoside is converted into glucose and methanol by a beta-glucosidase enzyme described by reaction rxn09979:

beta-Methylglucoside[c] + H_2O [c] <=> D-Glucose[c] + Methanol[c]

Methanol is converted to Formaldehyde via the action of a methanol dehydrogenase, described by reaction rxn00430:

NAD[c] + Methanol[c] <=> H^+ [c] + NADH[c] + Formaldehyde[c]

Like the previous transport reaction, the one encoding methanol dehydrogenase was also added by gap filling during the reconstruction process of *i*Bsu1103V2 and has no GPR associated. After these 3 steps, formaldehyde enters the ribulose monophosphate described above, thus leading to the *in silico* phenotype observed when we knockout the transcriptional activator of this pathway. Even though the mechanisms described above seem to validate our *in silico* phenotype, we analyzed the additions to the model via gap filling. The analysis of the reactions added by gap filling revealed that the NAD dependent methanol dehydrogenase does not occur in *B. subtilis*. Researchers found no activity of this enzyme in *B. subtilis*, neither sequence homologs when compared with the methanol dehydrogenase of *Bacillus methanolicus* [114]. Removal of the methanol dehydrogenase in the same environmental conditions leads to a lethal phenotype prediction by our model. Due to this fact, we cannot validate this regulatory interaction predicted by our model.

In our *in silico* simulation, the knockout of gene *gabR* caused a lethal phenotype with γ-Aminobutyric acid (GABA) as sole carbon source. GabR is described in the literature as the regulator of the GABA utilization pathway. GabR regulates the activity of two enzymes (GABA aminotransferase and succinic

semialdehyde dehydrogenase) that provide an alternative route of glutamate biosynthesis using GABA. This route via aminotransferases is represented in Figure 4.8.

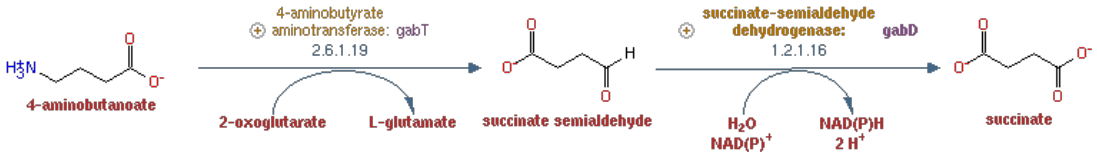


Figure 4.8 GABA degradation pathway (Pathway representation from BSubCyc [117]).

The knockout of the transcriptional activator GabR was found to be lethal, with GABA as sole nitrogen source [97]. The same work reported the inability of both the wild type and GabR mutant to grow with GABA as sole carbon source, the medium condition in which our simulation was performed and achieved growth in wild type. We inspected our model for the reactions encoded by *gabT* and *gabD* and found they are represented as described in Figure 4.8. Our model is able to grow with GABA as sole carbon and nitrogen source due to the use of the succinate that is produced by the reaction encoded by *gabD*. The most common path for GABA degradation involves transamination to succinate semialdehyde, followed by oxidation to succinate. Many organisms can grow on GABA as the sole carbon and nitrogen source, but *B. subtilis* can only grow with GABA as sole nitrogen source [118]. The reason why this happens in *B. subtilis* is unknown. With no basis, researchers theorized a possible regulatory toxicity effect due to the accumulation of GABA or succinate semialdehyde, or both [97], to be the cause of this behavior.

To assess if our model properly represents the regulation by GabR, as described in the literature, we simulated the knockout of GabR in a media with GABA as sole nitrogen source and glucose as carbon source. The result of this simulation also predicted a lethal phenotype. As our model appears to accurately represent the metabolism of the GABA degradation pathway, no suggestions are proposed to introduce changes in the model.

4.5 CONCLUSIONS

In this chapter, we present the first attempt to model the metabolism and regulation of *B. subtilis* at genome-scale. This was achieved by combining the comprehensive regulatory network reconstruction, manually curated by us and presented in chapter 3, a large dataset of high quality expression data with multiple growth conditions and the published genome-scale metabolic model *i*Bsu1103V2. We chose the PROM algorithm for the integrated simulation of metabolism and regulation, but we introduced a couple of changes to its original formulation.

We introduced pFBA into the formulation to avoid redundant alternative optimal solutions generated by FBA. Another change introduced to the methodology removed the ability of the flux bounds to be violated making the regulatory constraints hard constraints. By using hard regulatory constraints, we were able to better assess the impact of the flux restrictions in our *in silico* knockout studies, and noted an inability to proper represent the regulatory interactions for the global regulators TnrA and PhoP. We also modified the formulation to account for isoenzymes activity. We verified the impact of this change in our phenotype simulation for the isoenzymes PutC and RocA. The isoenzymes are induced by different substrates, a notion that is not being captured by our regulatory constraints leading to an inability to properly represent the knockout of an individual isoenzyme. To solve this issue we suggest that additional regulatory constraints accounting for the presence of induction substrates in the medium formulation could be added to the model. This approach is adopted by simulation methods that apply manually curated Boolean regulatory rules such as rFBA and SR-FBA.

