RECONSTRUCTION AND ANALYSIS OF GENOME-SCALE METABOLIC MODELS OF PHOTOSYNTHETIC ORGANISMS

A Dissertation in

Chemical Engineering

by

Rajib Saha

© 2014 Rajib Saha

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

August 2014
The dissertation of Rajib Saha was reviewed and approved* by the following:

Costas D. Maranas  
Donald B. Broughton Professor of Chemical Engineering  
Dissertation Advisor  
Chair of Committee

Reka Albert  
Professor of Physics and Biology

Howard M. Salis  
Assistant Professor in Chemical Engineering and Agricultural and Biological Engineering

Andrew Zydney  
Walter L. Robb Chair and Professor of Chemical Engineering  
Head of the department of Chemical Engineering

*Signatures are on file in the Graduate School
The scope and breadth of genome-scale metabolic reconstructions has continued to expand over the last decade. However, only a limited number of efforts exist on photosynthetic metabolism reconstruction. Cyanobacteria are an important group of photoautotrophic organisms that can synthesize valuable bio-products by harnessing solar energy. They are endowed with high photosynthetic efficiencies and diverse metabolic capabilities that confer the ability to convert solar energy into a variety of biofuels and their precursors. However, less well studied are the similarities and differences in metabolism of different species of cyanobacteria as they pertain to their suitability as microbial production chassis. Here we assemble, update and compare genome-scale models (iCyt773 and iSyn731) for two phylogenetically related cyanobacterial species, namely *Cyanothece* sp. ATCC 51142 and *Synechocystis* sp. PCC 6803. Comparisons of model predictions against gene essentiality data reveal a specificity of 0.94 (94/100) and a sensitivity of 1 (19/19) for the *Synechocystis* iSyn731 model. The diurnal rhythm of *Cyanothece* 51142 metabolism is modeled by constructing separate (light/dark) biomass equations and introducing regulatory restrictions over light and dark phases. Specific metabolic pathway differences between the two cyanobacteria alluding to different bio-production potentials are reflected in both models. In addition to these cyanobacterial species we also develop a genome-scale model for a plant with direct applications to food and bioenergy production (i.e., maize). The metabolic model *Zea mays* i1563 contains 1,563 genes and 1,825 metabolites involved in 1,985 reactions from primary and secondary maize metabolism. For approximately 42% of the reactions direct evidence for the participation of the reaction in maize was found. We describe results from performing flux balance analysis under different physiological conditions, (i.e., photosynthesis, photorespiration and respiration) of a C4 plant and also explore model predictions against experimental observations for two naturally occurring mutants (i.e., bm1 and bm3). Recently, we develop a second-generation genome-scale metabolic model for the maize leaf to capture C4 carbon fixation by modeling the interactions between the bundle sheath and mesophyll cells. Condition-specific biomass descriptions are introduced that account for amino acids, fatty acids, soluble sugars, proteins, chlorophyll, lingo-cellulose, and nucleic acids as experimentally measured biomass constituents. Compartmentalization of the model is based on proteomic/transcriptomic data and literature evidence. With the incorporation of the information from MetaCrop and MaizeCyc databases, this updated model spans 5824 genes, 8484 reactions, and 8918 metabolites, an increase of approximately five times the size of the earlier iRS1563 model. Transcriptomic and proteomic data is also used to introduce regulatory constraints in the model to simulate the limited nitrogen condition and glutamine synthetase gln1-3 and gln1-4 mutants. *In silico* results have achieved over 62% accuracy in predicting the direction of change in the metabolite pool under each of the mutant conditions compared to the wild-type condition with 82% accuracy determined in the limited nitrogen condition. The developed model corresponds to the largest and more complete to-date effort at cataloguing metabolism for any plant tissue-type.
TABLE OF CONTENTS

LIST OF FIGURES .................................................................................................................. x

LIST OF TABLES ..................................................................................................................... xii

ACKNOWLEDGEMENTS .......................................................................................................... xiii

Chapter 1 RECENT ADVANCES IN THE RECONSTRUCTION OF METABOLIC MODELS AND INTEGRATION OF OMICS DATA .......................... 1

1.1 Introduction .................................................................................................................... 1

1.2 Metabolic model reconstruction approaches ................................................................. 3

1.3 Integration of omics data in metabolic models ................................................................. 6

1.4 Concluding remarks ....................................................................................................... 10

Chapter 2 RECONSTRUCTION AND COMPARISON OF THE METABOLIC POTENTIAL OF CYANOBACTERIA CYANOTHICE SP. ATCC 51142 AND SYNECHOCYSTIS SP. PCC 6803 ......................................................................................... 11

2.1 Introduction .................................................................................................................... 11

2.2 Materials and methods .................................................................................................. 14

  2.2.1 Measurement of biomass precursors ........................................................................ 14

    2.2.1.1 Growth conditions ............................................................................................ 14

    2.2.1.2 Pigments ........................................................................................................... 14

    2.2.1.3 Amino Acids ..................................................................................................... 15

    2.2.1.4 Other cellular components .............................................................................. 15

2.2.2 Model simulations ...................................................................................................... 16
2.3 Results and Discussion

2.3.1 Model components

2.3.1.1 Biomass composition and diurnal cycle

2.3.1.2 Identification and correction of network gaps

2.3.2 GPR associations and elemental and charge balancing

2.3.3 Comparison of iSyn731 model predicted flux ranges against experimental measurements

2.3.4 iSyn731 model testing using in vivo gene essentiality data

2.3.5 Model comparisons

2.3.5.1 Synechocystis 6803 model comparisons

2.3.5.2 Cyanothece 51142 model comparisons

2.3.5.3 iSyn731 and iCyt773 models comparison

2.3.6 Using iSyn731 and iCyt773 to estimate production yields

2.4 Conclusion

Chapter 3 SYNTHETIC BIOLOGY OF CYANOBACTERIA: UNIQUE CHALLENGES AND OPPORTUNITIES

3.1 Introduction

3.2 Genetic modification of cyanobacteria

3.3 Genetic modification in cis: chromosome editing

3.4 Genetic modification in trans: foreign plasmids

3.5 Unique challenges of the cyanobacterial lifestyle

3.5.1 Life in a diurnal environment
3.5.2 Redirecting Carbon Flux by decoupling growth from production..... 56
3.5.3 RNA-based regulation ................................................................. 56
3.6 Parts for Cyanobacterial Synthetic Biology........................................ 57
3.6.1 Inducible Promoter........................................................................ 58
3.6.2 Reporters ...................................................................................... 66
3.6.3 Cultivation systems ........................................................................ 68
3.7 Genome-scale modeling and fluxomics of cyanobacteria .................... 68
3.7.1 Challenges .................................................................................... 70
  3.7.1.1 Incorporating photoautotrophy into metabolic models .............. 70
  3.7.1.2 Incompleteness of genome annotation .................................... 71
  3.7.1.3 Fewer mutant resources to test model accuracy ....................... 72
3.7.2 Recent advances ........................................................................... 73
  3.7.2.1 Detailed genome-scale models ............................................... 73
  3.7.2.2 $^{13}$C MFA analysis .................................................................. 74
3.8 Conclusions ...................................................................................... 75

Chapter 4  
ZEAS MAYS iRS1563: A COMPREHENSIVE GENOME-SCALE 
METABOLIC RECONSTRUCTION OF MAIZE METABOLISM .............. 76
4.1 Introduction ....................................................................................... 76
4.2 Results .............................................................................................. 79
  4.2.1 Construction of Auto & Draft models ...................................... 79
  4.2.2 Generation of computations-ready model ................................. 81
  4.2.3 Network connectivity analysis and restoration ......................... 86
Chapter 4 NITROGEN USE EFFICIENCY IN MAIZE (ZEA MAYS L.): FROM “OMICS” STUDIES TO METABOLIC MODELING ........................................ 105

4.1 Introduction ........................................................................................................................................ 105

4.2 Why improve nitrogen use efficiency in a crop such as maize? ........................................ 106

4.3 Nitrogen use efficiency: from “omics” studies to systems biology
   approaches .......................................................................................................................................... 108

4.4 Transcriptome studies ...................................................................................................................... 109

4.5 Proteome studies .............................................................................................................................. 114

4.6 Metabolome studies ........................................................................................................................ 115

4.7 Integrating “omics” data ............................................................................................................... 117

4.8 Metabolic modeling as a tool to unravel the limiting steps in NUE ........................................ 120

4.9 Incorporating transcriptome, proteome, and metabolome data into models
   .......................................................................................................................................................... 125

Chapter 5 NITROGEN USE EFFICIENCY IN MAIZE (ZEA MAYS L.): FROM “OMICS” STUDIES TO METABOLIC MODELING ........................................ 105
### 5.10 Concluding remarks

**Chapter 6 ASSESSING THE METABOLIC IMPACT OF NITROGEN AVAILABILITY USING A COMPARTMENTALIZED MAIZE LEAF GENOME-SCALE MODEL**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1 Introduction</td>
<td>129</td>
</tr>
<tr>
<td>6.2 Results and Discussion</td>
<td>133</td>
</tr>
<tr>
<td>6.2.1 Effect of Nitrogen Conditions on Biomass Components</td>
<td>133</td>
</tr>
<tr>
<td>6.2.2 Development of the second-generation maize leaf model</td>
<td>134</td>
</tr>
<tr>
<td>6.2.3 Incorporation of ‘omics’ data in the model</td>
<td>136</td>
</tr>
<tr>
<td>6.2.4 Flux range variations among conditions</td>
<td>140</td>
</tr>
<tr>
<td>6.3 Concluding remarks</td>
<td>142</td>
</tr>
<tr>
<td>6.4 Materials and Methods</td>
<td>143</td>
</tr>
<tr>
<td>6.4.1 Plant Material</td>
<td>143</td>
</tr>
<tr>
<td>6.4.2 Yield Components Analysis</td>
<td>144</td>
</tr>
<tr>
<td>6.4.3 RNA and DNA Preparation</td>
<td>145</td>
</tr>
<tr>
<td>6.4.4 Gene Expression Profiles using Maize cDNA Microarrays</td>
<td>145</td>
</tr>
<tr>
<td>6.4.5 Statistical Analysis of Maize cDNA Microarray Data</td>
<td>147</td>
</tr>
<tr>
<td>6.4.6 Total Protein Extraction, Solubilization, and Quantification</td>
<td>147</td>
</tr>
<tr>
<td>6.4.7 Two-dimensional Electrophoresis, Gel Staining, and Image Analysis</td>
<td>148</td>
</tr>
<tr>
<td>6.4.8 Protein Identification by LC-MS/MS</td>
<td>149</td>
</tr>
<tr>
<td>6.4.9 Metabolite Extraction and Analyses</td>
<td>149</td>
</tr>
</tbody>
</table>
6.4.10 Metabolome Analysis ................................................................. 151
6.4.11 Model Development and Curation............................................. 153
6.4.12 Incorporation of Transcriptomic, Proteomic and Metabolomic Data 158

References............................................................................................... 161

Appendix.................................................................................................... 190
LIST OF FIGURES

FIGURE 1-1: OUTLINE FOR THE DEVELOPMENT OF A HIGH-QUALITY METABOLIC MODEL. ..... 4

FIGURE 2-1: COMPARISON OF MODEL DERIVED AND EXPERIMENTALLY MEASURED [112] FLUX RANGES FOR *SYNECHOCYSTIS* 6803 UNDER THE MAXIMUM BIOMASS CONDITION. 24

FIGURE 2-2: COMPARISON OF GENE ESSENTIALITY/VIABILITY DATA WITH PREDICTIONS BY A NUMBER OF *SYNECHOCYSTIS* 6803 MODELS. ................................................................. 31

FIGURE 2-3: VENN DIAGRAM DEPICTING (COMMON AND UNIQUE) REACTIONS AND METABOLITES BETWEEN: (A) *iJN678* [109] AND *iSYN731*, (B) *iCCE806* [102] AND *iCYT773*, AND (C) *iSYN731* AND *iCYT773* MODELS......................................................... 34

FIGURE 2-4: SCHEMATICs THAT ILLUSTRATE THE THERMODYNAMICALLY INFEASIBLE CYCLES AND SUBSEQUENT RESOLUTION STRATEGIES......................................................... 35

FIGURE 2-5: LIST OF ADDED REACTIONS ACROSS PATHWAYS. (A) *iSYN731* COMPARED TO *iJN678* [109], AND (B) *iCYT773* COMPARED TO *iCCE806* [102]. ................................. 39

FIGURE 2-6: EXAMPLES OF PATHWAYS THAT DIFFER BETWEEN THE TWO CYANOBACTERIA. ......................................................................................................................... 41

FIGURE 3-1: DIFFERENT METHODS FOR CONSTRUCTING CYANOBACTERIAL MUTANTS. .... 49

FIGURE 3-2: DNA ASSEMBLY METHODS. ........................................................................ 53

FIGURE 3-3: USING FLUXOMICS AND GENOME SCAE MODELS TO LINK GENOTYPE TO METABOLIC PHENOTYPE. ........................................................................ 69

FIGURE 4-1: SPECIES ORIGIN OF NEWLY ADDED REACTIONS IN THE DRAFT MODEL........ 81

FIGURE 4-2: EXAMPLE OF CONNECTIVITY RESTORATION FOR PHENYLACETALDEHYDE. .... 88
Figure 4-3: Distribution of metabolites based on their number of appearance in different organelles

90

Figure 4-4: Compartment and localization information for Zea mays IRS1563.

93

Figure 4-5: Venn diagram for genes, reactions and metabolites. (A) between Zea mays IRS1563 and A. thaliana IRS1597, and (B) between Zea mays IRS1563 and Arabidopsis thaliana IRS1563 and Arabidopsis thaliana IRS1597, and (C) between Zea mays IRS1563 and maize C4GEM.

95

Figure 4-6: Maximum theoretical yields of (A) glucose and (B) galactose for wild-type vs bm1 mutant and wild-type vs bm3 mutant, respectively

99

Figure 5-1: Changes in metabolite content, enzyme activities and biomass-related components in leaves of nineteen selected maize lines covering their genetic diversity (Camus-Kulandaivelu et al., 2006) at two key stages of plant development.

119

Figure 5-2: Iterative process of genome-scale model building.

122

Figure 5-3: Flux balance analysis using a metabolic model.

124

Figure 6-1: Number of metabolic and transport reactions distributed between compartments in the bundle-sheath and mesophyll cell types.

135

Figure 6-2: Number of metabolites in each condition that statistically vary from the WT condition at the vegetative stage.

140

Figure 6-3: Effect of “omics” based regulation on the flux-sum prediction compared to the experimental trend in metabolite concentration.

141

Figure 6-4: Model development and curation schematic.

154
**LIST OF TABLES**

| Table 2-1: *Synechocystis* 6803 *i*Syn731 and *Cyanothece* 51142 *i*Cyt773 model statistics | 18 |
| Table 2-2: Summary of connectivity restoration in *Synechocystis* 6803 *i*Syn731 and *Cyanothece* 51142 *i*Cyt773 models | 22 |
| Table 2-3: Comparison of $^{13}$C MFA flux measurements [112] vs. model-predicted flux ranges | 25 |
| Table 3-1: Model strains of cyanobacteria for synthetic biology | 48 |
| Table 3-2: Inducible promoters used in cyanobacterial hosts | 59 |
| Table 4-1: Model size after each reconstruction step | 81 |
| Table 4-2: Maize gene annotation via bidirectional BLASTp homology searches against NCBI non-redundant protein database | 83 |
| Table 4-3: Biomass component list in *i*RS1563 | 85 |
| Table 4-4: Definition of three different physiological states | 86 |
| Table 4-5: Restoration of network connectivity using GapFill [36] | 87 |
| Table 4-6: Change in content of cell wall components in *bm1* and *bm3* Maize mutants | 97 |
| Table 5-1: Transcripts exhibiting significant increase following transfer from limiting to non-limiting N feeding conditions in different studies and across different species | 111 |
| Table 6-1: Number of reactions after each model creation and curation step | 137 |
ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Costas Maranas for all his guidance. The vision for the current project, the context of the research and the content of this dissertation have, in large part, been possible thanks to him. His broad and deep scientific knowledge combined with his scholarly way of student advising has been exceptionally inspiring which is why I am, indeed, indebted to him for teaching me how to perform high quality research. In addition, his incessant emphasis on improving presentation and communication skills as well as technical writing has played an important role towards achieving my academic and professional career goals. I would also like to thank to Drs. Andrew Zydney, Howard Salis and Reka Albert for agreeing to serve on my doctoral committee. Also I would like to extend special thanks to past and present members of Costas Maranas’ group specially Dr. Ali Zomorrodi for invaluable guidance as well as Anupam Chowdhury, Akhil Kumar, Dr. Anthony Burgard, Ali Khodayari and Satyakam Dash for insightful discussions. I would like to convey my sincere gratitude to my wife Dr. Shudipto Dishari whose love, encouragement and persistent confidence in me helped me to stick to my goal. Last but not least, I must thank my parents Bela and Debdash for their sacrifice, love and continuous support throughout my life and in all my pursuits.
Chapter 1

RECENT ADVANCES IN THE RECONSTRUCTION OF METABOLIC MODELS AND INTEGRATION OF OMICS DATA

This chapter has been previously published in modified form in Current Opinions in Biology (Saha, R.*, Chowdhury, A.* and C.D. Maranas (2014), "Recent advances in the reconstruction of metabolic models and integration of omics data", Current Opinion in Biotechnology, 29, 39-45) (*Authors contributed equally).

1.1 Introduction

A metabolic network captures the inter-conversion of metabolites through chemical transformations catalyzed by enzymes. To this end, a metabolic model describes reaction stoichiometry and directionality, gene to protein to reaction associations (GPRs), organelle-specific reaction localization, transporter/exchange reaction information, transcriptional/translational regulation and biomass composition [1]. By defining the metabolic space, a metabolic model can assess allowable cellular phenotypes under specific environmental and/or genetic conditions [2, 3]. The number of metabolic models developed in the past several years is a testament to their increasing usefulness and penetration in many areas of biotechnology and biomedicine [4-6]. Initially, metabolic models have been used to characterize biological systems and develop non-intuitive strategies to reengineer them for enhanced production of valuable bioproducts [7]. More recently, models have been developed and applied for a variety of goals ranging from metabolic disease drug target identification, study of microbial pathogenicity and parasitism (as highlighted in [5]).
The validation of high-quality [8] models is critical for not only recapitulating known physiological properties but also improving their prediction accuracy. Towards this end, strategies have been developed to incorporate other cellular processes such as gene/protein expression to better understand the emergence of complex cellular phenotypes [9, 10]. For example, genome-scale metabolic models of pathogens have been reconstructed to develop novel drugs for combating infections and also minimize side effects in the host [11]. An integrated model [12] of E. coli has been developed by combining Metabolism with gene Expression (i.e., ME model) to increase the scope and accuracy of model-computable phenotypes corresponding to the optimal growth condition. In addition, by combining all of the molecular components as well as their interactions, a whole-cell model [13] has been developed for Mycoplasma genitalium, a human pathogen, to study previously unexplored cellular behaviors including protein-DNA association and correlation between DNA replication initiation and replication itself. Tissue specific models have also been developed for eukaryotic organisms, such as Homo sapiens [14] and Zea mays [2], to scope out novel therapeutic targets and characterize metabolic capabilities, respectively. Moving beyond the single cell/tissue level, multi-cell/multi-tissue type metabolic models have been reconstructed for higher organisms. For example, Homo sapiens [14, 15] models have been employed for biomedicine applications and a Hordeum vulgare [16] model has been deployed for studying crop improvement and yield stability.

With rapid improvements in sequencing (and annotating) tools and techniques, the number of complete genomes (and annotations) is increasing at an exponential pace [17]. Metabolic models can greatly facilitate the assessment of the potential metabolic phenotypes attainable by these organisms. Therefore, rapid development of high-quality metabolic models and algorithms for analyzing their content are of critical importance. The recent genome-scale metabolic models, their automated generation, improvements and applications have been reviewed elsewhere [4, 18-20] and will not be covered in detail in this review. Rather, in this mini-review we will critically evaluate the available
repositories, model-building and data integration techniques and existing challenges related to rapid reconstruction of high-quality metabolic models.

1.2 Metabolic model reconstruction approaches

Metabolic network/model reconstruction process follows three major steps (as highlighted in Figure 1-1). Initially, upon sequencing and annotating a genome of interest, literature sources and/or homology searches are used to assign function to all the Open Reading Frames (i.e., ORFs). For every function with a metabolic fingerprint a specific chemical transformation is assigned. Therefore, by iteratively marching along the entire genome, a compilation of reactions encompassing the entire chemistry repertoire of the organism can be achieved. It must be noted that these models are not necessarily predictive but instead have a scoping nature by allowing us to assess what is metabolically feasible. Regulatory constraints on reaction fluxes are incorporated based on the thermodynamic (i.e., reaction reversibility) and omics (i.e., transcriptomic/proteomic) data that can further sharpen predictions.

One of the most critical steps of metabolic model building is to establish GPR information of a specific organism from biological databases and/or literature sources. To this end, biological databases (as highlighted in [1]) such as KEGG, SEED, Metacyc, BKM-react, Brenda, Uniprot, Expasy, PubChem, ChEBI and ChemSpider provide information about reactions/metabolites and associated enzymes and genes. However, as illustrated by Kumar et al.[1], incompatibilities in data representation such as metabolites with multiple names/chemical formulae across databases, stoichiometric errors (i.e., elemental or charge imbalances) and incomplete atomistic detail (e.g. absence of stereo-specificity, and presence of R-group(s)) are key bottlenecks for rapid reconstruction of new high-quality metabolic models by combining information from these databases. Recently, databases such as MetRxn [1] have been developed to address these issues by integrating information (of metabolites and reactions) from eight such databases and 90
published metabolic models. Overall, MetRxn (as of Dec, 2013) contains over 44,000 unique metabolites and 35,000 unique reactions that are charge and elementally balanced.

**Figure 1-1:** Outline for the development of a high-quality metabolic model.

The first step involves retrieving data from different biological databases, physiology and biochemistry of the organism as well as published literature. In the next step, GPR associations are established, the biomass equation is described based on experimental measurement and the model is represented in the form of a stoichiometric matrix. Furthermore, gaps in the model are identified and reconciled based on established gap filling techniques. Finally, in the third step high-throughput experimental measurements such as transcriptomic, proteomic and fluxomic data are utilized to improve the model accuracy.
In addition to the GPR information, subcellular localization of metabolic enzymes/reactions is critical to develop the metabolic model of any eukaryotic organism. In this regard, there exist protein localization databases such as PPDB [21] and SUBA [22] for plant species (e.g., Arabiopsis thaliana and Zea mays). There are also computational algorithms [23] to predict enzyme/reaction localization (when limited amount of localization data is available) by utilizing the embedded metabolic network and parsimony principal to minimize the number of transporters. However, none of these databases or algorithms is complete or error-proof, which necessitates manual scrutiny before making any final reaction assignments in one or multiple intracellular organelle(s).

In addition to databases of metabolic functionalities, there exists a number of knowledgebases of regulation (e.g., RegulonDB [24] (for E. coli) and Grassius [25] (for grasses)) and kinetic parameters (e.g., Sabio-RK [26], Ecocyc [27] and Brenda [28] for E. coli). Nevertheless, such information is largely incomplete and unavailable for all but a few model organisms, which emphasizes the need to thoroughly refer to primary published literature sources.

By making use of data from different biological databases, draft metabolic models can be reconstructed in an automated [18, 29-33] or semi-automated [34-37] fashion. Automated methods are fast and require minimal user input while semi-automated methods are slower and require user feedback and inspection. Automated methods such as SEED [29], BioNetBuilder [30] and ReMatch [31] can integrate data from several databases. However, the user is responsible for assessing the accuracy of the network gap filling step, removing thermodynamically infeasible cycles (e.g. using loopless FBA method [38]) and customizing the biomass composition to the organism of interest. Semi-automated methods (e.g., RAVEN [34], MicrobeFlux [35] and other works, as highlighted in [2, 3, 36]), make use of not only available databases but also published models of closely related species. These methods allow for user-driven gap filling and growth-discrepancy reconciliation measures and use biomass compositions based on experimental measurements whenever available. However, the existing semi-automated algorithms often create thermodynamically infeasible cycles while reconciling any
network gaps [39] or fixing growth inconsistencies [40, 41]. Overall, the automated methods are very useful for developing initial draft models, whereas semi-automated methods can refine these draft models to bring them to a required completion level.

Free energy reaction change estimates are frequently used to impose thermodynamic constraints on reaction fluxes, metabolite concentrations and kinetic parameters as highlighted elsewhere [42, 43]. The group contribution method [44] or recently improved group contribution method [45] can be utilized to estimate the reaction Gibbs free energy and ultimately predict the reaction direction. Furthermore, Hamilton et al. have developed TMFA [46] (Thermodynamics-based Metabolic Flux Analysis) to quantify metabolite concentrations and reaction free energy ranges and examine the effect of thermodynamic constraints on the allowable flux space (of the iJR904 E. coli model) that improve model performance such as gene essentiality prediction. Although TMFA provides some idea about directionality, the thermodynamic constraints can be too wide. Therefore, as shown in iAF1260 [47] E. coli model, literature survey still remains the best source for assigning reaction directionality.

1.3 Integration of omics data in metabolic models

In this section we review recent developments in integrating high-throughput omics data with metabolic models and critically analyze their contribution towards improving the genotype-phenotype prediction and metabolic network properties. Due to the underdetermined nature of genome-scale metabolic models, a lot of effort has been expended at improving the accuracy of estimation for the reaction fluxes. Metabolic Flux Analysis (MFA) [48] is a unique resource for quantifying internal metabolic fluxes by using relative enrichment of substrate labels from isotope labeling experiments (ILE) as additional information [49-51]. Detailed atom-transition information for each reaction involved in the MFA network is collected either from literature (for well-studied pathways), databases [52, 53], or from motif-searching optimization algorithms [54, 55]. Subsequently, the fluxes and their confidence intervals in the network are estimated by
minimizing the sum-squared error between experimental and simulated mass isotopomer distribution (MID) data using different optimization frameworks. Recent advances in the systematic identification of the input substrate labels [56] and the design of labeling experiments (e.g. [57]) has improved the accuracy and scope of flux estimation. The inferred flux data could then be integrated into metabolic models for further sharpening the allowable flux ranges of the remaining reactions in the model (using Flux Variability Analysis). A key impediment of MFA is that it is generally applied to core models of metabolism [58] spanning less than 5% [59] of a genome-scale metabolic model. As a result, flux information is generally available for the central carbon metabolism of the organism with limited information on flux redirections in other parts of metabolic network in response to genetic or environmental perturbations. In addition, the results are sensitive to the selection of the metabolic network used to fit the labeling data [59]. Even though recent attempts have been made at constructing large-scale MFA networks using the flux coupling method [60] and the elementary carbon modes approach [61], flux analysis at a genome-scale level has not been attempted yet.

Metabolic model prediction accuracy can further be improved by incorporating transcriptomic/proteomic data as regulatory constraints (see Figure 1-1). Thus, condition- and tissue-specific metabolic models can be developed to simulate specific phenotypes [62]. The main approaches for integrating omics data to abstract regulation can be broadly classified into two categories [62]: (a) the switch approach (e.g., GIMME and iMAT): on/off reaction fluxes based on threshold expression levels, and, (b) the valve approach (e.g. E-Flux and PROM): regulate reaction fluxes based on relative gene/protein expressions. To circumvent the problem of using arbitrary cutoffs for gene expression, recent approaches [63, 64] use absolute gene expression levels as a penalty metric such that the sum of squared error between the gene expressions and their encoded reaction fluxes is minimized. Overall, all of these approaches make the underlying assumption that transcription of genes is linearly correlated with the flux of the reactions they encoded, which is not necessarily accurate [65]. However, faced with a lack of detailed mechanistic information between transcription and enzyme activity, these
frameworks provide a “first-guess” type estimate for correlating genotype with phenotype.

Regulatory signaling and transcription networks have been integrated as separate modules with metabolic networks [66, 67]. Generally, these are simulated as boolean networks where information from signaling molecules and transcription factors is carried as on-off signals to target proteins. Similar frameworks have also been constructed for translation and post-translational regulation [68, 69]. Besides providing a mechanistic basis for correlating the genotype with the observed phenotype, these integrated frameworks have the added advantage of being dynamic in nature. Each module is assumed to be in an independent quasi-steady state during a specific time interval, and is updated for the next interval by solving a system of ordinary differential equations (ODEs) of the variables (e.g. enzyme and metabolite concentrations, transcription factors, and mRNA abundances) interacting at the interface of two modules. More recently, this framework has been extended to construct the first whole-cell model of *M. genitalium* [13], where 28 cellular functions designed as distinct modular networks have been integrated into a whole-cell dynamic framework interacting at the edges with ODEs of eight types of common variables. The whole-cell network couples metabolic and non-metabolic functions, as well as temporal information of protein localization and cell replication. However, it requires a detailed mechanistic approach to accurately describe transcription, translation and regulation of the enzyme activities, which is seldom available. In addition, the assumption that each module is at a quasi-steady state within the same time interval may not be universally applicable. Nevertheless, the whole-cell model framework is a major landmark in the reconstruction of integrated metabolic networks.

Integrated frameworks for metabolic model development discussed so far do not use detailed mechanistic relations to link gene expressions with reaction fluxes. The ME model framework [12, 70] has been developed to provide a detailed mechanistic basis to quantify transcription of mRNA, translation of proteins, formation of protein complexes,
catalysis of reactions and formation of macromolecules. Similar to flux balance analysis (FBA), simulations using ME models minimize the cellular machinery required to sustain an experimentally observed growth rate, where protein dilution is coupled with the growth of the organism. This framework can predict gene and protein expressions with reasonable accuracy along with an improved prediction for reaction fluxes. The ME model is also able to drive discovery of protein regulation. Despite not accounting for any post-transcriptional regulation, the ME framework provides a significant step towards a systems-wide quantitative description of biological processes.

Several attempts have also been made to link the enzyme activity and metabolite concentrations with the reaction fluxes of detailed mechanistic networks. Detailed kinetic models have been constructed using steady-state phenotype information for the wild-type organism and several of its mutants. For example, Cotton et al. [71] have constructed the kinetic model of central metabolism for E. coli, where the kinetic parameters are identified by minimizing the error between the experimental and model-predicted values of metabolite concentrations and enzyme activity for the wild-type and several of its single gene mutants [72]. The kinetic expressions are imported from an earlier kinetic model for E. coli [73]. Likewise, Vital Lopez et al. [74] have constructed a kinetic network for E. coli central metabolism spanning over 100 reactions using mass action kinetic expressions derived from transcriptomic and fluxomic data. The major restrictions in these models are either the size of the network (for the first one [71]), or the accuracy of the kinetic expressions (for the latter one [74]). Such limitations could be resolved by using the ensemble modeling approach [75] where each reaction is decomposed into its elementary steps (with detailed regulations, available from databases (e.g. BRENDA [28]), and the ensemble of kinetic models is filtered using fluxomic data for mutants. Genome-scale kinetic models using approximate mass action [76] or lin-log kinetics [77] have also been developed. In the latter approach [77], the kinetic parameters are estimated from metabolomic information and FBA. While these methods require significantly more refinement, especially in the construction of the kinetic expressions,
they delineate a strategy for future construction of high confidence, integrated metabolic networks linking the genome to the observable phenotypes.

### 1.4 Concluding remarks

Metabolic network models play an important role in quantitatively assessing the allowable metabolic phenotype of an organism and thereby can be deployed to guide metabolic engineering, synthetic biology and/or drug targeting interventions. Through the coordinated use of biological databases, model building strategies and high-throughput omics-data integration techniques both the quality and scope of metabolic models is increasing. However, significant knowledge gaps and a lack of best-practice methodologies require additional scrutiny. For example, delineating the effect of different levels of regulation (transcriptional, translational and/or post-translational) on metabolic flux would help establish the connectivity and directionality of regulation in metabolic models. In addition, the design of labeling protocols that will enable the elucidation of metabolic fluxes beyond core metabolism in a high-throughput manner for a number of genetic and/or environmental perturbations will provide the basis for the parameterization and construction of more predictive metabolic models. Finally, the adoption of common standards in metabolite and reaction description will speed up sharing of information across database resources. By integrating “best-practice” lessons learned from model organisms the development of systematic workflows will facilitate the construction of high-quality metabolic models for less studied organisms.
2 Chapter 2

RECONSTRUCTION AND COMPARISON OF THE METABOLIC
POTENTIAL OF CYANOBACTERIA CYANOTHECE SP. ATCC 51142
AND SYNECHOCYSTIS SP. PCC 6803


2.1 Introduction

Cyanobacteria represent a widespread group of photosynthetic prokaryotes [78]. By contributing oxygen to the atmosphere, they played an important role in the precambrian phase [79]. Cyanobacteria are primary producers in aquatic environments and contribute significantly to biological carbon sequestration, O$_2$ production and the nitrogen cycle [80-82]. Their inherent photosynthetic capability and ease in genetic modifications are two significant advantages over other microbes in the industrial production of valuable bioproducts [83]. In contrast to other microbial production processes requiring regionally limited cellulosic feedstocks, cyanobacteria only need CO$_2$, sunlight, water and a few mineral nutrients to grow [83]. Sunlight is the most abundant source of energy on earth. The incident solar flux onto the USA alone is approximately 23,000 terawatts which dwarfs the global energy usage of 3.16 terawatts [84]. Cyanobacteria perform photosynthesis more efficiently than terrestrial plants (3-9% vs. 2.4-3.7%) [85]. The short life cycle and transformability of cyanobacteria combined with a detailed understanding of their biochemical pathways are significant advantages of cyanobacteria as efficient platforms for harvesting solar energy and producing bio-products such as short chain alcohols, hydrogen and alkanes [83].
The genus *Cyanothece* includes unicellular cyanobacteria that can fix atmospheric nitrogen. *Cyanothece* sp. ATCC 51142 (hereafter *Cyanothece* 51142) is one of the most potent diazotrophs characterized and the first to be completely sequenced [86]. Studies show that it can fix atmospheric nitrogen at rates higher than many filamentous cyanobacteria and also accommodate the biochemically incompatible processes of photosynthesis and nitrogen fixation within the same cell by temporally separating them [87]. *Synechocystis* sp. PCC 6803 (hereafter *Synechocystis* 6803), the first photosynthetic organism with a completely sequenced genome [88], is probably the most extensively studied model organism for photosynthetic processes [89]. It is also closely related to *Cyanothece* 51142 and shares many characteristics with all *Cyanothece* [86]. The genome of *Cyanothece* 51142 is about 35% larger than that of *Synechocystis* 6803 mostly due to the presence of nitrogen fixation and temporal regulation related genes in *Cyanothece* 51142 [86]. *Synechocystis* 6803 has been the subject of many targeted genetic manipulations (e.g., expression of heterologous gene products) as a photo-biological platform for the production of valuable chemicals such as poly-beta-hydroxybutyrate, isoprene, hydrogen and biofuels [89-97]. However, genetic tools for *Cyanothece* 51142 are still lacking thus hampering its wide use as a bio-production strain even though it has many attractive native pathways. For example, *Cyanothece* 51142 can produce (in small amounts) pentadecane and other hydrocarbons while containing a novel (though incomplete) non-fermentative pathway for producing butanol [98, 99].

A breakthrough in solar biofuel production will require following one of two strategies: 1) obtaining photosynthetic strains that naturally have high-throughput pathways analogous to those in known biofuel producers, or 2) creating cellular environments conducive for heterologous enzyme function. Despite its attractive capabilities including nitrogen fixation and H₂ production [96], unfortunately genetic tools are not currently available to efficiently test engineering interventions directly for *Cyanothece* 51142. Therefore, a promising path forward may be to use *Synechocystis* 6803 as a “proxy” (for which a comprehensive genetic toolkit is available) and subsequently transfer knowledge gained during experimentation with *Synechocystis* 6803 to *Cyanothece* 51142. This requires high quality metabolic models for both organisms. Comprehensive genome-wide
metabolic reconstructions include the complete inventory of metabolic transformations of a given cyanobacterial system. Comparison of the metabolic capabilities of *Cyanothece* 51142 and *Synechocystis* 6803 derived from their corresponding genome-scale models will provide valuable insights into their niche biological functions and also open up new avenues for economical biofuel production.

Genome-scale models (GSM) contain gene to protein to reaction associations (GPRs) along with a stoichiometric representation of all possible biotransformations known to occur in an organism combined with a set of appropriate regulatory constraints on each reaction flux [100, 101]. By defining the global metabolic space and flux distribution potential, GSMs can assess allowable cellular phenotypes under specific environmental conditions [100, 101]. The first genome-scale model for *Cyanothece* 51142 was recently published [102]. The authors addressed the complexity of the electron transport chain (ETC) and explored further the specific roles of photosystem I (PSI) and photosystem II (PSII). In contrast, *Synechocystis* 6803 has been the target for metabolic model reconstruction for quite some time [89, 103-109]. Most of these earlier efforts for *Synechocystis* 6803 focused on only central metabolism [103-105]. Knoop et al. [89] and Montagud et al. [107, 108] developed genome-scale models for *Synechocystis* 6803, analyzed growth under different conditions, identified gene knock-out candidates for enhanced succinate production and performed flux coupling analysis to detect potential bottlenecks in ethanol and hydrogen production. A more recent model describes in detail the photosynthetic apparatus, identifies alternate electron flow pathways and highlights the high photosynthetic robustness of *Synechocystis* 6803 during photoautotrophic metabolism [109]. All these efforts have brought about an improved understanding of the metabolic capabilities of *Synechocystis* 6803 and cyanobacterial systems in general.

