Metabolic Network Reconstruction and Modeling of Microbial Communities

by

Yu Chen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Chemical Engineering)
in the University of Michigan
2012

Doctoral Committee:

Assistant Professor Xiaoxia Nina Lin, Chair
Professor Erdogan Gulari
Professor Lutgarde M. Raskin
Peter J. Woolf, Foodwiki LLC

ACKNOWLEDGEMENTS

My first acknowledgment goes to Dr. Xiaoxia Lin as my doctoral advisor. She patiently provided the vision, encouragement and advice for me to proceed through the doctorial program and my dissertation. Her wide knowledge and critical thinking have been of great value for me and will sitll guide me during my lifetime studies. Special thanks to my committee members, Dr. Erdogan Gulari, Dr. Lutgarde M. Raskin and Dr. Peter J. Woolf, for their support, guidance and helpful suggestions. Their guidance has served me well and I owe them my heartfelt appreciation.

Thanks should be given to all the members in the group of Dr. Lin. Dr. Jihyang Park, Dr. Marjan Varedi, Dr. Fengming Lin, Jeremy Minty, Alissa Kerner, Mike Nelson, Steven Wang, and Mathieu Rossion have given me useful suggestions on my studies, as well as preparing this dissertation. Marc E. Singer also helped me in organizing TransportDB datasets.

I wish to thank all the faculty members and staff in the Chemical Engineering Department. Without their help, I would never finish my PhD study. I owe my thanks to many graduate students in the department, who have given me kind help for both my class and GSI work.

I thank Dr. Manhong Dai and Dr. Fan Meng at the University of Michigan for providing technical support on a high-performance computer cluster. Dr. James Cavalcoli and Mr. Zachary Wright at the UM Center for Computational Medicine and Bioinformatics helped develop the web interface of PEER. I also need to thank Dr. Jill Banfield and Daniela Goltsman for providing some of the datasets and valuable suggestions along the work. I also thank Dr. Gregory Dick for helpful discussions about the Acid Mine Drainage(AMD) microbial community. Dr. Ian T. Paulsen at Macquarie University helped me in predicting transporters and transmembrane proteins from AMD metagenomic sequences. This project was partly supported by a University of Michigan Rackham Faculty Research Grant.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF FIGURES vii

LIST OF TABLES xii

LIST OF APPENDICES xiv

ABSTRACT xv

CHAPTER

I. Introduction 1

1.1 Genome-scale Metabolic Network Reconstruction 1
1.1.1 *In silico* Reconstructed Metabolic Networks and Their Applications 1
1.1.2 Metabolic Network Reconstruction Methods 8
1.2 Metabolic Network Modeling of Microbial Communities . . . 13
1.2.1 Meta-genomics on Microbial Communities 13
1.2.2 Metabolic Modeling of Microbial Communities . . . 15
1.3 Dissertation Overview 18

II. Bioinformatic Pipeline for Automated Genome-scale Metabolic Network Reconstruction 22

2.1 Introduction and Background 22
2.2 Pipeline for Metabolic Network Reconstruction (PEER) . . . 25
2.2.1 Preliminary Network Generation 27
2.2.2 Mixed Integer Linear Programming (MILP) Based Network Curation 30
2.2.3 Probability of Metabolic Network Prediction 32
2.2.4 Online Tool for Automated Metabolic Network Reconstruction 34

2.3 Automated Metabolic Network Reconstruction for Model Organism *Escherichia coli* 34
2.3.1 Automated Metabolic Network Reconstructions of *Escherichia coli* from Three Datasets 35
2.3.2 Comparison of Automated Metabolic Network Reconstruction With Reference Metabolic Networks 37
2.4 Automated Metabolic Network Reconstruction for Twelve Strains of *Prochlorococcus marinus* 40
2.4.1 Pan and Core Metabolic Networks of *P. marinus* . . 43
2.5 Discussion and Conclusions 43

III. Metabolic Network Reconstruction of Acid Mine Drainage Biofilms 50

3.1 Introduction 50
3.2 Metabolic Network Reconstruction and Modeling Methods . . 53
3.2.1 Metabolic Network Reconstruction for Individual Organisms in AMD Biofilm 53
3.2.2 Prediction of Pathway Probability 54
3.2.3 Multiple-organism Metabolic Model of Microbial Community 55
3.3 Reconstructed Metabolic Networks of Individual Organisms in AMD Biofilm 56
3.3.1 Metabolic Network of the Five Organisms 56
3.3.2 Network Gap Filling in the Reconstructed Networks 58
3.3.3 Carbon Flows in the AMD Biofilm 62
3.3.4 Pathway Distribution in the AMD Biofilm 65
3.4 Multiple-Organism Metabolic Modeling of AMD Biofilm . . . 67
3.4.1 Two-species Model for Major Organisms in the AMD Biofilm 67
3.4.2 Five-species Model for All the Organisms in the AMD Biofilm 73
3.4.3 Incorporation of Proteome Data in Metabolic Network Reconstruction 75
3.5 Discussion and Conclusions 78

IV. Metabolic Network Modeling of Gastrointestinal Microbial Communities 84

4.1 Introduction 84
4.2 Metabolic Network Modeling of a Two-species Model Microbial Community 88
4.2.1 Background 88

4.2.2 Three-step Metabolic Network Reconstruction and Modeling Method for Two-Species Model 93
4.2.3 Results 99
4.3 Metabolic Network Modeling of a Ten-species Model Microbial Community 108
4.3.1 Background 108
4.3.2 Three-step Metabolic Network Reconstruction and Modeling Method 110
4.3.3 Results 114
4.4 Discussion and Conclusions 123

V. Metabolic and Regulatory Network Design for Biochemical Production 128

5.1 Introduction 128
5.1.1 Metabolic and Regulatory Network Design 128
5.1.2 Fatty Acids As Potential Biofuel Precursors 131
5.2 Bi-level Optimization Framework for Metabolic and Regulatory Network Design 133
5.2.1 Assumptions and Simplifications 134
5.2.2 Bi-level optimization Framework 137
5.3 Strain Network Design for Fatty Acids Derived Hydrocarbons 141
5.3.1 Metabolic and Regulatory Networks of *Escherichia coli* 141
5.3.2 Experimental Verification of Designed Strain for Fatty Acid Production 148
5.4 Discussion and Conclusions 153

VI. Concluding Remarks and Future Directions 156

6.1 Concluding Remarks 156
6.2 Future Directions 160
6.2.1 Future Directions for Metabolic Network Reconstruction and Modeling 162
6.2.2 Metabolic Network Modeling of Real Human Gut Microbiota 164

APPENDICES 170

BIBLIOGRAPHY 184

LIST OF FIGURES

Figure

1.1 Overview of the genome-scale metabolic network of *Escherichia coli*. Adapted from Kyoto Encyclopedia of Genes and Genomes (KEGG) database. Nodes represent metabolites and edges represent metabolic reactions. Cofactors and some small molecules such as water are omitted for simplicity. 3

1.2 Growth of *Escherichia coli* K-12 on malate (*Ibarra et al.*, 2002). a. The line of optimality (LO, in red) predicted by FBA for *Escherichia coli* under malate-oxygen condition. Open circles are data points collected in separate experiments. b. Three-dimensional representation of growth rates. The x and y axes represent the same variables as in a. The z axis represents the cellular growth rate (h^{-1}). OUR, oxygen uptake rate; MUR, malate uptake rate. 4

1.3 The procedure to iteratively reconstruct metabolic networks as described in the protocol suggested by *Thiele and Palsson*(2010). The iterative steps should continue until the model predictions are close to the experimental phenotypic characteristics of the organism. 9

2.1 Flow chart for Pipeline for Metabolic Network Reconstruction (PEER) 26

2.2 Two examples of problematic metabolic reactions that commonly happen during the metabolic reconstruction process. 29

2.3 Demonstration of the MILP model for metabolic network gap filling. Red: binary variables, blue: continuous variables, black: parameters. Complete model descriptions can be found in Appendix B. 31

2.4 Phenotype phase plane (PPP)(a) and lines of optimality (LO)(b) for *Escherichia coli* K12 MG1655 predicted by automated metabolic network reconstruction. 48

2.5 a) Metabolic reactions and b) pathway information for twelve strains of *P. marinus* predicted by automated metabolic network reconstruction. 49

3.1 The ecological roles of Acid Mine Drainage (AMD) biofilm in AMD formation. 51

3.2 Metabolic network of the species in a AMD microbial community. a) Darkness of each bar represents the probability of existence. Both the reaction (column) and the species (row) were grouped by dendrogram. b) Metabolic network of *Leptospirillum* Gp II with 283 intracellular reactions and 288 metabolites, drawn by Pajek (*Bategeli and Mrvar*, 2003). 59
3.3 The carbon flow of the two dominant species in the AMD microbial community. The seven potential metabolic reactions directly related to carbon fixation in *Leptospirillum* groups are highlighted. 63
3.4 The carbon flow of *Ferroplasma acidarmanus* I and II (a) and *Thermoplasmatales archaeon* Gp1 (b). The five potential metabolic reactions directly related to carbon fixation in *Ferroplasma acidarmanus* and three in *Thermoplasmatales archaeon* are highlighted (Continued on next page). 64
3.4 The carbon flow of *Ferroplasma acidarmanus* I and II (a) and *Thermoplasmatales archaeon* Gp1 (b). The five potential metabolic reactions directly related to carbon fixation in *Ferroplasma acidarmanus* and three in *Thermoplasmatales archaeon* are highlighted. 65
3.5 a)Number of active reactions in pathways of the AMD biofilm. b) Active metabolic pathway distribution in 5wayCG dataset of AMD microbial community. 67
3.6 Active metabolic pathway distribution in UBA dataset of AMD microbial community. 68
3.7 The interactions in AMD microbial community in 2-organism model. a) Summary of predicted interactions between the two dominant organisms and interactions between microorganisms with the environment. b) Pathways related to three predicted interactions, which can explain the mechanisms for these interactions. 71
3.8 The interactions in AMD microbial community in 2-organism model. Complete predictions of interactions between *Leptospirillum* Gp II and III. Left for UBA dataset and right for 5wayCG dataset. 72
3.9 Potential interactions among five organisms in the AMD microbial community of 5wayCG site under three penalty levels in 5-organism model. The darkness of colors represents the probability of the prediction. For example, the darkest interactions shown in the highest penalty level are those most likely interactions predicted by the 5-organism model. 74
4.1 Genes involved in carbohydrate metabolism and energy production for *B. thetaiotaomicron* and *E. rectale* (adapted from *Mahowald et al.*, 2009). Red, enriched; blue, depleted; darker color $P \leq 0.001$ and light color $P \leq 0.05$ relative to the average of all Firmicute genomes. 90

4.2 Flow chart of three-step modeling procedure for the 2-species microbial community. The methods used in the three steps are briefly described. Detailed information is provided in Section 4.2.2. 92

4.3 Demonstration of the multi-condition MILP model for metabolic network gap filling. Red: binary variables, blue: continuous variables, black: parameters 94

4.4 Pathway analysis of the reconstructed metabolic network of *Bacteroides thetaiotaomicron* and *Eubacterium rectale*. Ordered by the abundance of the pathways. 101

4.5 Growth phenotypes of *B. thetaiotaomicron* and *E. rectale* on different carbon sources. BTH: *Bacteroides thetaiotaomicron*, ERE: *Eubacterium rectale*. Two false positive predictions are highlighted. 102

4.6 False positive and negative predictions of the model. BT: *Bacteroides thetaiotaomicron*; ER: *Eubacterium rectale*; Mix: co-inoculated; Iso: mono-inoculated. 104

4.7 Pathway activities in both mono- and co- inoculated conditions. BTH: *Bacteroides thetaiotaomicron*; ERE: *Eubacterium rectale*; Mixed: co-inoculated; Iso: mono-inoculated. 105

4.8 Interactions between *E. rectale* and *B. thetaiotaomicron* predicted by 2-species model. The grayscale indicates the probability of the predictions. 106

4.9 Flow chart of three-step modeling procedure for the ten-species microbial community. The methods used in the three steps are briefly described. More details are in Section 4.3.2 111

4.10 The most abundant metabolic pathways in the metabolic networks of the ten species. Putative reactions (metabolic gaps) are also included. The species are ordered by the total number of reactions in these pathways. 117

4.11 Fractions of predicted active reaction in the total active reactions accroding to expression data. The results predicted under four ratio parameters were shown. The results without gene expression data are provided for comparison. 118

4.12 Active metabolic pathways in the metabolic networks of the ten species. The pathway active fraction is the fraction of active reactions in all the reaction associated with the pathway. The 20 pathways with most active reactions were listed. 119

4.13 Fractions of active reaction in the reaction predicted to be active accroding to expression using MinExchange method under four ratio parameters and compared with complete mixed condition. Error bars were calculated from the selected suboptimal solutions. 120

4.14 Predicted uptake fluxes under four settings of ratio parameter in 10-species model. The complete predictions are reported for ratio = 10, and for the rest three settings, only molecules found in gut environment were reported, which are almost constant under all three settings. 122

4.15 Predicted minimal byproducts in the ten-species microbial community. The uptake fluxes predicted under 0.1 ratio setting were reported. 123

5.1 Example of boolean network and logic rules 135

5.2 Flow chart of regulatory Flux Balance Analysis(rFBA) 136

5.3 Structure of bi-level metabolic network and regulatory network design model. R: all reactions, M: all metabolites, Nutrient: nutrient uptake reactions, Inhibited: inhibited reactions, v:variable of flux, v_{Growth}: growth rate, $v_{product}$: rate of product synthesis, $I_{reactive}$: binary variables of re-activate reactions, $I_{knockout}$: eliminated reactions, S: stoichiometric matrix, F_{max}: maximum flux value, F_{supply}: nutrient uptake rates. 138

5.4 Metabolic network of *Escherichia coli* for the design model. The number behind the gene name is the number of transcription factors or signals regulating that reaction. 142

5.5 Designed strain for fatty acids production from glucose and glycerol mixed supply. 147

5.6 Model suggested Strain and Published Strain for Aerobic Succinate Production. a) Left is published strain for aerobic succinate production with 45% to 70% of theoretical yield. Our model predicts this strain with 52% of theoretical yield. b) Right is the optimal strain suggested by out model, with 58% of theoretical yield. 149

5.7 Model predictions of fatty acid production along optimal sequence of genetic manipulations. The results predicted by FBA and MOMA methods are shown in the optimal sequences calculated by the models.151

5.8 Profiling of gene modified strains based on the model prediction. The strains were cultured in M9 minimal medium with 2% glucose for 48 hrs. The total amount of fatty acids was quantified by GC-FID as well as the fatty acid composition was identified. The other fermentation parameters were determined, including the final concentration of by-products lactate and acetate, the growth (OD), and the final glucose concentration. 152

5.9 Total amount of fatty acid of strains suggested by the optimization model. Two culture mediums LB 1-2 and LB 5-10 were tested. All strains were cultured at 30?? for 48 hrs. 7Δ represents strain with ΔcyoA, Δadh, ΔnuoA, Δndh, Δpta, Δpta. 153

6.1 a) Secondary metabolites and byproducts synthesized by the core metabolic network of human gut microbiome. The numbers of metabolic gaps that are needed to fill are listed on the left. The compounds in the biomass composition were not counted; b) the corresponding active reactions for synthesis of the products (red bars). The frequencies of these reactions identified in samples are also listed. Putative: metabolic gaps filled by reaction not in pan metabolic network; Identified: metabolic gaps filled by reaction in pan metabolic network. The side bar indicates the sources of metabolic reactions. Blue: core metabolic networks; red: metabolic gaps. 166

6.2 Interaction results on Three-Phylum-Model of human gut microbiome. Higher penalty leads to more conservative prediction. . 168

LIST OF TABLES

Table

2.1 Metabolic network reconstructed from different input data of *Escherichia coli* K12 MG1655. 36

2.2 Metabolic gaps for metabolic network reconstruction *Escherichia coli* of based on KEGG dataset. 38

2.3 Automated metabolic network reconstructed for twelve strains of *P. marinus*. 41

3.1 Summary of metagenonic dataset of AMD biofilm. Data collected from work of *Tyson et al.* in 2004. 52

3.2 Summary of the individual organism metabolic network reconstruction in the AMD community in UBA site 57

3.3 Network gaps filling and the gene candidates of metabolic network reconstruction of individual organisms in the AMD community in UBA site 60

3.4 Examples of the network gaps and the gene candidates of the AMD microbial community in UBA site 61

3.5 Summary of community-wide metabolic network reconstructions of the AMD microbial community in 5wayCG site. The putative reactions and active reactions of suboptimal solutions within 5% optimality gap were reported. 76

3.6 Number of peptides identified in the proteome dataset for the active reactions 78

4.1 Summary of proteins detected by mass spectrometry and GeneChip for B. thetaiotaomicron and E. rectale. 91

4.2 Summary of the reconstructed metabolic networks of B. thetaiotaomicron and E. rectale. 99

4.3 Comparison of the gene annotations from author's annotation, KEGG annotation and RAST annotation. 100

4.4 Summary of the agreement of model prediction with expression data. BTH: *B. thetaiotaomicron* , ERE: *Eubacterium rectale.* Mixed: co-inoculation, mono: mono-inoculation. 103

4.5 Strains inoculated in the 10-species model community and the transcriptome coverages. The gene with more than 64 sequencing reads were considered as expressed. 109

4.6 Strains in TransportDB used for $Weight_{uptake}$ calculation in the ten-species model. If more than one strains was chosen, the $Weight_{uptake}$ was calculated according to all the transporters identified in the reference strains. 114

4.7 Summary of the reconstructed metabolic networks of the ten species in the model microbial community. 115

5.1 Summary of optimal designs for fatty acid production from D-glucose. The numbers of carbons in optimal fatty acid products are also listed. The results for wildtype strain are highlighted. Production rate is calculated based on 10mM Glucose/hr/g DCW. 145

5.2 Summary of optimal designs for fatty acid production from glycerol, glucose, xylose and their combinations in aerobic condition. The numbers of carbons in optimal fatty acid products are also listed. The results for wildtype strain are highlighted. Production rate is calculated based on 10mM Glucose/hr/g DCW. 146

5.3 Optimal production rate for designed strain from D-glucose. Production rate is calculated based on 10mM Glucose/hr/g DCW. 148

6.1 Potential Gene Classification Methods and Datasets for PEER. . . . 162

A.1 Biomass synthesis function assumed for *E. coli*. This composition was modified from iAF1260 model. Negative value represents requirment. 171

C.1 Minimal byproducts in the two-species microbial community. 178

C.2 Minimal byproducts in the ten-species microbial community. Only the byproducts predicted by all the solutions were listed. 179

D.1 Complete list of metabolites identified in relevant metabolomic studies.181

LIST OF APPENDICES

Appendix

A. Biomass Compositions Assumed for *E. coli* 171

B. Automated Curation Model in PEER 173

C. Byproducts in Model Gut Microbiota 178

D. Metabolites Identified in Metabolomic Studies 181

ABSTRACT

Metabolic Network Reconstruction and Modeling of Microbial Communities

by

Yu Chen

Chair: Xiaoxia Nina Lin

In nature, most microorganisms live in synergistic communities performing important biological functions and ecological roles, such as polysaccharide utilization in the gastrointestinal tract of mammalian hosts. In the past decade, metagenomics has advanced rapidly, providing detailed information of population structures and genetic sequences for microbial communities. In this dissertation, my goal is to develop community-wide metabolic network models to understand the cellular metabolic properties and inter-species relationships in microbial communities.

Metabolic network reconstruction at the (meta)genome scale requires tremendous efforts and time. To address this challenge, we started by developing a bioinformatic pipeline for automated reconstruction of high-quality genome-scale metabolic networks using annotated genomes. It was tested with model bacterium *Escherichia coli.* The results agreed well with a benchmark network manually curated for over a decade. Furthermore, we applied the pipeline to twelve strains of the most abundant cyanobacterium on earth, *Prochlorococcus marinus*, and defined pan and core metabolic networks of the species, demonstrating the utility of the tool.

Next, we extended our bioinformatic pipeline to community-wide metabolic network reconstruction and investigated two types of microbial communities. First, we studied the metagenomes of acid mine drainage biofilms, which cause water pollution in many mining areas. Both individual metabolic networks and community-wide metabolic networks were reconstructed to study the metabolism and inter-species interactions related to biofilm formation. Several essential interactions were predicted. For example, *Leptospirillun* Gp III was predicted to fix nitrogen for the whole community, which was supported by experimental data.

Second, we examined two synthetic gut microbiomes to explore their metabolic capabilities and microbe-microbe-host interactions. For each system, we reconstructed community-wide metabolic networks considering all the species using annotated genomes and transcriptomes through a three-step curation process. With these metabolic networks, we could explain mechanistically metabolic phenotypes and predict inter-species interactions. For instance, for a ten-species microbiome, a number of molecules, including urea, citrate and agmatine, were revealed to be cross-fed.

This dissertation demonstrates that metagenome-scale metabolic network reconstruction and analysis is a promising tool for studying intracellular metabolism and inter-species interactions of microbial communities, which can advance fundamental understanding and provide valuable hypothesis for experimental testing.

CHAPTER I

Introduction

1.1 Genome-scale Metabolic Network Reconstruction

1.1.1 *In silico* Reconstructed Metabolic Networks and Their Applications

Biochemists and biologists have long been occupied by the view of metabolic pathways, which provides researchers an intuitive perspective of cellular metabolisms. In the past decade, the emergence of genomic data has enabled a different approach, metabolic network modeling, for the study of cellular metabolism at a larger genome or even meta-genome scale. Rather than examining individual pathways, genome-scale metabolic network modeling considers cellular metabolism as a whole, which means metabolic reactions are no longer classified by defined pathways. Genome-wide metabolic networks can provide a more comprehensive representation of not only individual metabolic reactions in the organism but also their connections and interactions. In 1999, *Edwards and Palsson* developed the first genome-scale metabolic network for bacterium *Haemophilus influenzae*, the first free-living organism to have its whole genome sequenced (*Edwards and Palsson*, 1999) . In the years that followed, the number of genome-scale metabolic network reconstructions has increased quickly. To date, genome-scale metabolic networks have been manually reconstructed for 56 organisms (`http://gcrg.ucsd.edu/InSilicoOrganism/OtherOrganisms`, retrieved on July 27, 2012) These metabolic networks have greatly advanced system-level knowledge of cellular metabolism and have led to useful models and predictions that can be used to guide experimental studies, e.g. design of mutation strains for fermentation products (*Burgard et al.*, 2003) and prediction of metabolic fluxes in isotopic labeling experiments (*Wiechert*, 2001).