We were able to validate our model with TF knockouts described in the literature. For the mutant phenotypes reported by Subtiwiki, we were able to replicate the regulatory effects observed in the literature for the mutants of AlsR, CysL, GltC and PutR. It is important to note the importance of being able to replicate the mutant phenotypes for GltC and PutR, as these TFs regulate genes involved in glutamate biosynthesis, a key metabolite for both carbon and nitrogen metabolism. Being able to accurately predict TF knockouts can have added value for Metabolic Engineering strategies for strain optimization [119, 120].

We were also able to assess additional limitations of our methodology and point to possible unknown regulatory effects not represented by our model. We observed in multiple *in silico* knockouts that the expression data used failed to properly represent the regulatory interactions described to the literature. These observations were not totally unexpected as we attempted to simulate growth in many specific carbon sources not comprised in the experimental conditions of our dataset. This issue reveals a lack of better experimental datasets available for validation of this type of models, when compared with datasets available for metabolic model validation. As mentioned when we analyzed the mutant phenotype of PurR these results are direct consequence of our decision to use the PROM formulation to infer regulatory interactions from expression data in detriment of manual design of Boolean gene regulatory rules. Even with this limitation to properly represent some regulatory interactions, nevertheless, the overall results were encouraging. Our model was able to properly simulate multiple phenotypes described in the literature and was validated against a large dataset of multiple gene deletions.

The integration of the regulatory constraints with the metabolic model also demonstrated issues with the metabolic model with reactions that were added by gap filling. This fact points to the ability of using the integration of regulatory constraints as a tool to flag inconsistencies and curate the metabolic model Recently, studies have debated the role of transcriptional regulation in controlling metabolic fluxes [121] since other regulatory effects, such as allosteric regulation, thermodynamics and post transcriptional can also be a factor [122]. One of these effects may be responsible for the interesting result that was shown for growth with GABA as sole carbon source.

Additionally, work in progress for full implementation of these methods in the KBase of Systems Biology will allow for an easier access to these tools. Automatic genome-scale metabolic model reconstruction is available on KBase and the other necessary datasets can be uploaded to the KBase environment.

4.6 REFERENCES

1. Varner, J.D., *Large-scale prediction of phenotype: Concept.* Biotechnology and Bioengineering, 2000. **69**(6): p. 664-678.

2. Palsson, B., *The challenges of in silico biology.* Nature Biotechnology, 2000. **18**(11): p. 1147-1150.

3. Varma, A. and B.O. Palsson, *Metabolic Flux Balancing - Basic Concepts, Scientific and Practical Use.* Bio-Technology, 1994. **12**(10): p. 994-998.

4. Covert, M.W., C.H. Schilling, and B. Palsson, *Regulation of gene expression in flux balance models of metabolism.* Journal of Theoretical Biology, 2001. **213**(1): p. 73-88.

5. Tomita, M., *Whole-cell simulation: a grand challenge of the 21st century.* TRENDS in Biotechnology, 2001. **19**(6): p. 205-210.

6. Price, N.D., J.L. Reed, and B.O. Palsson, *Genome-scale models of microbial cells: Evaluating the consequences of constraints.* Nature Reviews Microbiology, 2004. **2**(11): p. 886-897.

7. Papin, J.A., et al., *Reconstruction of cellular signalling networks and analysis of their properties.* Nature Reviews Molecular Cell Biology, 2005. **6**(2): p. 99-111.

8. Feist, A.M., et al., *Reconstruction of biochemical networks in microorganisms.* Nature Reviews Microbiology, 2009. **7**(2): p. 129-143.

9. Covert, M.W., et al., *Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli.* Bioinformatics, 2008. **24**(18): p. 2044-50.

10. Lee, J.M., et al., *Dynamic analysis of integrated signaling, metabolic, and regulatory networks.* Plos Computational Biology, 2008. **4**(5).

11. Ovacik, M.A. and I.P. Androulakis, *On the Potential for Integrating Gene Expression and Metabolic Flux Data.* Current Bioinformatics, 2008. **3**(3): p. 142-148.

12. Edwards, J.S., M. Covert, and B. Palsson, *Metabolic modelling of microbes: the flux-balance approach.* Environ Microbiol, 2002. **4**(3): p. 133-40.

13. Covert, M.W., C.H. Schilling, and B. Palsson, *Regulation of gene expression in flux balance models of metabolism.* J Theor Biol, 2001. **213**(1): p. 73-88.

14. Shlomi, T., et al., *A genome-scale computational study of the interplay between transcriptional regulation and metabolism.* Mol Syst Biol, 2007. **3**: p. 101.

15. Buescher, J.M., et al., *Global network reorganization during dynamic adaptations of Bacillus subtilis metabolism.* Science, 2012. **335**(6072): p. 1099-103.

16. Nicolas, P., et al., *Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis.* Science, 2012. **335**(6072): p. 1103-6.

17. Chandrasekaran, S. and N.D. Price, *Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis.* Proc Natl Acad Sci U S A, 2010. **107**(41): p. 17845-50.

18. Tanaka, K., et al., *Building the repertoire of dispensable chromosome regions in Bacillus subtilis entails major refinement of cognate large-scale metabolic model.* Nucleic Acids Res, 2013. **41**(1): p. 687-99.

19. Goelzer, A., et al., *Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of Bacillus subtilis.* BMC Syst Biol, 2008. **2**: p. 20.