This paper introduces high-quality genome-scale models for *Cyanothece* 51142 iCyt773 and *Synechocystis* 6803 iSyn731 (as shown in Table 2-1) that integrate all recent developments [102, 109], supplements them with additional literature evidence and highlights their similarities and differences. As many as 322 unique reactions are introduced in the *Synechocystis* iSyn731 model and 266 in *Cyanothece* iCyt773. New
pathways include, among many, a TCA bypass [110], heptadecane biosynthesis [98] and detailed fatty acid biosynthesis in iSyn731 and comprehensive lipid and pigment biosynthesis and pentadecane biosynthesis [98] in iCyt773. For the first time, not only extensive gene essentiality data [111] is used to assess the quality of the developed model (i.e., iSyn731) but also the allowable model metabolic phenotypes are contrasted against MFA flux data [112]. The diurnal rhythm of Cyanothece metabolism is modeled for the first time via developing separate (light/dark) biomass equations and regulating metabolic fluxes based on available protein expression data over light and dark phases [113].

2.2 Materials and methods

2.2.1 Measurement of biomass precursors

2.2.1.1 Growth conditions

Wild-type Synechocystis 6803 and Cyanothece 51142 were grown for several days from an initial OD$_{730}$ of ~0.05 to ~0.4. Synechocystis 6803 was grown in BG-11 medium [114] and Cyanothece 51142 in ASP2 medium [115] with (+N) or without (-N) nitrate. All cultures were grown in shake flasks with continuous illumination of ~100 µmol photons/m$^2$/sec provided from cool white fluorescent tubes. Synechocystis was maintained at 30°C and Cyanothece at 25°C. For Synechocystis, the illumination was constant and doubling time was ~24 hours. Cyanothece alternated between 12 hours of light and 12 hours of darkness, with a doubling time of ~48 hours.

2.2.1.2 Pigments

1 mL of cells of both Synechocystis 6803 and Cyanothece 51142 (from light and dark phases) was pelleted and extracted twice with 5 mL 80% aqueous acetone and the extracts pooled. Spectra of this extract and of a sample of whole cells were taken on a
DW2000 spectrophotometer (Olis, GA, USA) against 80% acetone or BG-11 media as a reference. Chlorophyll a contents were calculated as reported [116] from the acetone extract. Total carotenoid concentrations were also calculated from the acetone extract according to a published method [117]. The relative amounts of different carotenoids included in the biomass equation were estimated according to known ratios [118]. Concentrations of phycocyanin were estimated from the spectra of intact cells [119]. All measurements were taken in triplicate.

2.2.1.3 Amino Acids

Total protein contents were measured using a Pierce BCA Assay kit. Amino acid proportions were determined according to published shotgun proteomics data for both *Cyanothece* 51142 and *Synechocystis* 6803 across a range of conditions [120] according to the following procedure: From peptide-level data, each mass spectral observation of a peptide was taken as an instance of a particular protein. The amino acid composition of each protein was taken from data in Cyanobase (http://genome.kazusa.or.jp/cyanobase) and thus the ‘proteome’ was taken to include all of the proteins whose peptides were observed in our data set, in proportion according to how often their peptides were observed. Amino acid frequencies were averaged across the proteome by a weighting factor of number of observations divided by the number of amino acids in the protein, similar to RPKM normalization for next-gen sequencing [121].

2.2.1.4 Other cellular components

The compositions of other cellular components of *Synechocystis* 6803 and *Cyanothece* 51142 were estimated based on values in the literature. DNA and RNA contents for *Synechocystis* 6803 were reported by Shastri and Morgan [104]. The remaining biomass components of *Synechocystis* 6803 (i.e., lipid, soluble pool and inorganic ions) were extracted from the measurements carried out by Nogales et al.[109]. For *Cyanothece* 51142, biochemical compositions of macromolecules such as lipids, RNA, DNA and
soluble pool were extracted from the measurements reported by Vu et al. [102].

2.2.2 Model simulations

Flux balance analysis (FBA) [122] was employed in both the model validation and model testing phases. *Cyanothec*ce iCyt773 and *Synechocystis* iSyn731 models were evaluated in terms of biomass production under several scenarios: light and dark phases, heterotrophic and mixotrophic conditions. Flux distributions for each one of these states were inferred using FBA:

\[
\text{Maximize } v_{\text{biomass}}
\]

Subject to

\[
\sum_{j=1}^{m} S_{ij} v_j = 0 \quad \forall \ i \in 1, \ldots, n \tag{1}
\]

\[
v_{j,\text{min}} \leq v_j \leq v_{j,\text{max}} \quad \forall \ j \in 1, \ldots, m \tag{2}
\]

Here, \( S_{ij} \) is the stoichiometric coefficient of metabolite \( i \) in reaction \( j \) and \( v_j \) is the flux value of reaction \( j \). Parameters \( v_{j,\text{min}} \) and \( v_{j,\text{max}} \) denote the minimum and maximum allowable fluxes for reaction \( j \), respectively. Light and dark phases in *Cyanothec*ce 51142 are represented via modifying the minimum or maximum allowable fluxes with the following constraints, respectively:

\[
v_{\text{Glyt}} = 0 \quad \text{and} \quad v_{\text{Glycr}} = 0 \tag{3}
\]

\[
v_{\text{CO2r}} = 0 \quad , \quad v_{\text{Glyt}} = 0 \quad , \quad v_{\text{light}} = 0 \quad \text{and} \quad v_{\text{cf}} = 0 \tag{4}
\]

Here, \( v_{\text{Biomass}} \) is the flux of biomass reaction and \( v_{\text{Glyt}} \), \( v_{\text{Glycr}} \) and \( v_{\text{CO2r}} \) are the fluxes of glycerol, glycogen and carbon dioxide transport reactions and \( v_{\text{light}} \) and \( v_{\text{cf}} \) are the fluxes of light reactions and carbon fixation reactions. For light phase, constraint (3) was included in the linear model, whereas for dark phase constraint (4) was included.

Once the *Synechocystis* iSyn731 model was validated, it was further tested for *in silico* gene essentiality. The following constraint(s) was included individually in the linear model to represent any mutant:
Here, $v_{\text{mutant}}$ represents flux of reaction(s) associated with any genetic mutation.

Flux variability analysis [123] for the reactions (for which photoautotrophic $^{13}$C MFA measurements [112] were available) was performed based on the following formulation:

Maximize/Minimize $v_j$

Subject to

$$\sum_{j=1}^{m} S_{ij} v_j = 0 \ \forall \ i \in 1, \ldots, n$$

$$v_{j,\text{min}} \leq v_j \leq v_{j,\text{max}} \ \forall \ j \in 1, \ldots, m$$

$$v_{\text{Biomass}}^{\text{Biomass}} \geq v_{\text{min}}$$

Here, $v_{\text{Biomass}}^{\text{Biomass}}$ is the minimum level of biomass production. In this case we fixed it to be the optimal value obtained under light condition for the Synechocystis iSyn731 model.

CPLEX solver (version 12.1, IBM ILOG) was used in the GAMS (version 23.3.3, GAMS Development Corporation) environment for implementing GapFind and GapFill [39] and solving the aforementioned optimization models. All computations were carried out on Intel Xeon E5450 Quad-Core 3.0 GH and Intel Xeon E5472 Quad-Core 3.0 GH processors that are the part of the lionxj cluster (Intel Xeon E type processors and 96 GB memory) of High Performance Computing Group of The Pennsylvania State University.
**Table 2-1:** *Synechocystis* 6803 iSyn731 and *Cyanothece* 51142 iCyt773 model statistics

<table>
<thead>
<tr>
<th></th>
<th><em>Synechocystis 6803 iSyn731 model</em></th>
<th><em>Cyanothece 51142 iCyt773 model</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Included genes</strong></td>
<td>731</td>
<td>773</td>
</tr>
<tr>
<td><strong>Proteins</strong></td>
<td>511</td>
<td>465</td>
</tr>
<tr>
<td>Single functional proteins</td>
<td>348</td>
<td>336</td>
</tr>
<tr>
<td>Multifunctional proteins</td>
<td>91</td>
<td>83</td>
</tr>
<tr>
<td>Isozymes</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Multimeric proteins</td>
<td>32</td>
<td>22</td>
</tr>
<tr>
<td><strong>Others</strong>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>36</td>
<td>23</td>
</tr>
<tr>
<td><strong>Reactions</strong></td>
<td>1,156</td>
<td>946</td>
</tr>
<tr>
<td>Metabolic reactions</td>
<td>972</td>
<td>761</td>
</tr>
<tr>
<td>Transport reactions</td>
<td>127</td>
<td>128</td>
</tr>
<tr>
<td><strong>GPR associations</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene associated (metabolic/transport)</td>
<td>827</td>
<td>686</td>
</tr>
<tr>
<td>Spontaneous&lt;sup&gt;b&lt;/sup&gt;</td>
<td>180</td>
<td>158</td>
</tr>
<tr>
<td>Nongene associated</td>
<td>59</td>
<td>16</td>
</tr>
<tr>
<td>(metabolic/transport)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No protein associated</td>
<td>90</td>
<td>86</td>
</tr>
<tr>
<td>Exchange reactions</td>
<td>57</td>
<td>57</td>
</tr>
<tr>
<td><strong>Metabolites</strong>&lt;sup&gt;c&lt;/sup&gt;</td>
<td>996</td>
<td>811</td>
</tr>
<tr>
<td>Cytosolic</td>
<td>862</td>
<td>675</td>
</tr>
<tr>
<td>Carboxisomic</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Thylakoidic</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>Periplasmic</td>
<td>59</td>
<td>62</td>
</tr>
<tr>
<td>Extracellular</td>
<td>57</td>
<td>57</td>
</tr>
</tbody>
</table>

<sup>a</sup>Others include proteins involve in complex relationships, e.g. multiple proteins act as protein complex which is one of the isozymes for any specific reaction.

<sup>b</sup>Spontaneous reactions are those without any enzyme as well as gene association.

<sup>c</sup>Metabolites represent total number of metabolites with considering their compartmental specificity.
2.3 Results and Discussion

2.3.1 Model components

2.3.1.1 Biomass composition and diurnal cycle

The biomass equation approximates the dry biomass composition by draining all building blocks or precursor molecules in their physiologically relevant ratios. Most of the earlier genome-scale modeling efforts [89, 106, 107] of Synechocystis 6803 contain approximate biomass equations completely or partially adopted from other species without direct measurements. This can adversely affect the accuracy of maximum biomass yield calculations, gene essentiality predictions and knockouts for overproduction.

Biomass composition for Synechocystis iSyn731 and Cyanothece iCyt773 models were generated by defining all essential cellular biomass content values by experimental measurement or collection from existing literature (see ‘Materials and Methods’ for detail). Macromolecules present in both cyanobacteria such as protein, carbohydrates, lipids, DNA, RNA, pigments, soluble pool and inorganic ions were assigned to their corresponding metabolic precursors (e.g., L-glycine, glucose, 16C-lipid, ATP, dGTP, beta-carotene, coenzyme A and potassium respectively). Based on the experimental measurements of precursor molecules needed to form a gram of the biomass, stoichiometric coefficients were assigned. For Synechocystis 6803 we measured compositions of proteins and pigments and extracted compositions of the remaining biomass macromolecules from the model by Nogales et al. [109]. Thereby we developed biomass equations for three different conditions: photoautotrophic, mixotrophic and heterotrophic. Experimental measurements (described in the Materials and Methods section) showed that biomass composition (i.e., mainly pigments) varies for Cyanothece 51142 between light and dark conditions and nitrogen supplementation. Since pigments such as chlorophyll, carotenoids and phycocyanobilin play important roles in photosynthetic processes their quantities are consequently higher under light conditions. In the presence of light Cyanothece 51142 uses photosynthesis to store solar energy in the
form of carbohydrates (i.e., glycogen), while in dark it expends that energy to fix nitrogen. Surprisingly, no significant change was measured in the carbohydrate pool between light and dark phases due to infinitesimal contribution of photosynthetically stored carbohydrates to total carbohydrate content in the biomass of *Cyanothece 51142*. Aggregate quantities of the remaining biomass macromolecules for *Cyanothece 51142* such as lipids, RNA, DNA and soluble pool were extracted from the most recent *Cyanothece 51142* model by Vu et al. [102] to develop biomass equations for light and dark phases.

An earlier characterization study for *Cyanothece 51142* revealed that 113 proteins are expressed in higher abundance in the light phase while 137 are expressed in higher abundance in dark conditions [113]. The constructed model spans 26 light-specific proteins, associated with 36 reactions mainly involved in fatty acid, pigment, and amino acid metabolism and 11 dark-specific proteins accounting for 16 reactions from glycolysis, purine, pyrimidine, pyruvate, and amino acid metabolism. Separate biomass equations as well as two regulatory structures for the model were derived in order to represent diurnal metabolic differences for *Cyanothece 51142*. In contrast, diurnal differences observed in *Synechocystis 6803* [124] are less pronounced (i.e., observed for only 54 genes) and less well functionally annotated (i.e., 32 genes with ‘unassigned’ functions). *When compared to existing biomass equations of Synechocystis 6803* [89, 107] we found significantly lower values for the percent weight contribution of proteins towards the biomass pool (i.e., 52% for *Synechocystis 6803* and 53% for *Cyanothece 51142* vs. 84% [89] and 66% [107], respectively). The new protein biomass contribution is in better agreement with the previously reported value of 55% for *Cyanothece 51142* [125].

2.3.1.2 Identification and correction of network gaps

Upon ensuring biomass formation, GapFind [39] was applied to assess network connectivity and blocked metabolites. By applying Gapfill [39] putative reconnection hypotheses were identified for blocked metabolites. Only the suggested modifications
that were independently corroborated using literature sources and also did not lead to the introduction of thermodynamically infeasible cycle were included in the model. For *Synechocystis* iSyn731 model, GapFind [39] identified 207 blocked metabolites. Note that there exist 125 blocked metabolites in the iJN678 model [109]. GapFill [39] identified unblocking hypotheses for 138 blocked metabolites. However, 88 of them led to the generation of infeasible thermodynamic cycles and thus were excluded. For only 5 blocked metabolites corroborating evidence for reconnection was obtained by adding 10 reactions (i.e., 2 metabolic, 4 transport and 4 exchange reactions). The added metabolic reactions have unknown gene associations while all 4 added transport reactions involve passive diffusion and thus are not associated with any specific gene(s) or protein(s). Ultimately, the 45 remaining blocked metabolites with GapFill suggested (but unconfirmed) reconnection mechanisms along with 69 blocked metabolites with no reconnection hypotheses were retained in the model iSyn731, while metabolites such as ubiquinone, which was proposed as an alternate substrate for succinate dehydrogenase [32] was excluded from iSyn731.

For the *Cyanothece* iCyt773 model, 74 blocked metabolites were found after applying GapFind [39]. Note that there are 66 blocked metabolites in iCce806 [102]. Two exchange reactions were added to allow the uptake of glucose and thyaminose ensuring biomass production under heterotrophic or mixotrophic conditions. Four blocked metabolites directly adopted from iCce806 (during the draft model creation phase) were linked to five reactions with spurious gene associations and thus both metabolites and reactions were removed from iCyt773. GapFill [39] suggested re-connection mechanisms for 52 blocked metabolites (out of a total of 70). However, for 12 blocked metabolites the re-connection model modifications led to the creation of thermodynamically infeasible cycles and thus were discarded. Corroborating evidence for the reconnection of 30 blocked metabolites was identified through the addition of 19 GapFill suggested reactions (i.e., 8 metabolic, 7 transport and 4 exchange reactions). Of the eight added metabolic reactions we found direct literature evidence for five, homology-based evidence for one while two reactions are spontaneous. All seven added transport reactions are through passive diffusion and thus are not connected with any
specific gene(s) or protein(s). Ten remaining blocked metabolites with GapFill suggested reconnection hypotheses (along with 22 with no reconnection hypotheses) were left blocked in iCyt773 as no information to corroborate the GapFill suggested changes was found in the published literature and databases. For example, biotin is produced in *Cyanobacterium* 51142; however, there is no literature evidence to support the presence of the initial step of the primary production pathway (i.e., conversion of pimeloyl-CoA from pimelate) and the intermediate step (i.e., biotransformation of 7,8-diamino-nonanoate from 8-amino-7-oxononanoate). This indicates that *Cyanobacterium* 51142 may utilize a currently unknown pathway for producing biotin. The six other blocked metabolites are involved in the nonfermentative alcohol production pathway (as explained in model comparison section) known to be incomplete in *Cyanobacterium* 51142. Table 2-2 summarizes the results related to connectivity restoration of *Synechocystis* iSyn731 and *Cyanobacterium* iCyt773 models.

**Table 2-2:** Summary of connectivity restoration in *Synechocystis* 6803 iSyn731 and *Cyanobacterium* 51142 iCyt773 models

<table>
<thead>
<tr>
<th></th>
<th><em>Synechocystis</em> 6803 iSyn731</th>
<th><em>Cyanobacterium</em> 51142 iCyt773</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of blocked metabolites</td>
<td>207</td>
<td>74</td>
</tr>
<tr>
<td>Number of metabolites with GapFill [39] suggested reconnection strategies</td>
<td>138</td>
<td>52</td>
</tr>
<tr>
<td>Number of metabolites whose reconnection forms a cycle</td>
<td>88</td>
<td>12</td>
</tr>
<tr>
<td>Number of metabolites with validated reconnection mechanisms</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>Number of added reactions to the model</td>
<td>10</td>
<td>19</td>
</tr>
</tbody>
</table>
2.3.2 GPR associations and elemental and charge balancing

GPR associations connect genotype to phenotype by linking gene(s) that code for the protein(s) that catalyze a particular reaction. They are important to trace correctly as they provide the means to target at the gene level any change in the network desired at the reaction level. This is critical because genes may catalyze multiple reactions in multiple pathways. Many earlier models for *Synechocystis* 6803 do not provide in detail complex GPR associations, rather list only gene(s) and enzyme(s) involved in a specific reaction [89, 107, 108]. For both iCyt773 and iSyn731 models, we included comprehensive GPR associations (see Table 2-1 for detail information). All four intracellular compartments (i.e., periplasm, cytosol, thylakoid lumen and carboxysome) were assumed to have the same pH (7.2) and subsequently, metabolites were assigned appropriate protonation states corresponding to this pH and each reaction was elementally and charge balanced.

Under high light intensity in photoautotrophic conditions, *Cyanothece* iCyt773 model produces 0.026 mole biomass/mole carbon fixed whereas *Synechocystis* iSyn731 yields 0.021 mole biomass/mole carbon fixed. These yields are almost identical to the ones calculated using the most recent models of *Cyanothece* 51142 [102] and *Synechocystis* 6803 [109]. Experimental measurements of biomass yields are in the same order of magnitude with model predictions for the two organisms (i.e., 0.072 [102, 126] and 0.082 mole biomass/mole carbon fixed [127]), respectively.

2.3.3 Comparison of iSyn731 model predicted flux ranges against experimental measurements

We superimposed photoautotrophic flux measurements [112] for *Synechocystis* 6803 onto iSyn731 model to assess if the measurements are consistent with the model and whether the biomass maximization assumption correctly apportions fluxes to the metabolic network. For each reaction that was assigned a flux we calculated the flux-range under the maximum biomass assumption.
Figure 2-1: Comparison of model derived and experimentally measured [112] flux ranges for *Synechocystis* 6803 under the maximum biomass condition.

Basis is 100 millimole of CO$_2$ plus H$_2$CO$_3$.

Perhaps the most informative discrepancy is for the CO$_2$ fixing RuBisCO (RBC) reaction, which has a measured flux range of (123.00 to 132.00) vs. the model-calculated range of (102.49 to 106.33). In both cases the increased RBC flux (in comparison to the basis of 100 millimole of CO$_2$ plus H$_2$CO$_3$ uptake) is needed to counteract the carbon loss due to the CO$_2$ releasing reactions such as isocitrate dehydrogenase (ICD) and pyruvate dehydrogenase (PDH).
Table 2-3: Comparison of $^{13}$C MFA flux measurements [112] vs. model-predicted flux ranges

<table>
<thead>
<tr>
<th>Reaction</th>
<th>Flux measurements by Young et al., 2011 [35]</th>
<th>Flux ranges predicted by $i$N678 model (With max biomass)</th>
<th>Flux ranges predicted by $i$Syn731 model (With max biomass)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>95% LB</td>
<td>95% UB</td>
<td>LB</td>
</tr>
<tr>
<td>RBC</td>
<td>123.00</td>
<td>132.00</td>
<td>109.02</td>
</tr>
<tr>
<td>PGK</td>
<td>219.00</td>
<td>237.00</td>
<td>187.11</td>
</tr>
<tr>
<td>$^{13}$PDG</td>
<td>219.00</td>
<td>237.00</td>
<td>187.11</td>
</tr>
<tr>
<td>GAPDH</td>
<td>90.00</td>
<td>99.00</td>
<td>74.98</td>
</tr>
<tr>
<td>FBA</td>
<td>53.00</td>
<td>66.00</td>
<td>-0.17</td>
</tr>
<tr>
<td>FBP</td>
<td>53.00</td>
<td>66.00</td>
<td>0.00</td>
</tr>
<tr>
<td>G6PD</td>
<td>15.00</td>
<td>24.00</td>
<td>0.68</td>
</tr>
<tr>
<td>6PGD</td>
<td>12.00</td>
<td>21.00</td>
<td>0.00</td>
</tr>
<tr>
<td>6PGL</td>
<td>12.00</td>
<td>21.00</td>
<td>0.00</td>
</tr>
<tr>
<td>PRK</td>
<td>123.00</td>
<td>132.00</td>
<td>109.02</td>
</tr>
<tr>
<td>SBGPL</td>
<td>29.00</td>
<td>43.00</td>
<td>-0.17</td>
</tr>
<tr>
<td>SBP</td>
<td>29.00</td>
<td>43.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TAL</td>
<td>-6.00</td>
<td>9.00</td>
<td>-36.74</td>
</tr>
<tr>
<td>TKT1</td>
<td>37.20</td>
<td>37.50</td>
<td>35.18</td>
</tr>
<tr>
<td>RPI</td>
<td>35.40</td>
<td>35.70</td>
<td>35.18</td>
</tr>
<tr>
<td>TKT2</td>
<td>35.40</td>
<td>35.70</td>
<td>37.25</td>
</tr>
<tr>
<td>RPE</td>
<td>75.50</td>
<td>76.20</td>
<td>73.83</td>
</tr>
<tr>
<td>PGM</td>
<td>22.90</td>
<td>23.60</td>
<td>26.83</td>
</tr>
<tr>
<td>ENO</td>
<td>23.40</td>
<td>23.80</td>
<td>26.84</td>
</tr>
<tr>
<td>PYK</td>
<td>7.90</td>
<td>11.10</td>
<td>0.00</td>
</tr>
<tr>
<td>PDH</td>
<td>11.50</td>
<td>12.00</td>
<td>0.00</td>
</tr>
<tr>
<td>CS</td>
<td>3.00</td>
<td>3.40</td>
<td>2.15</td>
</tr>
<tr>
<td>ACONT</td>
<td>3.00</td>
<td>3.40</td>
<td>2.15</td>
</tr>
<tr>
<td>ICD</td>
<td>3.00</td>
<td>3.00</td>
<td>2.15</td>
</tr>
<tr>
<td>SUCD</td>
<td>0.00</td>
<td>0.40</td>
<td>0.00</td>
</tr>
<tr>
<td>FUM</td>
<td>1.70</td>
<td>2.00</td>
<td>-5.44</td>
</tr>
<tr>
<td>MDH</td>
<td>1.90</td>
<td>5.20</td>
<td>5.35</td>
</tr>
<tr>
<td>ME1</td>
<td>3.70</td>
<td>6.90</td>
<td>0.00</td>
</tr>
<tr>
<td>ME2</td>
<td>3.70</td>
<td>6.90</td>
<td>-</td>
</tr>
<tr>
<td>PPC</td>
<td>9.90</td>
<td>13.20</td>
<td>11.74</td>
</tr>
</tbody>
</table>

Table 2-3 and Figure 2-1 summarize the obtained results for a basis of 100 millimole of CO$_2$ plus H$_2$CO$_3$ uptake [112]. In seven (out of thirty one) cases the measured flux is fully contained within the model predicted ranges obtained upon maximizing biomass formation implying model consistency with MFA measurements. In contrast, under the
maximum biomass assumption for thirteen fluxes the ranges underestimate and for four fluxes the ranges overestimate the experimentally deduced flux ranges while for seven fluxes the model derived flux ranges partially overlap with the experimental ones.

We find that flux ranges, under the maximum biomass production assumption, of reactions such as glucose 6-phosphate dehydrogenase (G6PD), 6-phosphogluconolactonase (6PGL) and phosphogluconate dehydrogenase (6PGD) in oxidative pentose phosphate (OPP) pathway are negligible (0.00 to 0.03). In contrast, the experimentally derived range for OPP is (12 to 21). This is approximately equal to the difference between the model-predicted vs. experimentally deduced RBC reaction range implying the persistence of OPP flux even under the photoautotrophic condition [112] despite the presence of a more efficient NADPH production route through photosynthesis as predicted by the model (under max biomass). The high values Young et al. [112] obtained for the OPP fluxes were surprising as OPP is not a very efficient route for cyanobacteria to generate reducing power. This may reflect some inherent biological constraint that is not captured by the optimality assumption.

Model predicted lower flux ranges for RBC are propagated to seven other reactions in the Calvin cycle (i.e., phosphoglycerate kinase (PGK), glyceraldehyde-3-phosphate dehydrogenase (13PDG), triose-phosphate isomerase (TPI), transketolase (TKT1), ribose-5-phosphate isomerase (RPI), ribulose 5-phosphate 3-epimerase (RPE) and phosphoribulokinase (PRK). The remaining six reaction fluxes with lower model predicted fluxes compared to measurements [112] are all in the TCA cycle (i.e., citrate synthase (CS), aconitase (ACONT), isocitrate dehydrogenase (ICD), succinate dehydrogenase (SUCD) and malic enzyme (ME1 and ME2) reactions). Even under the max biomass assumption, SUCD is not required to carry any flux due to the presence of other succinate dehydrogenases (as part of respiratory chain) in the iSyn731 model. Furthermore, in contrast with experimental observations, under the maximum biomass assumption, the model predicts no flux through the malic enzyme (ME) reactions presumably because it is a less energy-efficient route (i.e., phosphoenolpyruvate →
oxaloacetate → malate → pyruvate) for pyruvate generation than the pyruvate kinase (PYK) reaction [112].

There are nine reactions with experimentally derived ranges completely subsumed within the ones derived under the maximum biomass assumption. Five of them are in the Calvin cycle (i.e., fructose-bisphosphate aldolase (FBA), fructose-bisphosphatase (FBP), Sedoheptulose 1,7-bisphosphate D-glyceraldehyde-3-phosphate-lyase (SBGPL), sedoheptulose-bisphosphatase (SBP) and bidirectional transaldolase (TAL)). The first four reactions are essential with experimentally deduced flux ranges of (53.00 to 66.00) for FBA and FBP and (29.00 to 43.00) for SBGPL and SBP. In contrast, the calculated flux ranges (-0.08 to 73.17) for FBA and SBGPL and (0.00 to 73.17) for FBP and SBP imply that they are in silico non-essential. As depicted in Figure 2-1, these reactions are involved in the production of sedoheptulose 7-phosphate (S7P) from fructose 1,6-bisphosphate (FDP). An alternative production route for S7P is afforded in the model through the bidirectional transaldolase (TAL) reaction from fructose 6-phosphate (F6P) alluding to an explanation for the wider flux ranges derived using the model. Experimental and model predicted flux ranges for TAL are (-6.00 to 9.00) and (-35.93 and 37.32), respectively. Upon restricting the TAL flux ranges in the calculations to the ones found experimentally, the flux variability analysis shrinks the flux ranges for FBA and FBP to (28.22 to 43.27) and (28.22 to 43.33) and for SBGPL and SBP to (29.82 to 44.87) and (29.82 to 44.87), respectively which are very close to the experimentally measured ranges. This is indicative that in addition to the maximization of biomass formation, additional restrictions (e.g., photosynthetic efficiency and relative selectivity of RuBISCO for carboxylation over oxidation) limit the range of fluxes that the aforementioned glycolytic fluxes may span in vivo. Note that the presence of experimentally measured fluxes is important to test the model and the adopted maximization principle. We were fortunate in this case to have access to such data as for most organisms they are absent.

Phosphoglycerate mutase (PGM) and enolase (ENO) reactions have very similar model derived and experimentally obtained flux ranges. Model-predicted flux values of the
remaining two reactions, pyruvate kinase (PYK) and pyruvate dehydrogenase (PDH), could reach as low as zero due to the metabolic flexibility that the iSyn731 model possesses by having alternate enzymes with different cofactor specificities. The max biomass flux range of fumarase (FUM) is found to be (-7.26 to 1.49), compared to the experimentally measured (1.70 to 2.00). Therefore, it appears that under the photoautotrophic condition, the forward direction is kinetically favorable. By restricting the reaction to be irreversible the model predicted a FUM flux range of (0.00 to 1.49) which is close to the experimentally derived one (see Figure 2-1). However, contrary to MFA measurements these reactions (FUM and ME) are dispensable for in silico biomass production.

2.3.4 iSyn731 model testing using in vivo gene essentiality data

The quality of model iSyn731 for Synechocystis 6803 was tested using experimental data on the viability (or lack thereof) of single gene knockouts. We used the CyanoMutants database [111, 128] that includes in vivo gene essentiality data for 119 genes (i.e., 19 essential and 100 nonessential) with metabolic functions in iSyn731 model. Cases that were flagged with incomplete segregation in the database were omitted in iSyn731 model comparisons. We examined the feasibility of biomass production for the model iSyn731 by comparing the maximum biomass formation upon imposing the gene knockout with the maximum theoretical yield of the wild-type organism. A threshold of 10% of the maximum theoretical yield was used as a cutoff [41]. Comparisons between in vivo and in silico results led to four possible outcomes, as previously delineated by Kumar et al., GG, GNG, NGG and NGNG [41]. Initially, the model correctly predicted 18 out of 19 essential genes (i.e., 18 NGNG and 1 GNG) and 74 out of 100 non-essential genes (i.e., 73 GG and 27 NGG). We next explored the causes of these discrepancies and attempted to mitigate them whenever possible.

The single GNG case corresponds to mutant ΔchlA1 exhibiting no growth under aerobic conditions [129]. The ChlA1 system is a Mg-protoporphyrin IX monomethylester (MPE)
The cyclase system that is responsible for forming the isocyclic ring (E-ring) in chlorophylls under aerobic conditions [129]. The model allowed for the BchE and ChlA_{II} systems (alternate cyclase systems) to complement for the loss of the ChlA_{I} system leading to an \textit{in silico} viable mutant. However, the same literature source [129] suggested that both BchE and ChlA_{II} systems are unlikely to be active under aerobic conditions and thus rescue mutant \textit{ΔchlA}_{I}. This prompted the introduction of a regulatory restriction in \textit{iSyn731} model where only ChlA_{I} reactions were active under aerobic conditions as MPE while ChlA_{II} and BchE system reactions were deactivated. Using these regulatory restrictions resolves the single GNG inconsistency.

Twenty (out of 27) NGG cases were associated with Photosystem I (PSI), Photosystem II (PSII) and other photosynthesis reactions. While reconstructing the model, we assumed that all genes involved in photosynthetic reaction system were essential to the functioning of the overall system. Published literature [130-134] suggests that genes involved in photosynthetic reactions form complex interdependencies. We used NCBI COBALT multiple sequence alignment tool [135] to construct a phylogenetic tree of the genes associated with each photosystem along with BLASTp searches to identify putative complementation relationships between genes to explain the inconsistencies between the predicted \textit{in silico} and \textit{in vivo} growth. Genes deemed homologous (i.e., lie adjacent in the phylogenetic tree) were linked with “OR” GPR relations implying that the loss of one gene can be complemented by the other. Seven out of twenty NGG cases (i.e., \textit{psaD}, \textit{psal} and \textit{psbA2} for PSI and PSII and \textit{cpcC2}, \textit{cpcC1}, \textit{cpcD}, and \textit{apcD} for other photosynthesis reactions) were resolved by modifying the corresponding GPR using an OR relation [131, 132, 136-139]. However, no phylogenetically adjacent or related (or homologous) genes were found for the remaining 13 NGG cases (\textit{psaE}, \textit{psbD2}, \textit{psbO}, \textit{psbU}, \textit{psbV}, \textit{psb28}, \textit{psbX}, \textit{psb27}, \textit{petE}, \textit{cpcA}, \textit{cpcB}, \textit{apcE}, \textit{apcF}) [133, 134, 136, 139-143]. For these cases, the genes were deemed nonessential to the functioning of the reactions in question (i.e., photosynthesis reactions) and thereby the corresponding GPRs were modified to show an OR relation between each of these genes and an ‘\textit{unknown gene}’, similar to what was previously performed in the refinement of the \textit{iMM904} model [144].
The remaining seven NGG cases are associated with a variety of metabolic functions. One such case is the $\Delta modBC$ mutant corresponding to the sole ABC molybdate transporter in the model. Literature evidence [145] revealed that a related cyanobacterium, *Anabaena variabilis* ATCC 29413, could continue to grow despite the loss of its molybdate ABC transporter due to the presence of another low affinity molybdate transporter or an inducible sulfate transport system that can serve as a low affinity molybdate transporter when required. We found the same gene coding for the sulfate transporter in *A. variabilis* (*cysA*) in the *iSyn731* model allowing the resolution of the discrepancy by adding a *cysA*-linked alternate molybdate transporter. Another NGG case is mutant $\Delta crtO$ that cannot produce echinenone (a biomass component) in *iSyn731* with no effect on observed growth. Therefore, it appears that *iSyn731* cannot capture the flexibility of *Synechocystis* 6803 metabolism [146] when echinenone production is restricted. The remaining five NGG cases are spread across many metabolic pathways. The $\Delta actaA$ mutant eliminates the copper ABC transporter without affecting growth, which alludes to the existence of another unknown mode of copper uptake not present in *iSyn731* [147, 148]. The $\Delta menG$ mutant eliminates a reaction for the production of phyloquinone while mutant $\Delta ppd$ affects the production of homogentisate, a precursor for both tocopherols and plastoquinone. Finally, the $\Delta vte3$ mutant affects the production of both plastoquinone and $\alpha$-tocopherol [149] and the viable $\Delta ccmA$ mutant restricts the production of chorismate (a precursor to aromatic amino acids) and also restricts carboxysome formation [150-152]. These six inconsistencies between the model predictions and growth data imply that the cyanobacterium can co-opt another metabolic process to (partially) complement for the gene loss. Unlike the case of the $\Delta modBC$ mutant, we have found no plausible mechanism for the six remaining mutants.

After resolving the discrepancies, as described above, *iSyn731* correctly predicted all 19 essential genes (i.e., 19 NGNG and 0 GNG) and 94 (out of 100) non-essential genes (i.e., 94 GG and 6 NGG). Figure 2-2 shows our results and comparisons against two other available *Synechocystis* 6803 models by Knoop *et al.* [89] and Nogales *et al.* [109]. We used the CyanoMutants database [111] to identify 114 genes (i.e., 19 essential and 95 nonessential) having metabolic functions in the *iJN678* model by Nogales *et al.* [109].
Out of 114 genes the iJN678 model correctly predicted 18 essential genes (i.e., 18 NGNG and 1 GNG) and 69 non-essential genes (i.e., 69 GG and 26 NGG). The model by Knoop et al. [89] was tested for 51 mutants but we found that only 43 (i.e., 7 essential and 36 non-essential) of them were reported to have complete segregation [111]. Of these 43, Knoop et al.’s [89] model correctly predicted 5 essential genes (i.e., 5 NGNG and 2 GNG) and 32 nonessential genes (i.e., 32 GG and 4 NGG). The specificity and sensitivity of each of these three models were also calculated and displayed at the bottom of Figure 2-2.

**Figure 2-2:** Comparison of gene essentiality/viability data with predictions by a number of *Synechocystis* 6803 models.

(A) Tabulated growth (i.e., G) or non-growth (i.e., NG) predictions and experimental data. The first number denotes the number of GG, GNG, NGG and NGNG combinations...
whereas the second number signifies the number of experimentally observed lethal (or viable) mutants, and (B) Definition and comparison of specificity and sensitivity of all three models. Note that GG denotes both in silico and in vivo growth, NGG represents no growth in silico but in vivo growth. NGNG implies no growth for either in silico or in vivo, whereas GNG marks growth in silico but no growth in vivo.