Metabolic reactions link metabolites together to form a metabolic network (*Palsson*, 2006). Therefore, a genome-scale metabolic network represents comprehensive metabolic mechanisms in the cell at the molecular level, together with the associated components, including enzymes, substrates, and products. Figure 1.1 illustrates the genome-scale metabolic network of *Escherichia coli*. As shown in the figure, metabolites are represented by nodes and they are connected by metabolic reactions. The arrows indicate directions and reversibility of the metabolic reactions. If enzymes are involved, the corresponding metabolic reactions are labeled by the EC numbers or enzyme names in the figure. Reaction stochiometry together with reaction directions determine the primary topological properties of the metabolic network. The stochiometry of a metabolic reaction is usually invariant between organisms. In addition, this property should not change with conditions, including pressure, temperature and pH. There are only a few exceptions, including the variable proton efficiency in oxidative phosphorylation caused by proton leakage (*Chance and Williams*, 1955; *Divakaruni and Brand*, 2011). Special attention should be given to these reactions when they are included in metabolic networks.

A genome-scale metabolic network is an integration of several levels of knowledge about cellular metabolism, and has proven to be an effective tool for studying metabolic properties of organisms. Flux Balance Analysis (FBA) is one of

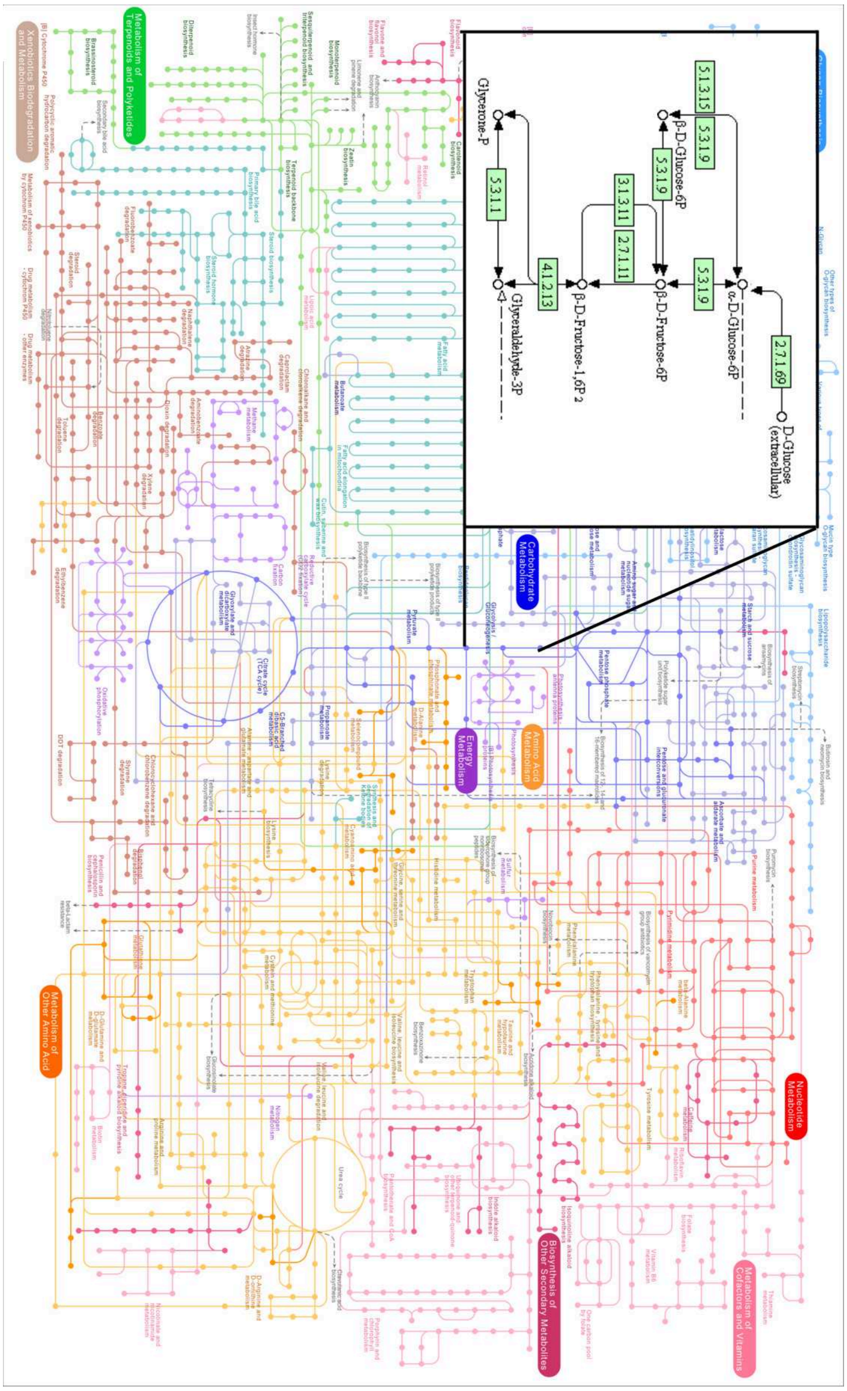


Figure 1.1: Overview of the genome-scale metabolic network of *Escherichia coli*. Adapted from Kyoto Encyclopedia of Genes and Genomes (KEGG) database. Nodes represent metabolites and edges represent metabolic reactions. Cofactors and some small molecules such as water are omitted for simplicity.

the tool that is commonly applied to genome-scale metabolic network, especially for bacteria. In 2001, *Edwards et al.* applied FBA to the first genome-scale metabolic network reconstruction of *Escherichia coli* and accurately predicted the growth rate under different conditions (Figure 1.2). In 2004, *Almaas et al.* used the genome-scale metabolic network of *Escherichia coli* and the FBA model to study the organization of global metabolic fluxes by predicting the flux distributions (*Almaas et al.*, 2004). FBA model have also been applied to predict the lethality (*Ghim et al.*, 2005) and robustness of *Escherichia coli* (*Edwards and Palsson*, 2000b). Besides *Escherichia coli*, FBA have been applied to other bacteria after their genome-scale metabolic networks were reconstructed, such as *Helicobacter pylori* (*Schilling et al.*, 2002), *Bacillus subtilis* (*Oh et al.*, 2007), and *Pseudomonas putida* (*Puchałka et al.*, 2008).

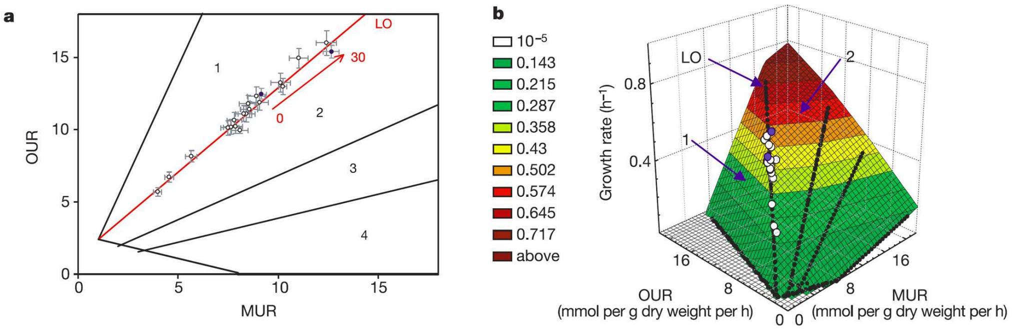


Figure 1.2: Growth of *Escherichia coli* K-12 on malate (*Ibarra et al.*, 2002). a. The line of optimality (LO, in red) predicted by FBA for *Escherichia coli* under malate-oxygen condition. Open circles are data points collected in separate experiments. b. Three-dimensional representation of growth rates. The x and y axes represent the same variables as in a. The z axis represents the cellular growth rate (h^{-1}). OUR, oxygen uptake rate; MUR, malate uptake rate.

FBA has been applied to eukaryotic microorganisms and even multicellular organisms. The metabolic network of *Saccharomyces cerevisiae* has been reconstructed and revised several times. Different from metabolic networks of bacteria, multiple compartments were considered, including cytosol, mitochondria, extracellular peroxisome, nucleus, golgi apparatus, endoplasmic reticulum and vacuole (*Duarte et al.*, 2004). All the metabolites and metabolic reactions were assigned to one of the eight compartments. Exchange reactions were included to enable the transport of metabolites across compartments. Based on these metabolic networks, FBA model were applied to predict the metabolic capabilities (*Förster et al.*, 2003) and *in silico* gene deletion analysis (*Duarte et al.*, 2004). Metabolic network of part of multicellular organisms has been reconstructed and studied. For example, metabolic network of *Homo sapiens* mitochondria were reconstructed and analyzed FBA framework to predict the candidate metabolic network states in human mitochondria under the impacts of diabetes, ischemia, and diet (*Thiele et al.*, 2005). The complete metabolic network of the *Homo sapiens* has already been reconstructed and revised (*Duarte et al.*, 2007; *Hao et al.*, 2012). Similar to *Saccharomyces cerevisiae*, eight compartments (vacuole was replaced by lysosome) were identified in the metabolic network of *Homo sapiens.* This metabolic network has been used to analyze high-throughput biological data sets, e.g. gene expression data. By integrating tissue-specific gene and protein expression data, *Shlomi et al.* predicted human tissue-specific metabolic behaviors under FBA framework. Tissue-specific uptake and secretion of metabolites can be predicted based on this metabolic network (*Shlomi et al.*, 2008).

Utilizing reconstructed genome-scale metabolic networks, Flux Balance Analysis (FBA) type of models can successfully predict cellular behaviors under various conditions. However, the accuracy of FBA models can be greatly improved by considering regulatory information. In 2004, *Covert et al.* analyzed metabolic network of *Escherichia coli* together with the regulatory network (rFBA) and made accurate predictions of growth phenotypes of this bacterium (*Covert et al.*, 2004). In the rFBA framework, the regulatory network was converted into Boolen equations which were used to determine the metabolic-regulatory steady state (MRS). Similarly, Christian Barrett et al., used metabolic network reconstruction with regulatory network of *Escherichia coli* to determine the different states of metabolism under different environmental conditions (*Barrett et al.*, 2005). Along the same direction, *Shlomi et al.* developed steady state regulatory FBA (SR-FBA) (*Shlomi et al.*, 2007). In the SR-FBA framework, the Boolen equations in the rFBA model were converted into linear equations and embeded into the FBA model. SR-FBA is used to predict the activity of reactions that are directly or indirectly controlled by the transcriptional factor. Different from the regulatory flux balance analysis (rFBA and SR-FBA), integrated flux balance analysis (iFBA) applies ordinary differential equations (ODEs) to represent signal pathways and integrate these ODEs into the constraints of FBA (*Covert et al.*, 2008). Another approach researchers have developed to model metabolic network with cellular regulation is probabilistic regulation of metabolism (PROM) (*Chandrasekaran and Price*, 2010). In this method, gene expression data collected from various conditions are used to predict the effects of cellular regulation. With the curation of gene expression data, PROM model can accurately predict the growth phenotypes of mutated *Escherichia coli* under different growth conditions.

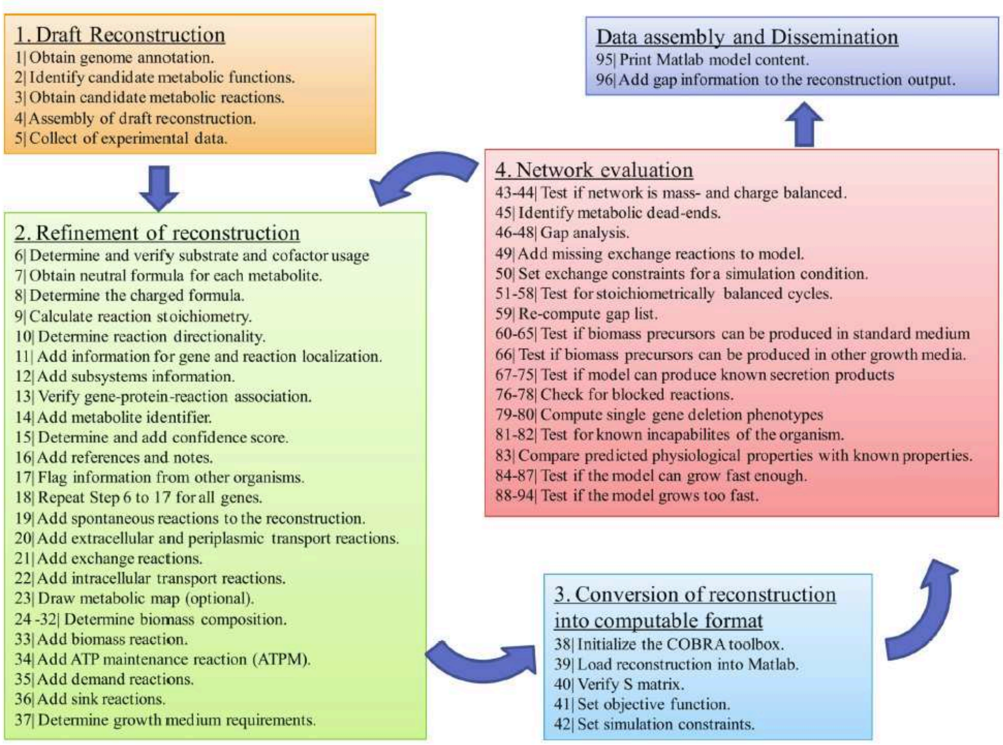
Based on the FBA approach, several other methods have been developed to utilize genome-scale metabolic networks. For example, the Minimization of Metabolic Adjustment (MOMA) method (*Segrè et al.*, 2002) was developed to predict metabolic fluxes after certain gene knockouts. Similar to FBA method, MOMA is also a constraint-based model, which assumes the metabolic states of mutated organisms are closed to the wild-type strains under the same condition. Therefore, in MOMA model, the target organisms are no longer assumed to fully adapted to maximize the growth, which is assumed in FBA model. In a lot cases, the mutated strains have not been evolved enough, so the MOMA model can provide better prediction than FBA. Dynamic FBA (dFBA) is another constraint-based model that can predict cell growth qualitatively (*Mahadevan et al.*, 2002). In the dFBA framework, FBA model is applied to predict the quasi steady growth rates in small time periods, which provide the change of both cell density and concentrations of nutrients and products. Flux variability analysis (FVA) is another method derived from FBA method. The FVA method have been applied to genome-scale metabolic networks of *Escherichia coli* to determine the variability of fluxes for the identification of reactions that are relatively important (*Reed and Palsson*, 2004). Flux Coupling Finder (FCF) framework has been developed to study the topological and flux connectivity features of genome-scale metabolic networks by classifying flux coupling that is indicated by predicted flux values (*Burgard et al.*, 2004).

There are certain applications of genome-scale metabolic networks that are not based on FBA methods. A method of singular value decomposition (SVD) of extreme pathways have been developed and applied to studied the regulation of a human red blood cell metabolism (*Price et al.*, 2003). *Handorf et al.* applied a method of network expansion to predict all possible metabolites that can be produced from defined compounds according to the structure of a metabolic network. Similar method have been utilized to predict possible environmental conditions of a organism based on its metabolic networks (*Handorf et al.*, 2008). Csaba Pal et al. examine the evolution of minimal metabolic networks by predicting contingency-dependent loss of alternative pathways of *Escherichia coli* using *in silico* metabolic reconstructions (*Pál et al.*, 2006). *Wunderlich and Mirny* utilized structure (topology) of genome-scale metabolic network of *Escherichia coli* to predict the viability of the mutant strains. By directly examining the co-occurrence of metabolites in metabolic networks, *Becker et al.* was able to prdict the metabolic relationships between metabolites. Very similarly, conservation relations between metabolites were predicted based on genome-scale metabolic networks *Escherichia coli* and applied to predict the novel growth media (*Imielinski et al.*, 2006).

1.1.2 Metabolic Network Reconstruction Methods

Before the genomic era, metabolic network reconstruction was mainly achieved manually and also required expertise on the knowledge of cellular metabolism for the specific organisms. Without genome-scale gene annotations, researchers tried to reconstruct the whole-cell-scale metabolic network of several organisms just based on primary literature and biochemical characterizations of identified enzymes. The earliest metabolic network reconstructions, including *Clostridium acetobutylicum* (*Papoutsakis*, 1984), *Bacillus subtilis* (*Papoutsakis and Meyer*, 1985a) and *Escherichia coli* (*Papoutsakis and Meyer*, 1985b; *Majewski and Domach*, 1990; *Varma et al.*, 1993b,a), all belong to this category. For some model organisms that are studied in this way, their metabolic reconstructions have been created and revised for a considerable period of time. One example is the genome-scale metabolic network of *Escherichia coli*, which was created and refined several times in the past ten years (*Edwards and Palsson*, 2000a; *Almaas et al.*, 2004; *Feist et al.*, 2007; *Orth et al.*, 2011). These manually reconstructed metabolic networks provide valuable datasets about cellular metabolism, such as collections of metabolic reactions that are being used when reconstructing metabolic networks for other organisms (*Henry et al.*, 2010). Beyond the metabolic reaction sets, researchers also accumulated experiences in reconstructing metabolic networks and summarized them into standard procedures (*Thiele and Palsson*, 2010).

After great amounts of annotated genome sequences became available, the reconstruction process mainly relies on these genomic data. Despite some recent developments, the reconstruction process is still labor intensive and time consuming. In a suggested protocol, there are 98 steps involved to reconstruct a high-quality metabolic network for the genome data (Figure 1.3), which might cost years for a well studied, medium sized bacterial genome (*Thiele and Palsson*, 2010). The situations will become even worse if the organism is not well studied and even not cultivable, for which only limited experimental data are available. Another limitation that constrains the application of manual reconstruction is the inconsistency between different databases and datasets. A important task in the manual reconstruction process is reconciliation of model predictions with experimental data. The inconsistency between different databases and datasets collected manually makes this task much more difficult.



1. Draft Reconstruction
1| Obtain genome annotation.
2| Identify candidate metabolic functions.
3| Obtain candidate metabolic reactions.
4| Assembly of draft reconstruction.
5| Collect of experimental data.

2. Refinement of reconstruction
6| Determine and verify substrate and cofactor usage
7| Obtain neutral formula for each metabolite.
8| Determine the charged formula.
9| Calculate reaction stoichiometry.
10| Determine reaction directionality.
11| Add information for gene and reaction localization.
12| Add subsystems information.
13| Verify gene-protein-reaction association.
14| Add metabolite identifier.
15| Determine and add confidence score.
16| Add references and notes.
17| Flag information from other organisms.
18| Repeat Step 6 to 17 for all genes.
19| Add spontaneous reactions to the reconstruction.
20| Add extracellular and periplasmic transport reactions.
21| Add exchange reactions.
22| Add intracellular transport reactions.
23| Draw metabolic map (optional).
24 -32| Determine biomass composition.
33| Add biomass reaction.
34| Add ATP maintenance reaction (ATPM).
35| Add demand reactions.
36| Add sink reactions.
37| Determine growth medium requirements.

3. Conversion of reconstruction into computable format
38| Initialize the COBRA toolbox.
39| Load reconstruction into Matlab.
40| Verify S matrix.
41| Set objective function.
42| Set simulation constraints.

4. Network evaluation
43-44| Test if network is mass- and charge balanced.
45| Identify metabolic dead-ends.
46-48| Gap analysis.
49| Add missing exchange reactions to model.
50| Set exchange constraints for a simulation condition.
51-58| Test for stoichiometrically balanced cycles.
59| Re-compute gap list.
60-65| Test if biomass precursors can be produced in standard medium
66| Test if biomass precursors can be produced in other growth media.
67-75| Test if model can produce known secretion products
76-78| Check for blocked reactions.
79-80| Compute single gene deletion phenotypes
81-82| Test for known incapabilites of the organism.
83| Compare predicted physiological properties with known properties.
84-87| Test if the model can grow fast enough.
88-94| Test if the model grows too fast.

Data assembly and Dissemination
95| Print Matlab model content.
96| Add gap information to the reconstruction output.

Figure 1.3: The procedure to iteratively reconstruct metabolic networks as described in the protocol suggested by *Thiele and Palsson*(2010). The iterative steps should continue until the model predictions are close to the experimental phenotypic characteristics of the organism.

To overcome the limitations of manual reconstruction, automated or semi-automated procedures that directly generate metabolic reconstructions from annotated genome are of great interest. A number of user-friendly resources have been developed to facilitate this demanding process. Some of these tools are designed for helping manual metabolic network reconstruction. For example, MetaFluxNet (*Lee et al.*, 2003) provides an interface for users to manually input information of metabolic networks and then carries out metabolic flux analysis (MFA). Similarly, MetNetMaker (*Forth et al.*, 2010) allow users to select metabolic reactions from reaction databases by EC number or other information and then generate a metabolic network. METANNOGEN (*Gille et al.*, 2007) and rBioNet (*Thorleifsson and Thiele*, 2011) are effective tools for data management for metabolic network reconstruction process. YANAsquare (*Schwarz et al.*, 2007) provide a interface that connect to KEGG database and allow user to select metabolic reactions in each pathways that are associated with annotated genes. By combined those selected reactions, the tool can generate draft genome-scale metabolic networks.

A number of methods and tools have been developed to facilitate metabolic network curation and gap filling. In 2004, *Green and Karp* designed a Bayesian based method for identifying missing reactions in database. In the same year, *Kharchenko et al.* published an algorithm that is able to fill metabolic gaps in a metabolic network using expression information. Along this line, *Kumar et al.* developed an optimization based procedure that can find and fill metabolic gaps by searching all downstream no-production metabolites and minimizing the number of added reactions. COBRA Toolbox (*Becker et al.*, 2007; *Schellenberger et al.*, 2011) also provides the metabolic gap-filling functions for metabolic network reconstruction. Recently, rBioNet, a COBRA toolbox extension for metabolic network reconstruction, was developed by *Thorleifsson and Thiele*. rBioNet enables users to combined metabolites and reaction database from different sources during the curation process. This function is important when the metabolic gaps are filled by reactions in another database or model. In 2009, *Kumar and Maranas* developed GrowMatch algorithm that can reconcile *in silico/in vivo* growth predictions and revise the metabolic reconstruction at the same time. GrowMatch has been implemented in COBRA toolbox v2.0 (*Schellenberger et al.*, 2011), which provides a more power platform for both metabolic network reconstruction and network analysis.

Several tools have been developed for network analysis and curation rather than creating a new metabolic network. BioMet (*Cvijovic et al.*, 2010), a web-based resource for stoichiometric analysis and integration of transcriptome and interactome data, has been developed. BioMet contains a tool that can convert metabolic networks written in system biology makeup language (SBML) into its own data framework. Acorn (*Sroka et al.*, 2011) is another web tool providing constraint based modeling and visualization for existing genome wide metabolic networks written in SBML. In addition to BioMet and Acorn, OptFlux (*Rocha et al.*, 2010) and SBRT (*Wright and Wagner*, 2008) both provide comprehensive software platform for *in silico* metabolic modeling and engineering.