20. Tenazinha, N. and S. Vinga, *A Survey on Methods for Modeling and Analyzing Integrated Biological Networks.* Ieee-Acm Transactions on Computational Biology and Bioinformatics, 2011. **8**(4): p. 943-958.

21. Machado, D., et al., *Modeling formalisms in Systems Biology.* AMB Express, 2011. **1**: p. 45.

22. Karlebach, G. and R. Shamir, *Modelling and analysis of gene regulatory networks.* Nat Rev Mol Cell Biol, 2008. **9**(10): p. 770-80.

23. Machado, D. and M. Herrgard, *Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism.* PLoS Comput Biol, 2014. **10**(4): p. e1003580.

24. Price, N.D., et al., *Genome-scale microbial in silico models: the constraints-based approach.* Trends Biotechnol, 2003. **21**(4): p. 162-9.

25. Llaneras, F. and J. Pico, *Stoichiometric modelling of cell metabolism.* J Biosci Bioeng, 2008. **105**(1): p. 1-11.

26. Smallbone, K., et al., *Towards a genome-scale kinetic model of cellular metabolism.* BMC Syst Biol, 2010. **4**: p. 6.

27. Bailey, J.E., *Complex biology with no parameters.* Nature Biotechnology, 2001. **19**(6): p. 503-504.

28. Price, N.D., et al., *Extreme pathways and Kirchhoff's second law.* Biophys J, 2002. **83**(5): p. 2879-82.

29. Papin, J.A., et al., *Comparison of network-based pathway analysis methods.* Trends Biotechnol, 2004. **22**(8): p. 400-5.

30. Covert, M.W. and B.O. Palsson, *Constraints-based models: regulation of gene expression reduces the steady-state solution space.* J Theor Biol, 2003. **221**(3): p. 309-25.

31. Shlomi, T., et al., *A genome-scale computational study of the interplay between transcriptional regulation and metabolism.* Molecular Systems Biology, 2007. **3**.

32. Edwards, J.S., M. Covert, and B. Palsson, *Metabolic modelling of microbes: the flux□balance approach.* Environmental Microbiology, 2002. **4**(3): p. 133-140.

33. Kauffman, K.J., P. Prakash, and J.S. Edwards, *Advances in flux balance analysis.* Curr Opin Biotechnol, 2003. **14**(5): p. 491-6.

34. Covert, M.W., et al., *Integrating high-throughput and computational data elucidates bacterial networks.* Nature, 2004. **429**(6987): p. 92-6.

35. Lee, J.M., et al., *Dynamic analysis of integrated signaling, metabolic, and regulatory networks.* PLoS Comput Biol, 2008. **4**(5): p. e1000086.

36. Covert, M.W. and B.O. Palsson, *Transcriptional regulation in constraints-based metabolic models of Escherichia coli.* J Biol Chem, 2002. **277**(31): p. 28058-64.

37. Hohmann, S., *Osmotic stress signaling and osmoadaptation in Yeasts.* Microbiology and Molecular Biology Reviews, 2002. **66**(2): p. 300-+.

38. Colijn, C., et al., *Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production.* PLoS Comput Biol, 2009. **5**(8): p. e1000489.

39. Navid, A. and E. Almaas, *Genome-level transcription data of Yersinia pestis analyzed with a new metabolic constraint-based approach.* BMC Syst Biol, 2012. **6**: p. 150.

40. Zur, H., E. Ruppin, and T. Shlomi, *iMAT: an integrative metabolic analysis tool.* Bioinformatics, 2010. **26**(24): p. 3140-2.

41. van Berlo, R.J.P., et al., *Predicting metabolic fluxes using gene expression differences as constraints.* Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 2011. **8**(1): p. 206-216.

42. Chandrasekaran, S. and N.D. Price, *Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(41): p. 17845-17850.

43. van Berlo, R.J.P., et al., *Predicting Metabolic Fluxes Using Gene Expression Differences As Constraints.* Ieee-Acm Transactions on Computational Biology and Bioinformatics, 2011. **8**(1): p. 206-216.

44. Shlomi, T., et al., *Network-based prediction of human tissue-specific metabolism.* Nat Biotechnol, 2008. **26**(9): p. 1003-10.

45. Lobel, L., et al., *Integrative genomic analysis identifies isoleucine and CodY as regulators of Listeria monocytogenes virulence.* PLoS Genet, 2012. **8**(9): p. e1002887.

46. Levine, D.M., et al., *Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways.* Genome Biol, 2006. **7**(10): p. R93.

47. Ibarra, R.U., J.S. Edwards, and B.O. Palsson, *Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth.* Nature, 2002. **420**(6912): p. 186-9.

48. Lewis, N.E., et al., *Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models.* Mol Syst Biol, 2010. **6**: p. 390.

49. Mahadevan, R. and C.H. Schilling, *The effects of alternate optimal solutions in constraint-based genome-scale metabolic models.* Metab Eng, 2003. **5**(4): p. 264-76.

50. Simeonidis, E., S. Chandrasekaran, and N.D. Price, *A guide to integrating transcriptional regulatory and metabolic networks using PROM (probabilistic regulation of metabolism).* Methods Mol Biol, 2013. **985**: p. 103-12.

51. Henry, C.S., et al., *iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations.* Genome Biol, 2009. **10**(6): p. R69.