All 114 genes tested for iJN678 were also present in the iSyn731 model. 26 NGG and one GNG cases present in iJN678 model correspond to NGG and GNG cases that were either fixed or still present in iSyn731 as discussed before. Lethal mutant $\Delta ppa$ is correctly predicted as NGNG in iSyn731 but deemed GNG in Knoop et al. [89] model. This was because $ppa$ in iSyn731 codes for the degradation of both triphosphate into diphosphate and diphosphate to phosphate. Only the latter activity is linked to $ppa$ in the Knoop et al’s model. Out of 4 NGG cases in [89], two involve $\Delta cmpA$ and $\Delta cmpB$ mutants. Both these genes are involved in the ABC transporter system for bicarbonate from periplasm to cytosol. iSyn731 avoids this inconsistency as it contains an alternate sodium and bicarbonate co-transport system.

### 2.3.5 Model comparisons

#### 2.3.5.1 Synechocystis 6803 model comparisons

The iSyn731 model integrates the description in the photosystems of the model presented by Nogales et al. [109] and adds additional detail. One notable difference is that iSyn731 uses a separate photon for each reaction center (i.e., PSI and PSII) as they are optimized for different ranges of wavelength [153], whereas iJN678 [109] uses a single photon shared by both photosystem reactions. As many as 322 new reactions (see Figure 2-3A), are added in iSyn731 distributed across many pathways. Most of the additions are in the lipid and fatty acid metabolism to support the synthesis of measured fatty acids and lipids present in the biomass equation. This list includes myristic acid (14-carbon saturated
fatty acid) and lauric acid (12-carbon saturated fatty acid). iJN678 [109] contained four reactions exhibiting unbounded flux (i.e., two duplicate glycine cleavage reactions and two duplicate leucine transaminase reactions). They form a thermodynamically infeasible cycle (see Figure 2-4A for leucine transaminase reactions) that was resolved in iSyn731 by eliminating redundant functions. In addition, the glycine cleavage system was recast in detail by abstracting the separate action of the four enzymes (named the T-, P-, L-, and H-proteins) that ultimately catalyze the demethylamination of glycine.

iSyn731 improves upon iJN678 [109] by eliminating lumped reactions whenever a multi-step description is available and expands the range of functions carried out with alternate cofactors. As many as twelve reactions with an enoyl-[acyl-carrier-protein] reductase function were linked with not only NADP but also with the more rare NAD cofactor specificity. Another important difference between iSyn731 and iJN678 [109] is the cellular location of the CO₂ fixation (i.e., ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) enzyme). Literature [154, 155] shows that cyanobacteria possess a micro-compartment (i.e., carboxysome) encapsulating RuBisCO and carbonic anhydrase (CA) enzymes. iSyn731 adds carboxysome as a cellular compartment and also all necessary transport reactions [154, 155].
Figure 2-3: Venn diagram depicting (common and unique) reactions and metabolites between: (A) iJN678 [109] and iSyn731, (B) iCce806 [102] and iCyt773, and (C) iSyn731 and iCyt773 models.

Recently, Zhang and Bryant [110] hypothesized the existence of a functional TCA cycle in most cyanobacterial species using a 2-ketoglutarate to succinate bypassing step. iSyn731 allows for a complete TCA cycle using the bypassing step. In addition, iSyn731 contains an intact heptadecane biosynthesis pathway as recently described [98] unlike earlier *Synechocystis* 6803 models [89, 106, 107, 109] (see Figure 2-5A for distribution of unique reactions in iSyn731)
2.3.5.2 *Cyanothece 51142* model comparisons

The *iCyt773* model for *Cyanothece 51142* improves upon the *iCce806* model [102]. *iCyt773* segregates reactions into the periplasm, thylakoid lumen, carboxysome, and cytoplasm compartments thus introducing an additional 60 transport reactions compared to *iCce806* [102]. Unlike *iCce806* [102], *iCyt773* does not track macromolecule synthesis for DNA, RNA, and proteins to maintain consistency with the *Synechocystis 6803* model.

![Diagram](image)

**Figure 2-4:** Schematics that illustrate the thermodynamically infeasible cycles and subsequent resolution strategies.

(A) Cycles present in *iJN678* [109], and (B) Cycles present in *iCce805* [102]. Blue colored lines represent the original reaction directionality whereas green ones denote modified directionality to eliminate cycle.
This difference accounts for 69 genes present in iCce806 [102] but absent from iCyt773. iCce806 [102] contained 15 reactions which formed five cycles that could carry unbounded metabolic flux (i.e., thermodynamically infeasible cycles). All these cycles were eliminated by restricting reaction directionality and eliminating reactions that were linear combinations of others (coded by the same gene) (see Figure 2-4B). iCyt773 contains 43 unique genes and 266 unique reactions (including transport and alternate cofactor utilizing reactions) as shown in Figure 2-3B. Figure 2-5B depicts the distribution of the new reactions across different pathways. Most of the additions are found in lipid and pigment biosynthesis pathways. The iCyt773 model captures in detail the lipid biosynthesis pathway composed of 73 reactions and links as many as 28 biomass precursor lipids (e.g., sulfoquinovosyldiacylglycerols, monogalactosyldiacylglycerols, digalactosyldiacyl-glycerols, and phosphatidylglycerols) directly to the biomass equation. The porphyrin and chlorophyll metabolism and carotenoid biosynthesis pathways were updated to include 24 reactions for the production of accessory pigments such as echinenone, an accessory pigment, and (3Z)-phycocyanobilin, a phycobilin. Accessory pigments donate electrons to chlorophyll rather than directly to photosynthesis. Phycobilins are adapted for many wavelengths not absorbed by chlorophyll thus broadening the spectrum useful for photosynthesis. The variety of pigments in cyanobacteria is well documented [156-158] providing so far untapped avenues for engineering increased efficiency in photosynthesis and control of electron transfer processes in biological systems. Another new function in iCyt773 is L-Aspartate Oxidase. L-Aspartate Oxidase allows the deamination of aspartate, forming oxaloacetate a key TCA-cycle metabolite and ammonia. The impact of this addition to iCyt773 is not evident under the photoautotrophic condition but becomes relevant for growth in a medium containing aspartate. iCyt773 also uniquely supports the synthesis of pentadecane as documented by Schirmer et al. [98] and contains an (almost) complete non-fermentative citramalate pathway as suggested by Wu et al. [99].

A number of lumped reactions in iCce806 [102] were recast in detail. For example, pyruvate dehydrogenase (PDH) is a three-enzyme complex that carries out the biotransformation of pyruvate to acetyl-CoA in three steps using five separate cofactors.
(i.e., TPP, CoA, FAD, lipoate, and NAD). Similar detail was used for lumped steps in the metabolism of glycine, histidine, and serine. All additions to the list of reactions in iCyt773 were corroborated using genome annotations [86] or published literature [97-99, 159] with the exception of ten enzymes, whose function in the lipid and pigment biosynthesis pathways was required for biomass production.

A shift in biomass composition was observed under light, dark, and nitrate supplemented (light and dark) conditions. These differences were captured in four separate biomass descriptions present in iCyt773. In addition, we used data from Stockel et al. [160] on the diurnal oscillations for approximately 20% of proteins in Cyanothece 51142 to identify regulatory reaction shutdowns in our metabolic model. Supplementary File S4 (which can be found in the online version of the published paper) lists the reactions that were inactivated under light and dark conditions, respectively. As expected, the nitrogenase genes cce_0559 and cce_0560, known to be active in the absence of light, exhibited low spectral counts under light conditions. In contrast, photosystem II gene cce_1526, showed no spectral count under dark conditions. Unexpectedly, the data suggested that the Mehler reactions associated gene (cce_2580), known to be active in Synechocystis 6803 [161] and expected to be active in Cyanothece 51142, exhibited lower expression in light than in dark conditions.

2.3.5.3 iSyn731 and iCyt773 models comparison

Figure 2-3C illustrates the total number of common and unique reactions and metabolites between iSyn731 and iCyt773 models. The Cyanothece 51142 genome [86, 162] is 1.5 times larger than the one for Synechocystis 6803 [88], nevertheless iCyt773 is smaller than iSyn731 due to differences in the level of detail of annotation and biochemical characterization. As many as 670 reactions and 596 metabolites are shared by both models corresponding to 47% and 63% of the total reactome and metabolome, respectively (see Figure 2-3C). The higher degree of conservation of metabolites (as opposed to reactions) across the two cyanobacteria suggests that lifestyle adaptations tend to usher new enzymatic activities that most of the time make use of the same metabolite
pool without introducing new metabolites. There are 486 reactions that are unique to iSyn731 with no counterpart in iCyt773. These reactions are not preferentially allotted to a handful of specific pathways. Instead they are spread over tens of different pathways. 

Primary metabolism reactions dispersed throughout fatty acid biosynthesis, lipid metabolism, oxidative phosphorylation, purine and pyrimidine metabolism, transport and exchange reactions account for 295 reactions. Secondary metabolism including chlorophyll and cyanophycin metabolism, folate, terpenoid, phenylpropanoid and flavonoid biosynthesis accounts for the remaining 191 iSyn731-specific reactions. Interestingly, the 276 iCyt773-specific reactions span the same set of diverse pathways implying that the two organisms have adopted unique/divergent biosynthetic capabilities for similar metabolic needs. Fifty-eight span primary metabolism pathways such as purine and pyrimidine metabolism, fatty acid and lipid biosynthesis, amino acid biosynthesis. The remaining 218 reactions describe secondary metabolism such as terpenoid biosynthesis, chlorophyll and cyanophycin biosynthesis, plastoquinone and phyloquinone biosynthesis. The much larger set of unique iSyn731-specific reactions compared to iCyt773 reflect more complete genome annotation and biochemical characterization rather than augmented metabolic versatility.
Figure 2-5: List of added reactions across pathways. (A) iSyn731 compared to iJN678 [109], and (B) iCyt773 compared to iCce806 [102].

A number of distinct differences in metabolism between the two organisms have been accounted for in the two models. For example, iCyt773 does not have the enzyme threonine ammonia-lyase, which catalyzes the conversion of threonine to 2-ketobutyrate and as a consequence lacks the traditional route for isoleucine synthesis. Instead it employs part of the alternative citramalate pathway for isoleucine synthesis with pyruvate
and acetyl-CoA as precursors. Follow up literature queries revealed the existence of this alternative pathway in *Cyanothece* 51142 [99]. Ketobutyrate, an intermediate in the citramalate pathway, can be readily converted to higher alcohols, such as propanol and butanol, via a non-fermentative alcohol production pathway. Using the iCyt773 model, we determined that only 2-ketoacid decarboxylase is missing from these three-step processes. In contrast, iSyn731 was found to have only the traditional route for isoleucine production with the citramalate pathway completely absent (see Figure 2-6A). In another example, the fermentative 1-butanol pathway is known to be incomplete in both organisms. By querying the developed models we can pinpoint exactly which steps are absent. Specifically, the conversion between 3-hydroxybutanoyl-CoA and butanal is missing in both models. In addition to higher alcohols, higher alkanes (C13 and above) are important biofuel molecules as the main constituents of diesel and jet fuel [98]. Recently reported [98] novel genes involved in the biosynthesis of alkanes in several cyanobacterial strains were incorporated in the models. Metabolic differences in *Cyanothece* 51142 and *Synechocystis* 6803 lead to the production of different alkanes (e.g., pentadecane in *Cyanothece* 51142 and heptadecane in *Synechocystis* 6803) (see Figure 2-6B).

Model iCyt773, in contrast to iSyn731, does not have a complete urea cycle as it lacks the enzyme L-arginine aminohydrolase catalyzing the production of urea from L-arginine. Literature sources [162, 163] support this finding and explain the absence of a functional urea cycle as a consequence of the nitrogen-fixation ability of *Cyanothece* 51142 [164, 165]. Because *Cyanothece* 51142 can fix nitrogen directly from the atmosphere and produce ammonium via the enzyme nitrogenase, genes corresponding to the activity of L-arginine aminohydrolase and urease (for breaking down urea) become redundant, explaining why they are not present in its genome [165]. In addition to nitrogen metabolism, iCyt773 and iSyn731 models reveal marked differences in anaerobic metabolic capabilities. Unlike iSyn731, iCyt773 includes a L-lactate dehydrogenase activity that enables the complete fermentative lactate production pathway. On the other hand, iSyn731 contains the anaerobic chlorophyll biosynthetic pathway using enzyme protoporphyrin IX cyclase (BchE) that is absent in iCyt773. Other differences in
metabolism include lipid and fatty acid synthesis, fructose-6-phosphate shunt and nitrogen fixation. Model $i$Syn731 traces the location of the double bond for unsaturated fatty acid synthesis pathways, as two separate isomers of unsaturated C$_{18}$ fatty acids are part of the biomass description. $i$Cyt773 allows for the shunting of fructose-6-phosphate into erythrose-4-phosphate along with acetate and ATP using the fructose-6-phosphate phosphoketolase activity. Finally, both $i$Syn731 and $i$Cyt773 contain multiple hydrogenases allowing both to produce hydrogen. However, only the latter has a nitrogenase activity that can fix nitrogen while simultaneously producing hydrogen.

**Figure 2-6:** Examples of pathways that differ between the two cyanobacteria.

(A) Nonfermentative alcohol production pathway highlighting the present and absent enzymes in *Cyanobacterium* 51142 and *Synechocystis* 6803, and (B) Alkane biosynthesis pathways in *Cyanobacterium* 51142 and *Synechocystis* 6803.
2.3.6 Using iSyn731 and iCyt773 to estimate production yields

We tested the recently developed models iSyn731 and iCyt773 by comparing the predicted maximum theoretical product yields with experimentally measured values for two very different metabolic products: isoprene and hydrogen. Isoprene, a volatile hydrocarbon and potential feedstock for biofuel, is mostly produced in plants under heat stress [90]. Cyanobacteria offer promising production alternatives as they can grow to high densities in bioreactors and produce isoprene directly from photosynthesis intermediates [90]. It was reported [90] that Synechocystis 6803 has all but one gene (encoding isoprene synthase) in the methyl-erythritol-4-phosphate (MEP) pathway for isoprene synthesis from dimethylallyl phosphate (DMAPP). Upon cloning the isoprene synthase from kudzu vine (Pueraria montana) into Synechocystis 6803 isoprene production was demonstrated using sunlight and atmospheric CO$_2$ of $4.3 \times 10^{-4}$ mole isoprene/mole carbon fixed [166]. We calculated the maximum isoprene yield using iSyn731 to be $3.63 \times 10^{-5}$ mole isoprene/mole carbon fixed upon adding the isoprene synthase activity to the model and simulating the conditions described in [127] under maximum biomass production. Similar isoprene yields were obtained with iJN678 [109] while earlier models of Synechocystis 6803 [89, 106-108] lack the MEP pathway (partially or completely) and thus do not support isoprene production. The underestimation of the experimentally observed isoprene yield by the model predicted maximum yield may be due to sub-optimal growth of the production strain, differences in the list of measured biomass components, missing isoprene-relevant reactions from the model or more likely a combination of the above factors.

Both Cyanothece 51142 and Synechocystis 6803 produce hydrogen by utilizing nitrogenase and hydrogenase activities, respectively [96]. Under subjective dark conditions [96] whereby (i) stored glycogen acts as a carbon source, (ii) photosynthesis harnesses light energy, and (iii) nitrogenase activity is not restricted, hydrogen production yield for Cyanothece 51142 was measured at 49.67 mole/mole glycogen consumed. Simulating the same conditions in iCyt773 and iCce806 [102] leads to maximum
theoretical yields for hydrogen production of 48.43 mole/mole glycogen and 102.4 mole/mole glycogen, respectively. The entire amount of hydrogen produced in iCyt773 is due to the nitrogenase activity. In contrast, the predicted doubling of the maximum hydrogen yield in iCce806 is due to the utilization of the reverse direction of two hydrogen dehydrogenase reactions without any nitrogenase activity. Utilization of the nitrogenase reaction requires the use and recycling of more ATP than simply running the dehydrogenase reactions in reverse. However, it has been reported that hydrogen production in Cyanothece 51142 is primarily mediated by the nitrogenase enzyme [96] in the dark phase. This lends support to the irreversibility of the dehydrogenase reactions (under dark condition) as present in the iCyt773 model. Experimental results for Synechocystis 6803 support up to 4.24 mole/mole glycogen consumed [96, 167] of hydrogen production. iSyn731 predicts a maximum hydrogen theoretical yield of 2.28 mole/mole glycogen consumed while iJN678 [109] yields a value of 2.00 mole/mole glycogen consumed. Again the factors outlined for isoprene production may explain the lower theoretical yields predicted by the two models. The small difference between the model predicted yields is due to the presence of one step lumped biotransformation between isocitrate and oxoglutarate via isocitrate dehydrogenase in iJN678 [109]. iSyn731 describes this biotransformation in two steps (isocitrate $\rightarrow$ oxalosuccinate $\rightarrow$ oxoglutarate) [168] generating an additional NADPH and subsequently more hydrogen via the hydrogenase reaction.

2.4 Conclusion

In this chapter, we expanded upon existing models to develop two genome-scale metabolic models (Synechocystis iSyn731 and Cyanothece iCyt773) for cyanobacterial metabolism by integrating all available knowledge available from public databases and published literature. All metabolite and reaction naming conventions are consistent between the two models allowing for direct comparisons. Systematic gap filling analyses led to the bridging of a number of network gaps in the two models and the elimination of orphan metabolites. Two separate biomass equations as well as two different versions of Cyanothece iCyt773 models were developed for light and dark phases to represent
diurnal regulation. The development of two separate models for *Cyanothece* 51142 (i.e., light and dark) provides the two “end-points” for the future development of dynamic metabolic models capturing the temporal evolution [113, 120, 169, 170] of fluxes during the transition phases DFBA [171]. Comparisons against available $^{13}$C MFA measurements for *Synechocystis* 6803 [112] revealed that the *iSyn731* model upon biomass maximization yields flux ranges that are generally consistent with experimental data. Discrepancies between the two identify metabolic nodes where regulatory constraints are needed in addition to biomass maximization to recapitulate physiological behavior. The ability of *iSyn731* to predict the fate of single gene knock-outs was further improved (specificity of 0.94 and sensitivity of 1.00) by reconciling *in silico* growth predictions with *in vivo* gene essentiality data [111]. Similar analyses could also be carried out for *Cyanothece* *iCyt773* model once such flux measurements and *in vivo* gene essentiality data become available.

It is becoming widely accepted that focusing on a single pathway at a time without quantitatively assessing the system-wide implications of genetic manipulations may be responsible for suboptimal production levels. By accounting for both primary and some secondary metabolism pathways, the *Cyanothece* *iCyt773* model can be used to explore *in silico* the effect of genetic modifications aimed at increased production of useful biofuel molecules. By taking full inventory of *Cyanothece* 51142 metabolism (as abstracted in *iCyt773*), and applying available strain optimization techniques [172, 173] optimal gene modifications could be pursued for a variety of targets in coordination with experimental techniques. In particular, the availability of a microaerobic environment in *Cyanothece* 51142 at certain times during the diurnal cycle can be exploited for the expression of novel pathways that are not usually found in oxygenic cyanobacterial strains that largely maintain an aerobic environment. However, the use of *Cyanothece* 51142 as a bio-production platform is currently hampered by the inability to efficiently carry out genetic modifications.

By systematically cataloguing the shared (and unique) metabolic content in *iSyn731* and *iCyt773*, successful genetic interventions assessed experimentally for *Synechocystis* 6803
can be “translated” to *Cyanothece* 51142. For example, it has been reported [174, 175] that overproduction of fatty alcohols can be achieved in *Synechocystis* 6803 upon cloning a fatty acyl-CoA reductase (*far*) from Jojoba (*Simmondsia chinensis*) and the over-expression of gene *slr1609* coding for an acyl-ACP synthetase. By using models *iSyn731* and *iCyt773* we can infer that in addition to cloning *far* from Jojoba, over-expression of gene *cce_1133* coding for a native acyl-ACP synthetase would be needed to bring about the same overproduction in *Cyanothece* 51142.
3 Chapter 3

SYNTHETIC BIOLOGY OF CYANOBACTERIA: UNIQUE CHALLENGES AND OPPORTUNITIES

This chapter has been previously published in modified form in Frontiers in Microbiology. The author has mainly contributed to the modeling section of the paper (Berla, B.M., Saha, R., Immethun, C.D., Maranas, C.D., Moon, T.S. and Pakrasi, H.B. (2013), "Synthetic biology of cyanobacteria: unique challenges and opportunities", Frontiers in Microbiology, 4).

3.1 Introduction

Cyanobacteria have garnered a great deal of attention recently as biofuel-producing organisms. Their key advantage over other bacteria is their ability to use photosynthesis to capture energy from sunlight and convert CO\textsubscript{2} into products of interest. As compared with eukaryotic algae and plants, cyanobacteria are much easier to manipulate genetically and grow much faster. They have been engineered to produce a wide and ever-expanding range of products including fatty acids, long-chain alcohols, alkanes, ethylene, polyhydroxybutyrate, 2,3-Butanediol, ethanol, and hydrogen. These processes have been reviewed recently [176] and will not be covered in detail in this review. Rather, we will look towards how the techniques of the emerging field of synthetic biology might bear fruit in improving the output of such engineered strains. Due to the low price of commodity goods like fuels and platform chemicals, it is critical to maximize the productivity of engineered strains to make them economically competitive. We believe that the tools of synthetic biology can help with this challenge.

Specifically, this review will cover systems, parts, and methods of analysis for synthetic biology systems. Synthetic biology requires a well-characterized host or ‘chassis’ strain that can be genetically manipulated with ease and predictability. Ideally, the host should grow quickly and tolerate a range of environmental conditions. The host should be simple to cultivate using readily available laboratory equipment and inexpensive growth media.
Simple, rapid, and high-throughput techniques should be available for procedures like DNA/RNA isolation, metabolomics, and proteomics. To achieve modular, ‘plug-and-play’ modification of the host strain, its metabolism and regulatory systems must be well-characterized under a wide variety of relevant conditions. Since cyanobacterial biofuel production processes will need to use sunlight as an energy source to be economically and environmentally useful, the day/night cycle will be particularly relevant; The intermittent nature of this energy source will be a key engineering challenge. We will discuss which cyanobacterial chassis have been used and their relative merits and unique traits. Ultimately, the hope is that one of these strains might be developed to become a ‘green E. coli’ for which a wide variety of genetic parts and systems are available for easy modification. Next, we will discuss the critical issue of how gene expression can be controlled in cyanobacteria. Compared with other systems, there are few examples of simple and effective controllable promoters in cyanobacteria. We will also discuss methods for analysis of gene expression using light-emitting reporters and for global analysis of metabolism using either constraint-based modeling or measurement of $^{13}$C labeling.

### 3.2 Genetic modification of cyanobacteria

Several strains of cyanobacteria are known which are readily amenable to genetic modification (See Table 3-1). Such modifications can be performed either in cis (through chromosome editing) or in trans (through plasmid addition) and synthetic biology experiments have used both approaches. We discuss advantages and disadvantages of each approach, as well as recent technical developments below. While even the best cyanobacterial model systems are still far from being a ‘green E. coli’, many tools are already available and more are being developed. The future holds great promise for this field.
Table 3-1: Model strains of cyanobacteria for synthetic biology

<table>
<thead>
<tr>
<th>Strain</th>
<th>Genetic Methods</th>
<th>Ideal Growth Temp (°C)</th>
<th>Doubling Time (hours)</th>
<th>Metabolisms</th>
<th>Genome-Scale models?</th>
<th>Notes</th>
<th>Selected Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>conjugation, natural transformation, Tn5 mutagenesis, fusion PCR</td>
<td>30</td>
<td>6-12</td>
<td>mixotrophic, autotrophic</td>
<td>Yes</td>
<td>Extensive systems biology datasets are available</td>
<td>[177]</td>
</tr>
<tr>
<td><em>Synechococcus elongatus</em> PCC 7942</td>
<td>conjugation, natural transformation, Tn5 mutagenesis</td>
<td>38</td>
<td>12-24</td>
<td>autotrophic</td>
<td>No</td>
<td>A model strain for the study of circadian clocks</td>
<td>[178]</td>
</tr>
<tr>
<td><em>Synechococcus</em> sp. PCC 7002</td>
<td>conjugation, natural transformation</td>
<td>38</td>
<td>3.5</td>
<td>mixotrophic, autotrophic</td>
<td>Yes</td>
<td>Among the fastest-growing strains known</td>
<td>[179]</td>
</tr>
<tr>
<td><em>Anabaena variabilis</em> PCC 7120</td>
<td>conjugation, natural transformation</td>
<td>30</td>
<td>&gt;24</td>
<td>mixotrophic, autotrophic</td>
<td>No</td>
<td>Nitrogen-fixing, Filamentous</td>
<td>[180]</td>
</tr>
<tr>
<td><em>Leptolyngbya</em> sp. Strain BL0902</td>
<td>conjugation, Tn5 mutagenesis</td>
<td>30</td>
<td>~20</td>
<td>autotrophic</td>
<td>No</td>
<td>Filamentous, Grows well in outdoor photo-bioreactors in a broad range of conditions</td>
<td>[181]</td>
</tr>
</tbody>
</table>
3.3 Genetic modification in cis: chromosome editing

Cis genetic modification is the most common approach in cyanobacterial synthetic biology. This approach takes advantage of the capability of many cyanobacterial strains for natural transformation and homologous recombination (see Table 3-1) to create insertion, deletion, or replacement mutations in cyanobacterial chromosomes. Traditionally, strains have been transformed with selectable markers linked to any sequence of interest and flanked by sequences homologous to any non-essential sequence on the chromosome (See Figure 3-1).

![Figure 3-1](image)

Figure 3-1: Different methods for constructing cyanobacterial mutants.

(A) shows the traditional method using double homologous recombination to insert a suicide vector into the genome at a neutral site (NS, gold) with upstream (US, orange) and downstream (DS, magenta) flanking regions in the vector. The insert contains an arbitrary sequence of interest (ATGCATG, green) and a selectable marker (SM, blue).

(B) shows 2 methods of creating markerless mutants, either by selection-counterselection or by using a recombinase system such as FLP/FRT. The counter-selection method’s first
step is the same as for the method in panel a, except that the insert also contains a counter-selectable marker (CSM, purple) such as sacB. A second transformation is performed to create a markerless mutant. Alternatively, the insert can contain recombinase recognition sites (RRS, gray) that are controlled by an inducible recombinase at a second (or the same) site in the genome. While it erases the selectable marker, this method does leave a scar sequence behind. (C) shows genetic modification in *trans* via expression plasmids.

This strategy allows the creation of targeted mutations to the chromosome, but sometimes raises concerns about segregation in polyploid strains. However, once segregated, such mutations can be stable over long time periods even in the absence of selective pressure from added antibiotics [182, 183]. While such stability is desirable, systems that create major metabolic demand, by for example redirecting flux into biofuel-producing pathways, will face greater selective pressures for mutation or loss of heterologous genes.

Recently, several methods have been developed that allow the creation of markerless mutations in cyanobacterial chromosomes (Figure 1b). Two of these methods operate on a similar principle: First, a conditionally toxic gene is linked to an antibiotic resistance cassette and then inserted into the chromosome, with selection for antibiotic-resistant mutants. Next, a second transformation is carried out in which the resistance cassette and toxin gene are deleted, and markerless mutants are selected which have lost the toxic gene. This principle has been used in cyanobacteria with the *B. subtilis* levansucrase synthase gene sacB, which confers sucrose sensitivity [184] as well as with *E. coli* mazF, a general protein synthesis inhibitor expressed under a nickel-inducible promoter [185]. This latter system has advantages for cyanobacterial strains that are naturally sucrose-sensitive. Either method allows the reuse of a single selectable marker for making multiple successive changes to the chromosome. In addition to these methods, a third system operates on a similar principle - a cyanobacterial strain that is streptomycin resistant due to a mutation in the *rps12* gene can be made streptomycin-sensitive by expressing a second heterologous copy of wild type *rps12* linked to a kanamycin (or other antibiotic) resistance cassette as well as any sequence of interest. Streptomycin-
resistant, kanamycin-sensitive markerless mutants can be recovered in a second transformation [186]. Although this method can also be used to make successive markerless mutants, it requires a background strain that is streptomycin-resistant due to an altered ribosome. Thus, it may not be an ideal method for synthetic biology studies that seek to draw conclusions about translation in wild-type systems. For the ability to transfer any translated genetic parts or parts involved in translation (such as ribosome binding sites) to other strains, this mutation could be problematic. A possible advantage of this system is that both selections are positive selections, whereas the sacB or mazF systems require a negative selection in their second transformation. Care must be taken to ensure that sucrose resistance is due to loss, as opposed to mutation, of the counter-selectable marker. Recombinase-based systems including Cre-LoxP (in Anabaena sp. PCC7120, [180]) or FLP/FRT (in Synechocystis sp. PCC6803 and Synechococcus elongatus PCC7942, [187]) have also been used to engineer mutants that lack a selectable marker. However, these methods leave a scar sequence, meaning that the final chromosomal sequence is not completely user-specifiable and also that multiple mutations using this technique in the same cell line may potentially lead to undesirable crossover events or other unexpected results.

Until recently, it has been difficult to create mutants at high throughput in cyanobacterial strains, as transposon-based methods developed for use in other strains can work poorly in cyanobacterial hosts. However, libraries can be created in other strains and subsequently transferred to a cyanobacterial host via homologous recombination. A Tn7-based library containing ~10,000 lines was recently created to screen for strains with increased polyhydroxybutyrate (PHB) production [188] and a similar approach has been taken for finding mutants in circadian clock function in Synechococcus 7942 [189] and later extended to include insertions into nearly 90% of open reading frames in that strain [178]. Chromosomal DNA fragments were first cloned into a plasmid library in E. coli and then the library was mutagenized with Tn7 before homologous recombination back into the cyanobacterial host strain. This could be an especially valuable approach for the validation of genome-scale models of cyanobacterial metabolism (see below).
3.4 Genetic modification \textit{in trans}: foreign plasmids

Although transgene expression \textit{in cis} is the most common approach in cyanobacterial research, genes are also routinely expressed in cyanobacteria \textit{in trans} [190-192]. In synthetic biology and metabolic engineering of other prokaryotes, this is by far the more common approach, and has led to such standardized approaches as “Bio-Brick” assembly in which standardized genetic ‘parts’ such as promoters, ribosome binding sites, genes, and terminators can be readily swapped in and out of standard plasmids (http://partsregistry.org). This move towards standardization of genetic parts is a critical aim for synthetic biology, independent of the chassis organism or method of transformation. However, a limited number of plasmids are available for expression in cyanobacterial hosts. Plasmid assembly for expression \textit{in cis} or \textit{in trans} in cyanobacterial hosts has generally been performed in \textit{E. coli} because of the longer growth times that would be associated with assembling vectors in cyanobacterial hosts (Figure 3-2a). This requires broad host range plasmids. However, with the rise of \textit{in vitro} assembly methods such as SLIC [193], Gibson assembly [194], CPEC [195], fusion PCR [196], and Golden Gate [197], this limitation may become less important over time (Figure 3-2b). These next-generation cloning methods have been reviewed elsewhere [198] and will not be covered here. Fusion PCR has been used to construct linear DNA fragments for homologous recombination in cyanobacterial chromosomes [199], but to our knowledge replicative vectors for cyanobacteria have so far not been constructed without the use of a helper heterotrophic strain. Techniques for \textit{in vivo} assembly of plasmids that have been developed for yeast [200] may be adaptable to cyanobacteria because of their facility for homologous recombination (Figure 3-2c). Such an improvement could greatly speed up the process of making cyanobacterial mutant strains, either for modification \textit{in cis} or \textit{in trans}. The major technical challenge for such an approach is that the long time after transformation required to isolate cyanobacterial mutants (typically 1 week or more) means it is critical to have high-fidelity assembly methods to avoid a time-consuming screening process.
Although shuttle vectors do exist for cyanobacteria, there has been little characterization of their copy numbers in cyanobacterial hosts, and the lack of replicative vectors with varied copy numbers limits the valuable ability to control the expression level of heterologous genes by selecting their copy number [201, 202]. Plasmids derived from RSF1010 appear to have a copy number of 10-30 (or ~1-3 per chromosome) in *Synechocystis sp.* PCC 6803 [190, 203], but copy numbers of other broad host-range plasmids have not been quantified to date. Endogenous plasmids of cyanobacteria have also been used as target sites for expression of heterologous genes in *Synechococcus sp.* PCC 7002 [179]. This strain harbors several endogenous plasmids whose copy numbers range from ~1-8 per chromosome, with an approximate chromosome copy number of 6 per cell. *Synechocystis* sp. PCC 6803 also has plasmids whose copy numbers span a similar range (from ~0.4-8 per chromosome [204]).

**Figure 3-2:** DNA assembly methods.

Traditionally in cyanobacterial synthetic biology, plasmids are assembled *in vitro* and then propagated in *E. coli* before being transformed into cyanobacteria (a). More recently, methods have been developed for *in vitro* assembly and direct transformation via fusion PCR (b). In just the last year, a method has been developed for *in vivo* plasmid
assembly via homologous recombination in yeast which may also be applicable in certain cyanobacterial strains.

The origins of replication from these plasmids constitute a source of genetic parts that could be used to generate cyanobacterial expression plasmids having a range of copy numbers, and which could potentially be modified to create higher or lower-copy plasmids that are compatible with existing plasmids in various cyanobacterial systems. The range of shuttle vectors that have been used in cyanobacterial hosts has been recently reviewed [205]. While many tools are available for genetic modification of these biotechnologically promising strains, opportunities abound to develop new and improved tools that will allow research to proceed faster.

3.5 Unique challenges of the cyanobacterial lifestyle

Organisms that survive using sunlight as a primary nutrient face unique challenges. These must be better understood and addressed to fulfill the biotechnological promise of cyanobacteria through synthetic biology.

3.5.1 Life in a diurnal environment

A primary goal of synthetic biology in cyanobacteria is to use photosynthesis to convert CO$_2$ into higher-value products such as biofuels and chemical precursors. To make such a process economically and environmentally feasible will require using sunlight as a primary energy source. While some cyanobacteria are facultative heterotrophs, their key advantage over obligate heterotrophic bacteria is photosynthesis. Unlike heterotrophic growth environments where carbon and energy sources can be provided more uniformly both in space and time, sunlight will only be available during the day and will be attenuated as it passes through the culture. Under certain conditions, cultures may be able to take advantage of a ‘flashing light effect’ to integrate spatially uneven illumination by storing chemical energy when in bright light near the reactor surface and using that
energy to conduct biochemistry during time spent in the dark away from the reactor surface. This ability will depend on light intensity, mixing rates, reactor geometry, and likely other factors. Certain diazotrophic cyanobacteria can even use daylight to continue growth during the night. *Cyanothece* sp. ATCC 51142 (and several other strains [206-208]) is a unicellular diazotrophic cyanobacterium that performs photosynthesis and accumulates glycogen during the day, and then during the night breaks down its glycogen reserves to supply energy for nitrogen fixation. Thus, these strains spread out the energy available from sunlight over a 24-hour period. This process involves a genome-wide oscillation in transcription, with more than 30% of genes oscillating in expression between day and night [209]. To take full advantage of sunlight, synthetic systems must be created that are capable of responding appropriately to this challenging dynamic environment. It has recently been shown that biofuel-producing strains that dynamically tune the expression of heterologous pathways in response to their own intracellular conditions produce more biofuel and exhibit greater stability of heterologous pathways [210]. As challenging as the design of such a system was for batch heterotrophic cultures, it will be even more challenging in production environments that include a diurnal light cycle.

While not all strains exhibit as complete a physiological change between day and night as *Cyanothece* 51142, all cyanobacteria do have a circadian clock that adapts them to their autotrophic lifestyle. The cyanobacterial circadian clock is anchored by master regulators KaiA, KaiB, and KaiC, which act by cyclically phosphorylating and dephosphorylating each other [211]. While the circadian rhythm can be reconstituted *in vitro* using the three Kai proteins in the presence of ATP [212], the accurate maintenance of this clock *in vivo* depends on proper protein turnover [189], on codon selection in the kaiBC transcript [213], on transcriptional feedback [214], and on the controlled response of the entire program of cellular transcription to the output of the KaiABC oscillator. While disturbing rhythmicity can lead to strains that grow better under constant light, the circadian clock is adaptive for strains living in a dynamic environment [213, 215]. Therefore, integrating synthetic gene circuits such as biofuel production processes into the circadian rhythm of
cyanobacterial hosts will likely lead to both improved production and improved strain stability in outdoor production environments.

3.5.2 Redirecting Carbon Flux by decoupling growth from production

While redirecting carbon flux is a challenge in all metabolic engineering efforts, it has been suggested that stringent control of fixed carbon partitioning among central metabolic pathways poses a major limitation to chemical production especially in photosynthetic organisms [216]. During the growth phase, it may be true that carbon partitioning is tightly controlled by any number of mechanisms including metabolite channeling or simply high demand for metabolic intermediates. However, biofuel production during non-growth phases [182, 183, 217] demonstrates that under appropriate conditions, cyanobacterial hosts can produce biofuel compounds with higher selectivity, since biofuel can be produced by metabolically active cells even in the absence of growth. Enhancing their productivity in this phase is a major opportunity for cyanobacterial synthetic biologists to overcome these limits on carbon partitioning. Capturing this opportunity will require designing complete metabolic circuits that remain highly active during stationary phase.