All these tools support increasingly sophisticated network analyses, but rely largely on existing network models and have very limited capabilities for creating new networks. To our best knowledge, There are two tools available that can reconstruct metabolic network automatically from genome sequence or annotation, including Model SEED (*Henry et al.*, 2010) and GEMSiRV (*Liao et al.*, 2012), which is based on MrBac (*Liao et al.*, 2011). The Model SEED, based on an automated genome annotation tool RAST (*Aziz et al.*, 2008), first compares the annotated genome with a self-maintained database contain both metabolic reactions and associated genes to generate a draft metabolic reconstruction. This draft metabolic network is then refined with minimal modifications to meet the growth requirements pre-assumed in the model (*Satish Kumar et al.*, 2007). However, users cannot specify the growth condition and biomass compositions during the automated curation, which limits its application. The gene candidates of the filled metabolic gaps are not provided by Model SEED, which is important information in metabolic network reconstruction process.

Different form SEED, GEMSiRV (*Liao et al.*, 2012), which is based on MrBac (*Liao et al.*, 2011), can generate a new metabolic network by comparing the genome sequences to a known organism with reconstructed metabolic network. Therefore, those metabolic reactions in the reference metabolic network will be added into the new one if there are genes that are similar to the genes that associated with the metabolic reactions in the reference genome. Sequence alignment is applied to identify these genes by setting a threshold of sequence identify or E value. This method can only apply to those organisms that are close to model organisms, the metabolic network of which have been reconstructed. Otherwise, this method cannot provide reliable metabolic networks due to the diversity of cellular metabolism. Another limitation of this method is it cannot make use the efforts spending in the manual curation of gene annotation because only sequence alignment results are utilized to identify reactions.

All these methods and tools developed for different purposes enable researchers to generate metabolic networks automatically or semi-automatically. However, two major challenges need to be addressed before we can automatically reconstruct high-quality metabolic networks. The major challenge is current automated gap-filling methods can not satisfy the quality requirement. This is because the quality of metabolic networks can be easily affected by a few incorrect gap-filling. Therefore, new gap-filling methods should be developed when we develop automated tools for metabolic network reconstruction. Another challenge is the reconstruction method should be able to apply to not only model organisms but also other species. Therefore, both the assumptions that are true for all the organisms and the assumptions that can only be applied to specific organisms should be accepted and considered. The reconstruction methods must allow users to add their system-specific assumptions, which have not been incorporated in current tools. We are going to address these two issues in Section 2.2.

1.2 Metabolic Network Modeling of Microbial Communities

1.2.1 Meta-genomics on Microbial Communities

Our knowledge about microbial diversity has been explosively expanded by cultivation-independent sequencing methods, such as phylogenetic analysis of 16S rRNA sequences, in the past few decades. According to records of the National Center for Biotechnology Information (NCBI), more than 11,000 bacterial species have been identified, which has doubled in the past 10 years. Based on some estimation, the total number of microbial species is 10^6, which is almost 100 fold more than what has been identified. Most of these natural microorganisms live in various microbial communities, which perform important biological functions and ecological roles.

Sequence-based approaches, such as meta-genomics, have been widely utilized to study microbial communities and reveal the complexity of these systems. More and more meta-genomes have been sequenced for microbial communities from various environments and ecosystems, such as marine environment, soil environment, and even higher organisms as hosts. There are 334 complete or on-going metagenomic projects have been carried out (`www.genomesonline.org`, 08/29/12). More than one half of the meta-genomic projects focus on environmental samples, while another 30% is based on host related samples. 28 projects, less than 10% in total, study the engineered microbial communities, such as wastewater treatment plant microbial communities.

The marine environment is the largest habitat on Earth as 70% of the earth surface is covered by oceans. The marine environment is also extremely diverse, e.g. the temperature and pressure are quite different in tropical sunlit surface and ocean trenches 11,000 m deep. Marine microorganisms have been adapted to all of these divergent environments and are believed to carry up to 98% of marine primary productivity (*Sogin et al.*, 2006). To study the diversity and abundance of marine microorganisms, both 16S rRNA based and total meta-genomic analyses have been carried out. Different marine environments, including ocean surface water (*Venter et al.*, 2004; *Rusch et al.*, 2007), mesopelagic water (*Giovannoni et al.*, 1996), deep sea (*Sogin et al.*, 2006), water columns (*DeLong et al.*, 2006) and sea subfloor sediments (*Biddle et al.*, 2008), have been studied by these sequence-based methods. According to one of the earliest metagenomic sequences of surface waters of Sargasso Sea, the microbial community is composed by nine bacterial phyla (Proteobacteria, Actinobacteria, Cyanobacteria, Firmicutes, Bacteroidetes, Chloroflexi, Spirochaetes, Fusobacteria and Deinococcus-Thermus) and two archaeal phyla (Crenarchaeota and Euryarchaeota). One of the application for these microbial and metagenomic sources is to identify novel enzymes. Till now, a number of important marine enzymes have been identified from these sequences, including esterase, lipase, cellulose, chitinase, amidase, amylase, phytase, protease, xylanase and alkane hydroxylase (*Kennedy et al.*, 2010).

Meta-genomic methods are also widely applied to host related microbial systems, such as microbial communities live on various body sites of human, which may massively affect human health (*Turnbaugh et al.*, 2007). The diversity and complexity of these microbial are extremely high. For example, there are more than 600 prevalent taxa at the species level have been identified from human oral cavity and shown to cause diverse oral diseases (*Dewhirst et al.*, 2010). In human skin microbiome, the species number is up to 1000 (*Grice et al.*, 2009), which se both the functions of protection and infection (*Cogen et al.*, 2008). In a similar scale, human gut microbiota consists of about 500 species which carry out various functions related to human metabolism including nutrient digestion, development of immune system and repression of pathogenic microbial growth (*Gill et al.*, 2006b).

Fewer metagenomic projects focus on engineered microbial communities. Activated sludge in waste water treatment plants is one such system. For example, Hector Martin et al. used the metagenomic analysis to study two lab-scale enhanced biological phosphorus removal (EBPR) sludge communities (*Garcia Martin et al.*, 2006). Later, Mads Albertsen et al. applied metagenomic analysis to microbial community in a full-scale EBPR process (*Albertsen et al.*, 2012). Microbial communities are also widely used in food industry even before they are noticed. Ji Young Jung et al applied metagenomic analysis to microbes in Kimchi, a traditional Korean fermented food, to study temporal changes of the microbial community, including cell populations and metabolic potential (*Park et al.*, 2012).

In all these works, metagenomic analysis provides very detailed information about the microbial communities, which not only covers compositions of population that can be retrieved by 16s rRNA sequences, but also genetic information about these organisms. The genetic information enables researchers to study the metabolism and function of the microorganisms in community level, which sometimes is more interesting than knowing what organisms are there. One challenge for functional analysis based on metagenomic sequences is the quality of the data. As demonstrated by simulated datasets, the completion of metagenomic sequences of dominant species will be less than 80% for a microbiome with medium complexity, and this number will decrease significantly when the complexity increases (*Mavromatis et al.*, 2007). These significant gaps of genetic information lead us to develop alternative methods for functional analysis besides direct mapping of genes with reference pathways.

1.2.2 Metabolic Modeling of Microbial Communities

Taking advantage of high-throughput cultivation-independent methods for microbial community analysis, such as meta-genomics, meta-transcriptomics, and meta-proteomics, researchers are able to capture the community-wide genetic information efficiently. However, we still lack of methods to interpret the metabolic contributions of organisms in the microbial community, as well as the cross-species communications. As discussed in Section 1.1, metabolic network reconstruction can provide comprehensive metabolic models for microorganisms according to their genomic sequences and some other data. Therefore, community-wide metabolic network reconstruction in provides a possible solution to study the metabolisms and interactions of microbial community.

Recently, researchers already started to generate metabolic reconstruction for simple artificial microbial consortia. Metabolic network reconstruction for a mutualistic microbial community, composed by *Desulfovibrio vulgaris* and *Methanococcus maripaludis* (*Stolyar et al.*, 2007), was firstly introduced by Sergey Stolyar et al. by integrating the two individual metabolic network reconstructions that are generated separately. In this model, the two microorganisms are treated as separate compartments with proposed exchange fluxes. To predict the fluxes of metabolic reactions, a linear combination of individual growth rates of the two organisms was used as objective function in the FBA framework. Experimental data were used to constrain nutrient uptake rates and byproduct production rates. This model can successfully predict the mutualistic relationship between the two organisms, that is, *Methanococcus maripaludis* removes the byproducts of *Desulfovibrio vulgaris* that inhibits its growth by using these byproducts as nutrients. In addition, relatively accurate growth rates of the two organisms can also be predicted, which are only slightly affected by the ratio of the two growth rate in the objective function. The same method was applied to the co-culture of *Clostridium butyricum* and *Methanosarcina mazei*, which is designed for converting glycerol into 1,3-propanediol (*Bizukojc et al.*, 2010). The interaction between the two organisms are the same as the *Desulfovibrio vulgaris* and *Methanococcus* *maripaludis* system. That is, *Methanosarcina mazei* is able to utilized the byproducts of *Clostridium butyricum* that inhibits its growth.

Different to mutualistic interaction, some negative interactions, such as competition, cannot be modeled in FBA framework as there is no single objective function can describe this type of interaction. Kai Zhuang et al. made use of dynamic flux balance analysis (dFBA) (*Mahadevan et al.*, 2002) to model the competition between *Rhodoferax ferrireducens* and *Geobacter sulfurreducens* in an anoxic subsurface environment (*Zhuang et al.*, 2011). In this modified dFBA framework, the uptake rates of the nutrients for the two species are subjected to different dynamic equations, representing the efficency of the transporters. Then the growth rates were predicted from these uptake rates and the two independent metabolic networks. Ali Zomorrodi and Costas Maranas developed a different strategy to model both positive and negative interactions through a model called OptCom (*Zomorrodi and Maranas*, 2012). Bi-level optimization is applied in OptCom, in order to trade off optimization of individual organisms versus the whole microbial community. The inner level of the bi-level model is the common FBA model for each organism separately but with exchange fluxes to connect them. The outer level is the summation of all the growth rates which indicate the maximum growth of the whole community.

Researchers are also interested in microbial co-cultures of different mutants from of same organism. For example, Tzamali et al. applied a graph-theoretic approach to look for microbial communities of non-lethal *Escherichia coli* mutants, and simulated their growth phenotypes using dFBA method (*Tollis and Reczko*, 2009). In their following work, their methods were applied to describe the growth phenotypes of pairs of *Escherichia coli* mutants utilizing different carbon sources (*Tzamali et al.*, 2011). Wintermute and Silver tried to utilize MOMA model on different pairs of *Escherichia coli* auxotroph mutants, and identified mutualistic relationships between them (*Wintermute and Silver*, 2010). From a different view, Klitgord and Segre developed Search for Exchanged Metabolites (SEM) algorithm that is able to predict growth environments that will promote potential interactions between different organisms (*Klitgord and Segrè*, 2010). They applied the SEM algorithm to different mutants of *Escherichia coli* and some other organisms. Shiri Freilich et al. utilized this SEM algorithm to identify potential competitive and cooperative metabolic interactions between 6,903 bacterial pairs (*Freilich et al.*, 2011).

There are several metabolic models for microbial communities that are not based on FBA type model. Reed Taffs et al. published three different methods based on elementary model analysis (EMA) (*Taffs et al.*, 2009) to study the phototrophic mat communities containing three distinct microbial guilds: oxygenic phototrophs, filamentous anoxygenic phototrophs, and sulfate reducing bacteria. The three different methods, with different compartment strategies, can be applied to systems with different complexity levels. Using a different approach, Erwin Frey et al., integrated evolutionary game theory, nonlinear dynamics, and the theory of stochastic processes to develop mathematical tools that can model various properties of ecological systems, such as stability (*Frey*, 2010).

1.3 Dissertation Overview

Previous studies have demonstrated the potential of applying metabolic network reconstruction to microbial communities. However, there are some fundamental challenges that limit the application of these community-wide metabolic models in community-scale. In this dissertation, I will address some of the critical challenges and provide solutions that can solve or partly solve these issues.

To study metabolism and interaction in microbial communities, high-quality metabolic network reconstructions are needed, because these community-wide models are even more sensitive to the accuracy of the data than single organism models. To meet this requirement, most of current community-wide metabolic models utilize only existing metabolic networks of single organisms, which are limited for artificial microbial communities in many cases. An alternative solution researchers are using is to utilize existing metabolic network reconstructions of different strains of the same or similar species. This method disregards strain variations existing in the same species, which have been observed even in the same type of samples collected from close but different locations (*Lo et al.*, 2007). Another limitation of this method is that there have been only 56 organisms (http://gcrg.ucsd.edu/InSilicoOrganism/OtherOrganisms, retrieved on July 27, 2012) for which high-quality genome-scale metabolic network reconstructions have been curated. The limited number of high-quality metabolic network reconstructions means not all the organisms in a microbial community are with high-quality metabolic network. This difference in the quality of metabolic network reconstruction can lead to false predictions, especially when modeling interactions in microbial communities.

In this dissertation, I am interested in applying metabolic network reconstruction on microbial communities to model the metabolism and interaction. We will answer the following questions:

- How to automatically reconstruct high-quality metabolic networks from annotated genomes?

- How to reconstruct community-wide metabolic networks from metagenomic datasets?

- What are the essential interactions for the formation of the AMD biofilm, which causes severe water pollutions?

- How to integrate -omics datasets to model the intracellular metabolism and interspecies interactions?

- What are the metabolic capabilities and interspecies relationships in gut microbiome that is directly related to host health, e.g. polysaccharide harvest in diet?

In order to generate high-quality metabolic network reconstructions, we need to develop a tool that can handle the reconstruction processes automatically due to the complexity of microbial communities. In Chapter II, we will introduce a bioinformatic pipeline that can automatically generate high-quality genome-scale metabolic networks. We applied this tool to the model organism *Escherichia coli* K12 and compared the results with metabolic network reconstructions in the literature, which have been developed for a long time. We further applied this tool to cyanobacterium *Prochlorococcus marinus* and generated metabolic network reconstructions for its twelve strains. Pan and core metabolic network of *Prochlorococcus marinus* were defined and the biosynthesis functions of the core metabolic networks were studied.

In Chapter III, we further explore the community-wide metabolic network reconstruction for an Acid Mine Drainage (AMD) biofilm from its meta-genomic sequences. After finishing the metabolic network reconstruction for individual organisms, we developed multi-organism metabolic network reconstructions for both abundant species and the whole biofilm. Then the multi-organism metabolic network reconstructions were used to predict potential interactions among the species in the community, which are essential for biofilm formation. In addition, we incorporated meta-proteomic dataset to verify and improve the metabolic network reconstruction.

In Chapter IV, we are interested in developing metabolic models for host-related microbial communities, such as the gastrointestinal microbial community. First, we investigated a two-species system in mice designed to capture interactions of two major phyla in human gut microbiota, which was designed to study the plant polysaccharide utilization. Using meta-genomic sequences, meta-transcriptome and meta-proteome, we developed a multi-step metabolic network reconstruction process that can utilize and integrate these datasets. We examined the metabolism and interactions in this two-species community utilizing annotated genomes and microarray expression data. Then we studied a ten-species model community, which could more accurately mimic the real human gut microbiome on studying the community changes in response to diets. By reconstructing community-wide metabolic networks utilizing annotated genomes and sequence-based expression data, we were able to predict the metabolic potentials, cross-species interactions and metabolic responses in gut microbiota.

Besides building metabolic models, in Chapter V, we also explore the integration of metabolic and regulatory network. We introduced a computational framework for metabolic and regulatory network design. This frameworks was applied to model organism *Escherichia coli* for over-production of fatty acid derived hydrocarbons. Strains for different products under different growth conditions were designed. Some of the results can be verified by literature data and strains for fatty acid overproduction have been partly implemented experimentally in our lab.

CHAPTER II

Bioinformatic Pipeline for Automated Genome-scale Metabolic Network Reconstruction

2.1 Introduction and Background

As mentioned in the Chapter I, a metabolic network reconstruction contains potential molecular mechanisms (metabolic reactions) in the cell, and the associated molecular components, such as enzymes, substrates, and products. As the development of genome sequencing methods, researchers are no longer satisfied with simplified or small scale metabolic network reconstructions. Genome-scale metabolic network reconstructions have gained more attentions because they can provide more comprehensive sketches about cell metabolism and bring more accurate phenotypic predictions. In the Section 1.1.2, we mentioned two tools that generate genome-scale metabolic networks automatically from annotated genomes. In this section, we will review the two methods with more details.

Model SEED (*Henry et al.*, 2010), one of the best automated metabolic network reconstruction tools, can generate genome-scale metabolic networks on the basis of genomic sequences in the aid of automated gene annotation server RAST (Aziz RK, et al., 2008). A draft model will be firstly generated by comparing the gene annotations with the genes associated with metabolic reactions that are collected from different resources. After this draft metabolic network is generated, an optimization based curation process (Satish Kumar, et al., 2007) will be applied to the draft. The basic assumption made in this curation process is the cell can synthesize all the biomass components from the compounds in the medium, and the curation with minimal number of changes that can enable this biosynthesis will be accepted. The number of the changes is weighted in the algorithm according to the type of changes. Therefore, model SEED can provide functional metabolic network reconstructions. However, researchers are not totally satisfied by the model SEED. One of the reasons is it does not enable user to customize the biomass components and mediums used in the model, which are sometimes critical in the reconstruction process. Another limitation of model SEED is the accuracy of the metabolic network reconstruction. Take the *Escherichia coli* for example, the metabolic network automated reconstructed without manual curation cannot correctly predict the phenotype-phase-plane of the strains, which is an important property of a metabolic network. We also noticed that the metabolic reactions contained in the self-maintained dataset were collected from different databases or metabolic network reconstructions of model organisms. This integration definitely makes the dataset more complete but also takes the risk due to the inconsistency among different databases. Model SEED does not enable recursive refinements for the metabolic reconstruction, which is normally applied in the manual reconstruction process. All these issues limit a wider application of this tool because high-quality metabolic network reconstructions are required in a lot of scenarios.

GEMSiRV (*Liao et al.*, 2012) is a software platform for genome-scale metabolic simulation, reconstruction and visualization developed after model SEED. In this software, a automated metabolic network reconstruction tool is embedded, which is the same as MrBac (*Liao et al.*, 2011). Different for model SEED and some other methods, it reconstructs metabolic network by comparing the genome of the strain with a genetically close strain with a metabolic reconstruction already. Sequence alignment is applied to identify these genes by setting a threshold of sequence identify or E value. This method can only apply to those organisms that are close to model organisms, the metabolic network of which have been reconstructed. Otherwise, this method cannot provide reliable metabolic networks due to the diversity of cellular metabolism. Furthermore, the quality of the new metabolic network reconstruction largely depends on the referenced metabolic network and normally the quality decreases due to the lacks of curation processes. The quality of reconstructed metabolic networks also affected significantly by the parameter settings, which are arbitrary. Another limitation of this method is it cannot make use the efforts spending in the manual curation of gene annotation because only sequence alignment results are utilized to identify reactions. Besides draft metabolic network reconstruction, GEMSiRV can convert the reconstructed metabolic network into mathematical models and provide FBA type analysis based on this runnable model.

Automated metabolic network reconstruction tool is necessary for community-wide metabolic modeling for microbial communities, because there are a large number of organisms existing in one microbial community, and it is impossible to manual reconstruct metabolic networks for all of them. Due to the limitation of current automated metabolic network reconstruction tools, we want to develop a bioinformatics pipeline for automatically reconstruct high-quality metabolic network from annotated genomes. Towards this goal, we will discuss the following questions in the next several sections.

- How can we automatically reconstruct high-quality metabolic network from annotated genomes?
- How accurate is the metabolic network for model organism *Escherichia coli*

reconstructed by the tool?

- Whether the tool can work with a large number of organisms? And if yes, what benefits we can get from the high throughput bioinformatic pipeline?

2.2 Pipeline for Metabolic Network Reconstruction (PEER)

To reconstruct metabolic network efficiently, we developed an automated bioinformatics pipeline to generate complete and reliable metabolic reconstruction based only on the genome and gene annotations of the organisms. In addition, we design this tool for community-wide metabolic network reconstruction from meta-genomic data. Different from single organism genome project, meta-genomic sequences cannot provide complete genomes for all the organisms. Only for the dominant organisms, nearly complete genome sequences can be derived. The incomplete genome sequences require this pipeline be able to fill the metabolic gaps which have not been identified in the genomes. Another challenge is the quality of gene annotation. Because most of the species in the environmental samples are not cultivable, there is litter information available for the enzymes and proteins in these cells besides the sequences data. Thus, the gene annotations have not been verified as those in model organisms. Further, due to the scale of meta-genome, the gene annotations are always only based on automated annotation methods, without manual curations. All these factors make the quality of gene annotation for meta-genomes not as good as in single genome projects. To overcome these problems, PEER must be able to bare the deficiency of the genomic sequence and annotation method, which may takes place in all the steps along the sampling, sequencing, assembling, and annotating process. Also, the pipeline needs to be flexible enough to deal with the organism that is not well studied and with limited information. We designed the pipeline (as shown in Figure 2.1) with three major components: automated metabolic network pre-reconstruction, probability of metabolic network prediction, and automated curation with function completion inspection.