52. Mader, U., et al., *SubtiWiki–a comprehensive community resource for the model organism Bacillus subtilis.* Nucleic Acids Res, 2012. **40**(Database issue): p. D1278-87.

53. Duarte, N.C., M.J. Herrgard, and B.O. Palsson, *Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model.* Genome Res, 2004. **14**(7): p. 1298-309.

54. Renna, M.C., et al., *Regulation of the Bacillus subtilis alsS, alsD, and alsR genes involved in post-exponential-phase production of acetoin.* J Bacteriol, 1993. **175**(12): p. 3863-75.

55. Yamamoto, H., M. Murata, and J. Sekiguchi, *The CitST two-component system regulates the expression of the Mg-citrate transporter in Bacillus subtilis.* Mol Microbiol, 2000. **37**(4): p. 898-912.

56. Tannler, S., et al., *CcpN controls central carbon fluxes in Bacillus subtilis.* J Bacteriol, 2008. **190**(18): p. 6178-87.

57. Guillouard, I., et al., *Identification of Bacillus subtilis CysL, a regulator of the cysJI operon, which encodes sulfite reductase.* J Bacteriol, 2002. **184**(17): p. 4681-9.

58. Bohannon, D.E. and A.L. Sonenshein, *Positive regulation of glutamate biosynthesis in Bacillus subtilis.* J Bacteriol, 1989. **171**(9): p. 4718-27.

59. Belitsky, B.R., *Indirect repression by Bacillus subtilis CodY via displacement of the activator of the proline utilization operon.* J Mol Biol, 2011. **413**(2): p. 321-36.

60. Xiao, Z. and P. Xu, *Acetoin metabolism in bacteria.* Crit Rev Microbiol, 2007. **33**(2): p. 127-40.

61. Chen, T., et al., *Engineering Bacillus subtilis for acetoin production from glucose and xylose mixtures.* J Biotechnol, 2013. **168**(4): p. 499-505.

62. Wang, M., et al., *Metabolic engineering of Bacillus subtilis for enhanced production of acetoin.* Biotechnol Lett, 2012. **34**(10): p. 1877-85.

63. Cruz Ramos, H., et al., *Fermentative metabolism of Bacillus subtilis: physiology and regulation of gene expression.* J Bacteriol, 2000. **182**(11): p. 3072-80.

64. Nicholson, W.L., *The Bacillus subtilis ydjL (bdhA) gene encodes acetoin reductase/2,3-butanediol dehydrogenase.* Appl Environ Microbiol, 2008. **74**(22): p. 6832-8.

65. Hoch, J.A. and T.J. Silhavy, *Two-component signal transduction*. Vol. 2. 1995: ASM press Washington, DC:.

66. Boorsma, A., et al., *Secondary transporters for citrate and the Mg(2+)-citrate complex in Bacillus subtilis are homologous proteins.* J Bacteriol, 1996. **178**(21): p. 6216-22.

67. Warner, J.B. and J.S. Lolkema, *Growth of Bacillus subtilis on citrate and isocitrate is supported by the Mg2+-citrate transporter CitM.* Microbiology, 2002. **148**(Pt 11): p. 3405-12.

68. Krom, B.P., et al., *Complementary metal ion specificity of the metal-citrate transporters CitM and CitH of Bacillus subtilis.* J Bacteriol, 2000. **182**(22): p. 6374-81.

69. Lensbouer, J.J. and R.P. Doyle, *Secondary transport of metal-citrate complexes: the CitMHS family.* Crit Rev Biochem Mol Biol, 2010. **45**(5): p. 453-62.

70. Servant, P., D. Le Coq, and S. Aymerich, *CcpN (YqzB), a novel regulator for CcpA-independent catabolite repression of Bacillus subtilis gluconeogenic genes.* Mol Microbiol, 2005. **55**(5): p. 1435-51.

71. Henkin, T.M., *The role of CcpA transcriptional regulator in carbon metabolism in Bacillus subtilis.* FEMS Microbiol Lett, 1996. **135**(1): p. 9-15.

72. Licht, A., S. Preis, and S. Brantl, *Implication of CcpN in the regulation of a novel untranslated RNA (SR1) in Bacillus subtilis.* Mol Microbiol, 2005. **58**(1): p. 189-206.

73. van der Ploeg, J.R., M. Barone, and T. Leisinger, *Functional analysis of the Bacillus subtilis cysK and cysJI genes.* FEMS Microbiol Lett, 2001. **201**(1): p. 29-35.

74. Wacker, I., et al., *The regulatory link between carbon and nitrogen metabolism in Bacillus subtilis: regulation of the gltAB operon by the catabolite control protein CcpA.* Microbiology, 2003. **149**(Pt 10): p. 3001-9.

75. Picossi, S., B.R. Belitsky, and A.L. Sonenshein, *Molecular mechanism of the regulation of Bacillus subtilis gltAB expression by GltC.* J Mol Biol, 2007. **365**(5): p. 1298-313.

76. Commichau, F.M., et al., *A regulatory protein-protein interaction governs glutamate biosynthesis in Bacillus subtilis: the glutamate dehydrogenase RocG moonlights in controlling the transcription factor GltC.* Mol Microbiol, 2007. **65**(3): p. 642-54.

77. Belitsky, B.R., P.J. Janssen, and A.L. Sonenshein, *Sites required for GltC-dependent regulation of Bacillus subtilis glutamate synthase expression.* J Bacteriol, 1995. **177**(19): p. 5686-95.