3.5.3 RNA-based regulation

Recently, regulation of gene expression through RNA mechanisms has received great attention across bacterial clades [218-220]. While these mechanisms of regulation may be important in all bacteria, their prominence is perhaps the greatest in the cyanobacteria and may help these diurnal organisms adapt to their highly dynamic environment: in a recent dRNA-seq study, many of the most highly expressed RNAs belonged to families of non-coding RNAs which are present in nearly all sequenced cyanobacteria, but not in any other organisms [220, 221]. While their high expression in Synechocystis 6803 suggests functional importance for non-coding RNAs, few have clearly elucidated functions to date. syr1 overexpression has been shown to lead to a severe growth defect in
Synechocystis 6803 [220]. Another small RNA, isiR, has a critical function in stress response in Synechocystis 6803. isiR binds to the mRNA (isiA) for the iron-stress inducible protein, which when translated, forms a ring around trimers of photosystem I, preventing their activity and thus oxidative stress in the absence of sufficient iron [222]. The binding of isiR to isiA appears to result in rapid degradation. This particular arrangement allows a very rapid and emphatic response to iron repletion in cyanobacteria, since a large pool of isiA transcripts can be quickly silenced and marked for degradation by transcription of the antisense isiR. Although little is so far known about the generality of this type of regulation, the dynamics of this response might also be effective to use for synthetic systems in cyanobacteria that live in the presence of light as an intermittently available but critical nutrient.

While non-coding RNA has received a lot of recent attention, two-component systems make up the most widely studied family of environmental response regulators in cyanobacteria. Many of these systems have known functions in response to diverse environmental stimuli such as nitrogen, phosphorous, CO₂, temperature, salt, and light intensity and quality [223, 224]. Many of the most widely-used systems in the construction of synthetic biological devices (such as the ara and lux clusters) use 2-component systems, and even combine 2-component systems with non-coding RNA to control system dynamics [225]. As synthetic biology advances into the construction of more and more complex systems, there will be a growing need to understand and use all of the different mechanisms available for control of gene expression and enzyme activity in cyanobacteria.

3.6 Parts for Cyanobacterial Synthetic Biology

While cyanobacteria are promising organisms for biotechnology, synthetic biology tools for these organisms lag behind what has been developed for E. coli and yeast [177]. Furthermore, synthetic biology tools developed in E. coli or yeast often do not function as designed in cyanobacteria [190]. Here, we discuss inducible promoters and reporters in cyanobacteria, and cultivation systems that will allow their testing at increased
throughput. Refining such systems will make cyanobacterial synthetic biology more user-friendly, a central goal for developing the ‘green *E. coli*.

### 3.6.1 Inducible Promoter

Creation of synthetic biology systems that predictably respond to a specific signal often depends upon inducible promoters for transcriptional control. An ideal inducible promoter will have the following properties: (1) It will not be activated in the absence of inducer. (2) It will produce a predictable response to a given concentration of inducer or repressor. This response may be digital (i.e., on/off) or graded change with different concentrations of inducer/repressor. (3) The inducer at saturating concentrations should have no harmful effect on the host organism. (4) The inducer should be cheap and stable under the growth conditions of the host. Finally, (5) the inducible system should act orthogonally to the host cell’s transcriptional program. Ideal transcriptional repressors should not bind to native promoters and if non-native transcriptional machinery is used (such as T7 RNA polymerase) it should not initiate transcription from native promoters. Promoters must perform as ideally as possible in order to be used in the construction of more complex genetic circuits [226].

Many common inducible promoters in cyanobacteria respond to transition metals. These have often been the basis of metal detection systems [227-231]. Cyanobacteria balance metal intake for the organisms’ needs against potential oxidative stress and protein denaturation [230, 232] via tightly regulated systems. As shown in Table 3-2, cyanobacteria’s metal-responsive promoters frequently show greater than 100-fold dynamic range. For example, the promoter for the *Synechocystis* sp. PCC 6803 gene, *coaA*, was induced 500 fold by 6 µM Co$^{2+}$ [233], and *P_smt* from *Synechococcus elongatus* PCC 7942 was induced 300-fold by 2 µM Zn$^{2+}$ [227]. The most responsive cyanobacterial promoters reported were *P_nrsB* from *Synechocystis* sp. PCC 6803, responding 1000-fold to 0.5 µM Ni$^{2+}$[231], and *P_isiAB* also from *Synechocystis* sp. PCC 6803, repressed 5000-fold by 30 µM Fe$^{3+}$ following depletion [234].
**Table 3-2: Inducible promoters used in cyanobacterial hosts**

<table>
<thead>
<tr>
<th>Promoter</th>
<th>Source</th>
<th>Inducer/Repressor &amp; Concentration</th>
<th>Expression Host</th>
<th>Expressed Gene</th>
<th>Dynamic Range</th>
<th>Measure of Expression</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Metal-Inducible promoters:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>arsB</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>Inducer AsO2-720 µM</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>arsB</td>
<td>100 fold</td>
<td>RT-PCR</td>
<td>Blasi et al., 2012</td>
</tr>
<tr>
<td>ziaA</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>Inducer Cd2+2 µM</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>ziaA</td>
<td>10 fold</td>
<td>RT-PCR</td>
<td>Blasi et al., 2012</td>
</tr>
<tr>
<td>coat</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>Inducer Co2+6 µM</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>gene encoding EFE from <em>Pseudomonas syringae</em></td>
<td>500 fold</td>
<td>48 nl ethylene/ml h</td>
<td>Guerrero et al., 2012</td>
</tr>
<tr>
<td>coat</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>Inducer Co2+6.4 µM</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>coaR+luxAB</td>
<td>70 fold</td>
<td>70 (relative luminescence)</td>
<td>Peca et al., 2008</td>
</tr>
<tr>
<td>coat</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>Inducer Co2+3 µM</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>coaT</td>
<td>10 fold</td>
<td>RT-PCR</td>
<td>Peca et al., 2007</td>
</tr>
<tr>
<td>nrsB</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>Inducer Co2+3 µM</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>nrsB</td>
<td>10 fold</td>
<td>RT-PCR</td>
<td>Peca et al., 2007</td>
</tr>
<tr>
<td>coat</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>Inducer Co2+1 µM</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>coaT</td>
<td>10 fold</td>
<td>RT-PCR</td>
<td>Blasi et al., 2012</td>
</tr>
<tr>
<td>petE</td>
<td><em>Synechocystis</em> sp. PCC</td>
<td>Inducer Cu2+</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>gene encoding EFE from</td>
<td>5 fold</td>
<td>28 nl ethylene/ml h</td>
<td>Guerrero et al., 2012</td>
</tr>
<tr>
<td>Genes</td>
<td>Species/Strain</td>
<td>Inducer</td>
<td>Concentration</td>
<td>Target</td>
<td>Function</td>
<td>Fold Change</td>
<td>Comments</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------</td>
<td>------------</td>
<td>---------------</td>
<td>-----------------</td>
<td>-----------------------------------------------</td>
<td>-------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>petE</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>Inducer</td>
<td>0.5 µM</td>
<td><em>Pseudomonas syringae</em></td>
<td>hetP</td>
<td>4.5 fold</td>
<td>8% heterocyst frequency</td>
</tr>
<tr>
<td>petE</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>Inducer</td>
<td>Cu2+ 3µM</td>
<td><em>Anabaena</em> sp. PCC 7120</td>
<td>hetN (prevents heterocyst formation)</td>
<td>qualified</td>
<td>0% heterocysts from 10% uninduced</td>
</tr>
<tr>
<td>isiAB</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>Repressor</td>
<td>Fe3+ 30 µM</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>isiAB+gfp</td>
<td>5000 fold</td>
<td>From 5000 GFP fluorescence (relative units)</td>
</tr>
<tr>
<td>idiA</td>
<td><em>Synechococcus elongatus</em> PCC 7942</td>
<td>Repressor</td>
<td>Fe2+ 0.043mM</td>
<td><em>Synechococcus</em> sp. PCC 7942</td>
<td>luxAB</td>
<td>170 fold</td>
<td>Luminescence 5.3 x 10^6 cpm</td>
</tr>
<tr>
<td>isiAB</td>
<td><em>Synechococcus</em> sp. strain PCC 7002</td>
<td>Repressor</td>
<td>Fe3+ 100 nM</td>
<td><em>Synechococcus</em> sp. strain PCC 7002</td>
<td>luxAB from <em>Vibrio harveyi</em></td>
<td>2 fold</td>
<td>From 1200 relative luminescence units cell-1 x 10^5 s-1</td>
</tr>
<tr>
<td>nrsB</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>Inducer</td>
<td>Ni2+ 0.5 µM</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>nrsB</td>
<td>1000 fold</td>
<td>RT-PCR</td>
</tr>
<tr>
<td>nrsB</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>Inducer</td>
<td>Ni2+ 5 µM</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>nrsB</td>
<td>400 fold</td>
<td>RT-PCR</td>
</tr>
<tr>
<td>nrsB</td>
<td><em>Synechocystis</em> sp. PCC</td>
<td>Inducer</td>
<td>Ni2+</td>
<td><em>Synechocystis</em> sp. PCC 6803</td>
<td>nrsR+luxAB</td>
<td>50 fold</td>
<td>50 (relative luminescence)</td>
</tr>
<tr>
<td></td>
<td>Zn2+ (µM)</td>
<td>Inducer</td>
<td>Gene Encoding</td>
<td>fold</td>
<td>Relative Luminescence</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---------</td>
<td>---------</td>
<td>---------------</td>
<td>------</td>
<td>----------------------</td>
<td>-----------</td>
<td></td>
</tr>
<tr>
<td>6803</td>
<td>6.4</td>
<td><em>Synechococcus elongatus</em> PCC 7942</td>
<td><em>luxCDABE from Vibrio fisheri</em></td>
<td>300</td>
<td>325,000 cps</td>
<td>Erbe et al., 1996</td>
<td></td>
</tr>
<tr>
<td>Smt</td>
<td>2</td>
<td><em>Synechococcus elongatus</em> PCC 7942</td>
<td><em>ziaA</em></td>
<td>40</td>
<td>25 (relative luminescence)</td>
<td>Peca et al., 2007</td>
<td></td>
</tr>
<tr>
<td>ziaA</td>
<td>5</td>
<td><em>Synechocystis sp. PCC 6803</em></td>
<td><em>ziaA</em></td>
<td>40</td>
<td>RT-PCR</td>
<td>Blasi et al., 2012</td>
<td></td>
</tr>
<tr>
<td>coat</td>
<td>2</td>
<td><em>Synechocystis sp. PCC 6803</em></td>
<td><em>coaR+luxAB</em></td>
<td>25</td>
<td>25 (relative luminescence)</td>
<td>Peca et al., 2008</td>
<td></td>
</tr>
<tr>
<td>coat</td>
<td>3.2</td>
<td><em>Synechocystis sp. PCC 6803</em></td>
<td><em>coaT</em></td>
<td>10</td>
<td>RT-PCR</td>
<td>Peca et al., 2007</td>
<td></td>
</tr>
<tr>
<td>coat</td>
<td>4</td>
<td><em>Synechocystis sp. PCC 6803</em></td>
<td><em>coaT</em></td>
<td>8</td>
<td>RT-PCR</td>
<td>Blasi et al., 2012</td>
<td></td>
</tr>
<tr>
<td>7002</td>
<td>2</td>
<td><em>Synechococcus elongatus</em> PCC 7002</td>
<td><em>hydA1 from Chlamydomonas reinhardtii</em></td>
<td>2</td>
<td>109 nmol H₂</td>
<td>Guerrero et al., 2012</td>
<td></td>
</tr>
<tr>
<td>6803</td>
<td>3.5</td>
<td><em>Synechocystis sp. PCC 6803</em></td>
<td><em>hydA1 from Chlamydomonas reinhardtii</em></td>
<td>qualified but not quantified</td>
<td>109 nmol H₂</td>
<td>Berto 2011</td>
<td></td>
</tr>
<tr>
<td>Metabolite -Inducible Promoters</td>
<td>Inducer</td>
<td>Inducer</td>
<td>Inducer</td>
<td>Inducer</td>
<td>Inducer</td>
<td>Inducer</td>
<td>Inducer</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>tetR&lt;sup&gt;3&lt;/sup&gt;</td>
<td>E. coli</td>
<td>aTc</td>
<td>10&lt;sup&gt;^-3&lt;/sup&gt; ng/ml</td>
<td>Synchocystis sp. strain ATCC27184</td>
<td>EYFP</td>
<td>290 fold over 10,000 emission/cell (a.u.)</td>
<td>Huang and Lindblad, 2013</td>
</tr>
<tr>
<td>trp-lac</td>
<td>E. coli</td>
<td>IPTG</td>
<td>100 µM</td>
<td>Synchococcus elongatus PCC 7942</td>
<td>invA and glf genes from Zymomonas mobilis</td>
<td>160 fold for fructose + 30 fold for glucose</td>
<td>160 µM fructose + 30 µM glucose</td>
</tr>
<tr>
<td>trc</td>
<td>E. coli</td>
<td>IPTG</td>
<td>1mM</td>
<td>Synchococcus elongatus PCC 7942</td>
<td>uidA from E. coli</td>
<td>36 fold</td>
<td>340 nmol MU min&lt;sup&gt;-1&lt;/sup&gt; (mg protein)&lt;sup&gt;-1&lt;/sup&gt; (β-Glucuronidase activity)</td>
</tr>
<tr>
<td>A1lacO-1</td>
<td>E. coli</td>
<td>IPTG</td>
<td>1 mM</td>
<td>Synchocystis sp. PCC 6803</td>
<td>gene encoding EFE from Pseudomonas syringae</td>
<td>8 fold</td>
<td>170 nl ethylene/ml h</td>
</tr>
<tr>
<td>trc20</td>
<td>E. coli</td>
<td>IPTG</td>
<td>2 mM</td>
<td>Synchocystis sp. PCC 6803</td>
<td>gene encoding GFPmut3B</td>
<td>4 fold</td>
<td>12 (units relative to promoter activity for lacI)</td>
</tr>
<tr>
<td>trc10</td>
<td>E. coli</td>
<td>IPTG</td>
<td>2 mM</td>
<td>Synchocystis sp. PCC 6803</td>
<td>gene encoding GFPmut3B</td>
<td>1.6 fold</td>
<td>101 (units relative to promoter activity for lacI)</td>
</tr>
<tr>
<td>Promoter</td>
<td>Organism</td>
<td>Inducer</td>
<td>Inducible Gene/Expression</td>
<td>Relative Activity</td>
<td>References</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>----------</td>
<td>---------</td>
<td>---------------------------</td>
<td>-------------------</td>
<td>------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LlacO1²</td>
<td><em>E. coli</em></td>
<td>IPTG 1 mM</td>
<td><em>Synechococcus elongatus</em> PCC7942</td>
<td>alsS (<em>B. subtilis</em>), alsD (<em>A. hydrophila</em>), and adh (<em>C. beijerincki</em>)</td>
<td>1.6 fold</td>
<td>1.6 (relative activity of sADH and ALS)</td>
<td>Oliver et al., 2013</td>
</tr>
<tr>
<td>Tre⁴</td>
<td><em>E. coli</em></td>
<td>IPTG 1 mM</td>
<td><em>Synechocystis</em> sp. PCC6803</td>
<td>gene encoding EFE from <em>Pseudomonas syringae</em></td>
<td>no significant difference</td>
<td>170 nl ethylene/ml h</td>
<td>Guerrero et al., 2012</td>
</tr>
<tr>
<td>Macronutrient-Inducible Promoters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>psbA2</td>
<td><em>Synechocystis</em> sp. PCC6803</td>
<td>light 500 µmol photons m⁻² s⁻¹</td>
<td><em>Synechocystis</em> sp. PCC6803</td>
<td><em>ispS</em> from <em>Pueraria montana</em> (kudzu)</td>
<td>qualified but not quantified</td>
<td>~50 mg isoprene per g dry cell weight per day</td>
<td>Lindberg et al., 2010</td>
</tr>
<tr>
<td>psbA2¹</td>
<td><em>Synechocystis</em> sp. PCC6803</td>
<td>light 50 µEm⁻² sec⁻¹</td>
<td><em>Synechocystis</em> sp. PCC6803</td>
<td><em>hydA1</em> from <em>Chlamydomonas reinhardtii</em></td>
<td>qualified but not quantified</td>
<td>130 nmol H₂ mg Chl⁻¹ min⁻¹</td>
<td>Berto et al., 2011</td>
</tr>
<tr>
<td>psbA1</td>
<td><em>Anabaena</em> sp. PCC7120</td>
<td>light 30 µEm⁻² sec⁻¹</td>
<td><em>Anabaena</em> sp. PCC7120</td>
<td><em>hetR</em> from <em>E. coli</em></td>
<td>17% heterocyst frequency</td>
<td></td>
<td>Chaurasia and Apte, 2011</td>
</tr>
<tr>
<td>nirA</td>
<td><em>Synechococcus elongatus</em></td>
<td>Inducer/Repressor NO3⁻/NH4⁺</td>
<td><em>Synechocystis</em> sp. PCC6803</td>
<td>gene encoding p-</td>
<td>25 fold</td>
<td>250 ng tocopherol/m</td>
<td>Qi et al., 2005</td>
</tr>
</tbody>
</table>
1: In the presence of 5 µM DCMU, which inhibits the PSII-dependent oxygen evolution.
2: Leaky production of 2,3-butanediol, no IPTG and 1 mM IPTG similar.
3: Grown in the dark on 5 mM glucose.
4: Plac variants had differential expression early in growth phase but dynamic range was reduced as growth proceeded.
5: in the presence of 5 µM DCMU, which inhibits the PSII-dependent oxygen evolution.
While the sensitivity of these promoters to low concentrations of ions may seem like an advantage, in practice it can make them difficult to use. Glassware must be thoroughly cleaned according to special protocols to remove trace metals and cells often have to be starved for extended periods, inducing stress responses, to use such inducible systems. Additionally, promoters endogenous to a chassis strain are woven into a complex, incompletely understood regulatory system. In this system, promoters are activated by multiple inducers, such as PcoaT (Co$^{2+}$ and Zn$^{2+}$) and PziaA (Cd$^{2+}$ and Zn$^{2+}$), both from *Synechocystis* sp. PCC 6803 and inducers can also activate multiple promoters, such as Cd$^{2+}$ inducing ziaA and isiA [229]. Thus, these promoters fall short according to criteria 2, 3, and 5 described above.

While few good choices have so far been available for inducible promoters in cyanobacteria, it will be helpful to understand the differences in the cellular machinery of *E. coli* and cyanobacteria in order to adapt existing systems for use in a cyanobacterial ‘green *E. coli*’. First, RNA polymerase (RNAP) is structurally different between *E. coli* and cyanobacteria. In cyanobacteria the β’ subunit of the RNAP holoenzyme is split into two parts, as opposed to one in most eubacteria, creating a different DNA binding domain [235]. Being photosynthetic, circadian, and sometimes nitrogen-fixing, cyanobacteria also employ three sets of interconnected σ factors that are different than those used by *E. coli* [235]. Guererro et al. (2012) looked at the variation in the -35 and -10 regions of P$_{A1lacO-1}$ and P$_{trc}$. P$_{trc}$ is not inducible in *Synechocystis* sp PCC 6803 and had the “standard” bacterial structure in these regions while P$_{A1lacO-1}$, which produced an eight fold response to IPTG in the same host, had a different structure in both regions. They postulated that *Synechocystis* 6803’s sigma factors had different selectivity for these two regions. In fact, by systematically altering the bases between -10 and the transcription start site, a library of TetR-regulated promoters with improved inducibility were created in *Synechocystis* sp. strain ATCC27184 (a glucose-tolerant derivative of *Synechocystis* 6803). The best performing promoter induced a 290-fold change in response to 1 ug/ml aTc [192]. This work demonstrates the improvements that can be seen when modifying parts to work in a particular chassis. However, the light-sensitivity of the inducer aTc
required the use of special growth lights that may have had other effects on photoautotrophic metabolism. Further studies that follow in this vein of using well-characterized synthetic biology parts and modifying them to function optimally in a particular cyanobacterial chassis are likely to bear fruit.

The lack of inducibility seen in lac-derived promoters in cyanobacteria could also be a function of inadequate transport of IPTG into cells. Concentrations of IPTG above 1 mM have been shown to induce lac-derived promoters in organisms without an active lactose permease, like many cyanobacteria. By introducing an active lactose permease into *Pseudomonas fluorescens*, inducibility was boosted 5 times at 0.1 mM IPTG [236]. Evolving the Lac repressor for improved inducibility is another strategy. Gene expression improved ten times with 1 µM IPTG through rounds of error prone PCR and DNA shuffling [237]. Strength of expression and inducibility may also vary between different cyanobacterial strains. IPTG caused as much as a 36-fold response using the trc promoter in *Synechococcus elongatus* PCC 7942, but little or no response in *Synechocystis* sp. PCC 6803 (See Table 3-2). Phylogenetic analysis of σ factors from six different cyanobacterial strains, including *Synechocystis* 6803, showed *S. elongatus* 7942 to be distinctive. *S. elongatus* 7942 has σ factors that are unique to marine cyanobacteria as well as a group 3 σ factor similar to those from the heterocyst-forming *Anabaena* sp. PCC 7120 [235]. Understanding these strain-specific differences will enhance the synthetic biologist’s ability to design promoters with ideal characteristics in their chassis of choice. This relates to the ability to take up inducers as well as the optimal characteristics of inducers (as in the light-sensitivity of aTc) as described above.

### 3.6.2 Reporters

Characterization of synthetic biological circuits depends on a reporting method to track the expression, interaction and position of proteins. Preferably the reporter should be detected without destruction of the organisms or additional inputs. Bacterial luciferase and fluorescent proteins are the most common non-invasive reporters. The *lux* operon is
frequently used for reporting in cyanobacteria [230, 232, 238] and is well suited for real
time reporting of gene expression due to the short half-life of the relevant enzymes [239].
The superior brightness of fluorescent proteins makes them more ideal for subcellular
localization via microscopy or for cell-sorting methods. Fluorescent proteins are
produced in an array of colors and also do not require additional substrates. Their use in
cyanobacteria is somewhat complicated by the fluorescence of the organism’s
photosynthetic pigments, but Cerulean, GFPmut3B (a mutant of green fluorescent
protein) and EYFP (enhanced yellow fluorescent protein) have all been used successfully
in cyanobacteria as reporters of gene expression [177, 190-192].

Bacterial luciferase luminesces upon oxidation of reduced flavin mononucleotide [240].
Fluorescent proteins also require oxygen to correctly fold and fluoresce [241]. The light-
dark cycle of nitrogen-fixing cyanobacteria provides temporal separation of the oxygen-
sensitive nitrogenase from oxygen-evolving photosynthesis [242]. During the dark
cycle, respiration reduces intra-cellular oxygen levels so that nitrogenase can function.
Therefore, neither bacterial luciferase nor traditional fluorescent proteins can likely be
used to study cyanobacteria in their dark cycle or to report on synthetic biology systems
that operate in these oxygen-depleted conditions. Using blue light photoreceptors from
*Bacillus subtilis* and *Pseudomonas putida*, oxygen-independent flavin mononucleotide-
binding florescent proteins have been devised [243]. With an excitation wavelength of
450 nm and an emission wavelength of 495 nm, they should perform well in
cyanobacteria, although no data supporting this has been published yet. Functionality of
these new fluorescent proteins was also improved by replacing a phenylalanine suspected
of quenching with serine or threonine, resulting in a doubling of the brightness [244].
This expanding variety of easily readable reporter systems will be extremely valuable for
cyanobacterial synthetic biology.
3.6.3 Cultivation systems

To date, most synthetic biology and metabolic engineering work in cyanobacteria has been performed using simple, low-tech cultivation methods such as shake flasks or bubbling tubes grown under standard fluorescent light sources. Often, laboratory incubators have simply been retrofitted by the addition of fluorescent light sources available in home improvement stores. However, as light and CO₂ are major nutrients for cyanobacteria, it is critical to properly standardize the inputs of these resources to reliably characterize biological parts. It is also critical to increase the throughput of cyanobacterial growth systems to be able to screen the large numbers of variants that can be generated by combinatorial methods, as is routinely performed by growing heterotrophic bacterial cultures in 96-well plate format. Growth of cyanobacteria in 6-well plates can be routinely performed in our lab and by others [192] along with 24-well plates [245], but growth in 96-well plates is poor, limiting assay throughput and requiring more space in lighted chambers under consistent illumination, which is often a limitation. Simple, low-cost systems to reproducibly grow many cyanobacterial cultures in parallel are necessary.

3.7 Genome-scale modeling and fluxomics of cyanobacteria

A primary aim of cyanobacterial synthetic biology is the production of particular metabolites as biofuels or platform chemicals. As such, better understanding the metabolic phenotypes of wild-type and synthetic strains is a critical aim. While cyanobacterial metabolomics have been recently reviewed [246], here we describe recent progress in genome-scale modeling and fluxomics of cyanobacteria. These approaches can help guide the creation of synthetic strains with desirable metabolic phenotypes such as biofuel overproduction via in silico prediction or in vivo measurement of metabolic fluxes (See Figure 3-3). Specific to cyanobacterial systems, we highlight a number of challenges including complexity of modeling the photosynthetic metabolism and performing flux balance analysis, poor annotations of important metabolic pathways, and
unavailability of *in vivo* gene essentiality information for most cyanobacteria. Finally, we focus on recent advancements in this area.

**Figure 3-3:** Using fluxomics and genome scale models to link genotype to metabolic phenotype.

From an annotated genome sequence, a stoichiometric model of metabolism can be constructed. That model can be solved via either prediction of an optimal flux phenotype (FBA) or measurement of actual flux phenotype (13C-MFA). These results can help suggest modifications for altering the phenotype of the cell in a desired manner. In this way, a synthetic biologist can design new strains, build them using genetic modification
methods, and test their phenotypes before designing new modifications in an iterative fashion.

3.7.1 Challenges

3.7.1.1 Incorporating photoautotrophy into metabolic models

Flux balance analysis (FBA) is a tool to make quantitative *in silico* predictions about metabolism [247-250]. An FBA model incorporates the stoichiometry of all genome-encoded metabolic reactions and assumes steady-state growth, such as during exponential phase. This assumption leads to a model that consists of a system of algebraic equations, which state that the rate of producing any given metabolite is equal to the rate of consuming that metabolite. A solution to this system of equations is a possible answer to the question “what are all the metabolic fluxes in this system?” Since there are usually more reactions than metabolites, this system of equations is underdetermined and has many possible solutions. Therefore, one has to pick a solution that satisfies a biological objective, such as maximal growth, energy production, or byproduct formation [251]. For this purpose, a model will also include upper and lower bounds of fluxes that constrain the model to produce physically and biologically reasonable solutions.

Success of FBA greatly depends on the quality of the metabolic network reconstruction as well as the availability of regulatory constraints under a given environmental or growth condition. For instance, constraints can be added that disable or limit fluxes due to known regulatory constraints or substrate availability [20]. For cyanobacteria, the major challenges to develop a genome-scale metabolic model and subsequently perform FBA are the same ones faced by these organisms in their diurnal environment: how to incorporate light and how to differentiate light and dark metabolisms. Although it has been nearly a decade since publication of the first study applying flux balance analysis to cyanobacteria, it is only recently that models have incorporated complete descriptions of
the light reactions of photosynthesis [109]. In so doing, these authors were able to
highlight the critical importance of alternate electron flow pathways to growth under
diverse environmental conditions, and to identify differences in metabolism during
carbon-limited and light-limited growth. However, debate remains among photosynthesis
researchers about the exact form of the light reactions [252, 253]. This uncertainty about
the exact stoichiometry of metabolism is a challenge for the predictive power of FBA in
photosynthetic systems. While FBA requires the assumption of a pseudo-steady state, all
cyanobacteria must alternate between day and night metabolisms during a diurnal cycle.
A recent model [3] of Cyanothece sp. ATCC 51142 utilizes proteomic data to model the
diurnal rhythm of this strain, which fixes carbon during the day and nitrogen during the
night (see section 4 above).

3.7.1.2 Incompleteness of genome annotation

Genome scale models are built starting with an annotated genome sequence (see Figure
3), which allows prediction of which metabolic reactions are available in a given strain.
However, genome annotation is constantly evolving, and open questions remain about
important metabolic reactions in cyanobacteria.

The understanding of several key pathways in cyanobacteria has been recently revised.
Zhang and Bryant [110] identified enzymes from Synechococcus 7002 that can complete
the TCA cycle in vitro and have homologues in most cyanobacterial species, which were
previously thought to possess an incomplete TCA cycle. Based on this information,
Synechocystis 6803 model iSyn731 [3] allows for a complete TCA cycle including these
reactions. However, using flux variability analysis [254, 255] it was determined that this
alternate pathway is not essential for maximal biomass production (unpublished results,
[3]). Fatty acid metabolism in cyanobacteria has unique properties that have been
recently uncovered due to increased interest in these pathways for biofuel production.
Both Synechocystis sp. PCC 6803 and Synechococcus elongatus sp. PCC 7942 contain a
single candidate gene annotated for fatty acid activation. While in both organisms the
gene is annotated as acyl-CoA synthetase, it shows only acyl-ACP synthetase activity instead [256]. Further analysis also shows the importance of acyl-ACP synthetase in enabling the transfer of fatty acids across the membrane [257]. Quinone synthesis is another pathway with conflicting annotations. Cyanobacteria contain neither ubiquinone nor menaquinone [258]. Despite the lack of ubiquinone within cyanobacteria, a number of cyanobacterial genomes contain homologs for six *E. coli* genes involved in ubiquinone biosynthesis [259]. Given these homologous genes it is probable that plastoquinone, a quinone molecule participating in the electron transport chain, is produced in cyanobacteria using a pathway very similar to that of ubiquinone production in proteobacteria. Wu et al. [99] showed that *Cyanothece* 51142 contains an alternative pathway for isoleucine biosynthesis. Threonine ammonia-lyase, catalyzing the conversion of threonine to 2-ketobutyrate, is absent in *Cyanothece* 51142. Instead, this organism uses a citramalate pathway with pyruvate and acetyl-CoA as precursors for isoleucine synthesis. An intermediate in this pathway, namely ketobutyrate, can be converted to higher alcohols (propanol and butanol) via this non-fermentative alcohol production pathway. These active areas of research will help to better define cyanobacterial metabolism and allow the generation of models that can more accurately predict cellular phenotypes. While newer fluxomics techniques can yield powerful results in well-characterized strains, developing a ‘green *E. coli*’ will also require expanded knowledge of biochemistry that to date can only come from older methods of single gene or single protein analysis.

### 3.7.1.3 Fewer mutant resources to test model accuracy

The quality or accuracy of any genome-scale metabolic model can be tested by contrasting the *in silico* growth phenotype with available experimental data on the viability of single or multiple gene knockouts [8]. Any discrepancies between model predictions and observed results can aid in model refinement [41]. For model strains besides cyanobacteria, concerted efforts to create complete mutant libraries have led to improvements in metabolic modeling. To the best of our knowledge, extensive *in vivo*
gene essentiality data are available only for Synechocystis 6803 among the cyanobacteria in the CyanoMutants database [111, 128], but only for ~119 genes, compared with 731 genes associated with metabolic reactions in a recent genome-scale model [3]. Thus, only a small subset of the model predictions on gene essentiality can be evaluated using available data for Synechocystis 6803, and the proportion is much less for any other strain. While a genome-wide library of knockout mutants has been created in Synechococcus 7942 [178] segregation (and thus essentiality) has only been checked for a small selection of these mutants and its not available in any large-scale public database to date. Unavailability of such mutant information limits model validation and in turn hurts the value of computational predictions from FBA. Efforts to create complete mutant libraries in model cyanobacterial strains would improve the fidelity of genome-scale metabolic models, leading to testable hypotheses about how to alter metabolism for metabolite overproduction.

3.7.2 Recent advances

3.7.2.1 Detailed genome-scale models

Genome-scale models contain detailed Gene-Protein-Reaction associations, a stoichiometric representation of all possible reactions occurring in an organism, and a set of appropriate regulatory constraints on each reaction flux. They are differentiated from more basic FBA models simply by their completeness – they span all or nearly all of the metabolic reactions encoded in a genome. Thus, these models can have greater predictive value than those of only central metabolism. Cyanothece 51142 is one of the most potently diazotrophic unicellular cyanobacteria characterized and the first diazotrophic cyanobacterium to be completely sequenced [86]. The first genome-scale model for Cyanothece 51142, iCce806, is recently developed [102], while another more recent genome-scale model iCyt773 contains an additional 266 unique reactions spanning pathways such as lipid, pigment and alkane biosynthesis [3]. iCyt773 also models diurnal
metabolism by including flux regulation based on available day/night protein expression data [160] and developing separate (light/dark) biomass equations. These models greatly enhance the ability to make computational predictions about this unique and promising diazotrophic organism.

Since *Synechocystis* 6803 is a model cyanobacterial strain, it has long been the target for modeling of photosynthetic central metabolism [104, 105]. More recent models [89, 108] analyze growth under different conditions and detect bottlenecks and gene knock-out candidates to enhance metabolite production (e.g., ethanol, succinate, and hydrogen). A recent model represents the photosynthetic apparatus in detail, detects alternate flow pathways of electrons and also pinpoints photosynthetic robustness during photoautotrophic metabolism [109]. iSyn731, the latest of all *Synechocystis* 6803 models, integrates all recent developments and supplements them with improved metabolic capability and additional literature evidence. As many as 322 unique reactions are introduced in iSyn731 including reactions distributed in pathways such as heptadecane and fatty acid biosynthesis [3]. Furthermore, iSyn731 is the first model for which both gene essentiality data [111] and MFA flux data [112] are utilized to assess the predictive quality. Additionally, genome scale modeling has been extended to include another model cyanobacterium, *Synechococcus* sp. PCC 7002 [260]. Other model strains highlighted in Table 3-1 have not yet had genome-scale models generated for their metabolism. Thus, stoichiometric models are emerging as a valuable tool for use across model cyanobacterial systems.

### 3.7.2.2 $^{13}$C MFA analysis

While *in silico* models are great tools for generating hypotheses on how to use synthetic biology interventions to alter metabolism, they need to be complemented by fluxomics methods that allow *in vivo* measurement of metabolic fluxes to assess these interventions. Such a suite of tools allows the closure of the design-build-test engineering cycle in synthetic biology. To this end, Young et al. (2011) have developed a method to measure
fluxes in autotrophic metabolism via dynamic isotope labeling measurements. In this approach, cultures are fed with a step-change from naturally labeled bicarbonate to NaH\(^{13}\)CO\(_3\) and the labeling patterns of metabolic intermediates are followed over a time-course to determine relative rates of metabolic flux. Previous studies [105] have also assessed metabolic fluxes under mixotrophic growth conditions, using a pseudo-steady-state approach in which cells are fed with \(^{13}\)C labeled glucose and metabolic fluxes are inferred from labeling patterns of proteinogenic amino acids. These studies have been extremely useful in identifying fluxes that exist \(\textit{in vivo}\), but have previously been regarded as wasteful or futile cycles, such as the oxidative pentose phosphate pathway and RuBP oxygenation. Comparisons between flux measurements [112] and flux predictions [3] for \textit{Synechocystis} 6803 have revealed the necessity of additional regulatory information for accurate \(\textit{in silico}\) predictions of phenotype. These modeling and fluxomics efforts have resulted in deeper understanding of the metabolic capabilities of the modeled strains and of cyanobacteria in general.

### 3.8 Conclusions

Cyanobacterial synthetic biology offers great promise for enhancing efforts to produce biofuels and chemicals in photoautotrophic hosts. While several cyanobacterial chassis strains have been used in synthetic biology efforts, the tools for their manipulation and analysis need greater development to unlock this potential and develop a ‘green \textit{E. coli}\’. Metabolic modeling is a complementary tool that can help guide the creation of synthetic strains with desirable phenotypes. By developing the tools for strain manipulation and control, synthetic biologists can unlock a bright future for the biotechnological use of abundant light and CO\(_2\).
Chapter 4  

**ZEA MAYS iRS1563: A COMPREHENSIVE GENOME-SCALE METABOLIC RECONSTRUCTION OF MAIZE METABOLISM**


4.1 Introduction

*Zea mays*, commonly known as maize or corn, is a plant organism of paramount importance as a food crop, biofuel production platform and a model for studying plant genetics [261]. Maize accounts for 31% of the world production of cereals occupying almost one-fifth of the worldwide land dedicated for cereal production [262]. Maize cultivation led to 12 billion bushels of grain in the USA alone in 2008 worth $47 billion [263]. Maize is the second largest crop, after soybean, used for biotech applications [262]. In addition to its importance as a food crop, 3.4 billion gallons of ethanol was produced from maize in 2004 [263]. Maize derived ethanol accounts for 99% of all biofuels produced in the United States [263]. However, currently nearly all of this bioethanol is produced from corn seed [264]. Ongoing efforts are focused on developing and commercializing technologies that will allow for the efficient utilization of plant fiber or cellulosic materials (e.g. maize stover and cereal straws) for biofuel production. Maize is the most studied species among all grasses with respect to cell wall lignification and digestibility, which are critical for the efficient production of cellulosic biofuels [265]. A thorough evaluation of the metabolic capabilities of maize would be an important resource to address challenges associated with its dual role as a food (e.g., starch storage) and biofuel crop (e.g., cell wall deconstruction).