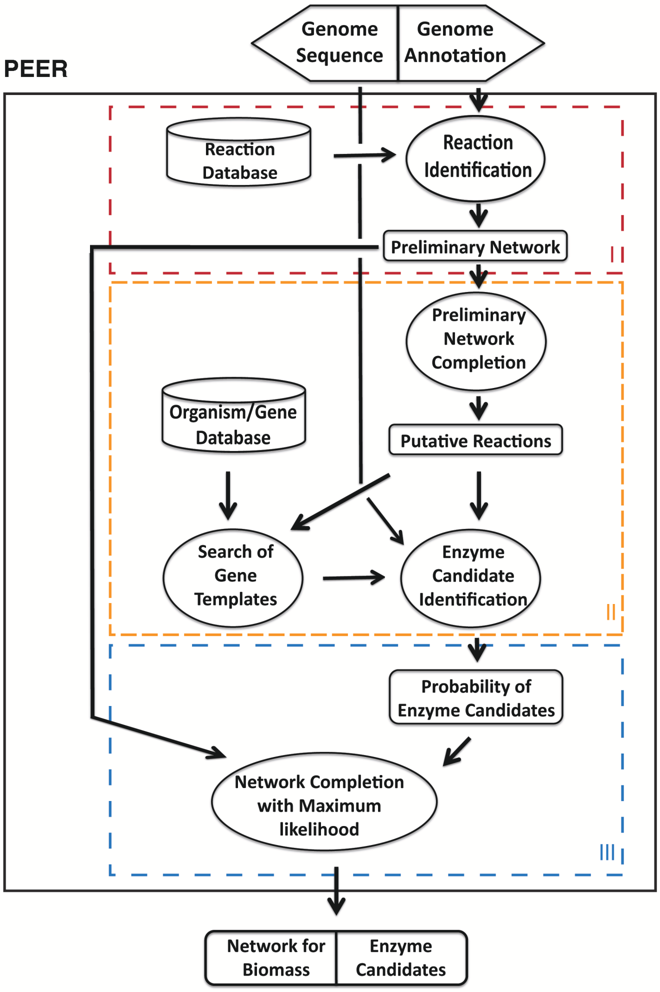


Figure 2.1: Flow chart for Pipeline for Metabolic Network Reconstruction (PEER)

As shown in the figure, to generate the output metabolic networks, three steps must be finished. First, the pre-reconstruction will provide the metabolic reactions whose corresponding enzymes have been identified. This pre-construction will be further revised and results with some defects are still acceptable, so this pre-reconstruction step is more flexible when inputting data from different databases. Second, automated curation process, is based on the assumption that any individual organisms (species) have to accomplish pre-defined functions, such as producing biomass and generating energy to sustain life. By imposing mass balance constraints and other information, such as reversibility of reaction and growth conditions, the automated curation process can provide a set of alternative reconstructions for the organisms, each of which contains a set of reactions that have not been identified in the pre-reconstruction. Evaluating the probability of the existence of these added reactions in the organisms would provide a quantitative measurement of the overall risk for the alternative reconstructions. Third, by defining an objective function which contains the discrete decision variables representing the revision in the solution, we can establish an optimization based algorithm to minimize the risk of making revisions in the reconstruction, which can finally return the most possible reconstruction according to the known information. This optimization process is under the mixed integer linear programming framework, because all the constraints and objective function can be written in linear format, including mass balance, reaction reversibility (thermodynamics), nutrients availability, and defined function completeness. Also the information that come from searching from gene candidates are also considered in the model by adjusting the weights of the discrete decision variables in the objective function, which is also linear. In addition to the reconstruction, the enzyme candidates for the added reactions are also implied. If this pipeline is applied to microbial communities, the microbe-microbe and microbe-environment interactions will also be considered and predicted. The details about all the steps will be discussed in the next three sections.

2.2.1 Preliminary Network Generation

The major goal of the first step, automated metabolic network pre-reconstruction, is to generate a draft metabolic network. Provided the genomic sequence and annotation of genes, the pipeline is able to identify the metabolic reactions existing in each organism by identifying the corresponding enzymes/proteins in the genome or annotations. In this matching process, we use EC number as the primary information. For example, reaction R00220 (L-serine ammonia-lyase) can be catalyzed by two enzymes, L-serine ammonia-lyase (EC:4.3.1.17) and L-serine dehydratase (EC:4.3.1.19). If any of the enzymes are identified in the species by matching the two EC number, reaction R00220 is added into the draft metabolic network. Because we only match the EC number, we can avoid the effects caused by the inconsistency of enzyme names used in different databases and annotation tools. However, if a gene annotated with correct enzyme name but not assigned EC number in the annotation, we will miss the corresponding metabolic reactions. This common problem takes place frequently when using gene annotations without manual curation. In our pipeline, the automated gap filling process in step II and step III can greatly reduce the risk of incorrect/inconsistent annotations and we will discuss this function later.

The transport reactions are another essential component in the metabolic network which represents the communication and interaction between organism and environment. However, it is different to identify a transport reaction compared to common enzymatic reactions because the methods for identification of the transporter in a community are less understood. TransportDB (*Ren et al.*, 2007) provides reasonable estimations for the transport reactions, but they are always not enough. To solve the problem, we will also add the transport reactions for the known nutrient uptakes or byproduct secretes, if they are not included in the substrate list of transporters.

After that, the stoichiometric matrix of the entire set of metabolic reactions and transport reactions is defined through matching the reactions with a reaction set, in which all the reactions have detailed and balanced reaction formulas. Both the reaction pool and the enzyme information of the reactions used in this project are collected from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Goto, Nishioka and Kanehisa, 1999; Goto, Nishioka and Kanehisa, 2000; Goto et al., 2002). Reactions whose reaction formulas are not specified or cannot be balanced are not included in the reaction set, and manual revisions have also been applied to the reaction formula and their reversibility. We spent great efforts to refine the problematic reactions in this reaction set as just a few errors in it may ruin the whole metabolic network reconstruction in certain analysis. In Figure 2.2, we demonstrate two simplified cases in which one mistake can cause significant changes in the whole metabolic network with either infinite energy (ATP) or reduced force (NADH) generation cycles. These problematic reactions sets can be much more complex than the two examples and difficult to identify. To avoid these problems, we looked for all the possible reaction sets that can either generate infinite energy or other resources. Then we manually corrected these reaction sets. Besides the common enzymatic reactions, there are still a certain number of non-enzymatic reactions and spontaneous reactions, which are also included in the reaction pool. These reactions are automatically added into the stoichiometric matrix of the organisms without requirement for enzymes.

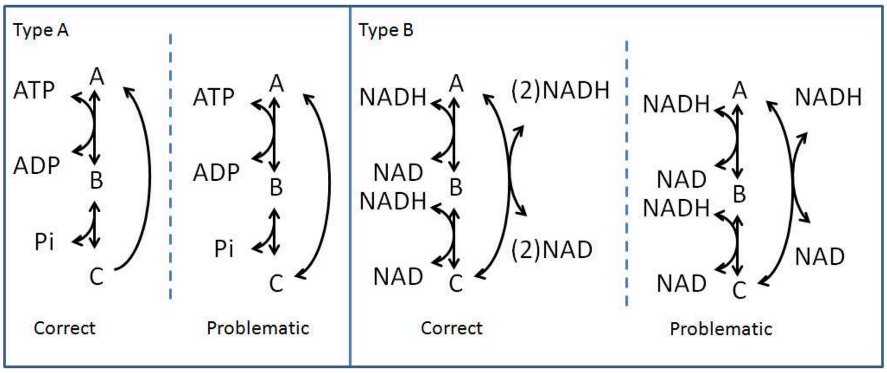


Figure 2.2: Two examples of problematic metabolic reactions that commonly happen during the metabolic reconstruction process.

After these three steps, we are able to generate a draft metabolic network with both metabolic reactions and exchange reactions. The gene associations for all the enzymatic reactions are also provided. As mentioned, this pre-reconstruction cannot be perfect due to the errors or format issues in gene annotations. Another potential error source is those metabolic reactions that have not been well identified, such as reactions catalyzed by orphan enzymes (*Lespinet and Labedan*, 2005). To correct these errors and fill the metabolic gaps, an optimization based automated curation process is introduced, which aims to generate the most plausible metabolic networks.

2.2.2 Mixed Integer Linear Programming (MILP) Based Network Curation

The optimization based metabolic network gap filling is employed in both step II and step III to find the potential metabolic gaps under different criteria. One basic assumption made in this gap filling model is the metabolic network should be able to achieve all of the pre-assumed metabolic functions. If not, revisions must be made to accomplish them. However, there is no restriction about the revision, which means any possible changes of the network which do not go against the part of network that has been confirmed are feasible. Two types of changes could easily be expected, including adding new reactions into the metabolic network and changing the reversibility of existing reactions in the network. By comparing all the feasible revisions, it is possible to identify the "best" one according certain criteria, such as minimal changes or most likelihood changes, which will be discussed in following section. Besides the enforced constraints about the assumed functions, other factors will also restrain the space of feasible revisions, such as the flux balance constraints and reversibility of reactions. We are employing a Mixed Integer Linear Programming (MILP) (*Floudas*, 1995) based model to predict the best metabolic network as described. Figure 2.3 demonstrated the major components in this model and detailed model descriptions were contained in Appendix B.

$$
\begin{aligned}
\min_{v,b_{add}} \sum_{r\in R} & b_{add}(r) * \text{Weight}(r) && \text{(Find a best network under constraints)}\\
\text{S.t.} \sum_{r\in R} & S(m,r) * v(r) = 0, \quad \forall m \in Metabolite && \text{(Mass balance of metabolites)}\\
& v(r) \geq \epsilon, \quad \forall r \in \text{Product} && \text{(Products must be synthesized)}\\
& F_{min} \leq v(r) \leq F_{max}, \quad \forall r \in \text{Exchange} && \text{(Exchange reactions with environment)}\\
& v(r) \geq 0, \quad \forall r \in \text{Irreversible} && \text{(Reversibility of reactions)}\\
& v(r) \leq F_{max} * b_{Active}(r), \quad \forall r \in \text{R} && \text{(Active reactions)}\\
b_{Active}(r) \leq & \; b_{\text{Exist}}(r) + b_{Spon}(r) + b_{Add}(r), \quad \forall r \in \text{R} && \text{(Existence of reactions)}\\
& \text{etc.} && \text{(Other constraints)}
\end{aligned}
$$

$v(r)$: Flux of reaction r

$b_{Add}(r)$: Whether r is added (putative reaction), binary variable

Figure 2.3: Demonstration of the MILP model for metabolic network gap filling. Red: binary variables, blue: continuous variables, black: parameters. Complete model descriptions can be found in Appendix B.

We assumed the probabilities of all metabolic gap filling are independent and can be evaluated by comparing the genome of target organism with the genes that were known to catalyze these putative reactions. In this model, all the revisions to the metabolic network will be assigned penalty parameters according to the corresponding probability. Further, a set of binary variables are assigned to represent these revisions. The MILP model is designed to search for a subset of the revisions that accomplish the defined functions, such as biomass syntheses, and also with the minimal sum of the penalties. The metabolic network with these revisions is the most plausible reconstruction under this framework. However, the mathematically optimal solution does not guarantee to be the biologically optimal one. Therefore, practically suboptimal solutions are also calculated and integrated with the optimal one to derive a reliable metabolic network.

There are certain optional constraints that can further improve the curation process. We can introduce a constraint that define the minimal growth rate $v_{growth} \geq v_{minimalgrowth}$. Along this line, we can define the maximum and minimum yields of elements in the biomass. These constraints can further reduce the solution spaces and provide more accurate prediction. For example, by setting minimal growth rate, alternative low efficient pathways will not be considered even if they may contain less metabolic gaps.

We implemented this model in both IBM Ilog Cplex platform and FICO XPRESS platform. In the IBM Ilog Cplex platform, branching method based on pseudo costs together with best-bound node searching method are chosen. The MIP emphasis indicator is set to emphasize feasibility over optimality. In the FICO XPRESS platform, the same branching method is used and the local first node selection strategy is chosen for node searching method. The parameters are chosen based on the best results in test runs and are maintained when solving all the problems.

2.2.3 Probability of Metabolic Network Prediction

To evaluate the uncertainty of a predicted metabolic network caused by the two types of curation, the probability of each individual change is first defined. If an enzymatic reaction is added into the metabolic network, this probability should represent the possibility that an enzyme that can catalyze the reaction exists in the species. If we assume the genome/metagenome used is correct or partly correct and all the sequences of the corresponding enzymes are known, this probability could be evaluated by the sequence alignment between the assigned enzymes sequences and the genome sequence. As the definition of the p-value in the common BLAST algorithm (Altschul et al., 1990), it is the probability (in the range of 0-1) of a given sequence occurring by random chance, which means the risk of taking the searching result as the target sequence. At the same time, for one enzymatic reaction, there may be more than one isoenzyme, and the sequences of them are varied in different organisms. So the overall probability of reaction (r) is defined as the minimal value of all the probabilities that the corresponding enzymes ($e \in E$) exist in the species, that is,

$$p(r) = \min_{e \in E(r)} p(e) = \min_{e \in E(r)} (1 - \text{P-value}(e))$$

in which E(r) represents the set of enzymes that catalyze reaction r and P-value(e) is the best p-value in the sequence alignment results for enzyme e in the genome of the species. After the probability of adding one reaction is defined, the probability of one reconstruction which contain a set of reactions ($r \in R$) can be easily calculated as

$$P = \prod_{r \in R} P(r)$$

In order to cooperate with the MILP frame work used in the automated curation process, the logarithm of the probability is employed in the model, that is,

$$\begin{aligned} \log P &= \log \prod_{r \in R} P(r) = \sum_{r \in R} \log P(r) \\ &= \sum_{r \in R} \log \min_{e \in E(r)} (1 - \text{P-value}(e)) \\ &\approx -\sum_{r \in R} \min_{e \in E(r)} \log \text{P-value}(e) \end{aligned}$$

However, the P-value from the blast is varied from 0.9 to 10^{-299} and in the annotation a threshold 10^{-30} is used to predict gene functions. Therefore, to consider the non-linear property and large scale distribution of P-value, another practical definition of overall risk of a reconstruction is applied, which scale the probabilities by their average value, making them center at one. This definition also fit the MILP framework, and the scaled risk of added reaction is taken as weight parameter in the optimization process, denoted as

$$weight(r) = \frac{n \min_{e \in E(r)} \log \text{P-value}(e)}{\sum_{r \in R} \min_{e \in E(r)} \log \text{P-value}(e)} \tag{2.1}$$

2.2.4 Online Tool for Automated Metabolic Network Reconstruction

An online tool, http://ccdu.ccmb.med.umich.edu/LinLab/, is designed to carry out this bioinformatics pipeline for metabolic network reconstruction. For simplification, the pre-defined function that the organism must achieve is biomass synthesis. However, other alternative functions that can be expressed in a similar form are also acceptable for this online tool, such as enforcing byproducts. Furthermore, the gene annotation and other knowledge of different organisms can vary dramatically, which means the input data have different qualities. Thus, this online tool is designed to adapt to these different situations by accepting optional input data. Because of the limitation of computational resources, the online tool in current version does not allow the submission of multi-organism problems (meta-genome), which will be provided in future.

2.3 Automated Metabolic Network Reconstruction for Model Organism *Escherichia coli*

Escherichia coli is one of the most studied organisms and the genome-scale metabolic network of *Escherichia coli* is the best metabolic network reconstruction existing currently. Similar to most other organisms, there is no perfect gene annotation for *Escherichia coli*, and we can obtain different versions of its gene annotations from various sources. This situation bring us a new question when applying automated metabolic network reconstruction; that is whether the datasets collected from different data sources have any effects and if yes, how significant the effects are. Another interesting question we will explore is that how good the quality of automated metabolic network reconstruction is when comparing with manual curated ones and whether it can satisfy the requirements of common applications. To address these issues, we generated metabolic network reconstructions of *Escherichia coli* based on three different datasets and compared them with manual curated metabolic networks.

2.3.1 Automated Metabolic Network Reconstructions of *Escherichia coli* from Three Datasets

To explore the effects of input datasets to the final metabolic network reconstructions, we apply the PEER to three datasets of *Escherichia coli* from different databases, including NCBI (U00096, 28-JUL-2009), Kyoto Encyclopedia of Genes and Genomes (KEGG, T00007, 15-Sep-2009) and Integrated Microbial Genomes (IMG, 637000106, 26-Sep-2009). All the three datasets are processed separately according to the algorithm mentioned. Besides the ORF sequences and function annotations, biomass composition of *Escherichia coli* K12 MG1655 is another parameter that may affect the final solution, because the pre-assumed function is defined as biomass synthesis. In this model, the biomass is composed of twenty basic amino acids, eight nucleotides (NTP and dNTP), four coenzymes (NAD, NADP, FAD, COA), and other seven metabolites. The detailed composition of biomass is listed in supplementary materials. The compositions of metabolites in the biomass used in PEER are the same as those in the *i*AF1260 model. Growth condition is another factor that needs to be considered; in this model we test the mediums with two different carbon sources (L-glucose and L-malate) separately in both anaerobic and aerobic conditions.

The input data from different databases are in different formats and with varied gene function annotation. These differences in input data may lead to artificial variances of the final results, which should be avoided as much as possible. However, we still observed slight disagreements among the metabolic networks reconstructed based on three input datasets. We show in Table 2.1 that the pipeline identified about fifty more metabolic reactions from the data from KEGG and IMG than those from GenBank. At the same time, the number of identified metabolic enzymes from KEGG is about one hundred more than those from IMG while the identified reactions of the two datasets are quite similar. This result may be explained by the quality of gene function annotation. Furthermore, the pipeline is focusing on the metabolic related enzymes, so the quality of these genes will affect the result the most. We also found that the reconstructed metabolic networks based data from IMG and KEGG are almost the same, even though results from IMG and GenBank have a similar number of putative reactions. Despite of the differences of metabolic networks, the active parts for biomass, especially putative reactions, are very conservative. All the six putative reactions from the network based on KEGG dataset are shared by the other two, and all the ten putative reactions from the network based on GenBank datasets are included in the results based on IMG dataset. Further, by investigating the other active reactions, we found that the different metabolic reactions are sometimes with the same functions among the three datasets but with different coenzymes (e.g. NADH vs. NADPH) or different forms (e.g. one-step reaction vs. multi-step reaction). Therefore, the differences among the three reconstructed networks are much less than observed directly if we consider these reactions with different equations but the same roles in metabolism.

Table 2.1: Metabolic network reconstructed from different input data of *Escherichia coli* K12 MG1655.

Source of input data	Number of protein genes	Number of genes encoding metabolic enzymes	Number of identified reactions	Number of putative reactions for biomass	Number of putative reactions with gene candidate
KEGG [1]	4148	1410	1147	6	4
NCBI [2]	4321	1106	1091	10	8
JGI/IMG [3]	4391	1311	1141	11	9

[1]: `http://www.genome.jp/kegg/`, retrieved in 2009/9.

[2]: `http://www.ncbi.nlm.nih.gov/genbank/`, retrieved in 2009/7.

[3]: `http://img.jgi.doe.gov/w/`,retrieved in 2009/9.

2.3.2 Comparison of Automated Metabolic Network Reconstruction With Reference Metabolic Networks

There are several versions of genome-scale metabolic networks of *Escherichia coli* K12 MG1655, which are in different scales and details of metabolic reactions. Here we are using the iJR904 and iAF1260 model as reference models to verify the metabolic network reconstructed by this pipeline. Three versions of results based on different input datasets have been generated, and here we mainly use the one based on the KEGG dataset to compare, which requires the least putative reactions. According to the results, 1153 (6 of them are putative) intracellular metabolic reactions have been identified, while there are 745 unique reactions in iJR904 model and 1339 (1187 cytoplasmic reactions) in iAF1260 model. In terms of scale, this automated reconstructed metabolic network is close to the iAF1260 model and better than the older one. Furthermore, all the reactions in the automated reconstructed metabolic network are either gene associated or no association required because of the mechanism of the pipeline, while 5% and 6% of the metabolic reactions are not associated with genes in iJR904 and iAF1260 model respectably.

The accuracy and reliability of putative reactions are also very important for the final reconstructions, even though the fraction of these reactions in this *Escherichia coli* model is very low. This is mainly due to the extensive research on the metabolism of this organism, which cannot be applied to other organisms. Table 2.2 demonstrates the metabolic gaps identified based on the KEGG dataset. There are six metabolic gaps in total and four of them can be assigned with certain gene candidates. R04292 is one of the two metabolic gaps without gene candidates, which is mainly due to lack of gene templates for this reaction. According to the latest version of annotation of *Escherichia coli* in KEGG, gene b0750 is annotated as 2.5.1.72, which is in agreement with the gene association in iAF1260 model. For the other reaction without gene candidate, R04457, enzyme lumazine synthase(EC:2.5.1.78) is required. In iAf1260, gene b1662, which is annotated as riboflavin synthase(EC:2.5.1.9), was assigned to this reaction.

Table 2.2: Metabolic gaps for metabolic network reconstruction *Escherichia coli* of based on KEGG dataset.

Reaction	Required Enzyme	Required Enzyme (EC)	Candidate Gene	P-value	Annotation in KEGG
R04457	Lumazine synthase	EC:2.5.1.78	NA	NA	NA
R04554	Transferases	EC:2.4.2.-	b2407	9.1E-185	ec:2.4.2.-
R04655	Carbon-carbon lyases	EC:4.1.3.-	b0352	4.6E-223	ec:4.1.3.-
R07280	Phosphoric monoester hydrolases	EC:3.1.3.-	b4016	0.0E+00	ec:2.7.11.5/ 3.1.3.-
R00188	Phosphoadenylate 3'-nucleotidase	EC:3.1.3.7	b4214	7.8E-159	NA
R04292	Quinolinate synthase	EC:2.5.1.72	NA	NA	NA

The four metabolic gaps that have been associated with gene candidates are mainly due to the non-specificity of the enzyme annotation or association to metabolic reactions (missing the last digit in EC number). Take reaction R07280 for example, Phosphoric monoester hydrolases (EC:3.1.3.-) is required to catalyze this reaction and gene b4016 is assigned as the gene candidate. This gene candidate b4016 can be annotated as multi-function enzyme (EC:2.7.11.5/3.1.3.-). For reaction R00188, enzyme phosphoadenylate 3'-nucleotidase (EC:3.1.3.7) is required and assigned with gene b4214 as gene candidate, which is annotated as EC:3.1.3.7 in the latest gene annotation. There is another type of metabolic gaps that are caused by in-consistency between different databases or datasets. For example, according to KEGG, the reaction that is catalyzed by 4-phospho-D-erythronate: NAD+ 2-oxidoreductase (EC: 1.1.1.290) is required but missing in the metabolic networks generated from IMG dataset before the gap filling. The suggested gene candidate in the IMG dataset is annotated as D-erythrose 4-phosphate dehydrogenase (EC: 1.1.1.-) but in KEGG database the gene is correctly annotated.

In conclusion, the metabolic gaps filled by our automatic gap filling algorithm can be either real metabolic gaps or caused by inconsistence/inaccuracy of gene annotation. And if the metabolic gaps are caused by annotation, gene candidates with very low p-value can be assigned to these gaps automatically by our bioinformatic pipeline, which indicates it can utilize inaccurate/incomplete input data and still can generate metabolic network reconstructions with high quality.

2.3.2.1 Phenotype Phase Plane (PPP) of the Reconstructed Metabolic Networks of *Escherichia coli* K12 MG1655

One application of reconstructed metabolic networks is to predict the phenotype phase plane, which represents the optimal growth conditions (*Ibarra et al.*, 2002). Phenotype phase planes provide the information of the growth rates of a organism under different fixed uptake rates of major nutrients. Here we test the metabolic network of *Escherichia coli* K12 MG1655 based on the KEGG dataset with two different growth conditions, malate-oxygen condition and glucose-oxygen condition. The lines of optimality (LO) for these two conditions are demonstrated in Figure 2.4. The same phenotype phase planes have been investigated before. By comparing our results with the reference line of optimality and their experimental data, we can find that the predicted lines of optimality (LO) are quite close to the published data. However, the growth rate in these predicted lines of optimality are slightly higher than the reference data, which is mainly due to the difference of biomass composition. These reasonable predictions of phenotype phase plane can further verify the metabolic networks reconstructed by the bioinformatics pipeline.