78. Huang, S.C., T.H. Lin, and G.C. Shaw, *PrcR, a PucR-type transcriptional activator, is essential for proline utilization and mediates proline-responsive expression of the proline utilization operon putBCP in Bacillus subtilis.* Microbiology, 2011. **157**(Pt 12): p. 3370-7.

79. Atkinson, M.R., L.V. Wray, Jr., and S.H. Fisher, *Regulation of histidine and proline degradation enzymes by amino acid availability in Bacillus subtilis.* J Bacteriol, 1990. **172**(9): p. 4758-65.

80. Calogero, S., et al., *RocR, a novel regulatory protein controlling arginine utilization in Bacillus subtilis, belongs to the NtrC/NifA family of transcriptional activators.* J Bacteriol, 1994. **176**(5): p. 1234-41.

81. Kobayashi, K., et al., *Essential Bacillus subtilis genes.* Proc Natl Acad Sci U S A, 2003. **100**(8): p. 4678-83.

82. Commichau, F.M. and J. Stulke, *Trigger enzymes: bifunctional proteins active in metabolism and in controlling gene expression.* Mol Microbiol, 2008. **67**(4): p. 692-702.

83. Beinert, H., M.C. Kennedy, and C.D. Stout, *Aconitase as Ironminus signSulfur Protein, Enzyme, and Iron-Regulatory Protein.* Chem Rev, 1996. **96**(7): p. 2335-2374.

84. Craig, J.E., et al., *A null mutation in the Bacillus subtilis aconitase gene causes a block in Spo0A-phosphate-dependent gene expression.* J Bacteriol, 1997. **179**(23): p. 7351-9.

85. Wray, L.V., Jr., et al., *TnrA, a transcription factor required for global nitrogen regulation in Bacillus subtilis.* Proc Natl Acad Sci U S A, 1996. **93**(17): p. 8841-5.

86. Belitsky, B.R., et al., *Role of TnrA in nitrogen source-dependent repression of Bacillus subtilis glutamate synthase gene expression.* J Bacteriol, 2000. **182**(21): p. 5939-47.

87. Liu, W., S. Eder, and F.M. Hulett, *Analysis of Bacillus subtilis tagAB and tagDEF expression during phosphate starvation identifies a repressor role for PhoP-P.* J Bacteriol, 1998. **180**(3): p. 753-8.

88. Pragai, Z., et al., *Transcriptional regulation of the phoPR operon in Bacillus subtilis.* J Bacteriol, 2004. **186**(4): p. 1182-90.

89. Halberg, R., V. Oke, and L. Kroos, *Effects of Bacillus subtilis sporulation regulatory protein SpoIIID on transcription by sigma K RNA polymerase in vivo and in vitro.* J Bacteriol, 1995. **177**(7): p. 1888-91.

90. Ollinger, J., et al., *Role of the Fur regulon in iron transport in Bacillus subtilis.* J Bacteriol, 2006. **188**(10): p. 3664-73.

91. Lindgren, V. and L. Rutberg, *Glycerol metabolism in Bacillus subtilis: gene-enzyme relationships.* J Bacteriol, 1974. **119**(2): p. 431-42.

92. Yoshida, K.I., et al., *Organization and transcription of the myo-inositol operon, iol, of Bacillus subtilis.* J Bacteriol, 1997. **179**(14): p. 4591-8.

93. Woodson, K. and K.M. Devine, *Analysis of a ribose transport operon from Bacillus subtilis.* Microbiology, 1994. **140 (Pt 8)**: p. 1829-38.

94. Burklen, L., F. Schock, and M.K. Dahl, *Molecular analysis of the interaction between the Bacillus subtilis trehalose repressor TreR and the tre operator.* Mol Gen Genet, 1998. **260**(1): p. 48-55.

95. Leyn, S.A., et al., *Genomic reconstruction of the transcriptional regulatory network in Bacillus subtilis.* J Bacteriol, 2013. **195**(11): p. 2463-73.

96. Sun, T. and J. Altenbuchner, *Characterization of a mannose utilization system in Bacillus subtilis.* J Bacteriol, 2010. **192**(8): p. 2128-39.

97. Belitsky, B.R. and A.L. Sonenshein, *GabR, a member of a novel protein family, regulates the utilization of gamma-aminobutyrate in Bacillus subtilis.* Mol Microbiol, 2002. **45**(2): p. 569-83.

98. Gartner, D., et al., *Regulation of the Bacillus subtilis W23 xylose utilization operon: interaction of the Xyl repressor with the xyl operator and the inducer xylose.* Mol Gen Genet, 1992. **232**(3): p. 415-22.

99. Ye, R., S.N. Rehemtulla, and S.L. Wong, *Glucitol induction in Bacillus subtilis is mediated by a regulatory factor, GutR.* J Bacteriol, 1994. **176**(11): p. 3321-7.

100. Beijer, L., et al., *The glpP and glpF genes of the glycerol regulon in Bacillus subtilis.* J Gen Microbiol, 1993. **139**(2): p. 349-59.

101. Greenblatt, J., J.R. Nodwell, and S.W. Mason, *Transcriptional antitermination.* Nature, 1993. **364**(6436): p. 401-6.