This decade we witnessed significant advancements towards mapping plant genes to metabolic functions culminating with the complete genome sequencing and partial
annotation of a number of plant species, namely, *Arabidopsis thaliana* [266], *Oryza Sativa* [267, 268], *Sorghum bicolor* [269], *Zea mays* [270] and *Theobroma cacao* [271]. Nevertheless, attempts to engineer plant metabolism for desired overproductions have been met with only limited success [272]. Genetic modifications seldom bring about the expected/desired effect in plant metabolism primarily due to the built-in metabolic redundancy circumventing the imposed genetic changes [273, 274]. This necessitates the development of genome-wide comprehensive metabolic reconstructions capable of taking account of the complete inventory of metabolic transformations of a given plant organism.

Genome-scale metabolic reconstructions are available for an increasing number of organisms [275, 276]. At least 40 bacterial, 2 archaeal and 15 eukaryotic reconstructions are available to-date [272, 275, 277, 278] while many others are under development. Recently Poolman et al (2009) and Dal’Molin et al (2010) independently constructed the first two genome-scale metabolic reconstructions for a plant organism (i.e., *Arabidopsis thaliana*). The model by Dal’Molin *et al* identifies the set of essential reactions, accounts for the classical photorespiratory cycle and highlights the significant differences between photosynthetic and non-photosynthetic metabolism. The model by Poolman *et al* includes ATP demand constraints for biomass production and maintenance and suggests strategies for the construction of metabolic modules as a consequence of variation in ATP requirement. Both models make a significant step forward towards assessing the metabolic capabilities of plants establishing production routes for key biomass precursors and major pathways of Arabidopsis primary metabolism. In addition, two recent efforts involved the reconstruction of plant models with an emphasis on specific physiological conditions or tissue types [279, 280]. Model C4GEM [280] focused on C4 plants such as maize, sugarcane and sorghum and investigated flux distributions in mesophyll and bundle sheath cells during C4 photosynthesis. Grafahrend-Belau *et al* developed a metabolic network of only primary metabolism in barley seeds and studied grain yield and metabolic fluxes under a variety of oxygen availability scenarios and genetic manipulations [279]. Pilalis *et al*. reconstructed a multi-compartmental model of the
central metabolism of *Brassica napus* (Rapeseed) and simulated seed growth during the stage of oil accumulation and subsequently studied network properties of seed metabolism via Flux Balance Analysis, Principal Component Analysis and reaction deletion studies [281].

In this chapter, we describe the construction of a genome-scale in silico model of maize metabolism (i.e., *Zea mays iRS1563*). This is, to the best of our knowledge, the first attempt of globally characterizing the metabolic capabilities (both primary and secondary metabolism) using a compartmentalized photosynthetic model of an important crop and energy plant species. The development of a genome-scale model for maize is a significant challenge due to its genome size which is 14 times larger [270] than that of *Arabidopsis thaliana* (157 million base pairs) [282]. The constructed model contains 1,563 genes and 1,825 metabolites participating in 1,985 reactions from both primary and secondary metabolism of maize. For 42% of the reaction entries direct literature evidence in addition to homology criteria for their inclusion to the model was identified. We found that as many as 676 reactions and 441 metabolites are unique to *Zea mays iRS1563* in comparison to the AraGEM model by Dal’Molin *et al.* We chose the AraGEM model as a basis of comparisons as at the onset of this study it was the most comprehensive genome-scale compartmentalized model of a plant species capable of recapitulating basic plant physiological states. In order to deduce the genuine differences between maize and Arabidopsis irrespective of annotation chronology we also reconstructed an up-to-date model of Arabidopsis, *A. thaliana iRS1597*. *A. thaliana iRS1597* contains 1597 genes, 1798 reactions and 1820 metabolites. In comparison to *A. thaliana iRS1597*, *Zea mays iRS1563* has 445 new reactions and 369 new metabolites. Notably, 893 reactions and 674 metabolites are included in *Zea mays iRS1563* that are absent from the maize C4GEM model. All reactions present in *Zea mays iRS1563* are elementally and charged balanced and localized into six compartments including cytoplasm, mitochondrion, plastid, peroxisome, vacuole and extracellular space. Provisions for accounting that photosynthesis in maize (i.e., a C4 plant) occurs in two separate cell types (i.e., mesophyll cell and bundle sheath cell) are included in the model. GPR associations are
delineated from the available functional annotation information and homology prediction accounting for monofunctional, multifunctional and multimeric proteins, isozymes and protein complexes. A biomass equation is established that quantifies the relative abundance of different constituents of dry plant cell biomass. Biomass production under three different physiological states (i.e., photosynthesis, photorespiration and respiration) is demonstrated and the model is tested against experimental data for two naturally occurring maize mutants (i.e., bm1 and bm3).

4.2 Results

The metabolic model reconstruction process follows three major steps: (1) Reconstruction of draft model via automated homology searches for the identification of native biotransformations; (2) Generation of a computations-ready model after defining biomass equation and system boundary and establishing GPR; (3) Model refinement via GapFind and GapFill [39] to unblock biomass precursors as well as reconnect unreachable metabolites. Upon construction of the model, key features such as physiological constraints, network connectivity, light reactions, carbon fixation and secondary metabolism and uniqueness compared to AraGEM and maize C4GEM are described. In addition, model predictions are contrasted against experimental observations.

4.2.1 Construction of Auto & Draft models

The B73 maize genome [270] has 32,540 genes and 53,764 transcripts in the Filtered Gene Set (FGS). Out of 32,540 genes, 30,599 (93%) are evidence-based [283], while the remaining 2,141 (7%) are predicted by the Fgenesh program [284]. 13,726 genes (42% of total) do not have any functional annotation information or are identified as proteins with no or hypothetical/putative functions. Of the remainder, 1,361 (7%) genes encode proteins that do not participate in specific metabolic transformations but rather are
involved in transcription, signal transduction, DNA repair, DNA binding, DNA/RNA polymerization, protein folding and adhesion. Because the B73 maize genome is not completely annotated we first established Gene-Protein-Reaction (GPR) mappings for the AraGEM genome-scale model of *A. thaliana* [272] to be used as a proxy. Using these GPRs as a point of comparison we next identified Arabidopsis gene orthologs in maize and transferred the corresponding GPRs via the AUTOGRAPH method [285]. This step was followed by annotation of the remainder maize genes by bidirectional protein BLAST (i.e., BLASTp) searches against the NCBI non-redundant (nr) database. Out of a total of 1,567 metabolic or transport reactions of AraGEM, GPRs were established for 1,254 reactions via 1,467 genes and 653 enzymes by making use of information from several online databases such as AraCyc, KEGG, Uniprot and Brenda. Bidirectional BLASTp searches for each one of the 1,467 genes included in AraGEM model were carried out against the B73 maize genome using a stringent cutoff value of $10^{-30}$. This fully automated process generated an initial model, termed as ‘Automodel’, containing 946 genes and 1,365 unique metabolites participating in 1,186 reactions (see Table 4-1) exclusively derived from AraGEM. Out of 1,186 reactions, 32 are inter-organelle transport reactions for which homologs were found in maize.

Genes not included in the automodel were scrutinized further by comparing them against the NCBI non-redundant protein database using the same BLASTp cut-off. This increased the model size to 1,485 genes and 1,703 unique metabolites involved in 1,667 reactions by pulling functionalities absent in AraGEM. This is referred to as the ‘Draft model’ (see Table 4-1). As described in Table 4-2, orthologous genes were found in *Oryza Sativa* (Rice), *Arabidopsis thaliana* (Arabidopsis), *Sorghum bicolor* (Sorghum) and less frequently in other plant species such as wheat, tobacco, spinach, soya bean, etc. Notably, 802 orthologous genes from *A. thaliana* were added in the model *Zea mays iRS1563* that were absent from AraGEM primarily due to recent annotation updates. Reactions associated with these genes were subsequently extracted from on-line databases such as KEGG and BRENDA. Table 4-2 shows the total number of reactions as well as the number of new reactions included in the draft model. Seven reactions
having KEGG reaction IDs R00379, R00381, R06023, R06049, R06082, R06138 and R06209 were excluded since they involve generic groups and were not elementally fully defined. Figure 4-1 shows the distribution of the newly added reactions in the draft model based on their orthologous gene of origin.

![Figure 4-1: Species origin of newly added reactions in the Draft model.](image)

### 4.2.2 Generation of computations-ready model

A computations-ready model requires a fully characterized biomass equation, assignment of metabolites to reactions, establishment of GPR associations, localization of reactions in compartment(s), and inclusion of intra- and extracellular transport reactions [36].

(i) **Establishing a fully characterized biomass equation**: A biomass equation that drains all necessary precursors present in maize was derived (see Table 4-3). We used the
Table 4-1: Model size after each reconstruction step

<table>
<thead>
<tr>
<th></th>
<th>Auto model</th>
<th>Draft model</th>
<th>Functional model</th>
<th>Final model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Included genes</strong></td>
<td>946</td>
<td>1,485</td>
<td>1,552</td>
<td>1,563</td>
</tr>
<tr>
<td><strong>Proteins</strong></td>
<td>472</td>
<td>714</td>
<td>774</td>
<td>876</td>
</tr>
<tr>
<td>Single functional proteins</td>
<td>178</td>
<td>322</td>
<td>381</td>
<td>463</td>
</tr>
<tr>
<td>Multifunctional proteins</td>
<td>92</td>
<td>150</td>
<td>153</td>
<td>170</td>
</tr>
<tr>
<td>Protein complexes</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Isozymes</td>
<td>21</td>
<td>36</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>Multimeric proteins</td>
<td>87</td>
<td>140</td>
<td>148</td>
<td>148</td>
</tr>
<tr>
<td>Othersa</td>
<td>94</td>
<td>62</td>
<td>62</td>
<td>55</td>
</tr>
<tr>
<td><strong>Reactions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolic reactions</td>
<td>1,154</td>
<td>1,635</td>
<td>1,739</td>
<td>1,900</td>
</tr>
<tr>
<td>Transport reactions</td>
<td>32</td>
<td>32</td>
<td>67</td>
<td>70</td>
</tr>
<tr>
<td><strong>GPR associations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene associated (metabolic/transport)</td>
<td>1,100</td>
<td>1,581</td>
<td>1,635</td>
<td>1,668</td>
</tr>
<tr>
<td>Nonenzyme associated (metabolic/transport)</td>
<td>86</td>
<td>86</td>
<td>86</td>
<td>86</td>
</tr>
<tr>
<td>Spontaneousb</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>41</td>
</tr>
<tr>
<td>Nongene associated (metabolic/transport)</td>
<td>0</td>
<td>0</td>
<td>78</td>
<td>175</td>
</tr>
<tr>
<td>Exchange reactions</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td><strong>Metabolitesc</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytoplasmic</td>
<td>1,309</td>
<td>1,643</td>
<td>1,689</td>
<td>1,744</td>
</tr>
<tr>
<td>Plastidic</td>
<td>91</td>
<td>102</td>
<td>114</td>
<td>115</td>
</tr>
<tr>
<td>Peroxisomic</td>
<td>67</td>
<td>69</td>
<td>92</td>
<td>93</td>
</tr>
<tr>
<td>Mitochondrial</td>
<td>60</td>
<td>82</td>
<td>86</td>
<td>86</td>
</tr>
<tr>
<td>Vacuolic</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Extracellular</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

aOthers include proteins involve in complex relationships, e.g. multiple proteins act as protein complex which is one of the isozymes for any specific reaction.

bSpontaneous reactions are those without any enzyme as well as gene association.

cUnique metabolites irrespective of their compartmental location.

biomass composition of young and vegetative maize plants as measured by Penningd et al. and expressed on a dry weight basis [286]. The amino acid and lignin composition were derived based on the data from [287, 288]. The composition of hemicellulose was approximated using data for Orchard Grass [289], another monocot grass species, as no
corresponding information was found for maize. Based on these compositions we also defined aggregate reactions such as ‘Amino acid synthesis’, ‘Protein synthesis’, ‘Carbohydrate synthesis’, ‘Hemicellulose synthesis’, ‘Lignin synthesis’, ‘Lipid synthesis’, ‘Material synthesis’, ‘Nitrogenous compound synthesis’, ‘Nucleic acid synthesis’ and ‘Organic acid synthesis’ to produce necessary biomass precursors (i.e., amino acids, protein, carbohydrates, hemicellulose, lignin, lipids, materials, nitrogenous compounds, nucleic acids and organic acids respectively). The biomass equation also contains a non-growth associated ATP maintenance as in the latest Arabidopsis model AraGEM [272].

Table 4-2: Maize gene annotation via bidirectional BLASTp homology searches against NCBI non-redundant protein database

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of orthologs</th>
<th>Number of associated reactions</th>
<th>Number of newly added reactions in draft model</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Oryza Sativa</em> (Rice)</td>
<td>4,109</td>
<td>312</td>
<td>145</td>
</tr>
<tr>
<td>Other plant species</td>
<td>833</td>
<td>214</td>
<td>185</td>
</tr>
<tr>
<td><em>Arabidopsis Thaliana</em></td>
<td>802</td>
<td>258</td>
<td>193</td>
</tr>
<tr>
<td><em>(Arabidopsis)</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Sorghum Bicolor</em> (sorghum)</td>
<td>47</td>
<td>20</td>
<td>11</td>
</tr>
</tbody>
</table>

(ii) *Assignments of genes, reactions, metabolites and compartments.* All metabolic and inter-organelle transport reactions in the draft model have full gene associations. During this step all reactions were elementally balanced and metabolites were assigned appropriate protonation states corresponding to a physiological pH of 7.2. We included an additional 86 reactions to the model without enzyme association information based on direct literature evidence [272]. For example, reactions with KEGG IDs R08053, R08054 and R08055 involved in chlorophyll metabolism are included in the model. Reaction localization information for maize can in some cases be found in database PPDB (a plant proteome database of maize and Arabidopsis) [290]. Because only limited reaction
localization information exists for maize, we adopted the compartment or organelle reaction location of the corresponding orthologous gene/enzyme in Arabidopsis using the Arabidopsis Subcellular Database, SUBA [291] and also PPDB [290]. As in AraGEM, reactions for which no such information is available we assumed that they are present only in the cytoplasm.

(iii) Identification of system boundary. The entire reaction network (i.e., system boundary) was distributed across five different intracellular organelles enveloped by the cytoplasmic membrane. Exchange reactions were added in the model to ensure that gaseous metabolites (i.e., carbon dioxide and oxygen), inorganic nutrient metabolites (i.e., nitrate, ammonia, hydrogen sulfide, sulfate, phosphate, potassium and chloride), sugar metabolites (i.e., glucose, fructose, maltose and sucrose), water and photons could enter and leave the system whenever necessary depending on the physiological state. As shown in Table 4-4, constraints on these exchange reactions as well as reactions involved with enzyme RuBisCO (Ribulose-1, 5-bisphosphate carboxylase oxygenase) were established to define three different physiological states (i.e., photosynthesis, photorespiration and respiration) by allowing the selective uptake/release of certain metabolites. Even though photorespiration is limited in C4 plants (i.e., maize, sorghum, etc.), literature evidence [292-294] alludes that it is still present. Therefore, we made sure that the model is capable of simulating this condition.
Table 4-3: Biomass component list in iRS1563

<table>
<thead>
<tr>
<th>Major components</th>
<th>Protein</th>
<th>Carbohydrates</th>
<th>Lipids</th>
<th>Ions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nitrogenous</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>compounds</td>
<td>L-alanine</td>
<td>ribose</td>
<td>glyceroltrilmitate</td>
<td>potassium</td>
</tr>
<tr>
<td>Carbohydrates</td>
<td>L-arginine</td>
<td>glucose</td>
<td>glyceroltristearate</td>
<td>chloride</td>
</tr>
<tr>
<td>Lipids</td>
<td>L-aspartic acid</td>
<td>fructose</td>
<td>glyceroltrioleate</td>
<td>RNA</td>
</tr>
<tr>
<td>Lignin</td>
<td>L-cystine</td>
<td>mannose</td>
<td>glyceroltrilinolate</td>
<td>ATP</td>
</tr>
<tr>
<td>Organic acids</td>
<td>L-glutamic acid</td>
<td>galactose</td>
<td></td>
<td>GTP</td>
</tr>
<tr>
<td>Ions</td>
<td>L-glycine</td>
<td>sucrose</td>
<td></td>
<td>CTP</td>
</tr>
<tr>
<td>Nitrogenous</td>
<td>L-histidine</td>
<td>cellulose</td>
<td></td>
<td>UTP</td>
</tr>
<tr>
<td>compounds</td>
<td>L-isoleucine</td>
<td>hemicellulose</td>
<td></td>
<td></td>
</tr>
<tr>
<td>amino acids</td>
<td>L-leucine</td>
<td>pectin</td>
<td>coniferyl alcohol</td>
<td>DNA</td>
</tr>
<tr>
<td>protein</td>
<td>L-lysine</td>
<td></td>
<td>sinapyl alcohol</td>
<td>dATP</td>
</tr>
<tr>
<td>nucleic acids</td>
<td>L-methionine</td>
<td></td>
<td></td>
<td>dGTP</td>
</tr>
<tr>
<td></td>
<td>L-phenylalanine</td>
<td></td>
<td></td>
<td>dCTP</td>
</tr>
<tr>
<td></td>
<td>L-proline</td>
<td></td>
<td></td>
<td>dUTP</td>
</tr>
<tr>
<td></td>
<td>L-serine</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>L-threonine</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>L-tryptophan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>L-tyrosine</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>L-valine</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Hemicellulose</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>arabinose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>xylose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mannose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>galactose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>glucose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>uronic acids</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Organic acids</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>oxalic acid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>glyoxylic acid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>oxalo-acetic acid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Malic acid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Citric acid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>aconitic acid</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The stoichiometric matrix of the draft model (see Table 4-1) contains 1,901 rows (i.e., total metabolites after taking account of their compartmental appearance) and 1,682 columns (i.e., metabolic reactions, inter-organelle transport reactions and exchange reactions). 970 reactions have one-to-one GPR associations whereas 712 map to more than one gene. 532 reactions map to both isozymes and protein complexes while 4 of them map to only protein complexes, 36 to only isozymes, and 140 to only multimeric proteins.
Table 4-4: Definition of three different physiological states

<table>
<thead>
<tr>
<th>Constraints</th>
<th>Photosynthesis (PS)</th>
<th>Photorespiration (PR)</th>
<th>Respiration (R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO₂ transport</td>
<td>Uptake</td>
<td>Uptake</td>
<td>Release</td>
</tr>
<tr>
<td>Sucrose transport</td>
<td>Disabled</td>
<td>Disabled</td>
<td>Uptake</td>
</tr>
<tr>
<td>Photon transport</td>
<td>Uptake</td>
<td>Uptake</td>
<td>Disabled</td>
</tr>
<tr>
<td>H₂O transport</td>
<td>Uptake</td>
<td>Uptake</td>
<td>Uptake</td>
</tr>
<tr>
<td>Inorganic nutrient transport</td>
<td>Uptake</td>
<td>Uptake</td>
<td>Uptake</td>
</tr>
<tr>
<td>O₂ transport</td>
<td>Release</td>
<td>Unconstrained</td>
<td>Uptake</td>
</tr>
<tr>
<td>RUBISCO: EC 4.1.1.39</td>
<td>Carboxylation</td>
<td>Carboxylation: Oxygenation = 3:1</td>
<td>Both disabled</td>
</tr>
</tbody>
</table>

4.2.3 Network connectivity analysis and restoration

The draft metabolic model inherently contained gaps, unreachable metabolites, omitted transport mechanisms and missing biomass components. We used the procedures termed GapFind and GapFill [295] to correct for these pathologies. We first concentrated on resolving problems with the participation of components in the biomass equation followed by network connectivity.

We found that 723 out of the 1,683 reactions in the draft model could not carry any flux (i.e., blocked reactions) under any of the relevant three physiological states (e.g. photosynthesis (PS), photorespiration (PR) and respiration (R)). As a result, these blocked reactions prevented the formation of some of biomass precursors. GapFind [295] revealed that only 21 out of 64 biomass components could be synthesized using the draft model. GapFill [295] was applied for bridging the gaps through the addition of metabolic and inter-organelle transport reactions and the relaxing of irreversible of existing reactions in the model. GapFill suggested the addition of 94 metabolic and 35 inter-organelle transport reactions in the model to unblock the production of all 64 biomass components. These putative additions to the model were tested by performing an additional round of BLASTp searches for the corresponding genes against the maize genome. We found that 54 (out of 93) metabolic reactions could be assigned to maize
gene(s) if the expectation value cut-off for BLASTp was lowered to $10^{-5}$. In light of the critical need of restoring biomass formation the less stringent cut-off for inclusion was accepted for these genes. Addition of these reactions ensured the production of biomass under all relevant physiological states validating the use of the term ‘Functional’ for the updated model (see Table 4-1).

Table 4-5: Restoration of network connectivity using GapFill [36]

<table>
<thead>
<tr>
<th>Number of metabolites</th>
<th>Number of blocked metabolites: before applying GapFill</th>
<th>Number of blocked metabolites: after applying GapFill</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytosolic (1744)</td>
<td>680</td>
<td>382</td>
</tr>
<tr>
<td>Plastidic (115)</td>
<td>28</td>
<td>11</td>
</tr>
<tr>
<td>Peroxisomic (93)</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Mitochondrial (86)</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Upon ensuring biomass formation GapFind was also applied to assess network connectivity and 715 blocked metabolites were found in the functional model. By applying GapFill connectivity of 322 (45%) blocked metabolites was restored through the addition of 159 metabolic and 3 inter-organelle transport reactions. Table 4-5 shows the distribution of blocked metabolites into four intracellular organelles before and after applying GapFill. BLASTp searches allowed us to assign 31 (20% of GapFill suggestions) metabolic reactions with specific maize genes. Biological evidence of the occurrence of such additional reactions in maize or other plant species was sought whenever possible. For example, as shown in Figure 4-2 phenylacetaldehyde appears to be a “no-consumption” [295] metabolite in the functional model as no reaction can consume it. Using GapFill we found a homolog in maize (i.e., BLASTp score of $10^{-24}$) and also literature evidence [296] that *Arabidopsis thaliana* has a aldehyde dehydrogenase activity that catalyzes the conversion of phenylacetaldehyde to phenylacetic acid. Hence, by adding this chemical transformation to *Zea mays* iRS1563 a
consumption pathway for phenylacetaldehyde is established. After adding these reactions to the functional model and following charge and elemental balancing and GPR association checking the ‘Final’ *Zea mays* iRS1563 model (see Table 4-1) is derived.

![Diagram showing the consumption pathway for phenylacetaldehyde](image)

**Figure 4-2:** Example of connectivity restoration for phenylacetaldehyde.

### 4.2.4 *Zea mays* iRS1563 model

The *Zea mays* iRS1563 metabolic reconstruction contains 1,825 unique metabolites and 1,985 reactions associated with 1,563 genes and 876 proteins. Of these reactions 1,898 are metabolic reactions, 70 are inter-organelle transport reactions and 15 are exchange reactions between intra- and extracellular environments. GPR associations are established for all entries (see Table 4-1). Notably, we identified that the fraction of multifunctional proteins (19% of the total number of proteins) in *Zea mays* iRS1563 is similar to the ratio found in *E. coli* [297]. *Zea mays* iRS1563 accounts for the metabolic functions for all three physiological states. Photosynthetic as well as photorespiration metabolism was modelled by including light mediated ATP and NADPH production via separate charged balanced reactions in the electron transfer system of the thylakoid membrane [103].
Furthermore, the ratio of fluxes for the carboxylation and oxidation reactions associated with enzyme RuBisCO was kept at 1:0 thus ensuring complete carbon fixation during photosynthesis. This ratio was shifted to 3:1 during photorespiration to model simultaneous carbon fixation and oxidation [298]. Because sucrose is the main growth substrate during respiration for higher plants [299], the aforementioned reactions were inactivated and the exchange reaction for sucrose uptake was activated. Under all these three conditions, inorganic nutrients required for plant growth, e.g. sulfate, nitrate, ammonia, hydrogen sulfide, phosphate, potassium and chloride, were allowed to be freely taken up from the environment via extracellular exchange reactions.

The participation of *Zea mays* iRS1563 metabolites across different compartments is shown in Figure 4-3. The five intracellular organelles differ notably in terms of mutual connectivity, metabolite uniqueness and number of metabolites. As shown in Figure 4-3a, approximately 90% of these metabolites are unique to cytoplasm. In addition, cytoplasm contains all metabolites shared between any two organelles because any metabolite needs to be transported through cytoplasm in order to be exchanged between organelles.
Figure 4-3: Distribution of metabolites based on their number of appearance in different organelles.

(a) cytoplasmic *Zea mays iRS1563* metabolites in cytoplasm and other organelles, and,
(b) non-cytoplasmic *Zea mays iRS1563* metabolite-organelle participation.

Among the remaining metabolites, cytoplasm shares the highest number with the plastid (i.e., 63) where photosynthesis and photorespiration occur. It also shares a significant number of metabolites with mitochondrion (i.e., 27) and peroxisome (i.e., 22) that are involved in energy production and fatty acid biosynthesis, respectively. Figure 4-3b
shows the distribution of other non-cytoplasmic *Zea mays iRS1563* metabolites in terms of how many organelles they participate.

### 4.2.5 Light reactions, carbon fixation and secondary metabolism

In plants photosynthesis reactions include light dependent and light independent or carbon fixation reactions [153]. *Zea mays iRS1563* includes charged balanced light reactions culled from a number of literature sources [103, 300-302]. The overall photosynthesis reaction cascade produces two NADPH, three ATP and one O$_2$ whenever nine photons are absorbed and fourteen H$^+$ are transferred via the electron-transport system. This defines the following overall balance equations:

$$
12 \text{H}^+[\text{c}] + 2 \text{H}_2\text{O}[\text{c}] + 2 \text{NADP}[\text{c}] + 9 \text{hv}\text{i}[\text{c}] \rightarrow 14 \text{H}^+[\text{p}] + 2 \text{NADPH}[\text{c}] + \text{O}_2[\text{c}] + 9 \text{hv}\text{o}[\text{c}]
$$

$$
3 \text{ADP}[\text{c}] + 14 \text{H}^+[\text{p}] + 3 \text{Pi}[\text{c}] \rightarrow 3 \text{ATP}[\text{c}] + 14 \text{H}^+[\text{c}]
$$

Here, [c] and [p] represent cytoplasm and plastid and hv$i$ and hv$o$ signify input and output photons respectively. Carbon fixation in maize (C$_4$ plant) is more complex compared to Arabidopsis or other C$_3$ plants [153]. *Zea mays iRS1563* captures these differences by accounting for (i) direct carboxylation of phosphoenol pyruvate and CO$_2$ fixation to form C$_4$ acids such as oxaloacetic acid [ATP: oxaloacetate carboxy-lyase (ocl)] and malic acid [Oxaloacetate: NADPH hydrogenase (oha)] in mesophyll cells, (ii) transport of malic acid from mesophyll cell to bundle-sheath cells, (iii) decarboxylation of malic acid [Malate:NADP$^+$ oxidoreductase (mor)] in bundle-sheath cells to produce pyruvic acid and CO$_2$, which enters the Calvin cycle, (iv) transport of pyruvic acid from bundle-sheath cells to mesophyll cells, and (v) production of phosphoenol pyruvic (i.e., C$_3$) acid [ATP:pyruvate,phosphate phosphotransferase (ppt)] from pyruvic acid [153]. Figure 4-4, pictorially shows the localization of reactions and organelles between mesophyll and bundle sheath cells. In addition, to differences in carbon fixation reactions, the peroxisome activity is primarily present in bundle-sheath cells and largely
absent from mesophyll cells [303]. Based on this localization information a standalone metabolic model can be developed for the photosynthetic tissue of maize. Because RuBisCO that operates in the Calvin cycle cannot come in direct contact with atmospheric oxygen during day time (see Figure 4-4), photorespiration is restricted providing an advantage for survival in hot and arid environments for maize and other C₄ plants. This comes at the expense of higher (ATP) requirements as C₄ carbon fixation involves additional steps [153].

In addition to photosynthesis, secondary metabolism plays a key role in the physiology of maize. For example, phenylpropanoid metabolism produces monolignols (i.e., p-coumaroyl alcohol, coniferyl alcohol and sinapyl alcohol) that are used in the generation of three major lignin subunits H-lignin, G-lignin and S-lignin, respectively [304].
Figure 4-4: Compartment and localization information for *Zea mays* iRS1563.

Mitochondrion and vacuole compartments are present in both cell types whereas peroxisome is only present in bundle-sheath cell [40]. Plastidic reactions are distributed between mesophyll and bundle-sheath cells.

Many of these enzymes such as hydroxycinnamoyl transferase (HCT), ferulate 5-hydroxylase (F5H) and caffeic acid 3-O-methyltransferase (COMT) along with their associated reactions are unique to C₄ plants and are not present in the lignin biosynthesis pathways of *A. thaliana* [304]. HCT is involved in the early stages of lignin biosynthesis by controlling the flux from *p*-coumaroyl-CoA towards caffeoyl-CoA while F5H and COMT regulate fluxes from coniferaldehyde and coniferyl alcohol to sinapaldehyde and sinapyl alcohol, respectively [304]. *Zea mays* iRS1563 contains all these enzymes and associated reactions thus providing a comprehensive lignin biosynthesis pathway for a C₄ plant.

In addition to phenylpropanoid metabolism, *Zea mays* iRS1563 provides a detailed description of flavonoid biosynthesis pathways. Flavonoids are pigments occurring in plant as secondary metabolites and mostly function in the recruitment of pollinators and/or seed dispersers [305]. For example, maize is known to produce 3-deoxyanthocyanins, which are a specialized class of flavonoids [306, 307]. *Zea mays* iRS1563 contains the dihydroflavonol 4-reductase (DFR) enzyme that catalyzes the reaction for flavan-4-ols biosynthesis that channels flux towards 3-deoxyanthocyanins production [307]. The model also accounts for isoflavone 7-O-glucosyltransferase (IF7GT) and associated reactions that are involved in the production of necessary intermediates for pterocarpin phytoalexin conjugates such as medicarpin 3-O-glucoside-6’-O-malonate (MeGM) and maackain 3-O-glucoside-6’-O-malonate (MaGM) involved in plant defense against fungal elicitation [308].
4.2.6 Comparing *Zea mays* iRS1563 with *Arabidopsis thaliana* and maize C4GEM models

Figure 4-5a compares the total number of genes, reactions and metabolites between *Zea mays* iRS1563 and the *A. thaliana* AraGEM genome-scale-models [272]. Approximately, only 61% of genes in *Zea mays* iRS1563 are present in AraGEM. This yields a surprisingly low degree of matching between these two models of 64% and 76%, respectively in terms of reactions and metabolites. In the interest of elucidating the true differences between maize and Arabidopsis irrespective of annotation chronology we constructed a more up-to-date genome-scale model for Arabidopsis by appending onto AraGEM newly annotated genes as well as full GPR annotations. We refer to this updated model containing 1,597 genes, 1,798 reactions and 1,820 metabolites as *A. thaliana* iRS1597. The newly added 228 reactions (absent from AraGEM) are involved in various pathways in primary (i.e., glycolysis, TCA, fatty acid and amino acid biosynthesis, starch and sucrose metabolism) and secondary (i.e., biosynthesis of steroid, ubiquionone, streptomycin, thiamin, riboflavin, terpenoid, brassinosteroid, phenylpropanoid, etc.) metabolism of Arabidopsis.
A direct comparison of *Zea mays* iRS1563 with *A. thaliana* iRS1597 reveals, as expected, an increased degree of matching of 72%, 76% and 80% in terms of genes, reactions and metabolites, respectively (see Figure 4-5b). We find that 445 reactions are unique to maize with no counterpart in *A. thaliana*. Secondary plant metabolism including
flavonoid, mono- and diterpenoid, brassinosteroid, phenylpropanoid, anthocyanin, zeatin biosynthesis, riboflavin and caffeine metabolism account for 185 of the maize-specific reactions. In addition, a variety of primary metabolism reactions dispersed throughout central metabolism, photosynthesis, amino acid and fatty acid biosynthesis account for the remaining 260 reactions. This comparison implies that about one third of the differences between Zea mays iRS1563 and AraGEM are caused by the incompleteness of AraGEM model especially in terms of secondary metabolism while the remaining two third reflect genuine differences between C₃ (i.e., Arabidopsis) and C₄ (i.e., maize) plant metabolism.

Figure 4-5c shows a similar comparison between Zea mays iRS1563 and maize C4GEM genome-scale-models. Degrees of matching between these two models are 39%, 53% and 63% in terms of genes, reactions and metabolites, respectively. This surprisingly low degree of matching is caused primarily due to the fact that maize C4GEM includes only metabolites and reactions in leaves during photosynthesis. Therefore, there are 893 reactions in Zea mays iRS1563 absent from maize C4GEM. 343 of these reactions describe secondary plant metabolism such as brassinosteroid, phenylpropanoid, carotenoid, flavonoid, mono- and diterpenoid, and glucosinolate metabolism. The remaining 550 reactions are found in a wide range of primary metabolism pathways such as central metabolism, photosynthesis, benzoate degradation, starch and sucrose metabolism, lipid metabolism, nitrogen metabolism amino acid and fatty acid biosynthesis. Conversely, 116 (out of 149) new reactions in maize C4GEM have untraceable EC numbers and gene loci.

4.2.7 Zea mays iRS 1563 model testing

Zea mays iRS1563 allows for the production of biomass under all three different physiological states. Due to limited photorespiration C4 plants usually have higher
photosynthetic efficiency [153]. Under higher light intensity and photosynthetic condition, *Zea mays* iRS1563 produces 0.0008 mole biomass/mole CO2 whereas *A. thaliana* iRS1597 yields 0.0006 mole biomass/mole CO2. Thus, the model predictions match with findings reported in literature [153]. We also investigated the model’s ability to predict the effect of suppressing genes in the lignin biosynthesis pathway observed in naturally occurring *brown midrib* (*bm*) maize mutants (i.e., *bm1*, *bm2*, *bm3* and *bm4*) [304, 309-311]. These maize mutants are Mendelian recessives that are characterized by brown vascular tissue in leaves and stems due to a changed lignin content and/or composition [312]. The specific genetic background for two of these mutants (*bm1* and *bm3*) was elucidated based on the analysis of cell wall composition [311]. Mutants *bm1* and *bm3* were found to have disrupted enzymatic activity for cinnamyl alcohol dehydrogenase (CAD) and caffeic acid 3-O-methyltransferase (COMT). Both of these enzymes are involved in the last stages of the monolignol pathway [311] that controls lignin synthesis and composition (i.e., the ratio of three major subunits, H-lignin, G-lignin and S-lignin) [313].

**Table 4-6:** Change in content of cell wall components in *bm1* and *bm3* Maize mutants

<table>
<thead>
<tr>
<th></th>
<th>Model findings vs Experimental observations</th>
<th><em>bm1</em> mutant</th>
<th><em>bm3</em> mutant</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-lignin</td>
<td>↓ / =</td>
<td>↓ / ↓</td>
<td></td>
</tr>
<tr>
<td>G-lignin</td>
<td>↓ / ↓</td>
<td>↓ / ↓</td>
<td></td>
</tr>
<tr>
<td>S-lignin</td>
<td>↓ / ↓</td>
<td>↓ / ↓</td>
<td></td>
</tr>
<tr>
<td>Total lignin</td>
<td>↓ / ↓</td>
<td>↓ / ↓</td>
<td></td>
</tr>
<tr>
<td>S-lignin/G-lignin ratio</td>
<td>= / =</td>
<td>= / =</td>
<td></td>
</tr>
<tr>
<td>Glucose</td>
<td>↓ / ↓</td>
<td>↓ / ↑</td>
<td></td>
</tr>
<tr>
<td>Mannose</td>
<td>↓ / ↓</td>
<td>↓ / ↓</td>
<td></td>
</tr>
<tr>
<td>Arabinose</td>
<td>↓ / ↓</td>
<td>↓ / ↓</td>
<td></td>
</tr>
<tr>
<td>Galactose</td>
<td>↓ / ↓</td>
<td>↓ / ↓</td>
<td></td>
</tr>
<tr>
<td>Xylose</td>
<td>↓ / ↑</td>
<td>↓ / ↑</td>
<td></td>
</tr>
<tr>
<td>Crude protein</td>
<td>-</td>
<td>↓ / ↓</td>
<td></td>
</tr>
</tbody>
</table>
We simulated mutants \textit{bm1} and \textit{bm3} using \textit{Zea mays} iRS1563 under photosynthetic conditions by restricting the flux of the reactions catalyzed by enzymes CAD and COMT to 10\% of the wild-type values. It is expected that the disruption of the activity for these genes will directly affect lignin content and composition. We were interested to see whether the \textit{Zea mays} iRS1563 metabolic model will be able to correctly propagate this disruption across the metabolic pathways and correctly predict the effect on other key metabolites. Table 4-6 contrasts experimental results by (Marita et al (2003), Vanholme et al (2008) and Sattler et al (2010)) with \textit{in silico} predictions for the maximum theoretical yield of lignins, sugars and crude protein in terms of whether they increased, decreased, or remained the same in the mutant strains. Out of 21 compared components \textit{Zea mays} iRS1563 correctly predicted the direction (or absence) of change for 17 cases.
**Figure 4-6:** Maximum theoretical yields of (a) glucose and (b) galactose for wild-type vs bm1 mutant and wild-type vs bm3 mutant, respectively.