2.4 Automated Metabolic Network Reconstruction for Twelve Strains of *Prochlorococcus marinus*

Prochlorococcus marinus is a marine cyanobacterium that dominates phytoplankton communities in most tropical and temperate open ocean area (*Partensky et al.*, 1999). These widely distributed cells are the smallest photosynthetic organisms known, and abundant efforts have been paid to investigate them. A number of strains of *Prochlorococcus marinus*, including two different ecotypes (high-light-adapted and low-light-adapted), have been collected throughout the world and sequenced. Here we applied the bioinformatics pipeline to twelve sequenced strains of *P. marinus* (*Kettler et al.*, 2007), including *P. marinus* AS9601, *P. marinus* MIT 9211, *P. marinus* MIT 9215, *P. marinus* MIT 9301, *P. marinus* MIT 9303, *P. marinus* MIT 9312, *P. marinus* MIT 9313, *P. marinus* MIT 9515, *P. marinus* NATL1A, *P. marinus* NATL2A, *P. marinus* marinus CCMP1375, and *P. marinus* pastoris CCMP1986. The genome sequences for these strains are also collected from IMG database and annotations are generated through RAST annotation server (*Aziz et al.*, 2008). Because there are no details about the biomass compositions for these twelve strains, we used a basic list of metabolites to represent the biomass, which contains amino acids, nucleotides, and certain common coenzymes. Some detailed information about the reconstructions can be found in Table 2.3.

From the table, the numbers of metabolic reaction in the twelve metabolic networks of *P. marinus* vary from 636 to 693, including 17 to 23 putative reactions. Further, the numbers of putative reactions show non-negative correlation with the numbers of identified reactions, which was not expected. The distribution of the metabolic reactions in the twelve strains, including putative reactions, is shown in Figure 2.5.a. The metabolic reactions that are related to biomass synthesis are also

Table 2.3: Automated metabolic network reconstructed for twelve strains of *P. marinus*.

Strains	Number of protein genes	Number of genes encoding metabolic enzymes	Number of identified reactions	Number of putative reactions for biomass
P. marinus AS9601	1939	537	618	21
P. marinus MIT 9211	1856	536	625	23
P. marinus MIT 9215	2014	552	631	22
P. marinus MIT 9301	1921	537	627	22
P. marinus MIT 9303	3075	611	669	23
P. marinus MIT 9312	1811	543	624	18
P. marinus MIT 9313	2275	602	673	22
P. marinus MIT 9515	1992	534	619	17
P. marinus NATL1A	2204	555	633	22
P. marinus NATL2A	1896	548	645	22
P. marinus marinus CCMP1375 (SS120)	1833	544	628	24
P. marinus pastoris CCMP1986(MED4)	1719	543	622	19

indicated in the figure, and take a significant fraction in the overall metabolic networks (38.5%), which might be explained by their compact genomes.

According to the figure 2.5, the major components of the metabolic networks are maintained in all the strains and the variations occur only on part of the network (34%). The clustering of the metabolic networks of the strains is also shown in the figure. Not surprisingly, the two ecotypes, high-light-adapted and low-light-adapted, are clearly separated by the structures of their metabolic networks, indicating their metabolisms have adapted to the environments. We also carried out t-test to identify those reactions that are either enriched in high-light adapted or low-light adapted strains. We found 16 reactions enriched in low-light adapted strains while 7 reactions enriched in high-light adapted strains ($P < 0.05$). For example, one reaction involved in converting sulfate to sulfite exists in almost all the in low-light adapted strains but none in high-light adapted strains. This enrichment can be explained by the low concentration of sulfate on the top layer of ocean (*Likens and Likens*, 1981). Another two reactions, involved in citric-acid cycle (CAC) converting 2-oxoglutarate to Succinyl-CoA, are enriched in low-light adapted strains. This observation may suggest that the low-light-adapted *P. marinus* are not obligate autotrophy as the high-light-adapted ones which only contain incomplete CAC (*Huynen et al.*, 1999). This hypothesis is in agreement with what Zubkov et al discovered in their experiments, which also suggests the low-light-adapted strains are mixotrophic *P. marinus* when comparing to high-light-adapted strains (*Zubkov et al.*, 2004). However, we did not observe any enrichment of reactions for nitrite utilization in low-light adapted strains. Researchers used to believe only low-light adapted strains can use nitrite while the major nitrogen source on top layer of ocean is ammonia. Recently, researcher found widespread metabolic potential for nitrite and nitrate assimilation among both two *Prochlorococcus* ecotypes (*Martiny et al.*, 2009), which agrees with our observations. Phosphorus resources also show difference for the water layers the two ecotypes live in. For high-light adapted strains organic phosphorus are the major sources while inorganic phosphate are the major sources for low-light adapted strains (*Rocap et al.*, 2003). We also did not observe any differentially existing metabolic reactions regarding this environmental change, which might be explained by that the phosphorus in organic phosphorus chemicals is still in phosphate form.

The corresponding pathway distributions are also shown in Figure3.b, and several pathways are enriched in the metabolic network of *P. marinus*, including purine metabolism, pyrimidine metabolism, porphyrin and chlorophyll metabolism, peptidoglycan and fatty acid biosynthesis and several amino acid synthesis pathways. All these pathways are involved in cell growth process. For instance, purine, pyrimidine, and amino acids are common compositions for biomass; porphyrin and chlorophyll are essential for photosynthesis; and peptidoglycan and fatty acid are important for membrane and cell wall formation.

2.4.1 Pan and Core Metabolic Networks of *P. marinus*

This bioinformatics pipeline enables us to reconstruct metabolic networks of sequenced organisms efficiently, which makes it possible to compare the metabolic networks of different strains belonging to the same species. Here we integrate the metabolic networks of these strains to predict the pan and core metabolic network of *P. marinus*, and further interpret the functions of them. The pan metabolic network, which contains 881 metabolic reactions, is mainly divided into two regions, conservative region and variable region (denoted as A, B in Figure3.a). Region B represents the variable metabolic reactions in strain level. From the Figure 3.b, there is only one pathway enriched in this region, xenobiotics metabolism, which represents the adaptation of strains to their local environments.

We consider the conservative region A of the pan metabolic network as the core metabolic network (588 metabolic reactions), which should contains the common features of *P. marinus*. The major function of the core metabolic network is still biomass synthesis, taking more than 44% metabolic reactions. To investigate more details of the functions and metabolic products of core metabolic network, we force a certain percentage of the core metabolic network to be connected and active, and then define the functions according to the byproducts of the active network. Certain metabolites, besides of biomass compositions, are predicted to be produced as byproducts by the core metabolic network, including indole, acetate, xanthine, nicotinamide and some other compounds.

2.5 Discussion and Conclusions

From the results of these metabolic reconstructions, we are able to evaluate the quality and reliability of this pipeline. The analysis of *Escherichia coli* K12 MG1655 indicates the metabolic reconstruction for a well studied organism from this automated pipeline can be comparable to some of the manually refined metabolic reconstructions. This is not surprising as the information for those well studied organisms is also well organized and recorded in databases. Furthermore, to make the metabolic network reconstruction more accurate, only the metabolic reactions without undefined parameters and not problematic are included in the pipeline, making the overall numbers of metabolic reactions contained in the networks smaller than it could be. One benefit of this filter is to improve the quantitative predictions, such as phenotype phase plane (PPP), which are sensitive to both false-negative and false-positive errors in the reconstruction.

We also demonstrated that the sources of input datasets have certain effects to the final results, even though these effects are alleviated greatly for *Escherichia coli* K12 MG1655. The metabolic reconstructions of twelve strains of *P. marinus* generated from annotations from different databases are separated clearly (Supplementary Materials 1), which indicates that the biological properties can be buried by the quality of annotation methods. Therefore, consistent inputs are essential for fair comparison of metabolic networks of organisms.

The automated gap filling process in this bioinformatics pipeline mainly introduces two types of revisions, revising incorrect or inaccurate function annotations and introducing additional metabolic reactions that are not well understood. The first type of revisions also includes introducing those reactions that are not captured in the pre-reconstruction process just because the pipeline does not recognize the annotation of corresponding enzymes. Some revisions in the second types can be explained by the existence of orphan enzymes, which are widely distributed (*Lespinet and Labedan*, 2005). Because the automated gap filling is an optimization based process, the optimal predictions may not be the biological correct ones. To improve the accuracy of the metabolic reconstructions, we can integrate multiple reconstructions including optimal and suboptimal solutions from the pipeline to further generate the final metabolic reconstructions.

Compared to previous automated metabolic network reconstruction methods, like SEED and GEMSiRV, PEER has three major advantages.

- PEER uses a more curated reaction data. Similar to SEED, we generate the draft metabolic network by matching the gene annotations of the target organism to a metabolic reaction database, which was developed based on KEGG. However, we want to eliminate potential problematic reactions in our metabolic reaction database (see Section 2.2.1) rather than including as many metabolic reactions as possible, which is how SEED generate its reaction database. One reason for this strategy is that the false positive errors in the reconstructed metabolic network are harder to remove from the draft metabolic network reconstruction when comparing to false negative errors.

- PEER contains gap filling methods considering gene candidates. The gap filling method used in PEER is different from that in SEED or COBRA toolbox, which considers sequence alignment results and biomass yield when fill the gaps. This improvement can increase the accuracy the gap filling, and provide the best gene candidates for those metabolic gaps if possible. Another advantage of this gap filling method is PEER is able to provide gene candidates for the metabolic gaps as soon as they are identified. Therefore, not only the putative reactions for these metabolic gaps are provided but also potential target genes in the genome are predicted, which is important for further manual curation.

- PEER allows customized input data. The interface of PEER allows user to customize the growth condition, biomass compositions and biomass yield, which cannot be changed in SEED. Therefore, users can utilize organism-specific growth condition, nutrients information and experimental biomass yield to improve the reconstructed metabolic networks using PEER.

One limitation of PEER is PEER does not provide genome annotation tools as SEED dose. Thus the quality of genome annotation provided by user can affect the quality of final results; even through the curation steps in PEER can mitigate some of errors caused by miss annotations. GEMSiRV or MrBac used BLAST results rather than gene annotation to avoid these issues, which makes the methods highly rely on the choice of reference organism. Therefore, unless there is one reference organism very similar to the target organism and the metabolic network of the reference organism has been generated accurately, the draft metabolic network derived by these methods is not trustable before extensive manual curations.

In this work, we demonstrate that through this bioinformatics pipeline we are able to reconstruct the metabolic networks of multiple strains of the same species, which can bring extra information about the metabolism of those organisms in comparison to single organism metabolic reconstructions. The pan and core metabolic networks of *P. marinus* are established and their metabolic capabilities are also predicted, which may provide another scope to investigate these organisms. Further, recent researches (Freilich et al., 2011) also indicate that by reconstructing metabolic networks with high-throughput we are able to anticipate potential interactions among species, including both cooperation and competition. Thus, automatic metabolic network reconstruction methods that are able to generate accurate and complete predictions would benefit these fields by providing high throughput.

In conclusion, we developed a bioinformatics pipeline that can automatically reconstruct metabolic networks for organisms or microbial communities based on their genome sequences and gene annotations. A web tool of this pipeline has been developed. We also test this pipeline with *Escherichia coli* K12 MG1655 and apply it to twelve strains of *P. marinus*. The reconstructed metabolic network of

Escherichia coli was compared with reference metabolic network and demonstrated the quality of the automated reconstructed metabolic network. The pan and core metabolic networks of *P. marinus* were established. By mapping the variable part of metabolic networks of the twelve strains with the environmental conditions, we identified several factors that shaped the metabolism of strains, including light and sulfur sources; as well as factors that have little effects on differentiating metabolic networks of the two ecotypes, including nitrogen sources and phosphorus source. Through these results, we illustrate that high-throughput metabolic reconstructions pipeline can bring extra information besides what is contained in the input genomes and annotations.

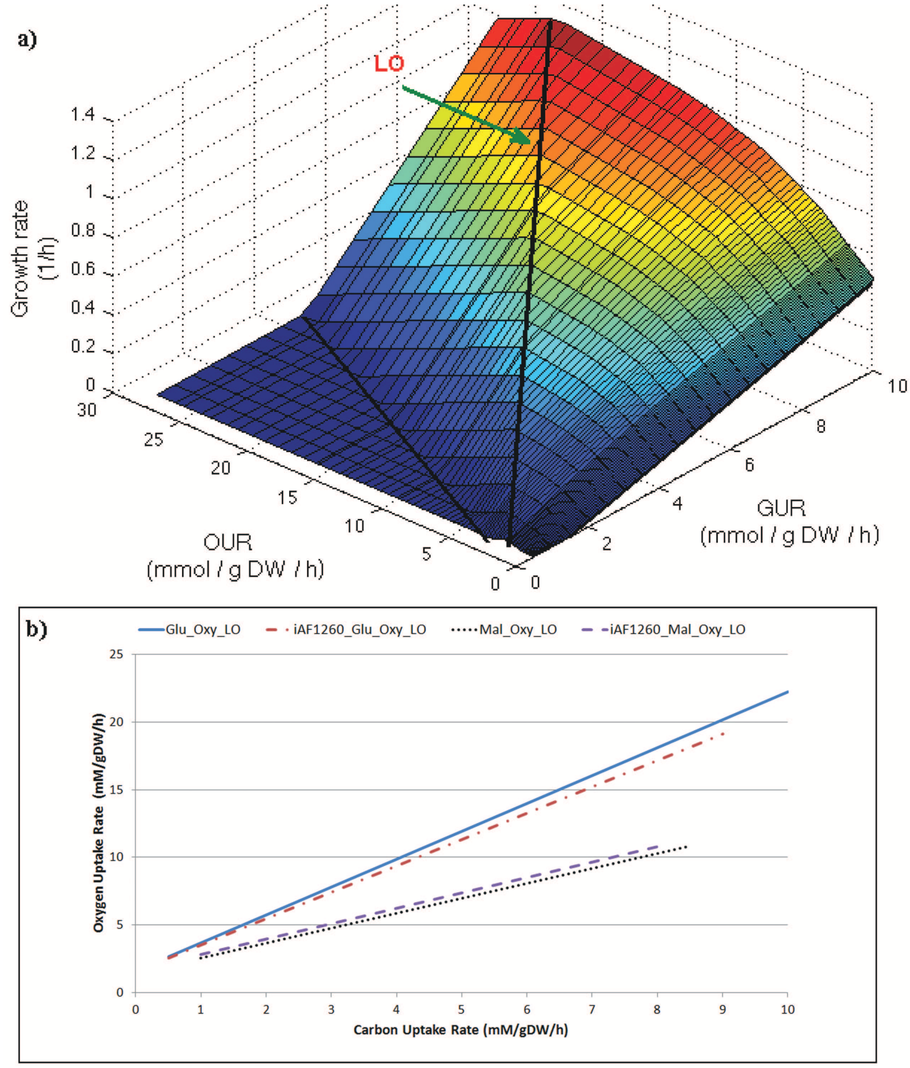


Figure 2.4: Phenotype phase plane (PPP)(a) and lines of optimality (LO)(b) for *Escherichia coli* K12 MG1655 predicted by automated metabolic network reconstruction.

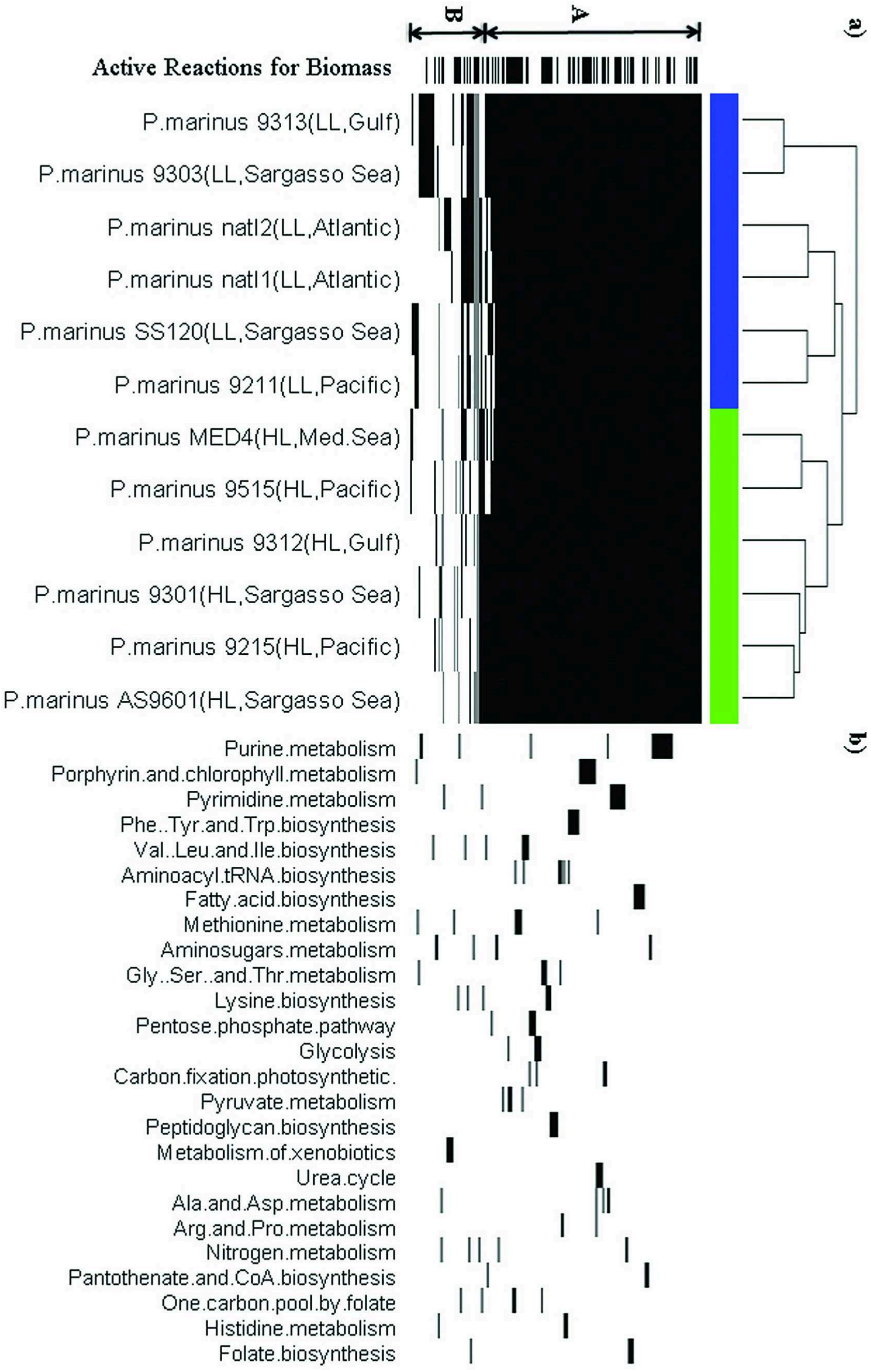


Figure 2.5: a) Metabolic reactions and b) pathway information for twelve strains of *P. marinus* predicted by automated metabolic network reconstruction.

CHAPTER III

Metabolic Network Reconstruction of Acid Mine Drainage Biofilms

3.1 Introduction

Acid mine drainage (AMD) is a worldwide environmental problem caused largely by the microbes in the biofilm (*Singer and Stumm*, 1970). Extensive efforts have been made to understand the role of this biofilm and the ways to eliminate it. Figure 3.1 demonstrates the interactions between the AMD biofilm with its environment. By utilizing the oxygen in air and the ions in the AMD solution, the biofilm is able to gain energy by oxidizing the Fe^{2+} to Fe^{3+}. The natural oxidization of Fe^{2+} to Fe^{3+} is much slower. Therefore the reaction taking place in the solution is accelerated by those regenerated Fe^{3+}. The energy harvested from the ions oxidization can be used to nitrogen fixation and carbon fixation, enabling the growth of AMD biofilm in this environment with poor nutrients.

Growing interests in microbial communities and the power of metabolic network modeling for single organisms naturally lead to a hypothesis that reconstructing the metabolic network of a whole microbial community might provide new systems-level insights of its complex metabolic interactions and functions (*Stolyar et al.*, 2007). On the other hand, emerging metagenomic sequencing data provide almost all the

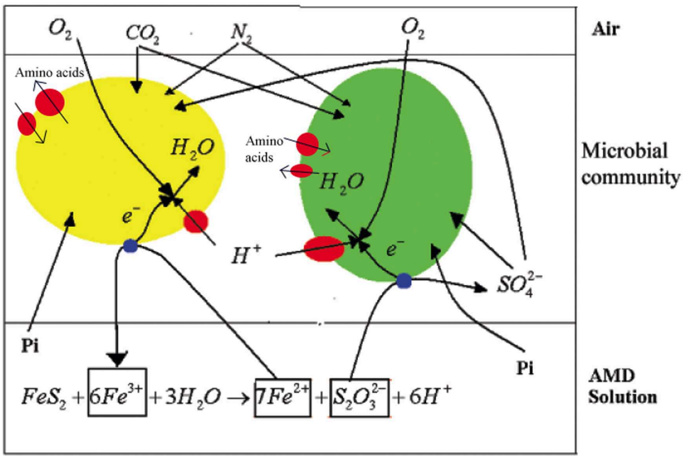


Figure 3.1: The ecological roles of Acid Mine Drainage (AMD) biofilm in AMD formation.

information necessary for the reconstruction. However, in contrast to genome data of individual organisms, environmental shotgun sequencing data might not provide complete genetic information for all the organisms in the communities, which poses tremendous challenges for network reconstruction based on the metagenomic sequence. According to the work of *Mavromatis et al.*, over 20% of all the genes in a dominant microorganism co-existing with others in a community could not be identified from the metagenomic sequences. For organisms that are not dominant in the community, even fewer genes can be identified.

One of the AMD biofilm communities (5wayCG site) was sequenced by *Tyson et al.* in 2004, and further revised later (*Goltsman et al.*, 2009). Five major species was identified: *Leptospirillun*Gp II (75%), *Leptospirillun* Gp III (10%), *Ferroplasma acidarmanus* I & II (10%), and *Thermoplasmatales archaeon* Gp(less than 5%). Table 3.1 summarizes this metagenomic dataset. Subsequent works also provided more information about the genomes and proteomes of organisms in this or closely related communities (UBA site) (*Ram et al.*, 2005; *Lo et al.*, 2007; *Goltsman et al.*, 2009). Due to the relatively simple structure, most of the metagenomic sequences can be classified into one of the five organisms, and nearly complete genomes of them were collected.

Table 3.1: Summary of metagenonic dataset of AMD biofilm. Data collected from work of *Tyson et al.* in 2004.