102. Yoshida, K., et al., *Identification of two myo-inositol transporter genes of Bacillus subtilis.* J Bacteriol, 2002. **184**(4): p. 983-91.

103. Yoshida, K.I., et al., *Interaction of a repressor and its binding sites for regulation of the Bacillus subtilis iol divergon.* J Mol Biol, 1999. **285**(3): p. 917-29.

104. Schock, F. and M.K. Dahl, *Expression of the tre operon of Bacillus subtilis 168 is regulated by the repressor TreR.* J Bacteriol, 1996. **178**(15): p. 4576-81.

105. Novichkov, P.S., et al., *RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes.* Nucleic Acids Res, 2010. **38**(Database issue): p. D111-8.

106. Novichkov, P.S., et al., *RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W299-307.

107. Weickert, M.J. and S. Adhya, *A family of bacterial regulators homologous to Gal and Lac repressors.* J Biol Chem, 1992. **267**(22): p. 15869-74.

108. Rodionova, I.A., et al., *Comparative genomics and functional analysis of rhamnose catabolic pathways and regulons in bacteria.* Front Microbiol, 2013. **4**: p. 407.

109. Mobley, H.L., et al., *Transport and incorporation of N-acetyl-D-glucosamine in Bacillus subtilis.* J Bacteriol, 1982. **150**(1): p. 8-15.

110. Reizer, J., et al., *Novel phosphotransferase system genes revealed by genome analysis - the complete complement of PTS proteins encoded within the genome of Bacillus subtilis.* Microbiology, 1999. **145 (Pt 12)**: p. 3419-29.

111. Ye, R. and S.L. Wong, *Transcriptional regulation of the Bacillus subtilis glucitol dehydrogenase gene.* J Bacteriol, 1994. **176**(11): p. 3314-20.

112. Watanabe, S., et al., *Mannitol-1-phosphate dehydrogenase (MtlD) is required for mannitol and glucitol assimilation in Bacillus subtilis: possible cooperation of mtl and gut operons.* J Bacteriol, 2003. **185**(16): p. 4816-24.

113. Perkins, A.E. and W.L. Nicholson, *Uncovering new metabolic capabilities of Bacillus subtilis using phenotype profiling of rifampin-resistant rpoB mutants.* J Bacteriol, 2008. **190**(3): p. 807-14.

114. Yasueda, H., Y. Kawahara, and S. Sugimoto, *Bacillus subtilis yckG and yckF encode two key enzymes of the ribulose monophosphate pathway used by methylotrophs, and yckH is required for their expression.* J Bacteriol, 1999. **181**(23): p. 7154-60.

115. Hanson, R.S. and T.E. Hanson, *Methanotrophic bacteria.* Microbiol Rev, 1996. **60**(2): p. 439-71.

116. Kato, N., H. Yurimoto, and R.K. Thauer, *The physiological role of the ribulose monophosphate pathway in bacteria and archaea.* Biosci Biotechnol Biochem, 2006. **70**(1): p. 10-21.

117. Caspi, R., et al., *The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.* Nucleic Acids Res, 2014. **42**(Database issue): p. D459-71.

118. Ferson, A.E., L.V. Wray, Jr., and S.H. Fisher, *Expression of the Bacillus subtilis gabP gene is regulated independently in response to nitrogen and amino acid availability.* Mol Microbiol, 1996. **22**(4): p. 693-701.

119. Burgard, A.P., P. Pharkya, and C.D. Maranas, *Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.* Biotechnol Bioeng, 2003. **84**(6): p. 647-57.

120. Rocha, I., et al., *OptFlux: an open-source software platform for in silico metabolic engineering.* BMC Syst Biol, 2010. **4**: p. 45.

121. Kochanowski, K., U. Sauer, and V. Chubukov, *Somewhat in control–the role of transcription in regulating microbial metabolic fluxes.* Curr Opin Biotechnol, 2013. **24**(6): p. 987-93.

122. Gerosa, L. and U. Sauer, *Regulation and control of metabolic fluxes in microbes.* Curr Opin Biotechnol, 2011. **22**(4): p. 566-75.

4.7 SUPPLEMENTARY MATERIAL

The supplementary material is available online at http://darwin.di.uminho.pt/jplfaria/phdthesis/Chapter_4_SupplMaterial.xlsx

The following tables comprise the supplementary material:

S4.1 Model validation for data set of multiple genes deletions

S4.2 Full list of regulator mutant phenotypes found in SubtiWiki

S4.3 Mutant phenotype simulations for the media conditions described in Table 4.13

S4.4 Mutant phenotype simulations results for the media conditions described in Table 4.13, with the original PROM formulation

S4.5 Mutant phenotype simulation results for 78 carbon sources

Additionally supplementary, material for all the data and simulations performed for the validation of the mutant phenotypes in section 4.4.1 is avaible on line in the KBase and can be found in the here:

https://narrative.kbase.us/functional-site/#/ws/objects/jplfaria:Chapter_4_SupplMaterial

All the details about the data used and simulation results are described on Table 4.16.