Here the numeric values represent reaction fluxes and have the unit of mM/gDW-h.

In Figure 4-6 we highlight two cases that describe the availability of glucose and galactose to cell wall for mutants bm1 and bm3, respectively. ‘Carbohydrate synthesis’ and ‘Hemicellulose synthesis’ are aggregate reactions that describe the utilization ratios of sugar molecules such as arabinose, fructose, galactose, glucose ribose, mannose, sucrose, and xylose for the production of carbohydrate and hemicellulose present in the plant cell wall. For simplicity, we have simulated the model under the photosynthetic condition where CO2 can be uptaken with a maximum allowable rate of 1000 mM/gDW-h along with photons in excess. In Figure 4-6a, wild-type and bm1 mutant flux values for reactions involving glucose as reactant including ‘Carbohydrate synthesis’, ‘Hemicellulose synthesis’, ‘Alpha,alpha-trehalose glucohydrolase’ [R00010], ‘Sucrose glucohydrolase’ [R00801], ‘Sn-Glycerol-3-phosphate: D-glucose 6-phosphotransferase’ [R00850] and ‘Cellobiose glucohydrolase’ [R00306], are highlighted. For the wild-type case, the maximum theoretical yield of glucose is predicted to be 1.66 moles/mole of CO2 but it is reduced to 0.93 moles/moles of CO2 for the bm1 mutant. The reduced capability of the bm1 mutant to direct flux towards ‘Carbohydrate synthesis’ and ‘Hemicellulose synthesis’ implies that less glucose is available for the formation of cell wall components which is consistent with the experimental finding of Table 4-6.

Figure 4-6b contrasts the wild-type and bm3 mutant maximum theoretical yields for all reactions involving galactose including ‘Hemicellulose synthesis’, ‘ATP: D-galactose 1-phosphotransferase’ [R01092] and ‘Galactosylglycerol galactohydrolase’ [R01104], ‘3-O-alpha-D-Galactosyl-1D-my-o-inositol galactohydrolase’ [R01194] and ‘alpha-galactosidase’ [R03634]. A reduction of the maximum theoretical yield of galactose from 0.81 to 0.65 moles/mole of CO2 for the bm3 mutant is observed. In addition, the maximum theoretical yield for reaction ‘Hemicellulose synthesis’ decreases by 4-fold.
compared to wild-type in line with the experimental finding. However, the experimentally observed increase of glucose availability in mutant bm3 and xylose availability for both bm1 and bm3 mutants are in contrast with the model predictions (see Table 4-6). As reported by Guillaumie et al (2007) several gene expression levels were changed during bm1 and bm3 mutations implying that additional regulatory constraints may be needed to capture these changes.

4.3 Discussion

Maize, apart from its central role a food crop, is also a promising plant biomass target for cellulosic biofuels production. Plant cell wall cellulose, hemicellulose and lignin polymers are major contributors of plant biomass [304, 314]. Therefore, controlling the amount and composition of cell wall polymers is important in developing cellulosic maize for biofuel production. In cell wall, lignin provides rigidity by forming a matrix where cellulose and hemicellulose are imbedded via cross-linking bonds [309, 315]. This makes digestion of cellulose and hemicellulose by microbial enzymes (i.e., cellulases) difficult during dilignification, one of the critical steps in cellulosic biofuel production [316]. Many genetic modification strategies have been explored to improve maize food crop and/or biofuel characteristics. For example, cellulosic biomass yield improvements have been pursued before by altering the lignin content and composition [317, 318], genetically manipulating the cellulose biosynthetic pathway [319] and over-expressing the gene encoding phosphoenolpyruvate carboxylase (PEPC) to improve CO₂ fixation rate [320]. At the same time, grain yield enhancements have been attempted by up-regulating ADP-glucose pyrophosphorylase (AGP) that catalyzes the rate limiting step in starch synthesis [321].

Unfortunately, existing genetic engineering strategies to reduce lignin content are problematic as lignin reductions are usually achieved at the expense of plant viability and fitness [316]. It is becoming widely accepted that focusing on a single pathway at a time without quantitatively assessing the system-wide implications of the genetic disruptions
may be responsible for not preserving the agronomic properties of the plant. By accounting for both primary and some secondary metabolism pathways of maize, *Zea mays iRS1563* can be used to explore *in silico* the effect of genetic modifications aimed at plant cell wall modification and/or starch storage on the overall metabolic state of the plant (e.g., biomass precursor availability, cofactor balancing, redox state, etc.). Moving a step further, the use of computational strain optimization techniques [7, 322] can be customized for engineering plant metabolism. By taking full inventory of plant metabolism optimal gene modifications could be pursued for a variety of targets in coordination with experimental techniques. These may include (i) increase cellulose and hemicellulose production, (ii) starch yield, (iii) tolerance against biotic stress (e.g., fungal elicitation), or (iv) disruption of the production of lignin subunits (H/G/S) while enhancing the production of easily digestible lignin precursor (e.g., rosmarinic acid, coniferyl ferulate, tyramine conjugates, etc).

In this chapter, we introduced the first comprehensive genome-scale metabolic model (*Zea mays iRS1563*) for maize metabolism. The model meets (or exceeds) the quality and completeness criteria set out [323, 324] for genome-scale reconstructions. In analogy to the human genome-scale model Recon 1 [325], *Zea mays iRS1563* can be viewed as a mathematically structured database enabling systematic studies of maize metabolism. Of unique to maize reactions accounting for a fraction of secondary metabolism were delineated. As a by product of this effort a more up-to-date version of AraGEM [272] was constructed including GPR associations. Comparisons between *Zea mays iRS1563* and maize C4GEM also revealed the detail in description of primary and secondary metabolism. Model predictions of *Zea mays iRS1563* for two widely occurring maize Mendelian mutants were tested against experimental observations with very good agreement in the direction of changes. By making use of high throughput enzymatic assays, proteomic and transcriptomic data across different parts of the maize plant, *Zea mays iRS1563* could serve as the starting point for the development of tissue-specific maize models [280, 326, 327]. Furthermore, *Zea mays iRS1563* could also serve as the stepping stone for the development of genome-scale models for other important C4 plants.
such as Sorghum and switch grass.

4.4 Materials and Methods

A number of recent publications [36, 275, 323] have outlined the general steps necessary for the metabolic reconstruction process. In the following section, we highlight the specific methods used in the reconstruction of Zea mays iRS1563 and subsequent model simulations in more detail.

4.4.1 Model reconstruction

The maizesequence database [270] provided the filtered gene set (FGS) which has been generated from the working gene set upon removing pseudogenes and low confidence hypothetical models. The FGS of B73 maize genome (release 4a.53) was downloaded from maizesequence database on February 17, 2010. Once maize genes were obtained, we used sequence comparison tools [328] such as stand-alone BLAST (version 2.2.22, NIH) and BLAST+ (version 2.2.22, NIH) for performing homology comparisons. Marvin (version 5.3.3, ChemAxon Kft) was used to calculate the average micro-species charge to determine the net charge of individual metabolites at pH 7.2 assumed for all organelles. In the final step of the model reconstruction, we implemented GapFind and GapFill [295] for analyzing and subsequently restoring metabolic network connectivity.

4.4.2 Model simulations

Flux balance analysis (FBA) [122] was employed both in model validation and model testing phases. Zea mays iRS1563 was evaluated in terms of biomass production under
three standard physiological scenarios: photosynthesis, photorespiration, and respiration. Flux distributions for each one of these states were approximated using FBA:

\[
\text{Maximize } v_{\text{Biomass}}
\]

Subject to

\[
\sum_{j=1}^{m} S_{ij} v_j = 0 \forall i \in 1, \ldots, n
\]  
(1)

\[
v_{j,\min} \leq v_j \leq v_{j,\max} \forall j \in 1, \ldots, m
\]  
(2)

Here, \(S_{ij}\) is the stoichiometric coefficient of metabolite \(i\) in reaction \(j\) and \(v_j\) is the flux value of reaction \(j\). Parameters \(v_{j,\min}\) and \(v_{j,\max}\) denote the minimum and maximum allowable fluxes for reaction \(j\), respectively. As mentioned in Table 4-4, the three physiological states were represented via modifying the relevant minimum or maximum allowable fluxes and the following constraints:

\[
v_{\text{oxi}} = 0
\]  
(3)

\[
v_{\text{carboxi}} \geq 3v_{\text{oxi}}
\]  
(4)

\[
v_{\text{carboxi}} = 0
\]  
(5)

where \(v_{\text{Biomass}}\) is the flux of biomass reaction and \(v_{\text{oxi}}\) and \(v_{\text{carboxi}}\) are the fluxes of carboxylation and oxidation reactions associated with enzyme RUBISCO. For photosynthesis and photorespiration, constraints (3) and (4) were respectively included in the linear model, whereas for respiration both constraints (3) and (5) were included.

Once the model was validated, it was further tested for two maize mutants (i.e., \(bm1\) and \(bm3\)) under the photosynthetic condition. The following two constraints were included individually in the linear model to represent the mutants:

\[
v_{\text{bm1}} \leq w \times WF_{\text{bm1}}
\]  
(6)

\[
v_{\text{bm3}} \leq w \times WF_{\text{bm3}}
\]  
(7)

Here, \(w\) represents the percent of residual activity of 10%. \(v_{\text{bm1}}\) and \(v_{\text{bm3}}\) are the fluxes of reactions catalyzed by CAD and COMT, respectively and \(WF_{\text{bm1}}\) and \(WF_{\text{bm3}}\) are the corresponding wild-type flux values under the photosynthetic condition.
CPLEX solver (version 12.1, IBM ILOG) was used in the GAMS (version 23.3.3, GAMS Development Corporation) environment for implementing GapFind and GapFill [295] and solving the aforementioned optimization models. All computations were carried out on Intel Xeon E5450 Quad-Core 3.0 GH and Intel Xeon E5472 Quad-Core 3.0 GH processors that are the part of the lionxj cluster (Intel Xeon E type processors and 96 GB memory) of High Performance Computing Group of The Pennsylvania State University.
5 Chapter 5

NITROGEN USE EFFICIENCY IN MAIZE (ZEA MAYS L.): FROM “OMICS” STUDIES TO METABOLIC MODELING


5.1 Introduction

Over the last decade, it has become possible to construct complete plant genome annotations for a wide variety of model and crop species (Jackson *et al*., 2011), and to use new high-throughput tools such as the transcriptomics, proteomics and metabolomics, to unravel the processes controlling plant productivity (Fukushima *et al*., 2009; Kusano and Fukushima, 2013). These processes have been shown to be a multitude of complex networks and interdependent pathways involving many genes, proteins, enzymes and metabolites, rather than distinct linear pathways (Fernie and Stitt, 2012). As in a large number of cases, this nonlinear complexity has hindered plant metabolic manipulation experiments focused on the original agronomic target of improving water use efficiency (WUE) (Ashraf, 2010) or nitrogen use efficiency (NUE) (Pathak *et al*., 2011; McAllister *et al*., 2012). Hence, the need to associate a biological function to a gene, the corresponding translation product, and finally the synthesis of a desired metabolite has led to the development of various systems biology approaches based on integrated “omics” studies. These studies have taken advantage of an increasing number of gene expression and metabolic databases (http://www.hsls.pitt.edu/obrc/index.php?page=metabolic_pathway). In particular, a comprehensive data collection called OPTIMAS Data Warehouse (OPTIMAS-DW) that includes transcriptomes, metabolomes, ionomes, proteomes and phenomes has recently
been released to support systems biology research in maize (Colmsee et al., 2012). The resource is available at http://www.optimas-bioenergy.org/optimas_dw.

One of these systems biology approaches consists of linking genes and metabolic functions to physiological or agronomic traits through the construction of whole genome-scale metabolic models (Ruppin et al., 2010). Such metabolic models at the interface between computation, biology, genetics and agronomy, are currently being developed to advance our ability to maximize phenotypic and agronomic traits in a rational manner. Construction of such metabolic models is generally conducted using as a working basis the most prominent models built from unicellular organisms such as bacteria and yeast and expanded in a stepwise manner to make them applicable to model plants such as Arabidopsis thaliana and then to crops (Seaver et al., 2012). The ultimate goal of developing such models is to provide a new tool for predicting crop yields, that will allow the selection of crops adapted to lower inputs and to particular environmental conditions. The knowledge gained from such modeling approaches could ultimately allow the identification of key developmental and metabolic components involved in the elaboration of complex agronomic trait such as NUE or WUE (Baldazzi et al., 2012; Shachar-Hill, 2013). The identification of such components and a better understanding of their regulation should provide additional tools for developing marker-assisted selection strategies for breeders, and for exploiting the possibilities offered by genetics, including natural variability, mutagenesis and genetic manipulation (Hirel et al., 2007).

5.2 Why improve nitrogen use efficiency in a crop such as maize?

Both from an agronomic and economic point of view, the main driver for crop improvement over the last century has been yield (Conant et al., 2013). During this period, the rate of yield improvement has accelerated due primarily to the introduction of an increasingly scientific approach to plant breeding, but also through the extensive use of fertilizers (Tilman et al., 2011; Andrews and Lea, 2013). Among these fertilizers,
nitrogen (N) is a major factor in agricultural production, where it can be supplied through chemical synthesis (Andrews et al., 2013), organic rotation (Tuomisto et al., 2012), or biological N fixation (Vitousek et al., 2013). However, this extensive use of N fertilizers has caused major detrimental impacts on the diversity and functioning of non-agricultural bacterial, animal and plant ecosystems (Erisman et al., 2013; Galloway et al., 2013). In addition, fertilizer-derived N oxide emissions into the atmosphere contribute to the depletion of the ozone layer, whilst volatilised ammonia is returned as wet or dry deposition, which can cause acidification and eutrophication (Cameron et al. 2013; Fowler et al. 2013). An excellent overview of the different possible strategies to optimize the use of N fertilizers worldwide for both economic and environmental benefits has recently been published by Good and Beatty (2011). This review emphasizes that implementing the best N management practices together with crop genetic improvement adapted for each country, can substantially reduce excess N fertilizer applications without compromising crop yields.

At present, a mixture of converging global factors is putting unprecedented pressure on agricultural productivity. These factors include increasing demand for human food and animal feed in developing nations with large populations, diminishing supplies and the rising cost of fossil fuel energy that is required for fertilizer production. Since cereals such as maize, wheat and rice are the basis of most human food in the world, improving their NUE is a major challenge for a sustainable agriculture (Hirel et al., 2007; Kant et al., 2011; McAllister et al., 2012). It is therefore necessary to select or release new varieties requiring less N-based fertilizer, whilst maintaining high yields and grain quality (protein content, in particular). For this reason, several public institutions and all major seed breeding companies are investing in crop genome research, and applying molecular marker and transgenic techniques to identify genes that can be used to improve NUE further (Edgerton, 2009; Xu et al., 2012; Fischer et al., 2013).

Improving NUE is particularly relevant for maize, as large amounts of N fertilizer are required to obtain the maximum yield and for which global NUE, as with other crops, has
been estimated on average to be less than 50% (Raun and Johnson, 1999). Recent studies have demonstrated that there are large differences in maize lines and hybrids in their ability to grow and yield well on soils with low mineral nutrient availability, which depends on both N uptake efficiency (NupE) and N utilization efficiency (NutE) (Hirel and Gallais, 2011). Maize is recognised, not only as a major crop, but also as a model species that is well adapted for fundamental research, especially for understanding the genetic basis of yield performance. Many tools are available in maize such as mutant collections, a wide genetic diversity, recombinant inbred lines (RILs), straightforward transformation protocols, physiological, biochemical and “omics” data as well as its genome sequence (Hirel and Lea, 2011) and more recently genome-scale metabolic models (Saha et al., 2011).

5.3 Nitrogen use efficiency: from “omics” studies to systems biology approaches

Due to the complexity of the biological systems involved in the control of NUE at the cellular, organ or whole plant levels, the emerging research field of systems biology was developed for both model and crop species. This allowed the researcher to focus on a holistic understanding of N-regulatory networks from the genomic to agronomic traits such as biomass production or yield. Such an approach consists of taking advantage of the various transcriptome, proteome, metabolome and fluxome data sets that can be further analysed in an integrated manner through the utilization of various mathematical, bioinformatic and computational tools (Gutiérrez, 2012). Ultimately such integrated analyses, possibly combined with whole plant physiology and quantitative genetic studies, may allow the identification of key individual or common regulatory elements that are involved in the control of complex biological processes (Saito and Matsuda, 2010).
5.4 Transcriptome studies

To identify some of the regulatory and structural elements representing the physiological changes associated with NUE, several studies have been carried out to evaluate modifications in gene expression under low and high N conditions. In an increasing number of model and crop species, transcriptome studies have highlighted the complexity of the regulatory mechanisms involved in the control of leaf or root gene expression under N limiting and non-limiting conditions (Wang et al., 2003; Krapp et al., 2011; Amiour et al., 2012; Wei et al., 2013). In mutants, transcriptome studies have also revealed deficiencies in key reactions or key regulatory proteins involved in primary N metabolism (Castaings et al., 2009, Beatty et al., 2009; Wang et al., 2009; Kissen et al., 2010). Several classes of N responsive genes have been identified, including those involved in a variety of metabolic and regulatory pathways. It is hoped that such a transcriptome approach could ultimately help to identify the genes and proteins required for a N-use-efficient phenotype under different environmental soil conditions (Ruzicka et al., 2010) up to the agronomic level (Tenea et al., 2012). For example, protein kinases such as AtCIPK8 (Hu et al., 2009) or transcription factors such as NLP7 (Marchive et al., 2013) identified following whole genome/transcriptome approaches, were shown to be key players in nitrate sensing and signalling in Arabidopsis. The TOR (Target of Rapamycin) signalling pathway seems also to be involved in the regulation of N assimilation in proliferating and conductive tissues, thus playing an important role in controlling long-distance and short-distance nutrient exchanges (Robaglia et al., 2012).

Improved NUE was obtained, when OsENOD93-1, a gene encoding another N-responsive transcription factor was overexpressed in rice (Bi et al., 2009), strengthening the finding that regulatory proteins are as important as enzymes in the control of N metabolism. In line with this finding, maize Dof1 (DNA-binding with One Finger) was expressed in Arabidopsis under the control of a maize pyruvate phosphate dikinase (PPDK) promoter. The transformed plants exhibited a considerable elevation in the concentration of soluble amino acids, especially glutamine, and increased growth under low N conditions (Yanagisawa et al., 2004). When the same maize ZmDof1 gene was
expressed in rice under the control of the ubiquitin promoter, increases in photosynthesis, N assimilation and growth were detected under low N conditions (Kurai et al., 2011). The latest advances in our understanding of N signalling in plants have been reviewed by Castaings et al. (2011), highlighting the roles of transcription factors, nitrate transporters and kinases in their interactions with hormones or N containing molecules in the regulation of N assimilation.

Plants are able to use both nitrate and ammonium ions as N sources as well as various organic sources (Andrews et al., 2013). Distinct signalling pathways and transcriptome response signatures for ammonium- and nitrate-supplied Arabidopsis have been identified. The data indicated that there is an ammonium- and a nitrate-specific pattern of gene expression, as well as a general inorganic N gene response (Patterson et al., 2010). Such observations suggest that the regulation of gene expression under agronomic conditions, when the two sources of inorganic N will be present in variable proportions, depending on the type of fertilizer used, is probably more complex than that occurring in plants grown under controlled conditions on a single N source.

Considering the agronomic importance and economic value of maize worldwide, an increasing number of whole genome transcriptome approaches have also been developed to identify genome-wide transcriptional circuits in various organs and tissues during maize development (Sekhon et al., 2011; Downs et al., 2013) and particularly those related to N-responsive genes. Depending both on the duration and intensity of the N-limiting stress applied, most of the studies in maize have ended up with a portfolio of genes involved in a variety of developmental, metabolic and regulatory functions (Amiour et al., 2013; Humbert et al., 2013). In some cases, a number of these N-responsive genes were also found in dicot species, but their level of response to the N feeding conditions appeared to be largely dependent on both the genotype and the experimental conditions (Table 5-1). Nevertheless among these genes, those encoding carbonic anhydrase, which plays an important role in the delivery of CO₂ for carbon assimilation (Moroney et al., 2001) and plastidic glutamine synthetase (GS2), which
assimilates or reassimilates ammonium (Hirel and Lea, 2001) were found to be down-regulated under N-deficient conditions. Those encoding germin-like proteins which are involved in various developmental and stress responses (Bernier and Berna, 2001; Wang et al., 2013) and those encoding peroxiredoxins which are known to play a major role in controlling organelle redox metabolism (König et al., 2012) were also found to be N-responsive. The developmental stage of the plant appears to be very important in some circumstances, since genes that responded to N-limiting stress at the vegetative stage of maize, were different from those that were responsive to N at the late grain filling stage (Amiour et al., 2012). Interestingly, a number of these genes were found to respond

**Table 5-1:** Transcripts exhibiting significant increase following transfer from limiting to non-limiting N feeding conditions in different studies and across different species

<table>
<thead>
<tr>
<th>Arabidopsis ID</th>
<th>Maize ID</th>
<th>Gene annotation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>At3g01500</td>
<td>TC259341</td>
<td>CA1 (CARBONIC ANHYDRASE 1); carbonate dehydratase/zinc ion binding</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>At5g20630</td>
<td>TC259932</td>
<td>GLP3 (GERMIN-LIKE PROTEIN 3); manganese ion binding/nutrient reservoir</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>At5g35630</td>
<td>TC271006</td>
<td>GS2 (GLUTAMINE SYNTHETASE 2); glutamate-ammonia ligase</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At1g03600</td>
<td>TC216153</td>
<td>Photosystem II family protein</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>At4g09650</td>
<td>TC249593</td>
<td>ATP synthase delta chain, chloroplast, putative/H+ - transporting two-sector ATPase</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>At4g28660</td>
<td>BM381938</td>
<td>PSB28, photosystem II reaction centre W (PsbW) family protein</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>At3g11630</td>
<td>TC262220</td>
<td>2-cys peroxiredoxin, chloroplast (BA51)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At5g15350</td>
<td>BG319827</td>
<td>Plastocyanin-like domain-containing protein</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>At5g62720</td>
<td>TC265222</td>
<td>Integral membrane HPP (Human Proteome Project) family protein</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The numbers on the right side of the panel correspond to: 1 (Wang et al., 2003 in Arabidopsis); 2 (Scheible et al. 2004 in Arabidopsis); 3 (Bi et al., 2007 in Arabidopsis); 4 (Peng et al., 2007 in Arabidopsis); 5 (Krapp et al., 2011 in Arabidopsis), 6 (Cai et al., 2012 in rice); 7 (Amiour et al., 2012 in maize). ID, gene identification number. A cross (x) indicates that the gene was identified in the majority of the seven studies. A number of transcripts for ribosomal proteins were also found in the different studies but did not exactly correspond to the gene annotation in Arabidopsis and in maize.

similarly to varying N nutrition conditions in different genotypes and under both controlled or field-growth conditions. This finding led Yang et al., (2011) to propose that
a small set of N-responsive genes could be used as biomarkers to monitor the *in planta* status of maize N. A number of these genes were also found in the study of Amiour *et al.* (2012), thus strengthening the idea that they could be used as agronomic tools for both breeding purposes and for optimizing fertilizer usage.

More recently, evidence showing the importance of microRNAs (miRNAs) in the regulation of a number of abiotic stresses has rekindled the interest of a number of research groups in the epigenetic regulation of NUE and its potential use for NUE improvement (Fischer *et al*., 2013). As revealed in studies performed on maize, the occurrence of miRNA-mediated control of gene expression could represent an important biological component of NUE that has hitherto been overlooked using standard transcriptome approaches (Trevisan *et al*., 2011; Zhao *et al*., 2013). Such a putative regulatory function mediated by the action of miRNAs was highlighted by the finding that significant differences in their accumulation were observed according to the level of N nutrition, as well as their spatiotemporal expression pattern in root tissues (Trevisan *et al*., 2012; Zhao *et al*., 2012). Although the genes targeted by miRNAs had various and ubiquitous functions, encompassing a variety of developmental and metabolic processes that were not necessarily directly linked to NUE (Xu *et al*., 2011), the genetic manipulation of the expression of miRNAs could be an alternative method of improving NUE in crops (Fischer *et al*., 2013).

In order to shed light on the dynamics of transcription in response to various environmental stimuli or stresses such as N limitation, tools such as MapMan were originally developed to visualize large gene expression data sets in *Arabidopsis* in order to search for similar global responses across large numbers of microarrays (Usadel *et al*., 2005). Such a tool, that has now been adapted for maize (Usadel *et al*., 2009) and solanaceous species (Urbanczyk-Wochniak *et al*., 2006; Ling *et al*., 2013), will provide information about the response of the expression of the whole genome to N nutrition, in relation to other cellular and metabolic processes. Several software and visualization tools have been developed to interpret more easily, large “omics” data sets and also to
identify genes, gene networks and regulatory hubs that control plant growth and development. For example, VirtualPlant (Katari et al., 2010), Geneinvestigator (Hruz et al., 2008) or Cytoscape (Killcoyne et al., 2009) have been used in an increasing number of studies, aimed at identifying gene regulatory networks involved in N metabolism in both model and crop species. These visualization tools have been extensively used to decipher the relationship between N responsive gene networks and other biological processes linked to C availability (Krouk et al., 2010; McIntyre et al., 2011), external signals such as light (Krouk et al., 2009), or internal signals such as hormones (Nero et al., 2009, Krouk et al., 2010). They have also proved particularly useful for investigating the natural variation of the response of Arabidopsis to N availability (Ikram et al., 2012).

An excellent review presenting the current knowledge of the regulatory components controlling the response of Arabidopsis to N, together with the response networks corresponding to metabolic, physiological, growth and development pathways has recently been published (Gutiérrez, 2012; Canales 2014). This review article highlights the function of receptors, transcription factors and other putative signalling components of N signalling pathways, deciphered by means of the integrated systems biology approach described above.

Although much more informative than conventional transcriptome studies, such whole genome expression approaches have remained confined to deciphering regulatory circuits at the transcriptional level, since only the steady state of transcripts has been considered. Such an approach, originally developed for Arabidopsis by virtue of the wealth of information available, when transferred to crops may help in identifying key master genes involved in the control of NUE (Gutiérrez, 2012). Nevertheless, transcriptome, metabolome and even fluxome coexpression network analyses will certainly be necessary in order to enhance our knowledge of the genes and metabolic pathways linked to NUE in crops such as maize (Saito and Matsuda, 2010).
5.5 Proteome studies

Although a number of proteome databases are now available on the world wide web (Jorrín-Novó et al., 2009), in comparison to the numerous transcriptome studies, there is much less information available on the proteome concerning NUE both in model and crop species, as time-consuming and difficult techniques are required. Moreover at best, less than a thousand proteins can usually be separated by two-dimensional (2-D) gel electrophoresis and identified using either the available databases, or mass spectrometry techniques (Jorrín et al., 2007). However, with second-generation quantitative proteome techniques, the coverage of the plant cell proteome has increased considerably (Jorrín-Novó et al., 2009). Under abiotic stress conditions, proteome studies are able to provide additional information on the quantity of expressed proteins and posttranslational modifications such as phosphorylation and glycosylation that cannot be identified by only determining mRNA transcription. Analysis of the proteome has identified protein response pathways shared by different plant species, as well as pathways that are unique to a given stress (Kosovà et al., 2011). With the improvements in MS-based proteome and phosphoproteome analyses, it is now possible to explore various areas of maize biology including the impact of N-deficiency stress on the plant proteome (Facette et al., 2013; Pechanova et al., 2013). The first proteome studies performed on wheat grown under N-deficiency stress conditions showed that the concentrations of enzymes and proteins involved in C metabolism were the most strongly reduced (Bahrman et al., 2004). Later on, changes in the protein profile were also examined in the roots and shoots of maize (Prinsi et al., 2009; Amiour et al., 2012), rice (Kim et al., 2009), barley (Møller et al., 2011) and Arabidopsis (Wang et al., 2012), when the plants were grown under a low or a high N supply. Results from these studies showed that the amounts of enzyme proteins that have a pivotal role in N assimilation such as GS and in C metabolism such as phosphoenolpyruvate carboxylase (PEPC) were higher when plants were fed with nitrate, in agreement with a previous study (Sugiharto and Sugiyama, 1992). Many other proteins involved in a number of photosynthetic reactions, in maintaining the energy and redox status of the cell, and signal transduction were also shown to be N-responsive.
Such data confirms the tight relationship that exists between N and other metabolism found at the transcriptional level (Gutiérrez, 2012). In the vast majority of the proteome investigations into the response of a plant to N-limitation, there was no direct relationship with transcriptome or metabolome studies. It was therefore difficult to tell if the regulation of protein synthesis occurred at the translational or post-translational level, or if the amount of protein was correlated with the amount of corresponding mRNA. In the study of Amiour et al. (2012) on maize, no simple and direct relationship between among transcript, protein and metabolite accumulation was found. In a similar manner to wheat, this finding suggests that posttranscriptional modifications may be positively or negatively regulated by the N metabolite concentration in the plant (Bahrman et al., 2005). In addition complex and still uncharacterized network interactions are probably occurring between gene transcription and protein and metabolite accumulation (Stitt and Fernie, 2012). It is likely that advanced proteome tools will be used more widely in the analysis of signalling and developmental processes in plants, as they are in medical research (Choudhary and Mann, 2010). When integrated with other “omics” data and with information from quantitative genetics, proteomes will be able to contribute to our understanding of complex regulatory networks underlying important phenotypic traits such as yield and nutrient perception and utilization (Kaufmann et al., 2011; Verma et al., 2013).

5.6 Metabolome studies

Over the last five years an increasing number of metabolome studies have been carried out for both model and crop plants, with the aim of identifying changes in metabolite concentrations under various biotic (Balmer et al., 2013) and abiotic stresses including N deficiency (Kusano et al., 2011; Obata and Fernie, 2012). These have also been valuable in improving our understanding of the interactions between C and N metabolism (Fait et al., 2011). Such approaches have allowed the identification of new compounds that accumulate in response to a given stress, as well as those sharing a common pattern of accumulation across various stress conditions. In addition, a number of plant metabolic
databases are now available that will facilitate the development of plant systems biology approaches (Fukushima and Kusano, 2013).

Up until now, the vast majority of metabolome studies have been carried out using Arabidopsis, but more recently these have been extended to a wider range of plants including cereals such as rice and maize (Kusano et al., 2011; Lisec et al., 2011; Amiour et al., 2012; Riedelsheimer et al., 2012a). Exhaustive metabolic profiling using Gas Chromatography coupled to Mass Spectrometry (GC/MS) based separation techniques provides information for plant phenotyping and the exploitation of genetic variability (Saito and Matsuda, 2010). Liquid chromatography coupled to mass spectrometry (LC-MS) has also been used frequently for metabolome analysis (Rohrmann et al., 2011; Tohge et al., 2011). In parallel, $^1$H-NMR (Nuclear Magnetic Resonance) spectroscopy approaches have been developed, which are less sensitive but non-invasive, compared to those requiring the extraction of plant material (Kim et al., 2010). $^1$H-NMR metabolomics appears to be an attractive technique for the development of mapping approaches (Graham et al., 2009) that could support breeding for improved NUE through the establishment of metabolite databases. $^1$H-NMR has also been used successfully to improve the characterization of GS deficient mutants of maize, indicating that in addition to the glutamine-derived amino acid biosynthetic pathways, lignin biosynthesis was also altered (Broyard et al., 2009). Such techniques have also been used to demonstrate that in tobacco, the enzyme glutamate dehydrogenase (GDH) does not assimilate ammonia, even when it is overexpressed several fold (Labboun et al., 2009).

The effect of N starvation on the plant metabolic profile has been examined in a few studies using Arabidopsis as a model species (Krapp et al., 2011), or maize as a crop (Amiour et al., 2012; Schlüter et al., 2012; 2013). In all these studies it was observed that in leaves, N deprivation caused a general decrease in most of the metabolites involved in both C and N primary assimilation, thus impacting on either biomass or grain production. An accumulation of starch and of a variety of stress-related carbohydrates was also a characteristic metabolic symptom induced by N deficiency. Interestingly, in the three
studies performed on maize it was observed that the accumulation of secondary metabolites, particularly those used as precursors for cell wall synthesis was strongly reduced in line with the work on the maize GS-deficient mutants (Broyard et al., 2009). This observation partly explains why N deficiency, or a perturbation of primary N assimilation, has a strong impact on maize growth and development through an altered synthesis of metabolites used as the precursors required for lignin and cellulose production. Moreover, Amiour et al. (2012) showed that the response of the leaf metabolite content of maize to N deficiency, varied according to the plant developmental stage. Thus it is essential that changes in the metabolome must be followed over a long developmental period, when studying environmental effects in field trials with genotypes, or transgenic plants with varied NUE (Asiago et al., 2012).

5.7 Integrating “omics” data

Although the main metabolic functions that were altered as a result of N deficiency were conserved across the different “omics”, there was very little correlation among mRNA transcript, protein and metabolite content. This would suggest that other regulatory elements such as uncharacterized genes or metabolites may have an important functions within the biological networks involved (Urano et al., 2010). Moreover, it is generally admitted that “omics” studies only provide a narrow and static picture of the physiological status of a given organ, at a particular stage of plant development (Fernie and Stitt, 2012). Thus, additional fluxomics studies based on the use of $^{15}$N- and $^{13}$C-labelled compounds may represent an interesting complementary approach to metabolomics, since these techniques can provide additional information on the metabolic fluxes occurring in mutants, genetically modified crops or genotypes exhibiting contrasting NUE (Kruger and Ratcliffe, 2012; Masakapalli et al., 2013). Moreover such fluxomics techniques are potentially able to provide additional information on the turnover and remobilization of metabolites in a given cellular compartment during the day/night cycle and at critical periods of plant development when N is required for optimal plant growth and development (Gauthier et al., 2010;
In addition to examining the impact of N deficiency, metabolome studies are becoming more and more extensively used for the high throughput phenotyping necessary for large scale molecular and quantitative genetic studies aimed at identifying candidate genes involved in the control of plant productivity (Kliebenstein, 2009), even when these studies are not necessarily focused on NUE (Meyer et al., 2007; Lisec et al., 2008). Figure 5-1, illustrates the genetic variability of leaf metabolite content, enzyme activities and biomass components in nineteen selected maize lines, which are representative of American and European plant diversity and used as a core collection for association genetic studies (Camus-Kulandaivelu et al., 2006). Such a large genetic variability could
**Figure 5-1:** Changes in metabolite content, enzyme activities and biomass-related components in leaves of nineteen selected maize lines covering their genetic diversity (Camus-Kulandaivelu *et al.*, 2006) at two key stages of plant development.

The top of the figure shows the variation coefficient (expressed in %) of the biomass-related components in red (C: total carbon; N: total nitrogen; WC: water content) including yield. Enzyme activities are in blue italics (PEPC: phosphoenolpyruvate carboxylase; GS: glutamine synthetase; PPDK: pyruvate Pi dikinase; MDH: NADP-dependent malate dehydrogenase; GDH: glutamate dehydrogenase, GOGAT: ferredoxin-dependent glutamate synthase; AspAT: aspartate aminotransferase; NR: nitrate reductase). Metabolites and classes of metabolites are in black (2-OG: 2-oxoglutarate; Cit: citrate; Suc: sucrose; Pyr: pyruvate; Glu: glutamate; Gln: glutamine; OA: total organic acids; AA= total amino acids; Fruc: fructose; Gluc: glucose; unknowns: unidentified metabolites. At the bottom of the figure is shown an overview representation of the average of variation coefficients (from 0 to 80%) for the main classes of metabolites, enzyme activities and biomass components.

be used to obtain a better understanding of the control of NUE, since we observed that in this core collection there was almost no difference in PEPC activity, whereas the asparagine content varied by up to 200-fold. In agreement with this observation, it is well known that the asparagine content can vary considerably from one plant species to the other and that its concentration can change dramatically depending on the physiological condition of the plant (Lea *et al.*, 2007). Interestingly, we also observed that of the different classes of metabolites analysed, the amounts of polyamines, secondary metabolites and unknown metabolites were the most variable (Figure 5-1). Such findings strengthen the idea that further work is required to identify their role in plant productivity in order to provide the necessary information required for a full characterization of the metabolic and regulatory networks involved. Moreover, the range of variation observed for both biochemical and biomass-related traits was different, depending on the developmental stage of the plant. These data confirm that “omics”-based studies should
be performed over a sufficiently long developmental period, to identify the critical physiological stages during which there is a progressive switch between N assimilation and N remobilization are included (Hirel et al., 2007). In addition, environment-dependent changes of the underlying metabolic networks need to be taken into account when investigating the relationship between plant metabolism and plant biomass production (Sulpice et al., 2013). Leaf metabolite profiling techniques have recently been successfully used to dissect complex traits in maize through the use of genome-wide association mapping both in maize lines (Riedelsheimer et al., 2012a) and hybrids (Riedelsheimer et al., 2012b). Thus, metabolome-assisted breeding techniques, in addition to genome-assisted selection of superior hybrids, are promising for narrowing the genotype/phenotype gap of complex traits such as NUE (DellaPenna and Last, 2008; Fernie and Shauer, 2008; Lisec et al., 2011).