Organism	Population fraction	Genome size	Gene number	Gene with functional annotation
AMD biofilm	100%	10.8M	12820	7095 (55%)
Leptospirillun Gp II	75%	2.2M	2573	1320 (52%)
Leptospirillun Gp III	10%	2.6M	2877	1564 (54%)
F. acidarmanus I	10%	1.5M	1702	1122 (62%)
F. acidarmanus II		1.8M	2588	1325 (51%)
T. archaeon Gp1	<5%	2.6M	3608	1578 (57%)

The aforementioned Pipeline for Metabolic Network Reconstruction (PEER) provides us powerful tools for high-quality metabolic network reconstruction from annotated genomes. To apply this tool for community-wide modeling, we want to use PEER to generate community-wide metabolic model for AMD biofilm from the metagenomic datasets, which can predict the intracellular metabolism and interspecies interactions in the AMD biofilm. This information might provide information for AMD treatments. More specifically, there are three major objectives we want to achieve.

- Reconstruct individual metabolic networks for the major organisms in the AMD biofilm from metagenomic datasets.
- Provide mechanisms about the metabolism of the five organisms in the AMD biofilm using reconstructed metabolic networks.
- Develop multiple-organism metabolic models to predict the interspecies interactions that are essential for AMD biofilm formation.

3.2 Metabolic Network Reconstruction and Modeling Methods

3.2.1 Metabolic Network Reconstruction for Individual Organisms in AMD Biofilm

For individual metabolic network reconstruction, we assumed all the organisms must be able to grow independently in the AMD environment. This assumption was the same as in PEER. Therefore, we could apply the PEER to the metagenomic data of the Acid Mine Drainage (AMD) microbial community to derive the metabolic networks of five major species. All three steps in the PEER were applied and briefly described here. First, a draft network is constructed using metabolic reactions whose corresponding enzymes have been identified in the existing metagenome annotation. At this step, the main data sources include existing functional annotations of genes and the reaction/pathway database from KEGG(*Goto et al.*, 1999, 2000, 2002). Because of the complexity of the metabolic network in a microbial community and the relatively low accuracy of annotation, this preliminary network is far from completeness, concerning the capability of fulfilling certain essential metabolic functions.

The subsequent automated curation step is based on the assumption that each species (or subdivision) requires biomass and energy for growth. Furthermore, all the metabolites involved must obey mass conservation. At the second step, by imposing these constraints and other information, such as reversibility of reactions, our automated curation process, based on the mixed-integer linear programming (MILP) framework (*Floudas*, 1995), provides a set of alternative metabolic networks, each of which contains specific putative reactions that have not been identified in the existing annotation. Then a comprehensive gene-search process based on BLAST (*Altschul et al.*, 1990) is carried out, to evaluate the probability of existence for these putative reactions in the genome . Here, nucleotide sequences of the (meta)genome and those from the KEGG database of previously known genes for each desirable putative reaction are utilized.

Finally, at the last step, the resulted probability estimates are incorporated in the MILP optimization model to generate the most reliable networks with the minimum risk of utilizing non-existing putative reactions. The model contains both binary decision variables, which determine whether to include a reaction in the network or not, and continuous variables representing the fluxes (reaction rates). Constraints include mass balances of all metabolites, reversibility of reactions, availability of nutrients from the environmental, and specifications of network functionality such as biomass generation. The objective is to minimize the overall risk of adding putative reaction in the network, which is represented by a function of the binary decision variables and the probability of each putative reaction estimated from the previous step.

3.2.2 Prediction of Pathway Probability

To calculate the probability of pathways, parallel solutions with different gap filling results are integrated. Equation 3.1 demonstrates the way different solutions integrate, in which n_r represents the number of occurrences of reaction r in all solutions. Applying this strategy, the probabilities of each pathway can be predicted, which provides more insight than only considering the number of reactions. The pathway information used for calculation comes from Kyoto Encyclopedia of Genes and Genomes (KEGG)(*Goto et al.*, 1999, 2000, 2002) (`http://www.genome.jp/kegg/`).

$$P_{pathway_i} = \prod_{r \in i} p_r^{1/n_r} \tag{3.1}$$

3.2.3 Multiple-organism Metabolic Model of Microbial Community

To integrate the metabolic networks of the five organisms into community-wide models, we assumed certain molecules can exchange between the organisms. To model these interactions, all organisms form compartments separately in the model just like single organism model. Then a list of metabolites can transfer across species through transport proteins or membrane, which are represented by exchange reactions in the model. However, only limited metabolites are allowed to exchange between organisms. This set of exchange fluxes are predicted based on the transporter prediction from Transporter Automatic Annotation Pipeline(TransAAP) in TransportDB (*Ren et al.*, 2007).

We assumed the interactions with accurate transporter predictions were more likely to take place. Therefore, we predicted the metabolic gaps and potential interaction simultaneously, and modified the objective function as in Equation 3.2, in which b_{add} is the binary variable indicates metabolic gap and $b_{exchange}$ is the binary variable indicates uptake flux. The objective function can be divided into two parts. The first part is the same as in PEER, reflecting the metabolic gap filling. The second part indicates the penalty for exchange fluxes. The penalty levels for exchange fluxes($Weight_{exchange}$) are determined by the specificity of the transporter prediction, and the absolute values do not affect the results as another parameter, overall penalty ratio, is used to balance the two parts in the objective function.

$$obj : \min_{v,b_{add},ve,b_{exchange}} \sum_{r \in R} b_{add}(r) \cdot Weight(r) - ratio \cdot \sum_{re \in Re} b_{exchange}(re) \cdot Weight_{exchange}(re) \tag{3.2}$$

We tested three different values for the penalty ratio, range from 1 to 20. For each setting, both optimal and suboptimal solutions were collected to predict the potential interactions. Because we only give penalties to uptake fluxes, the model emphasizes more on the uptake than secretion. Therefore, the uptake interactions predicted by the model were more accurate than secretion. However, for the two-species model, we only need to predict the uptake interactions.

3.3 Reconstructed Metabolic Networks of Individual Organisms in AMD Biofilm

3.3.1 Metabolic Network of the Five Organisms

The metabolic networks of the five organisms were generated by PEER. The pre-assumed function that all the species must achieve is the synthesis of biomass, which includes twenty amino acids, nucleotides, and some chosen coenzymes. Not surprisingly, the preliminary metabolic networks, based only on the metagenomic annotation, is not sufficient for this basic function, and none of them would be able to grow, which is obviously incorrect as certain strains have already been isolated and cultivated (*Tyson et al.*, 2005; *Baumler et al.*, 2005). The refined results from the automated curation process suggest that a significant number of metabolic reactions are missing in the annotations, which means either these reactions have been abandoned by the species or were not identified correctly during the sequencing and annotating process. Metabolic networks reconstructed from the bioinformatic pipeline provide alternative solutions to understand these missing reactions (gaps). The optimization based algorithm attempts to fill all the gaps in the metabolic network with the best putative reactions, of which gene candidates can be found in the genomes with the lowest p values. As a result, the final metabolic reconstruction is one with the lowest risk of adding non-existing reactions according to current knowledge of the AMD community.

The metabolic network of the whole community and the major species of both

UBA site (*Tyson et al.*, 2004) and 5wayCG site (*Goltsman et al.*, 2009) are generated by the pipeline, and the summary of the individual networks for UBA dataset is listed in the Table 3.2. From the results, it is clear that only part of the metabolic reactions (30%) are actively involved in biomass synthesis, which is similar to the observation made in the metabolic network model of *E .coli*. The metabolic networks generated from 5wayCG dataset are in the same situation. Even through the added reactions only take a small fraction of the overall reactions, they make up a significant proportion in the active reactions, which implies only some of the missing components in the whole network have been identified by our algorithm. These missing components suggest the sequence data and gene annotations still require careful review before directly usage, especially for these analysis that mainly rely on local information in the datasets.

Table 3.2: Summary of the individual organism metabolic network reconstruction in the AMD community in UBA site

Species	Proteome coverage	Reactions identified in genome	Putative reactions	Active reactions	Active reactions confirmed by proteome
Leptospirillum Gp II	64.6%	574	57	476	421 (88.4%)
Leptospirillum Gp III	44.9%	585	51	493	373 (75.7%)

The differences between the metabolic networks in different species could be easily observed, even if we only look at the overall or active reaction number. In addition, we found if two species are close phylogenetically, their metabolic network will also be similar, which can be further supported by detailed analysis of these metabolic networks. Figure 3.2a demonstrates the existence of these active reactions among the five organisms in 5wayCG dataset. There are four major groups of reactions that can be identified from the figure: reactions always exist in all the species, reactions only exist in the two dominant bacteria, reactions only exist in the three archaea organisms, and reactions that are missing in the dominant species (*Leptospirillum* Gp II) but exist in other species. About 34% reactions belong to the first group, and about half of those haven't been identified or specified in the draft annotations. The most conservative reactions in this group are the reactions that belong to the basic amino acids biosynthesis pathways, like transaminase reactions, and the reactions in the central carbon pathways such as phosphoglycerate kinase reactions. The reactions related to nucleotides biosynthesis also belong to the first type, such as nucleoside-diphosphate phophatransferase reactions. Therefore the first group of reactions may be marked as essential reactions. About 15% of the reactions exist in the two dominant species but not in the other species, including some reactions involved in the carbon fixation pathways and sulfur metabolism pathways. Besides that, certain reactions in the purine and urea metabolism pathways are completely missing in the *Leptospirillum* Gp II but exist in *Leptospirillum* Gp III or other species. It must be pointed out that reactions in the same pathways may belong to different groups, which might be caused by the variation of metabolism pathways or cooperation between species.

Figure3.2b is an example of a reconstructed metabolic network, which demonstrates the complexity of these metabolic networks. From the figure, it is clear that certain metabolites are involved in many more reactions than others, such as NADH, ATP and proton, whose concentrations are always well regulated. On the other hand, there are types of metabolites that only appear in some linear pathways, which are sometimes the precursors of biomass components.

3.3.2 Network Gap Filling in the Reconstructed Networks

In general, there is more than one network that can complete the required functions and more than one set of putative reactions could be used to fill the gaps.

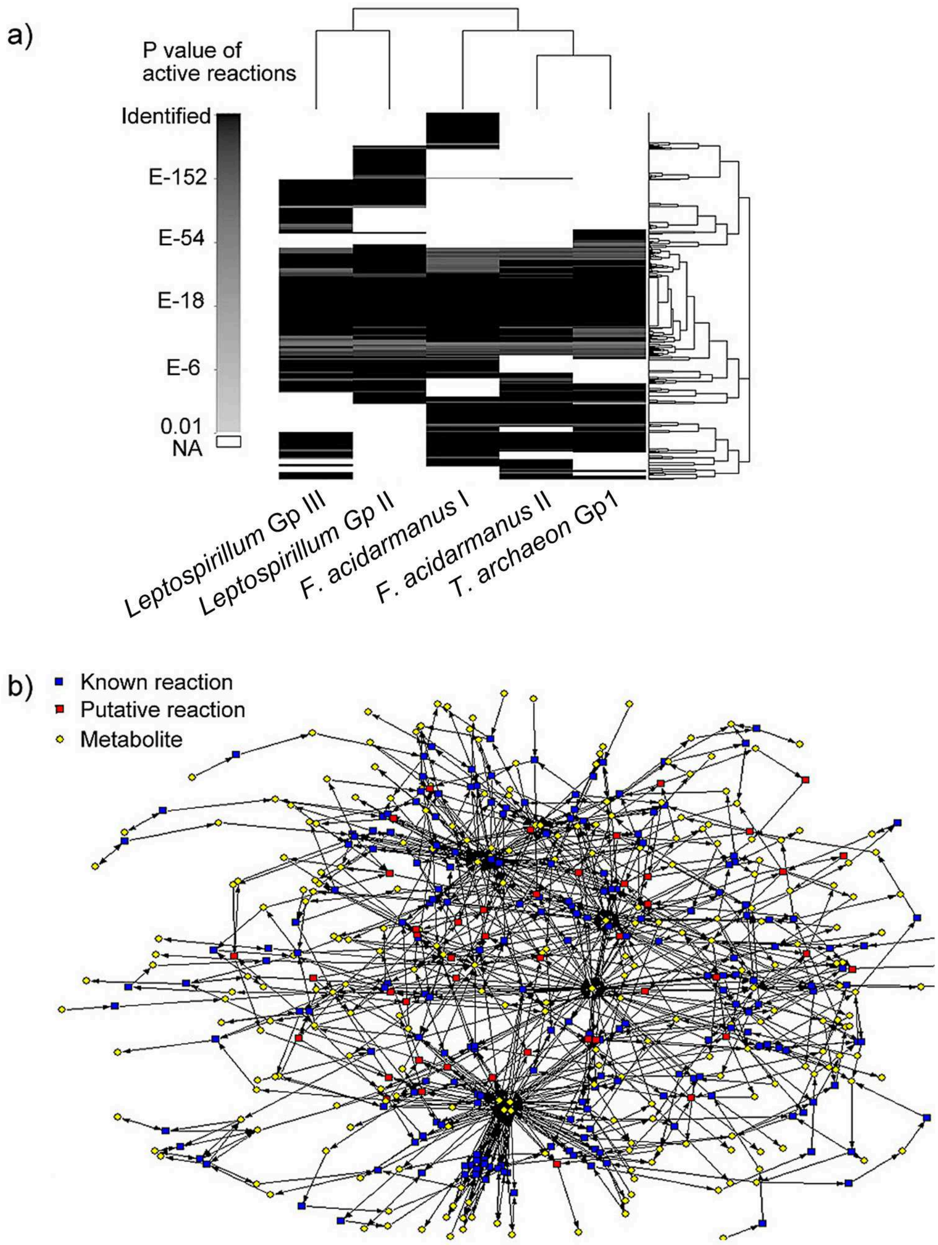


Figure 3.2: Metabolic network of the species in a AMD microbial community. a) Darkness of each bar represents the probability of existence. Both the reaction (column) and the species (row) were grouped by dendrogram. b) Metabolic network of *Leptospirillum* Gp II with 283 intracellular reactions and 288 metabolites, drawn by Pajek (*Bategeli and Mrvar*, 2003).

Furthermore, the gene candidates could be annotated differently, which means these candidates could be assigned different functions with different probabilities based on sequence similarities. In this situation, it is inaccurate to annotate these enzymes based only on sequence similarity as it treats all the target functions equally, which is not true when some functions are more desirable according to the functionality requirements. To deal with this situation, the final choices of the gene candidates are actually coupled with the choice of metabolic network gaps, representing the maximization of the overall conditional probability of the existence of the enzymes (reactions), given the occurrence of the other part of the metabolic network with required functions. All the network gaps are also demonstrated in Figure 3.2a, in which the brightness represents the probability of existence.

Table 3.3: Network gaps filling and the gene candidates of metabolic network reconstruction of individual organisms in the AMD community in UBA site

Organism	All missing reactions	New annotated gene	Better Annotations	Worse Annotations	
				$< 10^{-30}$	$> 10^{-30}$
Leptospirillum Gp II	57	14	15	14	10
Leptospirillum Gp III	51	8	18	15	7

Table 3.3 summarizes the network-gap-filling results for 5wayCG dataset and the gene candidates with low or comparable p-values (*Altschul et al.*, 1990) in the five major species in the community. As described above, it is still possible that because the new assigned functions are highly desired certain gene candidates are chosen whose p-value is larger than the one of the current functions and 10^{-30}. According to the results more than half of the putative reactions in *Ferroplasma acidarmanus* I, II and *Thermoplasmatales archaeon* Gp1 belong to this type, which means they are difficult to identify directly from the sequence alignment if the functionality of the whole network is not taken into account. In Table 3.4, some examples of the gene candidates are listed, and the details of all the network gaps and gene candidates can

be found in supplementary materials.

Table 3.4: Examples of the network gaps and the gene candidates of the AMD microbial community in UBA site

Network gap	Gene candidate	P-value of candidate	Previous annotation	P-value in previous annotation
Reduced ferredoxin:dinitrogen oxidoreductase	UBAL2_82410013	9.80E-73	Putative ATP binding protein, mrp like	NA
ATP:pantothenate 4'-phosphotransferase	UBAL2_80270023	5.90E-18	Transcriptional acitvator, Baf family	NA
L-Cystathionine L-homocysteine-lyase	UBAL2_82410105	2.50E-74	Cystathionine gamma-synthase	1.00E-84
L-arogenate:NAD+ oxidoreductase	UBAL2_81350086	1.80E-87	3-phosphoshikimate 1-carboxyvinyltransferase	1.0E-115
L-Valine:2-oxoglutarate aminotransferase	UBAL2_82410307	5.00E-92	Branched-chain amino acid aminotransferase	2.00E-77
ATP:nucleoside-phosphate phosphotransferase	UBAL2_82410600	8.40E-83	Uridylate kinase	8.00E-12

Three types of gene candidates have been found in the metagenome: genes which have not been assigned a function in the existing annotation, genes which could be assigned new functions with better sequence similarities, and genes that might be assigned alternative functions with comparable sequence similarities. One explanation for these three types of candidates is that the gene database used in the annotation process (e.g. COGs) only contains enzymes from limited organisms (26 in 2004), which may be phylogenetically far from some species in the community. In that case, a good sequence match does not guarantee similar enzyme function, which could also explain the existence of third type candidates. More enzymes and genomes have been sequenced and understood currently, and using a updated enzyme database in candidate searching could make the reconstructed metabolic network more reliable when dealing with the previously identified metagenomes and annotations. Certain gene candidates have been assigned similar functions as in the previous annotation but with different details, such as co-enzymes, which can specify the metabolic reactions. We also notice that certain gene candidates have been assigned the same functions as in previous annotation, which indicate these metabolic gaps identified through our pipeline might not be real gaps just not recognized by the algorithm. This can be caused by inconsistence of enzyme names and annotation across databases or annotation methods.

These results show that most of the metabolic network gaps can be filled with relatively high probability of existence (dark gray bars in Figure 3.2); however, there are certain selected metabolic reactions cannot be filled with good gene candidates, which are still reasonable. Unknown enzymes, unexpected mechanisms, and incorrect metagenome sequencing all may cause these low probability network gaps. Even though they are the best choices based on current knowledge, these filled gaps should be carefully reviewed in further studies.

3.3.3 Carbon Flows in the AMD Biofilm

Figure 3.3 illustrates central carbon flows in the dominant organisms, *Leptospirillum* Gp II and III, according to the metabolic network reconstructions. It must be noticed that more than one solution is used to generate these pathway maps for each organism and we are trying to capture those features that are consistent among solutions, even through some diversities are also reflected.

The major carbon fixation in *Leptospirillum* Gp II and III is completed by reductive carboxylate cycle, which agrees with some recent studies (*Goltsman et al.*, 2009). However, not all of the reactions are essential for them. There might be a break between succinyl-CoA and 2-Oxoglutarate, without which the carbon fixation

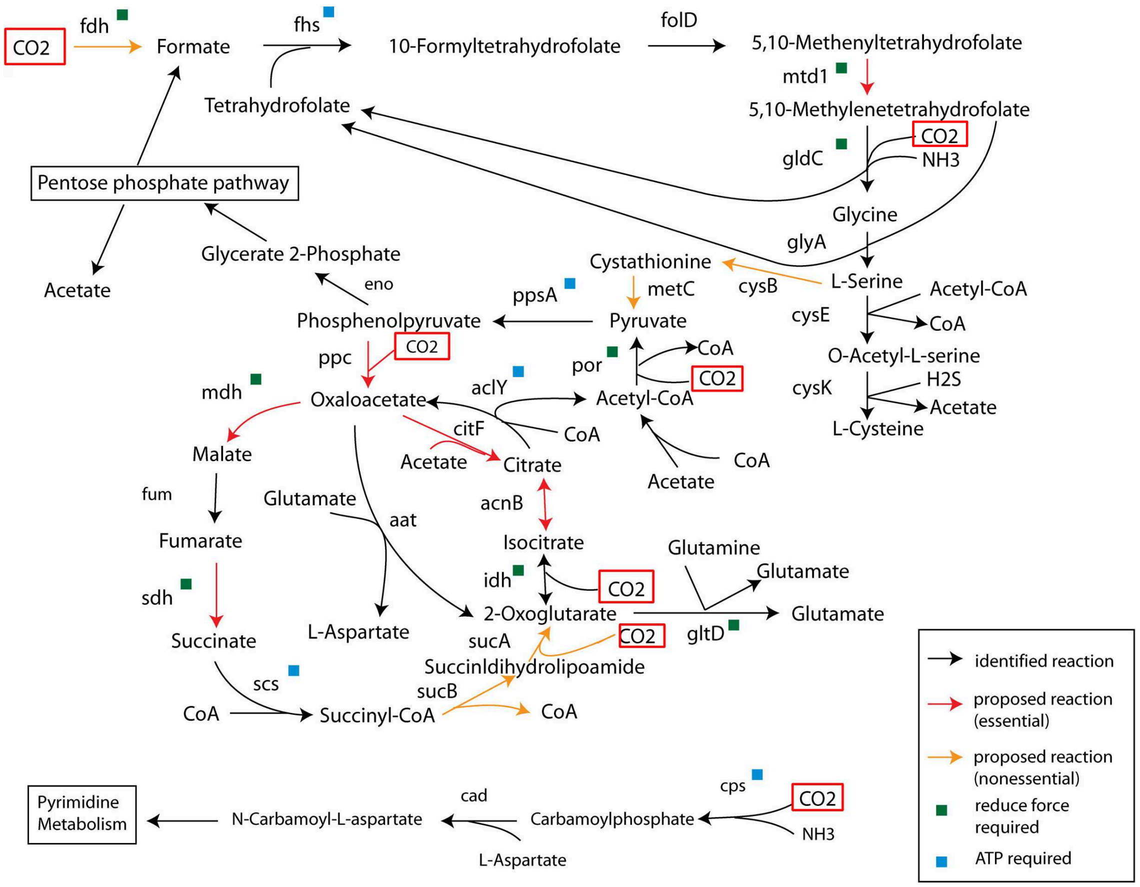


Figure 3.3: The carbon flow of the two dominant species in the AMD microbial community. The seven potential metabolic reactions directly related to carbon fixation in *Leptospirillum* groups are highlighted.

can also be made. An alternative carbon fixation may be carried out by formate dehydrogenase (fdh), which fixes the CO2 into formate and through a long pathway then fixes another CO2 into glycine. The fraction of this carbon fixation varies from 0 to 20% in different solutions, and the two pathways are connected by pyruvate.

Ferroplasma acidarmanus I and II have similar pathways regarding carbon fixation when we required them to do so (Figure 3.4.a), which might be inaccurate when considering interspecies interactions and will be discussed further in multiple-organism model section. Under the autotrophic assumption, one difference of *Ferroplasma acidarmanus* I and II carbon fixation is the carbon fixed through formate is essential for the organisms and can take up to 50% of the overall carbon fixation. Another difference is the gap between succinyl-CoA and 2-Oxoglutarate is not filled by any solutions, indicating they are abandoned by the organisms. There are also slight variations of the reactions that transfer acetate to acetyl-CoA. At the same situation, *Thermoplasmatales archaeon* Gp1 mainly fixes carbon through the incomplete reductive carboxylate cycle (Figure 3.4.b). The aceyl-CoA is generated through a set of reactions belong to lysine biosynthesis and threonine biosynthesis pathways, which are both essential to meet the amino acid requirement for biomass synthesis. Compared to the other four organisms in the microbial community, the efficiency of this type of carbon fixation is very low, which implies that *Thermoplasmatales archaeon* Gp1 should rely or partly rely on external organic carbon sources.