Table 4. 16 Data and simulations results on KBase

ΔalsR	
KBase Type Object	KBase Name
Model	iBsu1103V2
Media	LB
PROM constraint	PromConstraints
Simulation results	Wildtype_LB
	AlsR_KO_FVA_LB
ΔcitT	
Model	iBsu1103V2
Media	Carbon-Citrate
PROM constraint	PromConstraints
Simulation results	Wildtype_Carbon-citrate
	CitT_KO_Carbon-citrate
	citM_KO_Carbon-citrate
	citH_KO_Carbon-citrate
ΔccpN	
Model	iBsu1103V2
Media	Carbon-D-Glucose
PROM constraint	PromConstraints
Simulation results	CcpN_KO_Carbon-D-Glucose
ΔcysL	
Model	iBsu1103V2
	iBsu1103V3
Media	Sulfate-Sulfate
	Sulfate-Sulfite
	Sulfate-L-Cysteine
	Sulfate-L-Methionine
PROM constraint	PromConstraints
	AjustedPromConstraints_CysL
Simulation results	CysL_KO_Sulfate
	CysL_KO_Sulfite
	CysL_KO_L-Cysteine
	CysL_KO_-L-Methionine
	CysL_KO_Sulfatev2
	CysL_KO_Sulfitev2
	CysL_KO_L-Cysteinev2

	CysL_KO_-L-Methioninev2
ΔgltC	
Model	iBsu1103V2
Media	Carbon-D-Glucose
	Carbon-L-Glutamine
	LB
PROM constraint	PromConstraints
Simulation results	GltC_KO_Carbon-D-Glucose
	GltC_KO_Carbon-L-Glutamine
	GltC_KO_LB
ΔputR	
Model	iBsu1103V2
Media	Minimal-Media-Proline
	Carbon-D-Glucose-Proline
PROM constraint	PromConstraints
Simulation results	Wildtype_Minimal-Media-Proline
	Wildtype_Carbon-D-Glucose-Proline
	Carbon-D-Glucose-Proline_PutR_KO
	Proline_Minimal_PutR_KO
	Proline_Minimal_PutR_rocA_KO
	Carbon-D-Glucose-Proline_PutR_rocA_KO

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1 MAIN RESULTS AND CONTRIBUTIONS

The research conducted in this thesis had the overall objective of reconstructing and performing *in silico* phenotype simulations for integrated models of metabolism and regulation. We achieved that objective, proposing a genome-scale model for the metabolism and transcriptional regulation of the bacterium *Bacillus subtilis*. To achieve this goal, we set out to perform studies with the individual elements that are necessary for the reconstruction of integrated metabolic and regulatory models. Those elements include genome functional annotations, genome-scale metabolic models (GEMs), regulatory networks and gene expression data.

Here, we present the main results from those studies:

- **Genome annotations and GEMs –** Genome functional annotations are a key element in the reconstruction of GEMs as we add reactions to the metabolic network that correspond to metabolic functions in the genome. We performed a large-scale model reconstruction effort leading to the reconstruction of automatically generated GEMs for all prokaryotic genomes available in the SEED database [1]. Analysis of these models allowed us to assess the diversity and quality of automatically generated GEMs. In order to allow growth phenotype simulations, we used gap filling algorithms to complete pathways in the metabolic networks. Analyzing the impact of gap filling in the models led to the identification of inconsistencies in genome annotations.

  Inspired by those results, a protocol was developed for improving the functional annotations of a genus utilizing metabolic reconstructions as a measure of annotation consistency. This resulted in the production of more accurate and consistent annotations and inference of the metabolic network for the genus *Brucella* [2]. We also demonstrated that the use of a controlled vocabulary for the annotation of genomes leads to more consistent annotations, while annotation inconsistencies caused by sequencing and propagation errors still require manually curation.

- **Regulatory networks and gene of expression data –** We conducted an extensive survey of the available resources for gene expression and notable bacterial transcriptional regulatory network data [3]. This survey revealed the lack of detailed regulatory network information and expression data for most organisms. Prompted by the results of this survey, we presented a manually curated regulatory network for *B. subtilis,* compiling information from multiple databases with regulatory data. Our regulatory network reconstruction is more comprehensive when compared with others available in the public domain.

  In addition to the data survey, we also extensively reviewed methods for regulatory network inference. We introduced a new methodology for regulatory interaction inference from expression data called *Atomic Regulon Inference*. Reconciling the manually curated network with atomic regulons allowed us to expand our knowledge of the regulatory network. This analysis also demonstrated how atomic regulons could be a tool in genome annotation efforts. To make use of this potential, atomic regulons were integrated into the SEED database and are available as part of the annotation tools in the system.

- **Integrated models of metabolism and regulation –** We introduced the first genome-scale model for the metabolism and regulation of *B. subtilis* at genome-scale. This accomplishment was only possible by making use of the regulatory network introduced in this thesis, combined with a published GEM for *B sublitis*. Our model was validated with transcription factor (TF) knockouts described in *B. subtilis* dedicated databases and the literature for multiple environmental conditions. The accurate prediction of TF knockouts can make this model of use for methods addressing the discovery of gene deletions for strain optimization [4].

  The integration of the regulatory constraints with the metabolic model also flagged issues with the published metabolic model. As more data become available, the study of phenotypes for validation of regulatory interactions can be used as a tool to perform additional manual curation in the metabolic reconstructions. It is also important to note that the methodologies

necessary for this modeling effort were implemented in the KBase of Systems Biology (www.kbase.us).