5.8 **Metabolic modeling as a tool to unravel the limiting steps in NUE**

Due to the ever-accelerating pace of genome sequencing and annotation in the past few years, researchers have made significant advances in mapping plant genes to metabolic functions. Nevertheless, efforts to engineer plant metabolism, in general and N metabolism in particular, have on most occasions been met with limited success (Good et al., 2004; Hirel et al., 2007). Due to the built-in metabolic redundancy, genetic interventions often do not bring about the desired effect in plant metabolism (Gutiérrez et al., 2005; Sweetlove et al., 2003). Therefore, by taking into account the complete inventory of metabolic transformations of a given plant species, a genome-scale metabolic reconstruction has the potential to make valuable advances, such as improvement of yield, NUE, and nutritional quality in crops.

Unlike gene and protein regulatory networks that denote putative interactions (Cho et al., 2007), metabolic network models capture the inter-conversion of metabolites through chemical transformations catalyzed by enzymes. Therefore, their topology is generally better characterized than the one of regulatory networks. A genome-scale metabolic
model is constructed by encompassing the widest possible list of biotransformations present in the organism as supported by annotation and homology evidence. Thus these models attempt to map the entire chemical repertoire of a specific organism (see Figure 5-3).

Large amounts of data relating to metabolites, reactions and their associated enzymes/genes are currently available through generic databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2012), SEED (Devoid et al., 2013), Metacyc (Karp and Caspi, 2011), Brenda (Schomburg et al., 2013), Universal Protein Resource (Uniprot) (Consortium, 2012), PubChem (Wang et al., 2013), ChemSpider (Pence and Williams, 2010) and plant specific databases such as: MaizeCyc (Monaco et al., 2013), MetaCrop (Schreiber et al., 2012), Corncyc (Dreher, 2014) (for maize) and Plant Metabolic Network (PMN) (Dreher, 2014), which is a combination of 18 plant databases. However, incompatibilities of representation (e.g., metabolites with multiple names/chemical formulae across databases), stoichiometric errors (i.e., elemental or charge imbalances) and generic metabolite descriptions (e.g., absence of stereospecificity or use of generic side chains) are key bottlenecks for the rapid reconstruction of new high-quality metabolic models by combining information from these databases. Of late, a database called MetRxn was developed to address these issues by integrating information (of metabolites and reactions) from 8 such databases and 44 published metabolic models (Kumar et al., 2012). In addition to gene/protein/reaction information, knowledge of the subcellular localization of enzymes is critical for the development of plant metabolic models. Towards this end, there exists protein localization databases such as Plant Proteome DataBase (PPDB) (Hieno et al., 2014) and SUBcellular localization database for Arabidopsis proteins (SUBA) (Tanz et al., 2013) for the two plant species Arabidopsis and maize. As displayed in Figure 5-2, the combination of biological databases, localization databases, and literature evidence comprise the initial information required to create a genome-scale model.
**Figure 5-2:** Iterative process of genome-scale model building.

A combination of biological and localization databases with published literature provides information for the initial genome-scale model. This gene, protein, and reaction information is combined with a biomass equation, which contains user-specified stoichiometries, to give the stoichiometric matrix. Using flux balance analysis with an objective function, typically maximizing biomass, a set of simulated reaction fluxes is determined. Transcriptome, proteome, and metabolome data constrain the model either by turning on or off reactions in a “switch” approach or by modifying the allowable flux through a reaction in a “valve” approach. The model is validated using knockout data or fluxomics data and depending on these results, iterations continue until a high quality genome-scale model is developed.
Encouragingly, an increasing number of genome-scale metabolic models of microorganisms and multicellular organisms are emerging, and the applications of these models are expanding (de Oliveira Dal'Molin and Nielsen, 2013; McCloskey et al., 2013). These models are being developed in an iterative manner using literature evidence in tandem with genome, transcriptome, proteome, and metabolome data. By using flux balance analysis (FBA), combined with the pseudo-steady state assumption (Orth et al., 2010), metabolic fluxes (or feasible ranges thereof) can be calculated using a fitness optimization proxy, such as biomass yield maximization. Under the pseudo-steady state assumption, FBA assumes that each metabolite that is produced must be consumed at an identical rate. To investigate the growth phenotype of the cell via FBA, a reaction that contains all required precursors for cell growth in experimentally measured proportions is generated and added to the model. This reaction is known as the “biomass” reaction and its flux is an abstraction of the cell biomass composition. Constraints in FBA are represented: (i) as equations balancing metabolite production to consumption via all possible reactions in the model, and (ii) as inequalities imposing bounds (i.e. the maximum or minimum allowable fluxes) in the system, such as uptake or secretion of specific metabolites, or upper and lower bounds of reaction fluxes (based on reaction thermodynamics). All these balances and bounds determine the feasible solution space i.e. allowable flux distributions of the model and an optimal solution is found under the specific objective (e.g., biomass yield maximization), as displayed in Figure 5-3.

Genome-scale models can be tested by comparing the in vivo growth of knockout strains to in silico growth under corresponding conditions, or by comparing fluxome data to simulated fluxes (McCloskey et al., 2013). Saha et al. (2011) compared experimental results to in silico model predictions in 17 of 21 cases by comparing the directional change of the maximum theoretical yield of lignins, sugars and crude protein between the wild type and two mutant strains. The C4GEM model quantitatively showed a strong correlation between the differential expression of proteins or protein complexes and predicted flux differences between the bundle sheath and mesophyll cell types in 50 of 66 cases (de Oliveira Dal'Molin et al., 2010b). These models can then guide
physiological characterization, metabolic engineering and discovery (Oberhardt et al., 2009). Mintz-Oron et al. (2012) demonstrated the use of genome-scale models by suggesting the knockout of 71 target enzymes to increase vitamin E accumulation in Arabidopsis. Furthermore, tissue-specific (Jerby et al., 2010) and multi-tissue (Jerby et al., 2010; Thiele et al., 2013) type models have also been developed for Homo sapiens and employed for studying metabolic interactions and therapeutic applications.

**Figure 5-3:** Flux balance analysis using a metabolic model.

A simplified metabolic model is displayed encompassing multiple tissue types, cell types, and compartments, as well as the transporters between them. Flux balance analysis assumes that each metabolite is produced and consumed at equal rates, as required by the pseudo-steady state assumption. This constraint is imposed for every metabolite and, along with the flux bounds of each reaction determined by reaction thermodynamics, creates the feasible solution space. By optimizing an objective function, typically biomass production, a flux value is predicted for each reaction within the model.
Metabolic modeling of plants is a rapidly developing field. Models of *Arabidopsis* (Poolman *et al.*, 2009; de Oliveira Dal'Molin *et al.*, 2010a; Radrich *et al.*, 2010; Mintz-Oron *et al.*, 2012), rice (Poolman *et al.*, 2013), barley (Grafahrend-Belau *et al.*, 2009; Grafahrend-Belau *et al.*, 2013), rapeseed (Pilalis *et al.*, 2011), sorghum (de Oliveira Dal'Molin *et al.*, 2010b), sugarcane (de Oliveira Dal'Molin *et al.*, 2010b), and maize (de Oliveira Dal'Molin *et al.*, 2010b; Saha *et al.*, 2011) have already been developed. Although the majority of these available models focus on a single tissue, whole-plant models are beginning to emerge. For instance, Grafahrend-Belau *et al.* (2013) developed a whole-plant model of barley, including specific models for the leaf, stem, and seed tissues, to analyze seed development, crop improvement and yield stability. To this end, a whole-plant metabolic model of maize would also be useful to not only characterize metabolism, but also to highlight the limiting steps in NUE. Integration of high throughput “omic” information with metabolic models contributes towards improving the genotype-phenotype relationship and prediction accuracy. Plant genome-scale metabolic model development is currently exploring ways of incorporating such experimental data by adopting approaches developed for microbial organisms (Töpfer *et al.*, 2013).

5.9 Incorporating transcriptome, proteome, and metabolome data into models

Incorporating transcriptome, proteome and metabolome data can increase the predictive accuracy of genome-scale models. Utilizing such high-throughput “omics” information not only provides regulation (for tissue-specific models) but also ensures that the correct reactions and metabolites are represented within specific tissue-types (for multiple tissue/whole plant models). To this end, Jerby *et al.* (2010) developed the Model-Building Algorithm (MBA) to reconstruct a tissue-specific model from a generic model by combining “omics” (i.e., transcriptome, proteome, and metabolome) information with the published literature. Furthermore, many other approaches have been developed for capturing “omics” data in the form of regulation on the model (Blazier and Papin, 2012;
Two main modeling philosophies, that abstract regulation as either an on/off “switch” or a continuous flow “valve” have been put forth, as shown in Figure 5-2. The Gene Inactivity Moderated by Metabolism and Expression (GIMME) (Becker and Palsson, 2008), integrative Metabolic Analysis Tool (iMAT) (Shlomi et al., 2008; Zur et al., 2010), and Metabolic Adjustment by Differential Expression (MADE) (Jensen and Papin, 2011) algorithms use a “switch” approach to turn the reactions on or off, based on differential expression changes. The GIMME approach simply turns off reactions based on a user-specified threshold for expression data. The iMAT approach discretises the expression data into lowly, moderately, and highly expressed genes, and then utilizes an algorithm to turn on the smallest number of lowly expressed genes required to achieve a specified metabolic function (i.e. a user specified biomass objection function). The MADE algorithm employs multiple data sets from two or more related conditions to activate or repress appropriate reactions for simulating the progression of experimental conditions. Only statistically significant changes in expression levels will convert an activated reaction to a repressed state, or vice-versa, when comparing one experimental condition to another. All “switch” approaches require essential reactions to remain active in the model, regardless of their expression level, to ensure biomass is produced. Contrary to these “switch” approaches, E-FLUX (a combination of flux and expression data) (Colijn et al., 2009) and Probabilistic Regulation Of Metabolism (PROM) (Chandrasekaran and Price, 2010) algorithms adopt a “valve” approach, via modifying the allowable range of the flux of any reaction (i.e. the upper and lower flux bounds of the reaction) based on gene/protein expression data. The E-FLUX algorithm incorporates a single data set (i.e., one experimental condition) and requires a user-specified function to convert expression levels to flux constraints. The PROM algorithm incorporates multiple data sets to set maximum reaction flux levels based on the probability that the gene is active among all experimental data sets. Recently Lee et al. (2012) developed another “valve” approach by using absolute gene expression levels to regulate reaction fluxes. The integration of metabolome data with transcriptome or proteome data can further increase the accuracy of the genome-scale metabolic models. The Integrative Omics-Metabolome Analysis (IOMA) algorithm uses a Michaelis-
Menten-type rate equation to calculate an empirical reaction flux using metabolome and proteome data. Each empirical reaction flux includes an error correction that is used to account for missing experimental metabolite concentrations and errors in experimental measurements. An algorithm is then used to minimize the error in reaction flux predictions with an additional biomass constraint (Yizhak et al., 2010). Overall, these advancements to integrate “omics” data into microbial models have resulted in a collection of data integration techniques that have set the stage for similar implementations in plant models.

While many of these algorithms have not yet been applied to plants, Töpfer et al. (2013) applied the E-FLUX algorithm to the Arabidopsis model developed by Mintz-Oron et al. (2012) to predict the maximum flux through metabolic pathways altered under eight varying light and temperature conditions. Of the 167 metabolic functions or pathways studied, 37 functions resulted in a differential capacity in at least one of the eight conditions modelled, meaning that the flux through the pathway changed more than the expected flux variations from random chance. With the successful incorporation of E-FLUX in Arabidopsis, more plant models employing transcriptome, proteome, and metabolome data are expected to emerge. For a more in-depth review of integrating “omics” data with genome-scale models of model organism, see the recent review paper by Saha et al. (2014).

5.10 Concluding remarks

Understanding the complexity of the control of NUE of model crop species such as maize, requires a holistic understanding of N-flow and associated regulation at the cellular, organ, and whole-plant levels. In this review, we have highlighted the current status of plant “omics” and critically analyzed the importance of metabolic modeling in the study of NUE and other agronomic traits such as biomass and grain yield. While the integration of several biological databases, model-building strategies and high-throughput
“omics” procedures are already available, there is still no whole-plant model that has been developed for maize. Therefore, by applying a combinatorial semi-automated (Suthers et al., 2009; Jerby et al., 2010) model building workflow, a high quality whole-plant model could be developed for maize. Then, the “omics” data obtained by growing plants under varying N conditions and by analysing genetically modified plants and mutants altered for the expression of structural or regulatory genes for N uptake, assimilation and remobilisation can be incorporated in the model. They could provide a more accurate simulation of the effect of N on the metabolic interactions and flow throughout the plant and subsequently could identify the key reactions (i.e., genes) controlling NUE. Ultimately, an integrated model combined with quantitative genetic studies may identify possible genetic interventions to improve NUE. Exploiting natural and created genetic variability could then be experimentally tested to either verify the model or provide new information to resolve discrepancies of model predictions, thereby increasing the model fidelity in the future.

In addition, genome-wide association studies combined with metabolic and gene expression analyses are becoming more commonly implemented for screening large collections of genotypes and hybrids for their potential productivity (Riedelsheimer et al., 2012b). Such studies also focus on the effect of the environment on plant phenotypic plasticity under various N regimes and environmental conditions (Brunetti et al., 2013; Gifford, 2013). In a similar manner to gene expression studies, it is also now possible to study the relationship between measured metabolite contents, in order to interpret complex data sets and identify key network components for further practical metabolic engineering (Yonekura-Sakakibara et al., 2013; Toubiana et al., 2013). Thus, the next major challenge for plant biologists and breeders will consist of integrating full “omics” data sets into the modeling, population structure and selection strategies (Langridge and Fleury, 2011).
Chapter 6

ASSESSING THE METABOLIC IMPACT OF NITROGEN AVAILABILITY USING A COMPARTMENTALIZED MAIZE LEAF GENOME-SCALE MODEL


6.1 Introduction

Zea mays L., commonly known as maize or corn, is an essential dual use food and energy crop. Maize production is increasing at the greatest rate among all cereals with a worldwide trend of 0.06 t/ha/year (tons/hectare/year) [329], and a record 877 million tons produced in 2011-2012 fiscal year [330]. With the recent completion of the maize genome in 2009 along with the creation and curation of databases such as MaizeGDB in 2011 [331], MaizeCyc in 2013 [332], and MetaCrop 2.0 in 2012 [333], there is a need for an updated genome-scale metabolic (GSM) model [334] that will integrate all newly available information from diverse sources. The integration of this information with experimental transcriptomic data, proteomic data, and biomass composition measurements in a excess nitrogen (N^+ WT) condition, limited nitrogen (N^- WT) condition and two glutamine synthetase (GS) mutants, gln1-3 and gln1-4 mutants, [335] allows for more accurate assessment of the nitrogen (N) metabolism within the maize leaf.

Maize is a C_4 plant that overcomes the inefficiencies of RuBisCO, to capture oxygen over the preferred CO_2, by separating the carbon (C) fixation process into two cell types: the bundle sheath and mesophyll cell. In comparison to C_3 plants, this separation causes C_4 plants to decrease in photorespiration rate [336], increase in the photosynthetic nitrogen
use efficiency (NUE) [337], and increase in the net photosynthesis at high light intensities
(under standard air and temperature conditions) [338]. A C4 specific maize GSM can
yield insight into N metabolism and provide cues for improving NUE (i.e. the vegetative
biomass or grain yield produced per unit of N present in the soil). Since N is the major
limiting factor in agricultural production among mineral fertilizers [339, 340], improving
the NUE is essential in improving overall productivity in maize [341]. Amiour et al.
(2012) experimentally determined 150 gene transcripts, 40 proteins, and 89 metabolites
that are significantly different between the N+ WT and N- WT conditions during the
vegetative stage of growth. N utilization is strongly linked to the GS enzyme [335]
because all N, either in the form of nitrate or ammonium ions, are channeled through the
reaction catalyzed by the GS enzyme [335, 342]. The mesophyll cell specific GS1-3
isozyme is involved in synthesizing glutamine after nitrate reduction from the vegetative
state until the plant reaches maturity. Leaf aging induces the bundle sheath specific GS1-4
isozyme. Consequently, Martin et al. hypothesizes that this isoform is used in the
reassimilation of ammonium during protein degradation in senescing leaves [335].
During vegetative growth in the leaf tissue, DNA microarray data reveal that 243 gene
transcripts, 46 proteins, and 48 metabolites exhibit significant differences in the glnl-3
mutants and 107 gene transcripts, 14 proteins, and 18 metabolites display substantial
differences in the gln1-4 mutants (Table S1). In this second-generation maize model, we
explore the effect of the computational knockout of genes encoding for GS1-3 and GS1-4
isozymes using flux balance analysis (FBA) to elucidate the role of GS in N metabolism.

FBA of GSMs is used to model organism specific metabolism by simulating the internal
flow of metabolites. The number of GSMs for plant organisms is quickly increasing with
models for Arabidopsis [343, 344], barley seed [345], maize [334, 346], sorghum [346],
sugarcane [346], rapeseed [347], and rice [348]. These models rely on annotation
information to assemble comprehensive compilations of all reactions and metabolites
known to occur within the organism. Currently, whole-genome sequencing is completed
for approximately 40 vascular plants including Arabidopsis thaliana [349], Arabidopsis
lyrata [350], Glycine max [351], Oryza Sativa [352, 353], Populus trichocarpa [354],
Sorghum bicolor [355], Theobroma cacao [354], and Zea mays [356]. Gene annotations of the whole-genome sequences are used to determine the reactions within an organism and therefore build a GSM. FBA calculates all reaction fluxes in a metabolic network based on the optimization of an objective function (typically the maximization of the biomass yield). A quasi-steady state is assumed and flux constraints are set based on the specific media or the reversibility of reactions derived from thermodynamics. Incorporation of ‘omics’ data into GSMS is achieved through appropriate constraints on fluxes that restrict metabolic flows to only condition-relevant phenotypes.

Within the last few years, multiple methods have been developed to integrate ‘omics’ data into GSMS. Proteomic and transcriptomic data have been used to apply flux constraints on corresponding reactions determined by gene-protein-reaction (GPR) associations. The GIMME [357], iMAT [358], and MADE [359] algorithms use a “switch” approach to turn on/off reactions based on expression levels. The GIMME algorithm turns off reactions based on a user-specified threshold of the expression level. The iMAT algorithm turns on a minimal set of reactions associated with low expression data in order to achieve a user-specified metabolic function. The MADE algorithm incorporates related experimental datasets into the model to activate or repress reactions based on the progression of the experimental conditions. A different class of algorithms, known as the “valve” approach, was developed to incorporate proteomic and transcriptomic data by constraining the allowable flux ranges of reactions. The E-Flux method incorporates a user-specified function to convert gene expression data to flux constraints [360]. Finally, the PROM [361] algorithm uses multiple datasets to constrain flux bounds based on the probabilities associated with gene activity among all datasets. These advancements pertaining to the integration of ‘omics’ data with GSMS has allowed for more accurate flux predictions.

In this work, we describe the reconstruction of a second-generation maize leaf model and incorporation of ‘omics’ data to the model with the goal of improving the understanding of N metabolism. Both primary and secondary metabolism of maize are captured by
combining information from Metacrop [333], MaizeCyc [332], and the earlier iRS1563 [334] model. In comparison to the iRS1563 model, this second-generation model spans an additional 4,261 genes and 6,498 reactions. The increased number of accounted genes and reactions enables the inclusion of additional pathways such as fructan biosynthesis, siroheme biosynthesis, and ubiquinol-9 biosynthesis. The model accounts for two major cell types in the leaf (i.e., the bundle sheath and mesophyll cells). The bundle sheath cell contains seven compartments: the cytosol, mitochondrion, peroxisome, plastid, plasma membrane, thylakoid membrane, and vacuole. The mesophyll cell contains six compartments: the cytosol, mitochondrion, plastid, plasma membrane, thylakoid membrane, and vacuole. Compartmentalization is based on maize specific experimental proteomic and transcriptomic measurements [362-365] as opposed to the Arabidopsis-based compartmentalization adopted in the previous iRS1563 maize model [334]. Light reactions have been expanded from an aggregate reaction (as described in the iRS1563 model) to multiple reactions for each complex with the inclusion of a thylakoid membrane compartment. In contrast to the C4GEM maize model [346], which focuses exclusively on primary metabolism in maize, the developed model also spans secondary metabolism by including all reactions known to occur within the maize leaf tissue. The model includes as many as 763 secondary metabolism reactions (without including duplicate counting due to compartmentalization). Through the incorporation of ‘omics’ data, regulatory restrictions are introduced in the model to switch-off/on reactions under the N⁺ WT, N⁻ WT condition and two GS knockout mutants (gln1-3 and gln1-4) (Martin et al., 2006). Reactions linked to genes or proteins with significantly different expression levels between N⁺ WT and N⁻ WT condition, as well as the gln1-3 and gln1-4 mutants versus the N⁺ WT are conditionally turned on or off accordingly. The metabolite pool is simulated by maximizing the total flux through a metabolite (i.e. flux-sum) as a proxy for the metabolite turnover rate [366]. The directional change of flux-sum levels between the N⁻ WT condition and the N⁺ WT condition, as well as the GS mutant conditions and the N⁺ WT condition are qualitatively compared to the directional change in experimentally measured concentration levels. This analysis reveals similar trends as the recently developed Flux Imbalance Analysis [367] that makes use of dual variable values
associated with metabolite balances to infer the effect of concentration changes on the objective function value.

6.2 Results and Discussion

6.2.1 Effect of Nitrogen Conditions on Biomass Components

Biomass components were measured in N\textsuperscript{+} conditions, as well as in each N stress condition (N\textsuperscript{-}, \textit{gln1.3} and \textit{gln1.4} mutants). As expected, in the majority of cases, the N\textsuperscript{-} condition produces a smaller quantity of biomass components than the WT, \textit{gln1-3}, and \textit{gln1-4} mutant conditions. However, the amount of amino acids produced is about 5 times higher in the \textit{gln1-4} mutant than the \textit{gln1-3} mutant, resulting in comparable amino acid measurements between the \textit{gln1-4} mutant and the WT conditions, as well as between the \textit{gln1-3} mutant and N\textsuperscript{-} condition. The similar amino acid quantities between the \textit{gln1-4} mutant and WT condition in the vegetative stage help to confirm that the GS1-4 isozyme is essential in plant maturity and has a smaller effect compared to the GS1-3 isozyme in the vegetative stage. As expected, the amount of starch increases from the WT condition to the N\textsuperscript{-} condition. With the light conditions remaining constant for all conditions, we expect the photosynthesis rate to also remain constant producing a similar amount of starch. Under N\textsuperscript{-} conditions, the breakdown of starch is limited by the amount of N available (Tercé-Laforgue et al., 2004; Amiour et al., 2012). Due to the limited N available, the starch is stored rather than broken down to produce other biomass components as expected. The condition-specific biomass measurements have been incorporated in the maize leaf model to more accurately represent metabolism under each condition.
6.2.2 Development of the second-generation maize leaf model

The second-generation maize leaf model was developed using a combination of gene, protein, and reaction information from our previously developed maize model iRS1563, biological databases such as KEGG, MaizeCyc, and Metacrop, as well as published literature sources. The model contains 5,824 genes and 8,484 reactions, a significant increase to the iRS1563 model, which contained 1,563 genes and 1,985 reactions. The second-generation maize model is split into two cell types, (i.e. the bundle sheath and mesophyll cells). The bundle sheath cell is further divided into seven compartments, while the mesophyll cell contains six compartments, as displayed in Figure 6-1. Of the 8,484 reactions in the model, 3,997 reactions are unique, meaning duplicated counts due to compartmentalization are disregarded. Of these 3,997 unique reactions, 1,012 reactions were assigned localization information based on transcriptomic and proteomic data [362-365]. Light reactions were adjusted to model the flow of protons between the thylakoid membrane and chloroplast, represent the pH differential between compartments, and describe the conversion of light to energy [368]. The mitochondrial electron transport chain was similarly updated to include the proton exchange of ATP synthase between the intermembrane space and the mitochondrial matrix [369]. Finally, 303 specific reactions were added to model glycerolipid synthesis [370-376]. To the best of our knowledge, this is the first plant model to include specific glycerolipid synthesis. Aggregate reactions were then included to link specific two-tailed glycerolipids to the experimentally measured single lipids (see Table S3). Compiling transcriptomic and proteomic compartmentalization data with literature-based pathways yielded a model of 3,524 reactions leaving 2,985 unique reactions with unknown localization.
Figure 6-1: Number of metabolic and transport reactions distributed between compartments in the bundle-sheath and mesophyll cell types.

The number of metabolic and transport reactions are shown for each compartment. Integral membrane proteins are considered in metabolic reaction counts in the compartment in which the main biotransformations occur. For example the ATP synthase associated with the mitochondrial electron transport chain is counted as a metabolic reaction in the mitochondria, not the inner mitochondrial membrane (IMM).

Once reactions were compartmentalized based on transcriptomic data, proteomic data, and published literature, the reactions were divided into two groups. The first group
(core set) includes reactions with known localization, while the second group (non-core set) spans reactions known to occur within the maize leaf but with no localization evidence. Whenever possible, core reactions were unblocked by first adding reaction(s) from the non-core set to one or multiple compartment(s) and second by appending inter- or intra-cellular transporter(s) (see Materials and Methods section). By following this approach, 1,032 unique reactions with previously unknown localization were assigned to compartments and 732 transporters were added. The remaining 1,953 unique reactions were assigned to compartments based on available pathway information or assigned to the cytosol of both the bundle sheath and mesophyll cells.

With all reactions assigned to specific compartments, thermodynamically infeasible cycles that were generated due to the overly permissive inclusion of reactions in the model as well as lack of reaction directionality information were subsequently identified and eliminated. By first restricting the directionality of reactions and second removing reactions, we were able to eliminate all thermodynamically infeasible cycles in the model. In this process we restricted the directionality of 36 reactions and removed 665 reactions from the model (Table 6-1). Upon the resolution of thermodynamically infeasible cycles, we attempted to unblock the remaining blocked core reactions by adding reactions from similar organisms and model organisms (i.e. *Orzya sativa japonica*, *Brachypodium distachyon*, *Sorghum bicolor*, and *Arabidopsis thaliana*).

### 6.2.3 Incorporation of ‘omics’ data in the model

In order to more accurately model the N+ WT, N- WT, and GS mutant conditions in maize, GPR associations mapped the gene transcripts and proteins that are statistically expressed at a low level to reactions that were turned-off in the model. However, no essential reactions were altered. In this manner, the fluxes through 85 reactions in the N+ WT condition, 20 in the N- WT condition, 100 in the gln1-3 mutant, and 9 reactions in the gln1-4 mutant were restricted. The reactions restricted in the N+ WT mainly correspond to reactions known to only occur under stress. Nitrogen perturbations within
the leaf tissue were modeled by combining the incorporation of transcriptomic and proteomic data with the unique biomass compositions for each condition.

**Table 6-1:** Number of reactions after each model creation and curation step.

The original two data sets are the core set and gapfill set which combine to form the final model statistics. The total number of metabolic, transport, exchange, and biomass reactions are displayed after each process during model curation. Metabolic reaction totals include duplication from compartmentalization.

<table>
<thead>
<tr>
<th>Processes Performed</th>
<th>Metabolic Reactions</th>
<th>Transport Reactions</th>
<th>Exchange Reactions</th>
<th>Biomass Reactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Data</td>
<td>Core Set</td>
<td>3002</td>
<td>418</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>Core Set + Manually Created Pathways</td>
<td>3264</td>
<td>469</td>
<td>290</td>
</tr>
<tr>
<td></td>
<td>Non-core Set</td>
<td>18951</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Core Set + Manually Created Pathways</td>
<td>Modified Gapfill Compartmentalization</td>
<td>3971</td>
<td>1198</td>
<td>290</td>
</tr>
<tr>
<td></td>
<td>Manually Determined Compartmentalization</td>
<td>9005</td>
<td>1198</td>
<td>290</td>
</tr>
<tr>
<td></td>
<td>Thermodynamically infeasible cycles</td>
<td>7362</td>
<td>742</td>
<td>290</td>
</tr>
<tr>
<td></td>
<td>Similar Organism Gapfill</td>
<td>7367</td>
<td>742</td>
<td>290</td>
</tr>
<tr>
<td>Final Model</td>
<td>Second-generation GSM model</td>
<td>7367</td>
<td>742</td>
<td>290</td>
</tr>
</tbody>
</table>

The minimal set of reactions, whose regulation causes a decrease in the biomass yield, was determined for the N⁺ WT, N⁻ WT, gln1-3 mutant, and gln1-4 mutant conditions. Of the 85 reactions with restricted flux in the N⁺ WT condition, only 5 reactions (3 unique reactions irrespective of compartmentalization) were identified to affect the biomass yield. These reactions include ferredoxin oxidoreductase, GTP phosphohydrolase, and ethanol oxidoreductase. The GTP phosphohydrolase reaction is a contributing reaction to
purine metabolism in maize, but is not essential as GTP can be synthesized by other reactions, such as the ATP-GDP phosphotransferase reaction. In the N WT condition, none of the reactions have an effect on the biomass yield suggesting, as expected, that the decreased amount of nitrogen is the main limiting factor in biomass yield. In the gln1-3 mutant condition 12 of the 100 reactions, which are switched off based on “omics” data, affect the biomass yield. Of the 12 reactions 6 are unique, in regards to compartmentalization, including the ferredoxin oxidoreductase, GTP phosphohydrolase, ribose-5-phosphate isomerase, glyceraldehyde-3-phosphate dehydrogenase, fructose-bisphosphate aldolase, and ATP 6-phosphofructokinase reactions. The ribose-5-phosphate isomerase reaction loss can be rescued by the ATP dependent ribose-5-phosphate diphosphotransferase reaction. The gyceraldehyde-3-phosphate dehydrogenase reaction is not essential because glyceraldehyde-3-phosphate is synthesized during carbon fixation in photosynthesis and 3-phospho-D-glycerol phosphate can be synthesized through the 3-phospho-D-glycerate 1-phosphotransferase reaction using ATP. The additional ATP required in 3-phospho-D-glycerate 1-phosphotransferase results in a decrease in the biomass yield. The fructose-bisphosphate aldolase reaction, which is involved in the Calvin-Benson-Bassham cycle and glycolysis pathway, can be bypassed using the sedoheptulose 1,7-bisphosphate D-glyceraldehyde-3-phosphate-lyase reaction. The conversion of fructose 6-phosphate to fructose 1,6 bisphosphate can be completed with either ATP or UTP. With the conversion involving ATP restricted based on “omics” data, the UTP dependent conversion results in a decrease in biomass yield, suggesting UTP is more energetically expensive in the gln1-3 mutant case than ATP. Finally, the gln1-4 mutant condition involves only 9 reactions, of which 4 affect the biomass. These 4 reactions, however, are the ribose-5-phosphate isomerase reaction in multiple compartments. While only a subset of reactions affect the biomass levels in the N WT, gln1-3 mutant, and gln1-4 mutant conditions, the additional regulation will have an effect of the flux predictions within the model.

The metabolomic data was compared to flux predictions within the model in each of the various N background conditions. The increasing or decreasing trend of the metabolite
concentration, displayed in Figure 6-2, was qualitatively compared to the change in the maximum flux-sum determined by the model, as displayed in Figure 6-3. Therefore, an increase/decrease in the flux-sum (i.e. often interpreted as the maximum metabolite pool) of a metabolite between the N\(^{-}\) WT condition and the N\(^{+}\) WT condition and between the two GS mutants and the N\(^{+}\) WT condition were contrasted against metabolite level changes. Figure 6-3 demonstrates the importance of restricting fluxes based on transcriptomic and proteomic data. In the N\(^{-}\) condition the accuracy changes from 57% when the fluxes are not constrained to 80% when the flux constraints are incorporated. Similarly, the accuracy increases from 50% to 67% in the gln1-3 mutant and from 41% to 62% in the gln1-4 mutant. The identified flux-sum levels are included in Supplemental S4 (which can be found in the on-line version of the published paper). The flux-sum level is affected by the “omics” based constraints for all metabolites in all conditions with the exception of only four metabolites in the N\(^{-}\) WT condition. This suggests that limiting the amount of nitrogen is the primary factor in correctly predicting the metabolite pool decrease in galacturonate, phophethanolamine, and isoleucine. The final metabolite not affected by the “omics” based constraints is rhamnose, which is simulated as decreasing using the flux-sum concept, but is experimentally shown to increase in concentration, suggesting possibly a missing regulation in the N\(^{+}\) WT condition. An overall 62% accuracy in reproducing the experimental trends in three different N conditions is assessed for the model.
Figure 6-2: Number of metabolites in each condition that statistically vary from the WT condition at the vegetative stage.

The number of metabolites that experimentally significantly increase and decrease in comparison to the WT condition are displayed for each of the N conditions tested.

6.2.4 Flux range variations among conditions

The flux range of each reaction was determined under maximum biomass in the N⁺ WT, N⁻ WT, gln1-3 mutant, and gln1-4 mutant conditions. The reaction’s flux range in the N⁻ WT, gln1-3 mutant, and gln1-4 mutant conditions was compared to the flux range in the N⁺ WT reference condition to determine reactions whose flux ranges do not overlap. This indicates that the flux
Figure 6-3: Effect of “omics” based regulation on the flux-sum prediction compared to the experimental trend in metabolite concentration.

By restricting the reaction flux based on the transcriptomic and proteomic data, the accuracy of the qualitative trend in metabolite pool size between the N⁻ WT to N⁺ WT, \textit{gln1}-3 mutant to N⁺ WT, and \textit{gln1}-4 mutant to N⁺ WT increases. Before adding “omics” based constraints, the model was able to correctly predict the direction of change in 57% of the metabolites measured in the N⁻ WT compared to the N⁺ WT condition. By including the “omics” based constraints, the accuracy increased to 80%. The \textit{gln1}-3 mutant condition compared to the N⁺ WT condition and the \textit{gln1}-4 mutant compared to the N⁺ WT condition increases from 50% to 67% accuracy and 41% to 62%, respectively, in predicting the metabolite pool size with the inclusion of flux restrictions based on “omics” data.

through the reaction must change as a result of the limited nitrogen or mutant condition. Overall, the flux through 139 reactions must change between the N⁻ WT and N⁺ WT
condition, 665 reaction fluxes must change between the \textit{gln1-3} mutant and N\textsuperscript{+} WT, and 649 reaction fluxes must change between the \textit{gln1-4} mutant and N\textsuperscript{+} WT conditions. In all three nitrogen backgrounds (i.e. the N\textsuperscript{−} WT condition, \textit{gln1-3} mutant, and \textit{gln1-4} mutant conditions) the flux compared to the N\textsuperscript{+} WT reference condition decreases under maximum biomass through leucine biosynthesis in the bundle sheath cell, hexadecanoic acid metabolism, homoserine biosynthesis, and vitamin K biosynthesis pathways. In addition, the flux through adenosine nucleotide degradation, serine biosynthesis, stachyose biosynthesis, and guanine and guanosine salvage pathway must decrease in both of the GS mutant conditions compared to the N\textsuperscript{+} WT condition. However, the flux through alanine biosynthesis, γ-glutamyl cycle, histidine biosynthesis, xylose degradation, and glutathione degradation increase in the GS mutant conditions compared to the N\textsuperscript{+} WT condition. The flux through the γ-glutamyl cycle increases to synthesize glutamate through the consumption of two ATP. The increase in xylose degradation causes an increase in the formation of glyceraldehyde 3-phosphate. The change in the majority of these pathways is directly related to the differences in the proportion of the biomass components between the modeled conditions. For example, the glutathione degradation pathway yields glycine, which experimentally increases in proportion in both the GS mutant conditions compared to N\textsuperscript{+} WT.

\section*{6.3 Concluding remarks}

We have introduced a second-generation model that is specific to the leaf tissue with specificity between the bundle sheath and mesophyll cell types. This model with the addition of other maize tissue-specific models can be integrated into a whole-plant genome-scale model for maize. By determining a required metabolic function that is specific to each tissue, tissue-specific models can be created ensuring that only functional reactions are included in each tissue. Future efforts will focus on tissue-specific models for the seed, stalk, tassel, and root tissues. These tissue-specific models can then be linked using inter-tissue transport reactions with the stalk tissue acting as the central transporter among the various tissues. A whole-plant genome-scale model of maize will
help to elucidate the flow of nitrogen from the root to the other tissues in the plant. By modeling the entire plant, non-intuitive bottlenecks in nitrogen metabolism can be determined, which then can be used to suggest genetic interventions to increase the nitrogen use efficiency in maize. In addition, the flow of sugars to the seed tissue can be analyzed to suggest genetic interventions to increase the carbohydrate/sugar content of maize seed. Apart from its crucial role as a food crop, maize is also used for cellulosic biofuels. To this end, the amount and composition of cell wall polymers is important in developing cellulosic maize. Lignin not only provides rigidity to the maize plant [309, 315], but also makes digestion of cellulosic and hemicellulosic sugars difficult during dilignification [316]. The available genetic engineering strategies are not suitable since plant viability and fitness are affected upon lignin reductions [316]. Therefore, by utilizing the whole-plant genome-scale model a system-wide implication of these genetic disruptions can be quantitatively assessed, thus facilitating new strategies for reducing lignin content without affecting the mechanical integrity of the maize plant.