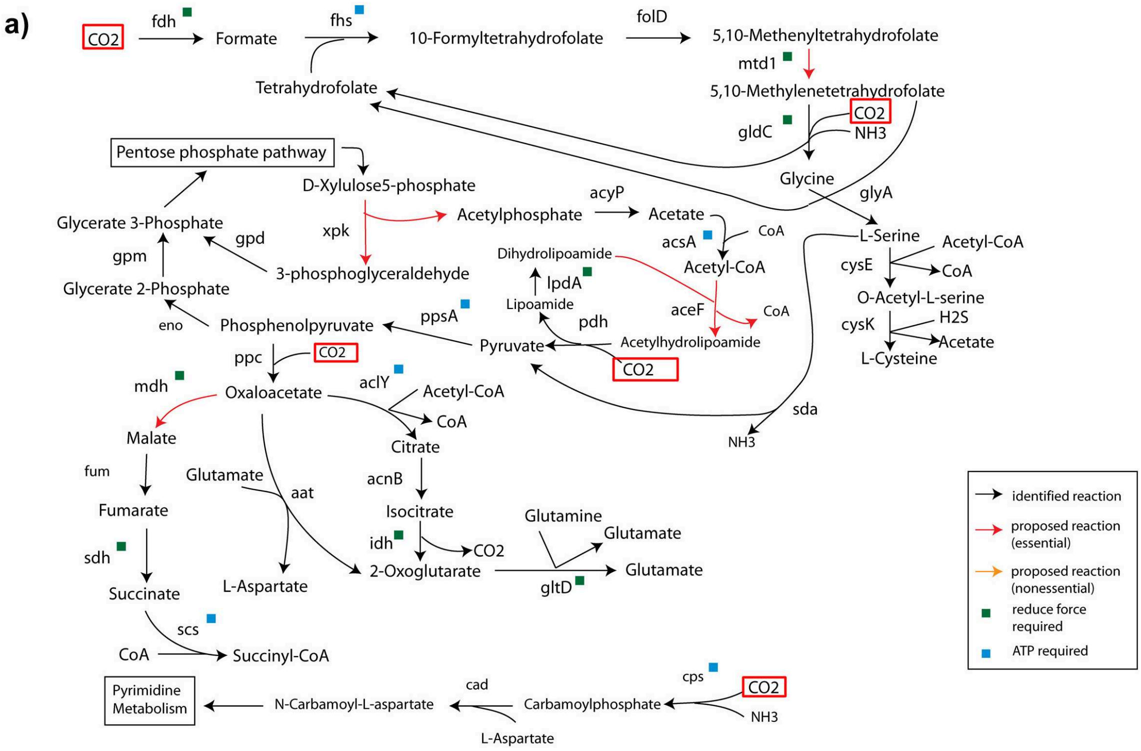


Figure 3.4: The carbon flow of *Ferroplasma acidarmanus* I and II (a) and *Thermoplasmatales archaeon* Gp1 (b). The five potential metabolic reactions directly related to carbon fixation in *Ferroplasma acidarmanus* and three in *Thermoplasmatales archaeon* are highlighted (Continued on next page).

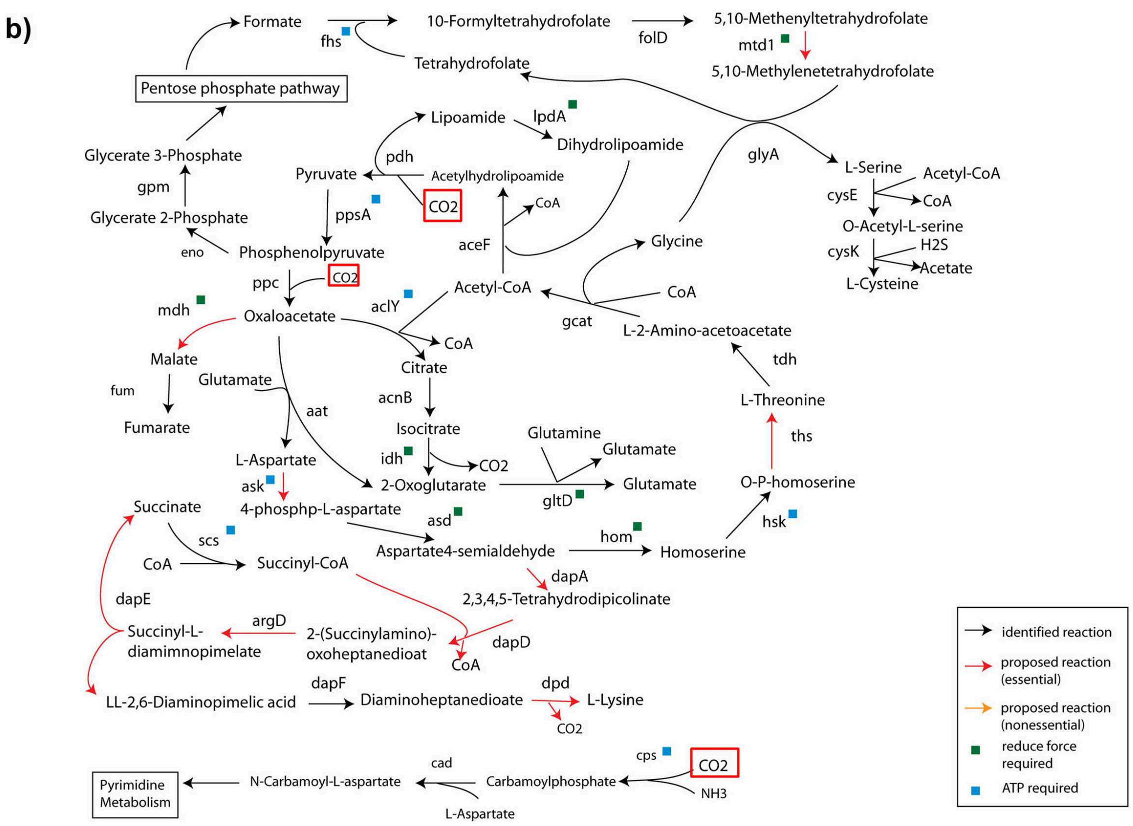


Figure 3.4: The carbon flow of *Ferroplasma acidarmanus* I and II (a) and *Thermoplasmatales archaeon* Gp1 (b). The five potential metabolic reactions directly related to carbon fixation in *Ferroplasma acidarmanus* and three in *Thermoplasmatales archaeon* are highlighted.

3.3.4 Pathway Distribution in the AMD Biofilm

Based on the metabolic network reconstructions of the five species, it is possible to predict the existence of each active pathway. Furthermore, the network reconstructions from this pipeline contain two types of reaction, identified and putative. The later ones may be assigned coefficients representing probabilities, which should also be considered. The information from multiple solutions of individual organisms can be integrated and then provides more reliable results. As shown in Figure 3.5, different pathways have different fractions that are active for biomass synthesis. Furthermore, the calculated probabilities of these active parts are quite diverse among species, implying potential interactions between them. All the pathways can be divided into three categories, pathways with high probabilities in all the organisms, pathways with high probabilities in part of the organisms while medium probabilities for the rest, and pathways with extremely low probabilities in part of the organisms but with medium or high for the rest. Glutamate metabolism belongs to the first group and most of essential reactions in this pathway are active and well identified. There is another type of pathways also belong to the first group, in which only a small fraction of reactions in the pathways are active. These active reactions are not likely be able to complete the functions of those pathways but provide necessary paths for other pathways, like reactions in porphyrin and chlorophyll metabolism pathway. Both citrate cycle and reductive carboxylate cycle pathway belong to the second type, which does not mean they are not active in certain organism, oppositely the result suggests that these pathways maybe active in all the organisms but some enzymes in this pathway need further identification or certain variations exist in the pathways. This finding is coincident with the carbon flow results. Histidine metabolism and valine, leucine and isoleucine metabolism pathway also belong to the second group, indicating there might be interactions taking place. The methionine metabolism and lysine biosynthesis pathway belong to the third group, which suggest one or more organisms might completely rely on the external supply of these amino acids.

It is interesting that the pathway distributions between the two dominant species in UBA dataset are much less diverse (Figure 3.6), indicating the major interspecies interactions in this microbial community may not take place between the dominant organisms, as their metabolic functions are very similar. The UBA dataset was collected in the downstream of 5wayCG site. This may explain the increasing similarity of the two organisms as they are co-evolving along the river.

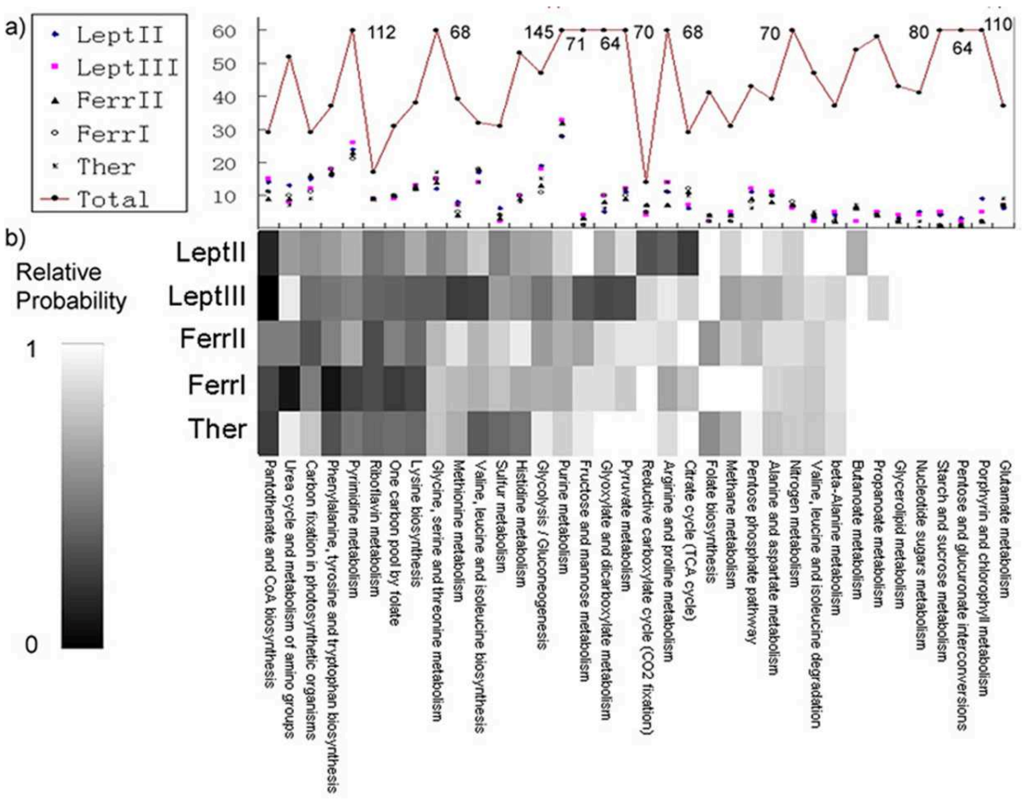


Figure 3.5: a)Number of active reactions in pathways of the AMD biofilm. b) Active metabolic pathway distribution in 5wayCG dataset of AMD microbial community.

3.4 Multiple-Organism Metabolic Modeling of AMD Biofilm

3.4.1 Two-species Model for Major Organisms in the AMD Biofilm

Even through biomass synthesis is a basic function that all the organisms should complete, it is possible that organisms may rely on external supplies of certain components from the biofilm, which is also suggested by the pathway distribution analysis. Mean-while, more information must be taken into account when predicting this type of interaction in addition to the intracellular metabolic network of each organism. To achieve this prediction, a multiple-organism metabolic network reconstruction model is developed, which simultaneously considers the intracellular metabolic network together with intercellular exchange fluxes among multiple organisms in the microbial community. One of the major difficulties in this model is

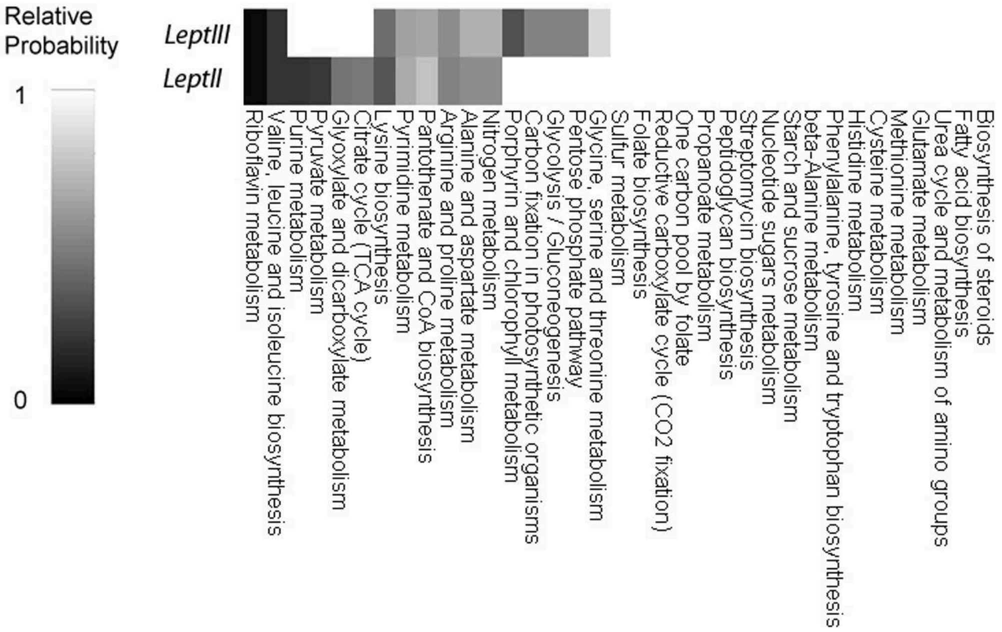


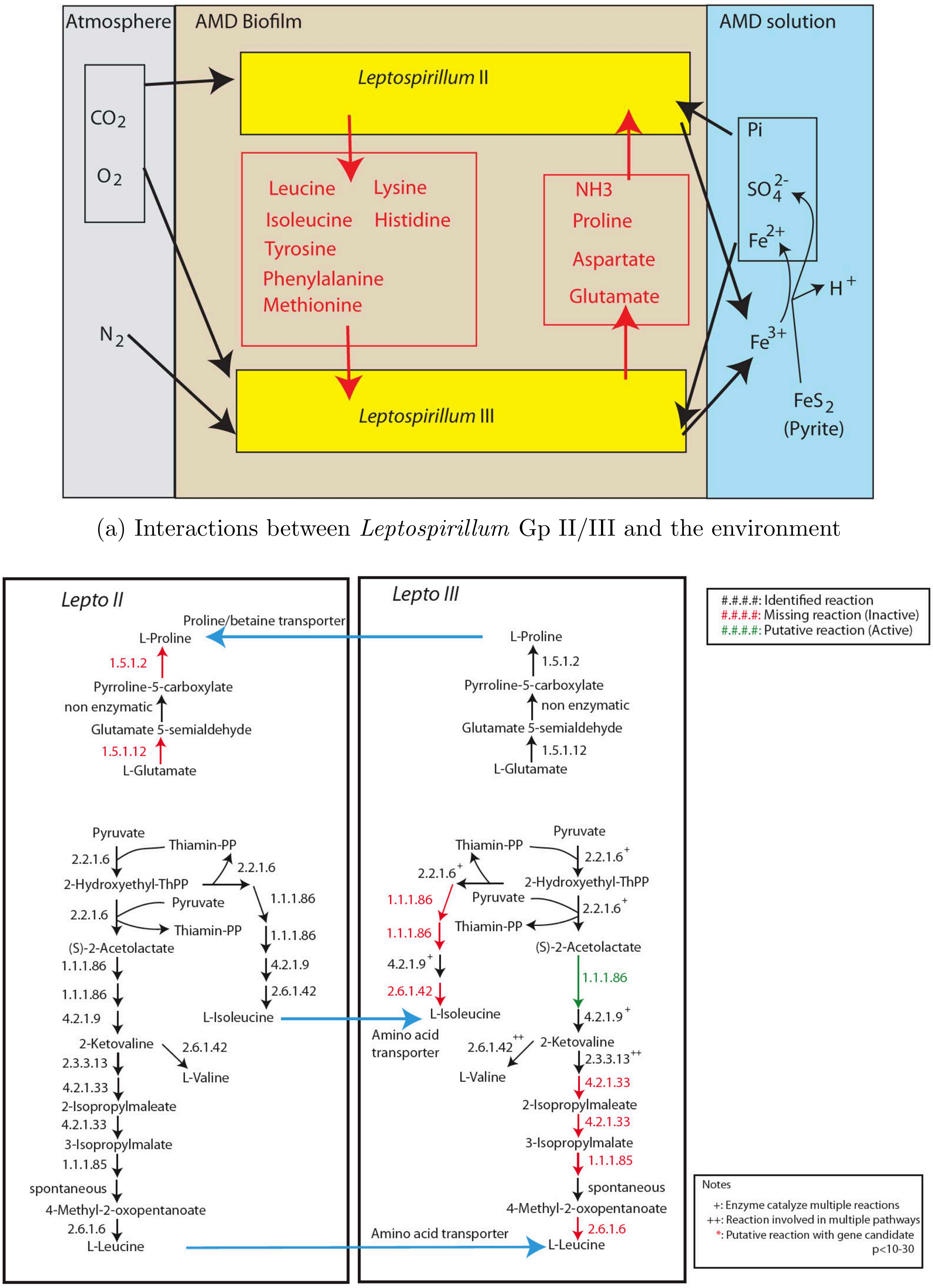
Figure 3.6: Active metabolic pathway distribution in UBA dataset of AMD microbial community.

to predict the potential transporter proteins and corresponding specific substrates list with reasonable confidence level. To get further support for these predictions, analysis of the transporters and transmembrane proteins in the organisms is made with specified method (Transporter Automatic Annotation Pipeline(TransAAP) in TransportDB (*Ren et al.*, 2007)). From the results, more detailed transporters are predicted, for example, uptake of glutamate, aspartate, and proline in *Leptospirillum* Gp II could be supported by finding the amino acid (glutamine/glutamate/aspartate) transporters and proline/betaine transporter. Also the cationic amino acid transporter in *Leptospirillum* Gp III is identified, which is coincident with the predicted uptake of lysine, arginine and histidine. The ammonium transporters are found in all the five species, which are necessary for the predicted relation in ammonium metabolism. Some other transporters whose substrates have not be specified with enough detail might also be correlated to the predicted interactions, such as the general amino acid transporters. By evaluating all this information, a set of numerical coefficients are defined for all the potential exchange fluxes (Supplementary Materials S4).

This multiple-organism model was first applied to the two bacterial organisms *Leptospirillum* Gp II and III, which account for 80% of biofilm, in both UBA dataset and 5wayCG dataset. Different from the single organism model, the intercellular exchange fluxes rather than the intracellular metabolic reactions are the most important determination variables with more interest. Therefore, the overall objective function in this method contains two parts, the penalty of adding uncertain metabolic reactions and the penalty of adding uncertain intercellular exchange fluxes. As a result, the optimization needs to balance these two parts of uncertainness and a separate parameter was assigned to represent this compromise. The prediction of exchange fluxes is partly dependent on this separate parameter, representing the overall penalty level of exchange fluxes. Thus this parameter should be able to reflect different factors, including the mass transfer resistance, the distance between organism clusters and the quality of transporter prediction compared with intracellular enzyme annotation. Because of the complexity, it is almost impossible to calculate this parameter accurately. To solve this problem, a set of penalty levels, distributed in three magnitudes, were arbitrarily set and tested.

Figure 3.7a demonstrates the interactions between the two organisms and the environment, the results are derived from one solution of 5wayCG dataset in 5 fold penalty level. To understand these predicted interactions, we also analysis the related metabolic pathways for these interactions, which can also testify the predictions. Two examples are shown in Figure 3.7b. From the figure, we can find that the predicted interactions are caused mainly by the diversities of the metabolic networks. If the pathway to synthesize one necessary metabolite is more complete in one organism than in the other one, and at the same time this metabolites can be exchanged between the organisms, the interaction may take place through this metabolite, which can reduce the uncertainty of the metabolic networks. Figure 3.8 is the prediction of interactions between *Leptospirillum* Gp II and III at three different penalty levels in both the two datasets. In all these solutions, every uptake flux requires secretion flux from the other organism but not every secretion flux has corresponding uptake flux, which is reasonable because we only consider dominant organisms in this model and there may be other heterotrophs exist. In addition, we found that the interactions predicted based on 5wayCG site is more stable than those based on UBA site crossing the different penalty levels. This finding is coincident with the previous results of pathway distributions. One explanation is the multiple-organism model will be more sensitive to the parameter of penalty level when the metabolisms of the different organisms are more similar, and as consequence, the predicted interactions will be less reliable as we cannot provide an accurate estimation of the penalty level in most cases.

Not surprisingly, as the increasing of the penalty levels, fewer interactions are found. Even through the interactions exist at high penalty level means more important or efficient for the microbial community, directly relating the penalty level with confidence level of prediction is incorrect, because ideally there is a “true” value for the penalty level in a specific system, which means either direction of the perturbation can introduce false negative or positive errors. In extremely cases, when the metabolism of two organisms are the same or highly similar, for example two strains of one genus, a relative high penalty level might still lead to certain number of putative interactions. According to the results, generally *Leptospirillum* Gp II will rely on *Leptospirillum* Gp III for organic nitrogen supply through ammonium or amino acids and *Leptospirillum* Gp III also takes in certain amino acids, such as tyrosine and lysine. This prediction is coincident with the knowledge that *Leptospirillum* Gp III is the only nitrogen fixation organism in the AMD microbial community(*Goltsman et al.*, 2009; *Tyson et al.*, 2004, 2005). Amino acids



(a) Interactions between *Leptospirillum* Gp II/III and the environment

(b) Examples of the interactions and related pathways

Figure 3.7: The interactions in AMD microbial community in 2-organism model. a) Summary of predicted interactions between the two dominant organisms and interactions between microorganisms with the environment. b) Pathways related to three predicted interactions, which can explain the mechanisms for these interactions.