Across all the results, we highlight the following main contributions:

- Identification and correction of multiple genome annotation inconsistencies in the SEED database, leading to the development of a protocol to make use of GEMs as tools for genome annotation curation.
- Reconstruction of a more comprehensive manually curated transcriptional regulatory network for *B. subtilis.*
- First genome-scale modeling study for the metabolism and transcriptional regulation of *B. subtilis*.

5.2 LIMITATIONS

All research subjects have inherent limitations. Acknowledging and discussing those limitations can bring insights into some of the conclusions drawn previously:

- **Genome annotations –** Genome functional annotations are a key element for GEM reconstruction. We used genome annotations for over 3000 organisms in the reconstruction efforts conducted. The increasing number of genomes available led to the need to annotate genomes with automatic pipelines that heavily rely in sequence homology, leading to several inconsistencies in the genome annotations [5]. Recent studies evaluating the performance of genome annotation tools have shown that this reliance on sequence homology can have lower accuracy when compared with other methods available [6]. Even the highly curated genome annotation for *B. subtilis*, which is continuously updated by the community still reports approximately 1/3 of unknown gene functions [7].

- **Genome-scale metabolic models** – We made use of automated reconstructed GEMs and a highly curated published model for *B. subtilis*. Even with highly curated models, there is still a lack of capabilities in the pursuit to accurately simulate cellular behavior with no inherent dynamic or regulatory predictions. The use of comprehensible kinetic information for all reactions should be able to allow a more accurate representation of the microorganisms that are undergoing any kind of model reconstruction processes. However, kinetic data are still scarce for most organisms, and the simulation of large-scale dynamic models brings additional computational challenges.

  Regulatory networks – We introduced a more comprehensive transcriptional regulatory network for *B. subtilis*. The network was then used for the integrated model of metabolism and regulation. Capturing only transcriptional regulation limits our understanding of complete regulatory effects in the cell. Other regulatory effects, such as allosteric regulation, post

transcriptional or pure thermodynamics, were shown to also be involved in the regulation of metabolic fluxes [8].

- **Phenotype simulation –** FBA was the method of choice for growth phenotype simulation with genome-scale metabolic models and PROM the algorithm for making predictions with integrated models of metabolism and regulation. It is important to note that PROM uses FBA for growth simulation with the additional regulatory constraints from gene expression data. Predictions based mostly on growth phenotypes represent a simplification of actual cell behavior. Both simulation methods perform their predictions under a steady-state assumption, not being able to perform model predictions in other phases of microorganism's life cycle. The inability of FBA to use kinetic parameters makes it unsuitable to predict metabolite concentrations. Additionally, the PROM algorithm assumes that transcription factors that are not knocked out are active. In the same manner, target genes that are not directly connected to a knocked out transcription factor are also active.

- **Expression data –** To infer regulatory interactions in our studies we made use of gene expression data. We pointed to the high quality of the *B. subtilis* gene expression data sets [9] used, but we were unable to infer proper regulatory interactions for growth in multiple environmental constraints. This fact tells us that increasing the variety of conditions for genome expression studies is still needed to increase the accuracy of methods that infer regulatory interactions from expression data.

5.3 ONGOING AND FUTURE WORK

The following objectives reflect the future research efforts that spawned from the work developed for this thesis:

- The reconstruction of GEMs for all genomes available in the SEED database paved the way for a new version of the ModelSEED reconstruction pipeline to be released in 2015.
- The development of an atomic regulon inference pipeline that will allow users to infer their own atomic regulons. This pipeline will only require users to submit a genome annotated with SEED subsystems and a normalized gene expression dataset.
- The implementation of the methodologies for the modeling of integrated metabolic and regulatory network in the KBase environment (www.kbase.us) will allow us to easily adopt the same strategy to develop integrated models for others organisms.

5.4 REFERENCES

1. Overbeek, R., et al., *The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).* Nucleic Acids Res, 2014. **42**(Database issue): p. D206-14.

2. Faria, J.P., et al., *Enabling comparative modeling of closely related genomes: example genus Brucella.* 3 Biotech, 2014: p. 1-5.

3. Faria, J.P., et al., *Genome-scale bacterial transcriptional regulatory networks: reconstruction and integrated analysis with metabolic models.* Brief Bioinform, 2014. **15**(4): p. 592-611.

4. Burgard, A.P., P. Pharkya, and C.D. Maranas, *Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.* Biotechnol Bioeng, 2003. **84**(6): p. 647-57.

5. Richardson, E.J. and M. Watson, *The automatic annotation of bacterial genomes.* Brief Bioinform, 2013. **14**(1): p. 1-12.

6. Radivojac, P., et al., *A large-scale evaluation of computational protein function prediction.* Nat Methods, 2013. **10**(3): p. 221-7.

7. Florez, L.A., et al., *A community-curated consensual annotation that is continuously updated: the Bacillus subtilis centred wiki SubtiWiki.* Database (Oxford), 2009. **2009**: p. bap012.

8. Kochanowski, K., U. Sauer, and V. Chubukov, *Somewhat in control–the role of transcription in regulating microbial metabolic fluxes.* Curr Opin Biotechnol, 2013. **24**(6): p. 987-93.

9. Chalancon, G., K. Kruse, and M.M. Babu, *Cell biology. Reconfiguring regulation.* Science, 2012. **335**(6072): p. 1050-1.