6.4 Materials and Methods

6.4.1 Plant Material

Maize (Zea mays L., genotype B73), wild-type (WT), gln1-3 and gln1-4 mutant seeds were first sown on coarse sand and after one week, when two to three leaves had emerged, individual plants were transferred to pots (diameter and height of 30 cm) containing clay loam soil with one plant per pot and 12 pots in total. Clay loam soil is composed of a mixture of loam (washed fine pit with no minerals) and loam balls of about 0.5 cm diameter that ensure sufficient aeration of the roots and allow growing the plant until maturity without lodging. Clay loam soil also allows a constant flow of the nutrient solution that is provided several times a day. Pots were placed in a glasshouse at the Institut National de la Recherche Agronomique, Versailles, France and grown from May to September 2004. Pots were moved every week to avoid shading and position
effects. Three individual plants of similar size and of similar developmental stage were randomly selected. They correspond to the three replicates used for the three different ‘omics’ experiments. At the 10 to 11 leaf stage, the three youngest fully expanded leaves were harvested and pooled for the vegetative stage (V) samples to obtain enough homogenous plant material representative of this plant developmental stage. The leaf below the ear was harvested at 55 days after silking and named the mature leaf developmental stage (M). The M stage corresponds to the end of the grain-filling period, during which under our experimental conditions, leaves did not show any visible symptoms of yellowing. The leaf, below the ear, was selected since it has been shown to provide a good indication of the source sink transition during kernel filling [377-379]. Leaf samples were harvested between 9 a.m. and noon and frozen in liquid N\textsubscript{2}, ground to a homogenous powder, and stored at -80°C. Choosing a single time point during the middle of the light period, is commonly used in a large number of physiological and molecular studies and was successfully used both in our previous quantitative genetic and ‘omics’ studies [377, 380]. In the glasshouse, WT, glnl-3, and glnl-4 mutant plants were watered daily with a complete nutrient solution containing 10 mM KNO\textsubscript{3} as the sole N source[381]. The complete nutrient solution contained 1.25 mM K\textsuperscript{+}, 0.25 mM Ca\textsuperscript{2+}, 0.25 mM Mg\textsuperscript{2+}, 1.25 mM H\textsubscript{2}PO\textsubscript{4}\textsuperscript{-}, 0.75 mM SO\textsubscript{4}\textsuperscript{2-}, 21.5 mM Fe\textsuperscript{2+}(Sequestrene; Ciba-Geigy), 23 mM B\textsuperscript{3+}, 9 mM Mn\textsuperscript{2+}, 0.3 mM Mo\textsuperscript{2+}, 0.95 mM Cu\textsuperscript{2+}, and 3.5 mM Zn\textsuperscript{2+}. For growing WT plants under low N-deficient conditions (N\textsuperscript{-}), NO\textsubscript{3}\textsuperscript{-} was supplied as 0.1mM KNO\textsubscript{3}, a N concentration that has previously been shown to provide N-deficiency stress for most plant species [377, 378, 382].

6.4.2 Yield Components Analysis

Kernel yield, its components, and the N content of different parts of the plant at stages of development from silking to maturity were determined according to the method described by Martin et al. (2005) and corresponded to the data described in Martin et al., 2006 and Amiour et al., 2012.
6.4.3 RNA and DNA Preparation

Total RNA was extracted as described by Verwoerd et al. (1989) [383] from leaves that had been stored at -80°C. 50 mg of total RNA were incubated at 37°C for 30 min with 40U RNase inhibitor and 25U RNase-free DNase (Promega, Charbonnieres, France) in 6 ml 10X buffer (Promega) with DEPC treated water added to a final volume of 60 ml. The DNase was removed by phenol/chloroform/iosamyl alcohol (25:24:1) extraction and total RNA was precipitated overnight at -20°C in 0.1 volume ammonium acetate (3M) and 2.5 volume ethanol (100%) and resuspended in DEPC treated water. Purification of mRNA from total RNA was performed using the Dynabeads mRNA Direct Kit (Dynal/Invitrogen, Cergy Pontoise, France). Genomic DNA was isolated from frozen leaves according to the micro-scale procedure [384]. Leaf dNTP and NTP concentrations were calculated using the maize genome sequence and the abundance of leaf maize transcript available at www.maizesequence.org. For the unidentified deoxynucleotides and nucleotides calculations were performed using the average composition of the identified ones.

6.4.4 Gene Expression Profiles using Maize cDNA Microarrays

Starting with 3 µg of total RNA, non-modified amplified antisense RNA (aRNA) products were prepared using the Amino Allyl MessageAmp™ aRNA Kit (Ambion, Foster City, CA, USA), according to the manufacturer’s instructions. Briefly, RNA was transcribed into cDNA using reverse transcriptase with a T7 primer that contains a promoter for DNA-dependent RNA polymerase[385]. After RNase H-mediated second-strand cDNA synthesis, the double-stranded cDNA (dscDNA) was purified and served as a template in the subsequent in vitro transcription (IVT) reaction. Following this, 2 µg- aliquots of aRNA were labeled using the SuperSript™ Indirect cDNA Labeling System Kit (Invitrogen, Carlsbad, CA, USA) as described by the manufacturer’s protocol, except that the purification steps were carried out using QIAquick® PCR columns (QIAGEN,
Hilden, Germany). The quantity and quality of each intermediate product, including total RNA, dscDNA, aRNA and labeled targets, were evaluated using a Nanodrop ND-1000 spectrophotometer and an Agilent Technologies 2100 Bioanalyzer [386, 387]. Whole-genome leaf transcript profiling was performed using the maize 46K arrays obtained from the maize oligonucleotide array project (http://www.maizearray.org/index.shtml). Transcript abundance in each of the three replicates for V and M leaves at low (N-) and high (N+) N supply was determined using a mixture of all the samples (12 in total, each with the same mRNA concentration) as a reference. Hybridizations between the maize oligonucleotide microarrays and fluorescently labeled samples were performed in MICROMAX Hybridization Buffer III (Perkin Elmer) using the manufacturer's hybridization and wash conditions and a GeneTac™ HybStation (Genomic Solutions, Ann Arbor, MI, USA). Before hybridization, 50 pmol Cy3- and 50 pmol Cy5-labelled targets were mixed, dried using compressed air, and reconstituted with 115 µl of hybridization buffer, followed by denaturing at 90°C for 3 min. Each hybridization mixture was placed on the maize 45K array slides mounted in the hybridization station and the hybridizations were performed for 3 h at 65°C, followed by 3 h at 55°C, then 12 h at 50°C with gentle agitation. Thereafter, the arrays were automatically washed with the GeneTac™ washing solutions (Genomic Solutions, Ann Arbor, MI, USA) using the program for multiple automatic washes, with a flow time of 40 s. Immediately after the completion of the final washing step, the arrays were removed from the station, briefly immersed in distilled water and air-dried with ozone-safe dry air. Hybridized microarrays were scanned using a GenePix 4000B Microarray Scanner (Molecular Devices, Sunnyvale, CA, USA) at 10-µm resolution and variable photomultiplier (PMT) voltage to obtain maximal signal intensities with <0.05% probe saturation. Subsequent image analysis was performed with the GenePix Pro (v6.0.1.26) software. Analysis included defining the spots, measuring the intensities, flagging spots when inadequate quality control parameters were found, and evaluating local background. The resulting files, containing all the scan data, were further processed using the statistical programming language R (http://www.r-project.org) together with the packages of the MAnGo project (Version 0.9.7, MicroArray Normalisation tool of GODMAP, CNRS BioInfome Team
The background level was calculated using morphological operators (a short closing followed by a large opening) and subtracted. Raw data were normalized using a global loess method [389]. Gene annotation was provided by the maize oligonucleotide array project mentioned above (http://www.maizearray.org/index.shtml).

### 6.4.5 Statistical Analysis of Maize cDNA Microarray Data

Statistical group comparisons were performed using multiple testing procedures to evaluate statistical significance for differentially expressed genes. Two gene selection approaches were applied, including the Significance Analysis of Microarrays (SAM;[390]) permutation algorithm, and a $p$-value ranking strategy using both $z$-statistics in ArrayStat 1.0 software (Imaging Research Inc.) and moderated t-statistics using a moderated t-test available in MAnGO tools [http://bioinfome.cgm.cnrs-gif.fr] and BRBArrayTools v3.2.3 packages [391]. For multiple testing corrections, the false discovery rate (FDR) procedure was used [392]. Statistical tests were computed and combined for each probe set using the log-transformed data and a probe set was declared to be significant when the adjusted $p$-value was less than the effective $\alpha$-level ($\alpha=0.05$) in at least one of these tests. A filtering procedure additionally excluded those data points considered biologically unreliable due to low signal intensities (Amean < 7.0).

### 6.4.6 Total Protein Extraction, Solubilization, and Quantification

A TCA/acetone protein precipitation was performed as described by Méchin et al. (2007)[393], from the leaves of N$^+$ and N$^-$ plants harvested at the V and M stage of development. The frozen leaf powder was resuspended in acetone with 0.07% (v/v) 2-mercaptoethanol and 10% (w/v) TCA. Proteins were allowed to precipitate for 1 h at -20°C. The pellet was then washed overnight with acetone containing 0.07% (v/v) 2-mercaptoethanol. The supernatant was discarded and the pellet dried under vacuum.
Protein resolubilization was performed according to Méchin et al. (2007) using 60 µL/mg of R2D2 buffer (5 M urea, 2 M thiourea, 2% CHAPS, 2% SB3-10, 20 mM dithiothreitol, 5 mM Tris(2-carboxyethyl)phosphine hydrochloride, 0.75% carrier ampholytes). After resolubilization, samples were centrifuged and the supernatant was transferred to an Eppendorf tube prior to protein quantification. Total protein content of each sample was evaluated using the 2-D Quant kit (Amersham Biosciences).

6.4.7 Two-dimensional Electrophoresis, Gel Staining, and Image Analysis

Solubilized proteins (300 µg) were separated on a pH 4–7 gradient Immobilized pH Gradient (IPG) strip (Amersham Biosciences) using a Protean Isoelectrofocusing (IEF) cell (Bio-Rad), as follows: Active rehydration was performed at 20°C for 13 h at 50 V; then the focusing itself was carried out. For improved sample entry, the voltage was increased step by step from 50 to 10,000 V (0.5 h at 200 V, 0.5 h at 500 V, 1 h at 1,000 V, then 10,000 V for a total of 84,000 Vh). After IEF, strips were equilibrated to improve protein transfer to the two-dimensional gel (2-D gel). The second separation was performed in an 11% SDS–PAGE gel. Separation was carried out at 20 V for 1 h and subsequently at a maximum of 30 mA/gel and 120 V overnight, until the bromophenol blue front had reached the end of the gel. After SDS-PAGE, the gels were subsequently stained with colloidal Coomassie blue. Scanning was carried out at 300 dpi with a 16-bit greyscale pixel depth using an image scanner (Amersham Biosciences), and then gel images were analyzed using the Progenesis and SameSpot software packages (Nonlinear Dynamics Ltd). The SAS package (specifically, the GLM procedure for one way ANOVA analysis) was used to examine modifications of individual protein spot volumes. For each protein spot, the mean normalized volume was then computed separately at the V and M plant developmental stages. A protein spot was selected if its variation had a p-value < 0.05.
6.4.8 Protein Identification by LC-MS/MS

Spot digestion and LC-MS/MS were performed as described by Martin et al., 2006. In-gel digestion was performed with the Progest system (Genomic Solution). Gel pieces were washed twice by successive separate baths of 10% acetic acid, 40% ethanol, and acetonitrile (ACN). The pieces were then washed twice with successive baths of 25 mM NH$_4$CO$_3$ and ACN. Digestion was subsequently performed for 6 h at 37°C with 125 ng of modified trypsin (Promega) dissolved in 20% methanol and 20 mM NH$_4$CO$_3$. The peptides were extracted successively with 2% trifluoroacetic acid (TFA) and 50% ACN and then with ACN. Peptide extracts were dried in a vacuum centrifuge and suspended in 20 mL of 0.05% TFA, 0.05% formic acid, and 2% ACN. HPLC was performed on an Ultimate LC system combined with a Famos Autosampler, and a Switchos II microcolumn switch system (Dionex). Trypsin digestion was declared with one possible cleavage. Cys carboxyamidomethylation and Met oxidation were set to static and variable modifications, respectively. A multiple-threshold filter was applied at the peptide level: Xcorr magnitude were up to 1.7, 2.2, 3.3, and 4.3 for peptides with one, two, three, and four isotopic charges, respectively; peptide probability lower than 0.05, ΔCn > 0.1 with a minimum of two different peptides for an identified protein. A database search was performed with Bioworks 3.3.1 (Thermo Electron). The TIGR maize gene index database v 16, 72047*6 EST sequences (http://compbio.dfci.harvard.edu/tgi/plant.html) was used.

6.4.9 Metabolite Extraction and Analyses

Lyophilized leaf material was used for metabolite extraction. Approximately 20 mg of the powder was extracted in 1ml of 80% ethanol/20% distilled water for an hour at 4°C. During extraction, the samples were continuously agitated and then centrifuged for 5 min at 15,000 rpm. The supernatant was removed and the pellet was subjected to a further extraction in 60% ethanol and finally in water at 4°C, as described above. All supernatants were combined to form the aqueous alcoholic extract.
Nitrate was determined by the method of Cataldo et al. (1975) [394]. Total free amino acids were determined by the Rosen colorimetric method with leucine as a standard [395]. Chlorophyll was estimated using 10 mg of fresh leaf material [396]. The total N content of 2 mg of lyophilized material was determined in a N elemental analyzer using the combustion method of Dumas (Flash 2000, Thermo Scientific, Cergy-Pontoise, France). Starch content was determined as described by Ferrario-Méry et al. (1998) [397].

Total lipids were extracted from frozen leaf material according to Miquel and Browse (1992) [398]. Individual lipids were purified from the extracts by one-dimensional thin layer chromatography on silica gel 60 [399]. Silica gel 60 plates were from Merck (Merck-Millipore, Molsheim, France). Lipids were located by spraying the plates with solution of 0.001% primuline (Sigma, Saint-Quentin Fallavier, France) in 80% acetone, followed by visualization under ultraviolet light. To determine the fatty acid composition and relative amounts of individual lipids, the silica gel from each lipid was transferred to a screw-capped tube with 1 ml of 2.5% (v/v) H$_2$SO$_4$ in methanol and an appropriate amount of C17:0 fatty acid (Sigma, Saint-Quentin Fallavier, France), as an internal standard. After heating for 90 minutes at 80°C, 1 ml of hexane and 1.5 ml of 0.9% NaCl$_2$ were added. Fatty acids were extracted in the upper organic phase by shaking and low-speed centrifugation. Samples (1 µl) of the organic phase were separated by gas chromatography on a 30-m x 0.53-mm ECTM-WAX column (Alltech Associates Inc., Deerfield, USA) and quantified using a flame ionization detector. The gas chromatograph was programmed for an initial temperature of 160°C for 1 min, followed by an increase of 20°C min$^{-1}$ to 190°C and a ramp of 4°C min$^{-1}$ to 230°C, with a 9 min hold of the final temperature.

Monosaccharide composition and linkage analysis of polysaccharides were performed on an alcohol insoluble material prepared as follows. Hundred mg (FW) of grounded leaf were washed twice in 4 volumes of absolute ethanol for 15 min, then rinsed twice in 4 volumes of acetone at room temperature for 10 min and left to dry under a fume hood.
overnight at room temperature. Neutral monosaccharide composition was measured on 5 mg of dried alcohol insoluble material after hydrolysis in 2.5 M trifluoroacetic acid for 1.5 h at 100 °C as described by Harholt et al. (2006) [400]. To determine the cellulose content, the residual pellet obtained after the monosaccharide analysis was rinsed twice with ten volumes of water and hydrolysed with H2SO4 as described [401]. The released glucose was diluted 500 times and then quantified using an HPAEC-PAD chromatography as described by Harholt et al (2006) [402].

For lignin quantification, 100mg (FW) of grounded leaf were washed twice in 4 volumes of absolute ethanol for 15 min, twice with 4 volumes of water at room temperature then rinsed twice in 4 volumes of acetone at room temperature for 10 min and left to dry under a fume hood overnight at room temperature. The following protocol is adapted from Fukushima and Hatfield (2001) [403]. Lignins from the prepared cell wall residue was solubilized in 1mL of acetyl bromide solution [acetyl bromide/acetic acid (1/3, V/V)] in a glass vial at 55°C for 2.5 hours under shaking. Samples were then let to cool down at room temperature and 1.2 mL of NaOH 2M/Acetic acid (9/50 V/V) was added in the vial. Hundred μL of this sample was transferred in 300 μL of 0.5M Hydroxylamine Chlorhydrate and mixed with 1.4 mL of Acetic Acid. The A280 absorbance of the samples was measured. Lignin content was calculated using the following formula:

\[
%\text{lignin} = \frac{100 \times (A_{280} \times V_{\text{reaction}} \times V_{\text{dilution}})}{20 \times V_{\text{sample}} \times m_{\text{sample}} \times mg}
\]

6.4.10 Metabolome Analysis

All steps were adapted from the original protocol described by Fiehn (2006) [404]. All extraction steps were performed in 2 ml Safelock Eppendorf tubes. The ground frozen leaf samples were resuspended in 1 ml of frozen (−20°C) Water:Chloroform:Methanol (1:1:2.5) and extracted for 10 min at 4°C with shaking at 1400 rpm in an Eppendorf Thermomixer. Insoluble material was removed by centrifugation and 900 μl of the
supernatant were mixed with 20 µl of 200 µg/ml ribitol in methanol. Water (360 µl) was then added and after mixing and centrifugation, 50 µl of the upper polar phase were collected and dried for 3h in a Speed-Vac and stored at -80°C. Four blank tubes were subjected to the same steps as the samples. For derivatization, samples were removed from -80°C storage, warmed 15 min before opening and Speed-Vac dried for 1 h before the addition of 10 µl of 20 mg/ml methoxyamine in pyridine. The reactions with the individual samples, blanks and amino acid standards were performed for 90 min at 28°C with continuous shaking in an Eppendorf thermomixer. 90 µl of N-methyl-N-trimethylsilyl-trifluoroacetamide (MSTFA) were then added and the reaction continued for 30 min at 37°C. After cooling, 50 µl of the reaction mixture were transferred to an Agilent vial for injection. For the analysis, 3 h and 20 min after derivatization, 1 µl of the derivatized samples were injected in the Splitless mode onto an Agilent 7890A gas chromatograph (GC) coupled to an Agilent 5975C mass spectrometer (MS). The column used was an Rxi-5SilMS from Restek (30 m with 10 m Integra-Guard column). The liner (Restek # 20994) was changed before each series of analyses and 10 cm of the column were removed. The oven temperature ramp was 70°C for 7 min, then increased to 10°C/min until 325°C, which was maintained for 4 min. Overall the total run time was 36.5 min. A constant flow of helium was maintained at 1.5231 mL/min. Temperatures in the GC were the following: injector: 250°C, transfer line: 290°C, source: 250°C and quadripole 150°C. Samples and blanks were randomized. Amino acid standards were injected at the beginning and end of the analyses for monitoring of the derivatization stability. An alkane mix (C10, C12, C15, C19, C22, C28, C32, C36) was injected in the middle of the analyses for external RI calibration. Five scans per second were acquired.

For data processing, Raw Agilent data files were converted into the NetCDF format and analyzed with AMDIS [http://chemdata.nist.gov/mass-spc/amdis/]. A home retention indices/mass spectra library built from the NIST, Golm, and Fiehn databases and standard compounds, was used for metabolite identification. Peak areas were then determined using the quanlynx software (Waters) after conversion of the NetCDF file into the masslynx format. Statistical analyses were carried out with TMEV [http://www.tm4.org/mev.html]. Univariate analyses by permutation (1-way ANOVA and
2-way ANOVA) were first used to select the metabolites exhibiting significant changes in their concentration. An order of magnitude for metabolite concentration on a fresh weight basis was calculated by a one point external calibration comparatively to a set of standards.

### 6.4.11 Model Development and Curation

Figure 6-4 outlines the workflow used for model development. Our previously developed maize model, iRS1563 [334] and biological databases such as MetaCrop (downloaded in December 2012) [333] and MaizeCyc (version 2.0.2) [332] provided information pertaining to the genes, proteins, reactions, and metabolites used to reconstruct the second-generation maize leaf genome-scale model. In addition, available proteomic and transcriptomic data, maize-specific biological databases, namely Metacrop and Maizecyc and published literature were used to assign cellular (i.e., bundle-sheath or mesophyll) and intra-cellular organelle specificity to the curated reactions.

When the gene expression level is reported in reads per kilobase per million mapped reads (RPKM) [363, 364] the cell specificity of any gene \(i\) can be calculated as:

\[
R_i = \frac{|m_i - b_i|}{\max(m_i, b_i)}
\]

Here, \(m_i\) and \(b_i\) are the RPKM abundance of gene \(i\) in the mesophyll and bundle sheath cells, respectively [364]. A gene that is only expressed in one cell type will have a \(R_i\) of 1, while a gene that is equally expressed in both cell types will have a \(R_i\) of 0. As suggested by Chang et al., a threshold of 0.8 or a 5-fold abundance difference is adopted to assign gene cell type specificity. In the absence of RPKM information, an adjusted spectral count (adjSPC) along with the fold change difference between the mesophyll and bundle sheath cells was used to determine gene cell type specificity [362]. adjSPC is the number of mass spectra identified for a protein normalized by the number of unique spectral counts. Since low counts are not statistically informative, a cutoff of 10 was used for adjSPC [405, 406]. Similar to the threshold used for RPKM data, a 5-fold difference
between the mesophyll and bundle sheath cell type normalized spectral abundance factor (nSAF) was used to determine cellular specificity of any gene [362]. nSAF is a weighted adjSPC based on the number of theoretical tryptic peptides with a relevant length [362, 407]. Additional intracellular compartmentalization was carried out based on the MetaCrop database [333], MaizeCyc database [332], and primary literature sources [364, 408].

**Figure 6-4:** Model development and curation schematic.

The workflow for the second-generation genome-scale metabolic model of the maize leaf is displayed. The data sources give three types of retrieved data (i.e. the raw reaction data, reaction directionality, and compartmentalization) that are then manipulated as shown to create the final model.
The intracellular compartmentalization was determined based first on the MetaCrop database [333], literature sources [362, 364], compartmentalization information in the MaizeCyc database, and finally the Plant Proteome DataBase (PPDB) [290]. An original set of intercellular and intracellular transporters was determined based on literature evidence [362, 409-417]. In the subsequent standardization step, the MetRxn knowledgebase [1] as well as manual curation was used to standardize the description of metabolites and reactions such as fixing stoichiometric errors (i.e., elemental or charge imbalances) and incomplete atomistic detail (e.g. absence of stereo-specificity, and presence of R-group(s)). Reactions and metabolites were given KEGG identifiers where available or were otherwise given new identifiers (in the form of MR or MC, respectively). Reaction directionality was adopted from the manually curated MetaCrop database, as available, and from the MaizeCyc database for the remaining reactions.

In the next step of model development, all reactions (including metabolic, intra- and extra-cellular transport reactions) were divided into two categories based on the evidence of their inter- and intra-cellular compartmental specificity. The core set contains all metabolic reactions with experimental or literature-backed evidence of intracellular or intercellular compartmentalization, as well as known intracellular and intercellular transporters. The non-core set contains reactions with partial or completely absent localization information. Barring any conflicting evidence, these reactions were provisionally placed in all compartments. An optimization formulation (as shown below) was developed by imposing flow though the maximal number of core reactions while including minimal intra- and inter-cellular transporters and minimal participation of non-core reactions in various compartments. A parsimony criterion was used to apportion non-core functions so as core functions could be restored. Furthermore, in order to restore a core function the resolution strategy was prioritized in the following order: (i) apportion non-core reaction(s) in one/multiple compartment(s), (ii) add intra-cellular transporter(s) and (iii) add inter-cellular transporter(s). To this end, an objective function was formulated by taking the weighted sum of number of non-core reactions, intra- and inter-
cellular transporters by providing weights of $10^4$ and $10^6$, respectively for these three groups of reactions. However, it is important to ensure that any resolution strategy does not cause thermodynamically infeasible cycles. Therefore, each of these solutions was further checked and only those that do not form such cycles were accepted. This approach is analogous to the one proposed by Mintz-Oron et al. (2012) [418] but does not rely on a complicated scoring system. It is also computationally less taxing as it activates one core reaction at a time. Furthermore, in contrast to the Mintz-Oron et al. approach, the method proposed here allows for the minimal number of transporters added, rather than potentially minimizing the flux through many transporters.

In order to allow flux through all reactions in the core set $C=\{1,\ldots,c\}$ we minimize the addition of reactions from the non-core set $NC=\{c+1,\ldots,g\}$, intracellular transporter set $T=\{g+1,\ldots,t\}$, and intercellular transporter set $IC=\{t+1,\ldots,m\}$. This encompasses an overall set of reactions $M=\{1,\ldots,m\}$ and a set of metabolites $N=\{1,\ldots,n\}$. In addition, binary variable $y_j$ is defined as:

$$y_j = \begin{cases} 1 & \text{if the reaction is added to the model from NC, I, or IC sets} \\ 0 & \text{otherwise} \end{cases}$$

The task of identifying the minimal set of additional reactions that enable flux through a core reaction $j^*$ is posed as the following mixed integer linear programming problem.

Minimize $c_1 \sum_{j=NC} y_j + c_2 \sum_{j=I} y_j + c_3 \sum_{j=IC} y_j \forall j \in C$ \hspace{1cm} (1)

Subject to:

$$\sum_{j=1}^{m} S_{ij} y_j = 0 \forall i \in 1,\ldots, n$$ \hspace{1cm} (2)

$$v_j^* \geq \epsilon \quad \forall j \in C$$ \hspace{1cm} (3)

$$v_{j,max} \geq v_j \geq v_{j,min} \forall j \in C$$ \hspace{1cm} (4)

$$v_{j,max} y_j \geq v_j \geq v_{j,min} y_j \forall j \in NC \text{ or } T \text{ or } I \text{ or } IC$$ \hspace{1cm} (5)

Here, $S_{ij}$ is the stoichiometric coefficient of metabolite $i$ in reaction $j$ and $v_j$ is the flux value of reaction $j$. Parameters $v_{j,min}$ and $v_{j,max}$ denote the minimum and maximum
allowable fluxes for reaction \( j \), respectively. \( v_{j*} \) represents the core reaction flux that is currently being unblocked and \( \varepsilon \) is a small value to ensure a threshold amount of flux through each core reaction. \( c_1, c_2, \) and \( c_3 \) represent weights associated with each set of reactions (i.e., non-core set, intracellular transporters set, and intercellular transporters set, respectively). In this formulation, the objective function (1) minimizes the number of added reactions (from three reaction sets as mentioned earlier) so as to restore flux flow through reaction \( j^* \). We chose values of 1, \( 10^4 \), and \( 10^6 \) for \( c_1, c_2, \) and \( c_3 \), respectively, so metabolic reactions without experimental or literature evidence for compartmental specificity are added to specific compartment(s) before including additional transport reactions with no literature evidence. Constraint set (2) represents the pseudo-steady state assumption, while constraint (3) determines the threshold amount of flux necessary through \( j^* \). Bounds on core reaction fluxes are imposed by constraint set (4), while constraint set (5) ensures that only reactions from those three sets having non-zero flow are added to the model. This algorithm is repeated for each core reaction \( j^* \) to ensure flux and, hence, provides compartmentalization assignments for 431 metabolic reactions by assigning them to at least one compartment, adding 1,032 total metabolic reactions to the model as shown in Table 6-1.

The reactions identified by the above-mentioned algorithm plus the reactions from the core set constituted two new sets, a set of reactions with resolved compartmental information and a set whose location still needs resolution as shown in Figure 6-4. Reactions from the latter set that are known to occur within the maize leaf tissue, but were not in the initial model were added to intra/inter-cellular compartments manually based on pathway localization or simply added to cytosol of bundle sheath and/or mesophyll cells. Thermodynamically infeasible cycles were resolved by changing the minimum number of reaction directionalities as possible and eliminating the smallest number of reactions from the model [419] while conserving biomass formation. An optimization procedure was iteratively run for each reaction in a thermodynamically infeasible cycle to determine the minimum number of directionality changes or removal of reactions required to fix the cycle. These results were then compared for each reaction
to determine the changes that resolve the largest number of reactions participating in thermodynamically infeasible cycles. The solutions found were manually inspected before the changes were applied to the model. The application of this optimization procedure led to restricting the directionality for 507 reactions that prevented 889 reactions from carrying unbounded fluxes thus eliminating the corresponding thermodynamically infeasible cycles.

In the final step, as shown in Figure 6-4, the GapFind/GapFill [295] procedure was applied to identify blocked/dead-end metabolites and subsequently restore their connectivity. A gapfilling database of reactions was created by combining reactions from phylogenetically close/model plant species (i.e. *Oryza sativa japonica*, *Brachypodium distachyon*, *Sorghum bicolor*, and *Arabidopsis thaliana*), non-core reactions without compartmental specificity (not identified by our aforementioned algorithm), and all possible intra/inter-cellular transporters. The gapfilling procedure was modified by prioritizing the addition of reactions from closely related/model plant species or non-core reactions over transporters to unblock the flow through metabolites while ensuring no new thermodynamically infeasible cycles are created. After completing this step, we added 5 reactions from closely related/model plant species, changed the directionality of 14 reactions and added 8 intracellular transporters.

6.4.12 Incorporation of Transcriptomic, Proteomic and Metabolomic Data

Significantly different gene transcripts and proteins were incorporated into the model by switching off corresponding reactions under a uniform WT condition, a limited N condition [377], *gln1*-3 mutant, and *gln1*-4 mutant [335] cases. The number of proteins, gene transcripts, and metabolites with abundances that are statistically differentially expressed in the various conditions are listed in Table 6-2. Reactions with GPRs associated with significantly lowered transcriptomic and proteomic expression are switched off under the corresponding conditions. Metabolite turnover rates were determined based on the flux-sum analysis method [366] and compared to the
metabolomic data. A minimum biomass level of 90% of optimal biomass under the N− WT condition was used in all conditions. The flux-sum or the flow through each metabolite with experimental measurements was maximized as follows:

$$\text{Maximize } 0.5 \sum_i |S_{ij}v_j| \quad \forall \ i \in E$$  \hspace{1cm} (6)

Subject to

$$\sum_{j=1}^{m} S_{ij}v_j = 0 \quad \forall \ i \in 1, \ldots, n$$  \hspace{1cm} (2)

$$v_{j,\text{max}} \geq v_j \geq v_{j,\text{min}} \quad \forall \ j \in 1, \ldots, m$$  \hspace{1cm} (4)

$$v_j = 0 \quad \forall \ j \in LE$$  \hspace{1cm} (7)

Here set $E$ represents the set of metabolites with experimental measurements and set $LE$ represents reactions with statistically lower expression of gene transcripts and/or proteins. The formulation was run for each individual condition ensuring the proper nutrients and simulated knockouts were considered. By linearizing the objective function the resulting formulation is a mixed integer linear programming problem similar to the description by Chung and Lee [366]. Therefore, the basic idea is to maximize the flux-sum of a metabolite (for which metabolomic data is available) under a given condition by switching off reaction fluxes corresponding to gene transcripts and/or proteins with lower expression levels. The flux-sum levels in the N− WT, $gln1-3$ mutant, and $gln1-4$ mutant condition were compared to the reference N+ WT condition to find the qualitative trend in the change of metabolite pool size between the conditions.

The WT condition for each study was combined to create one uniform WT condition. The number of gene transcripts, proteins, and metabolites that statistically vary are displayed below.
Flux variability analysis (FVA) was used to determine the flux range of each reaction under maximum biomass by subsequently maximizing and minimizing the flux through each reaction. The flux range of each reaction for the N+ WT, gln1-3 mutant, and gln1-4 mutant conditions were compared to the reference N+ WT condition. Flux ranges that do not overlap between one of the N background conditions and the reference condition were further analyzed. These are reactions that must change in response to the limited amount of nitrogen or the mutant conditions. Finally, for each condition, the minimum number of reactions which, when not regulated, will restore the biomass to the yield obtained when no “omics” based regulation is applied were determined. This was done by identifying the minimal set of reactions, which are included in the “omics” based regulation, that when active would allow for a biomass yield equivalent to the yield under no “omics” based regulation. This set of reactions represent the reactions whose restriction affects the biomass yield.

The CPLEX solver (version 12.3 IBM ILOG) is used in the GAMS (version 23.3.3, GAMS Development Corporation) environment to solve the optimization problems. The Python programming language is also used during model development (mainly for scripting and data analysis). All computations are carried out on Intel Xeon X5675 Six-Core 3.06 GHz processors constituting the lionxf cluster, which was built and operated by the Research Computing and Cyberinfrastructure Group of The Pennsylvania State University.

<table>
<thead>
<tr>
<th>Type Of Data</th>
<th>WT Condition</th>
<th>N- Condition</th>
<th>gln1-3 Mutant</th>
<th>gln1-4 Mutant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcriptom</td>
<td>256</td>
<td>76</td>
<td>102</td>
<td>53</td>
</tr>
<tr>
<td>Proteomic</td>
<td>38</td>
<td>14</td>
<td>29</td>
<td>-</td>
</tr>
<tr>
<td>Metabolomic</td>
<td>83</td>
<td>20</td>
<td>31</td>
<td>13</td>
</tr>
</tbody>
</table>

Table 6-2: Number of gene transcripts, proteins, and metabolites that significantly vary.
References


options and contents in BRENDA. Nucleic acids research 2013, 41(D1):D764-D772.


46. Hamilton JJ, Dwivedi V, Reed JL: **Quantitative assessment of thermodynamic constraints on the solution space of genome-scale metabolic models.** *Biophysical journal* 2013, **105:**512-522.

47. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.** *Molecular Systems Biology* 2007, **3**.


57. Leighty RW, Antoniewicz MR: **COMPLETE-MFA: Complementary parallel labeling experiments technique for metabolic flux analysis.** *Metabolic engineering* 2013, **20:**49-55.


75. Zomorrodi AR, Lafontaine Rivera JG, Liao JC, Maranas CD: Optimization-driven identification of genetic perturbations accelerates the convergence of


96. Bandyopadhyay A, Stockel J, Min H, Sherman LA, Pakrasi HB: High rates of photobiological H2 production by a cyanobacterium under aerobic conditions. **Nat Commun** 2010, **1**:139.


146. Fernandez-Gonzalez B, Sandmann G, Vioque A: A new type of asymmetrically acting beta-carotene ketolase is required for the synthesis of echinenone in
171


174


Bennett MD, Leitch IJ, Price HJ, Johnston JS: Comparisons with Caenorhabditis (approximately 100 Mb) and Drosophila (approximately 175 Mb) using flow cytometry show genome size in Arabidopsis to be approximately 157 Mb and thus approximately 25% larger than the Arabidopsis genome initiative estimate of approximately 125 Mb. *Ann Bot* 2003, **91**(5):547-557.


Kumar VS, Dasika MS, Maranas CD: Optimization based automated curation of metabolic reconstructions. *BMC bioinformatics* 2007, **8**:212.


185


400. Harholt J, Jensen JK, Sorensen SO, Orfíla C, Pauly M, Scheller HV: ARABINAN DEFICIENT 1 is a putative arabinosyltransferase involved in


Appendix

SBML files of all the genome-scale metabolic models (as presented in this dissertation) could be accessible via the web page of Maranas lab (http://maranas.che.psu.edu/models.htm). In addition, the supplementary files/info of each of the publications (related to this dissertation) can be obtained via each of these journals’ individual web pages.
VITA

Rajib Saha

Education

<table>
<thead>
<tr>
<th>Institute</th>
<th>Field of study</th>
<th>Degree</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penn State University</td>
<td>Chemical Engineering</td>
<td>PhD</td>
<td>2009-2014</td>
</tr>
<tr>
<td>Penn State University</td>
<td>Chemical Engineering</td>
<td>MS</td>
<td>2009-2011</td>
</tr>
<tr>
<td>Bangladesh University of Engineering &amp; Technology</td>
<td>Chemical Engineering</td>
<td>BS</td>
<td>2000-2005</td>
</tr>
</tbody>
</table>

Honors and Awards

- 2012 Genomic Sciences Meeting Student Travel Grant, Department of Energy (DOE), Bethesda, MD
- 2014 NSF N$_2$ Meeting (PI retreat) Student Travel Grant, San Francisco, CA

Publications


* These authors contributed equally.