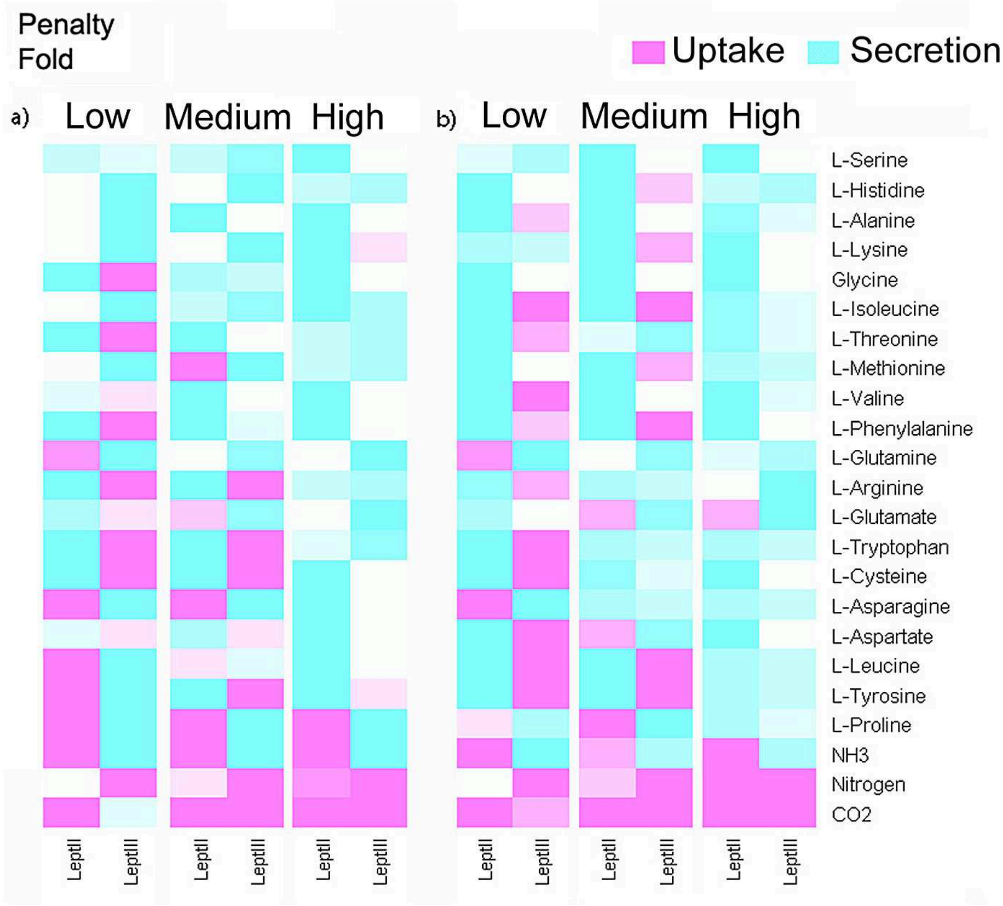


Figure 3.8: The interactions in AMD microbial community in 2-organism model. Complete predictions of interactions between *Leptospirillum* Gp II and III. Left for UBA dataset and right for 5wayCG dataset.

can also be potential carriers of organic carbon, which was also observed in the solutions. If these interactions are verified, the two dominant organisms form a cross-feeding mutualistic relation. Furthermore, according to the solutions, there are always more interactions than necessary, even through at relatively high penalty levels. This finding might be explained by the stability of mutualistic communities is positively related to the species connectivity(*Okuyama and Holland*, 2008) and the redundant interactions may improve the stability of the microbial community.

3.4.2 Five-species Model for All the Organisms in the AMD Biofilm

To fully capture the structural interactions in this AMD system, a model containing all the five identified organisms was made. Compared to the two-organism model, the five-organism model is more complex and requires more computational effort. On the other hand, because the less abundant organisms may also play important roles in natural microbial systems, it is still worth employing five-organism model to explore the interactions among the five organisms. Different from the simplified case, the interactions in the five organism model are asymmetric and more metabolites may get involved. A similar strategy was applied to determine the proper penalty level for interactions (shown in Figure 3.9). From the figure, it is clear that higher penalty level can reduce nonessential interactions, which is that same as two organism model. One difference is distinguishing the role of each organism becomes much more difficult due to the high connectivity between organisms, which might be true in real cases. Attentions should also be paid to the unstable interactions observed in the results, which suggests different phenotypes of one species might have similar roles in the microbial community and all of them have comparable uncertainness according to current information. This computational diversity suggests there may be more than one genotype of one species in the same microbial community, which have been observed in the AMD microbial community (*Allen and Banfield*, 2005). As same as found in two organism model, amino acids are used as carbon source and nitrogen source for some organisms, on the other hand ammonium is another nitrogen source for organisms other than *Leptospirillum* Gp III. The top choices of exchanged amino acids were predicted from these result, including lysine, tyrosine, and phenylalanine. The biosynthesis pathways of these amino acids are also distributed diversely in the microbial community, which are suggested in the pathway distribution analysis. However, only part of these pathways indicated in previous analysis are finally interacted in the multiple-organism model, which are more specific and trustable.

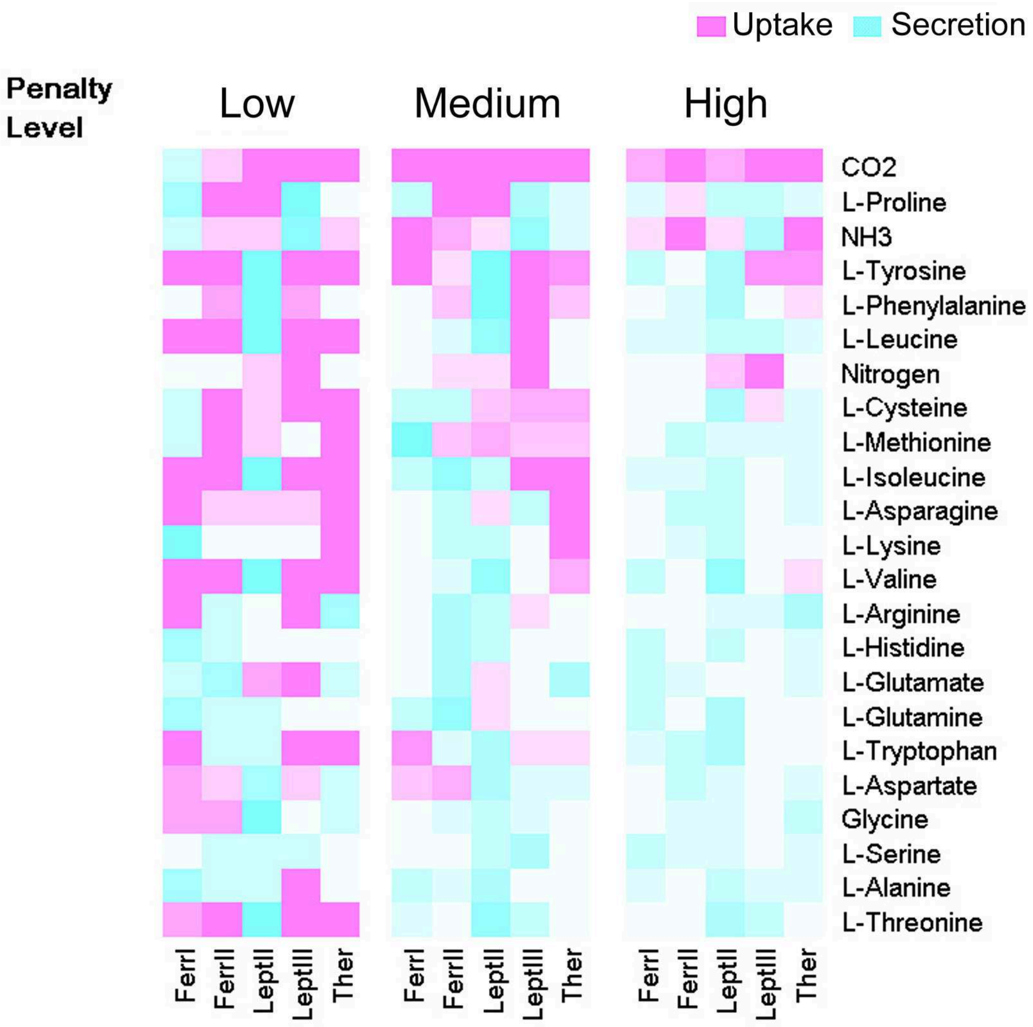


Figure 3.9: Potential interactions among five organisms in the AMD microbial community of 5wayCG site under three penalty levels in 5-organism model. The darkness of colors represents the probability of the prediction. For example, the darkest interactions shown in the highest penalty level are those most likely interactions predicted by the 5-organism model.

This five-organism model can provide more comprehensive intracellular metabolic network reconstructions for all the five organisms (Table 3.5). Different from the reconstructions in the single-organism model, these results take the intercellular interactions into account when making the reconstruction, during which the less probable reactions and functions would be eliminated. Thus these reconstructions may reflect the real situation better than those from single organism model. Furthermore, the objective function values in multiple-organism models are always less than the summation of five objective values in single organism model, which indicates less uncertainness of the putative reactions. Therefore, the metabolic networks of five individual species reported in the supplementary materials are derived from these reconstructions. The pathways that contain these metabolic reactions provide another scope to interpret the metabolic network at higher level. We found that certain pathways are conservative in the whole microbial community, including the purine metabolism and phenylalanine, tyrosine and tryptophan biosynthesis pathway. However, there are a few pathways that parts of the pathways are conservative while some other parts only exist in some species. For example, about one third of the valine, leucine and isoleucine biosynthesis pathway is mostly contained by the there archaea while the rest part is still conservative. There is another type of pathways that does not contain major conservative part in this microbial community, such as glycolysis, pentose phosphate, and Citrate cycle (TCA cycle) pathway. This also indicates the diversity of carbohydrate metabolism within this microbial community, which is consistent with the predicted carbon flow based on single organism model.

3.4.3 Incorporation of Proteome Data in Metabolic Network Reconstruction

Proteome data can provide further information about metabolism of microbial organisms, which can partly project the active metabolic reactions. The proteome of *Leptospirillum* Gp II has become available and more than 48% of the ORFs identified in the genome have been confirmed (*Ram et al.*, 2005; *Lo et al.*, 2007). In a more recently dataset, another AMD microbial community in a different location (UBA site) (*Goltsman et al.*, 2009) has been sequenced and its proteome dataset has also

Table 3.5: Summary of community-wide metabolic network reconstructions of the AMD microbial community in 5wayCG site. The putative reactions and active reactions of suboptimal solutions within 5% optimality gap were reported.

Species	Genome (Mbp)	Identified reactions	Putative reactions	Active reactions	Exchange reactions
Leptospirillum GP II	2.22	569	38(72)	258(306)	44
Leptospirillum GP III	2.66	549	35(65)	253(285)	44
F.acidarmanus I	1.48	454	60(91)	238(281)	33
F.acidarmanus II	1.82	468	58(92)	250(290)	46
T. archaeon	2.64	539	42(66)	251(272)	38

become available. In this proteome dataset, the coverage of *Leptospirillum* Gp II and Gp III have reached to 64.6% and 44.9%. These identified ORFs provide another perspective of the metabolism within the microbial community.

This meta-proteome data contains information about the two dominant organisms in the biofilm, and with even higher coverage of the proteins. Even through the two organisms are also belong to *Leptospirillun sp.* Gp II and *Leptospirillun sp.* Gp III, we still reconstruct the metabolic networks of the the dominant species based on the metagenome from the same location using the descried method, to derive more accurate results. Some details of the metabolic network reconstruction are listed in Table 3.2. From the results, we notice that the reconstructed metabolic network is with higher quality than the earlier one, as both the number of additional reactions and the penalty of adding these reactions are reduced. We believe this improvement is due to the better gene annotations in the latest dataset. The two organism model is also applied to this microbial community to predict the interactions. The reconstructed metabolic networks are compared with the proteome dataset, and part of the network is confirmed by the proteome data (Table 3.2). Compared to the coverage of the proteome dataset, the fraction of confirmed metabolic reactions is significantly higher (p values $< 10^{-200}$ in binomial test), indicating the confirmed reactions are enriched in the reconstructed networks. This enrichment of identified proteins or reactions partly validates the metabolic network reconstructed by the pipelines and provides some other indicators about the quality of the reconstruction, which is hard to evaluate before. In the proteome dataset the identified proteins have different confidence levels, for example the number of peptides identified is different. Not surprisingly, the corresponding enzymes of the active reactions are also distributed in different confidence levels, and some of these proteins that are highly recommended by the model might only have one or two peptides been identified, which might be able to provide another scope of protein identification in proteome dataset. We also summary the number of peptides that are confirmed in the dataset for those active reactions (Table 3.6). From the results, it is surprising that the fractions of confirmed putative reactions are not significantly lower than these of overall active reactions. One explanation can be there is no difference in difficulties between identifying putative enzymes and other proteins, even through the former ones are more difficulty in function annotation.

We investigate the proteins which were observed in the proteome data but not in the metabolic networks. These proteins may conduct other functions rather than the biomass synthesis and growth, for example, secondary metabolisms. Part of these metabolic enzymes are selected and forced to be active in the model, which will lead the model to predict the other functions besides growth, which are also interesting. By comparing the byproducts of the new results with the ones without these enzymes, we are able to predict a list of secondary metabolites or their precursors that are potentially produced by these organisms. This list contains some signal molecules, cofactors and even antibiotics. These results indicate another approach that can connect the proteome data with the metabolism of the organisms, especially for those pathways in which the enzymes are expressed in low level and difficult to identify.

Table 3.6: Number of peptides identified in the proteome dataset for the active reactions

Species	Active reactions for biomass	Confirmed Active reactions	Putative reactions	Confirmed putative reactions
Leptospirillum Gp II	476	421	57	47
Leptospirillum Gp III	493	373	51	40

3.5 Discussion and Conclusions

Metagenomic DNA sequencing as well as proteomic studies provides a huge amount of data about the metabolism and structure of microbial communities. However, analysis and synthesis of these data poses substantial challenges in practical applications. For instance, the accuracy and coverage of the data always limit further interpretations. Furthermore, the large scale of the dataset makes it almost impossible to curate manually. Therefore, the automatic curation steps in our metabolic network reconstruction pipeline are essential for successful reconstructions. The algorithms used in the pipeline and the multiple-organism model also allow modifications of the metabolic networks based on the functionality analysis and gene candidate search results. These two strategies make it easier to directly employ the metagenomic sequences and proteomic dataset, even though modifications suggested by these methods not necessarily be true. As discussed before, the mathematically optimal solution does not guarantee a best biological prediction, thus during this work a large amount of suboptimal solutions are also employed to make a final conclusion, which means even if none of these solutions are absolutely correct deriving a conclusion with high confidence is still possible. By analyzing a series of solutions rather than the theoretical optimal one, these methods are able to capture the major properties of the intracellular metabolism and intercellular interactions. At the same time, this diversity of metabolic networks of the same species suggested by the different solutions, which cannot be eliminated during natural selection, implies that certain variation of the metabolic network may not hurt the role of the microbe in the microbial community.

As observed in many works, metabolic network reconstruction may also indicate some inaccurate information within the original dataset, such as improper or uncertain annotations of gene functions. Different from directly manual curation of the metabolic network, we defined a rigid procedure for the automatic curation in the bioinformatic pipeline for metabolic network reconstruction. According to this algorithm, the pipeline will search for the most probable perturbations of the metabolic network and then evaluate the overall uncertainness of the new metabolic network, which is used as penalty in the following optimization process. As a result, the modifications calculated by the pipeline are the most necessary and likely ones. The probability of existence, which is used to calculate the uncertainness of additional reactions, is derived from the sequence alignment results. It is also possible to combine other gene annotation methods to define a better penalty parameter, which can be a direction for further research. As long as the evaluation of uncertainty can capture the major profiles, this algorithm will generate an acceptable metabolic network with reasonable modifications, which can be further improved by applying proteomic data.

One basic assumption made in the algorithm is that all organisms need to acquire all the components of the biomass, either by making them or ingesting them. However, the detailed biomass composition of natural microbial communities is not as available as other information, making it is necessary to predict a reasonable biomass composition. In the AMD project, the predicted biomass compositions simply include the common components, such as amino acids and nucleotides. However, lipids, another major component, are highly specific in different organisms and not included in current methods. Besides the biomass components, some other metabolites such as secondary metabolites, which are not required in the current model, should also be synthesized by the microbial community. Therefore, the assumed biomass requirement is only a conservative prediction of the essential functions that the organisms should achieve, which can be modified if any other functions have been observed and specified. On the other hand, this assumption does not consider the interactions among organisms, which is one of the drive forces for us to develop multiple-organism model. Meanwhile, these metabolic reconstructions under isolated conditions provide important information for the follow analysis. For example, large varied pathway distributions among species may indicate a highly interdependent interaction of the whole community.

Proteome data of organisms together with the metabolic network reconstruction process provides another approach that can integrate information at the whole cell (or community) level. Compared to genomic sequence data, the proteomic data are less accurate and with lower coverage, which limits their applications and makes it impossible to reconstruct whole cell scale metabolic network direct from proteomic data. Taking the proteomic data as supplementary in the method can avoid the risk of using incorrect data and provides a cross-examination approach to validate the predictions, which is important when the direct experimental validations are difficult. In the current model, we enforced some metabolic reactions that are verified by the proteomic data to be active in the network, which may still be problematic if the proteins are identified incorrectly. To avoid this risk, in this work only the proteins that are confirmed by multiple peptides are considered, which might be an underestimation of the real case. However, the perturbations allowed by the algorithm can retrieve some of those metabolic reactions (enzymes) that are ignored due to the quality of proteome data. As a result, a conservative prediction is made by the model when incorporating the proteomic data, which generates the metabolic network that are capable for both biomass synthesis and some secondary metabolites production.

In addition to the reconstructed metabolic networks of the microbial community, the interactions among the organisms have also been predicted. These predicted interactions are focus on the exchanges of major metabolites, which make the organisms interdependent. Even through the final predictions are partly dependent on the candidates of metabolites that can exchange, we still can capture some meaningful results based on the prediction of transporter proteins. A complex network among the organisms is suggested by the multiple-organism model, and it become more difficult to identify the roles of different species in the community. From the results, we also find that the *Leptospirillun sp.* groups are more like autotroph while the *Ferroplasma acidarmanus* groups and *Thermoplasmatales archaeon* are more like heterotroph in the view of carbon source. The heterotrophic growth of *Ferroplasma acidarmanus* I has been observed (*Baumler et al.*, 2005) and because there is no other organic carbon supply we can predict the major organisms in the microbial community *Leptospirillun sp.* groups to be autotrophic, which is supported by our prediction. In view of nitrogen usage, the solutions suggest *Leptospirillun sp.* Gp.III is the only organism that can fix nitrogen and all other organisms will make use of the ammonium or some amino acids as nitrogen sources. This results can also be supported by the experimental results (*Tyson et al.*, 2005), which demonstrated a strain belong to *Leptospirillun sp.* Gp.III is the key nitrogen fixer in the microbial community. The results from the multiple-organism model also suggest the metabolites that carry these interactions, such as amino acids. Due to the poor predictions of transporter proteins, the quality of these predicted interactions are not as good as reconstructed intracellular metabolic reactions, thus more solutions are required to make a fair conclusion. We also develop a strategy to analyze the quality of the prediction by comparing the results under different levels of penalty for exchanges. The most stable interactions suggested by multiple solutions can be the most conservative predictions, even through the solutions may become less connected under very high levels of penalty. We also observed that if two organisms with pretty similar pathway distributions, the interaction predicted by the multiple-organism model will become less stable due to the mathematic property of the model. Thus, we should pay more attentions to these results. Careful review is needed to make final conclusions about these interactions and experiments can be designed to examine them directly or indirectly.

In conclusion, we proposed a metabolic network based framework to investigate the metabolism of individual species and their interactions in microbial communities, by building metagenome-scale metabolic networks of the whole biofilm based on metagenomic sequences and annotations. This framework was applied to datasets of two similar Acid Mine Drainage (AMD) microbial communities from different locations. Both metabolic network of individual organisms and whole microbial community have been reconstructed and analyzed. These metabolic networks provide detailed mechanisms about the cellular metabolism of this organism. For example, three different types of carbon utilization were identified. The microbe-microbe and microbe-environment interactions were predicted based on a multiple-organism model, which is able to consider multiple species interactively at the same time. Extensive interactions have been observed according to the results, forming highly interdependent relationships. These results indicate a potential treatment for AMD formation by blocking these interactions because these interactions are predicted to be essential for the biomass synthesis of these organisms. According to the model prediction, amino acids and ammonia are the major molecules involved in these interactions, which are the potential targets when blocking the interactions. Proteomic data of the dominant species were also incorporated to verify the metabolic network reconstruction. The fractions of confirmed metabolic reactions in the metabolic reconstructions of the two dominant species are over 88.4% and 75.7% respectively, which are significant higher than the coverage of proteomic datasets.

CHAPTER IV

Metabolic Network Modeling of Gastrointestinal Microbial Communities

4.1 Introduction

Mammals, including human beings, have co-evolved with complex microbial communities inhabiting the surfaces and alimentary tract of the host (*Ventura et al.*, 2009). Researchers predict the number of these germ cells to be about ten fold of that of their host's cells. (*Hooper et al.*, 1999). For example, the human gut environment is considered one of the largest and most dense niches, supporting 10^{13} to 10^{14} microorganisms. The number of genes in these abundant microorganisms is estimated to be at least 100-fold more than that of human genome (*Gill et al.*, 2006a). This microbiota is believed to be critically important for many gut functions, including dietary energy harvest, regulation of host fat storage, vitamin and amino acid biosynthesis, stimulation of intestinal angiogenesis, inflammatory immune response and protection against pathogens (*Petrosino et al.*, 2009; *Greenblum et al.*, 2012). These mutually beneficial relationship between the microbiota and host have been co-evolved and contribute to the fitness of the host (*Hosokawa et al.*, 2006). However, many of these interactions and conclusions were taken directly from the macroscopic ecology, which are needed to be formulated more precisely at molecular level as well (*Brüls and Weissenbach*, 2011).

One of the major interests about the gut microbiota is its composition. In order to measure the composition of population, culture-based methods as well as culture-independent methods based on amplification or direct sequencing have been extensively applied to the intestinal microbiota (*Ventura et al.*, 2009). Metagenomics is one of the culture-independent approaches that is able to provide researchers with both ribosomal RNA and genome sequences in the niches. Currently, due to the importance and complexity of gut microbiota, a large number of metagenomic projects have been focused on this ecosystem (*Brüls and Weissenbach*, 2011). These metagenomic projects provide detailed information about the composition of the gut microbiota. According to these works, several anaerobic genera constitute the major part of the gut microbiome, including Bacteroides, Eubacterium, Bifidobacterium, Ruminococcus, Clostridium, and Faecalibacterium (*Eckburg et al.*, 2005; *Turroni et al.*, 2008). Studies have also shown that the proportion of Bacteroidetes and Bifidobacteria is relatively stable within individuals; however, the compositions of *Clostridium* group show much higher level of variations (*Lay et al.*, 2005). Researchers also classified the hosts according to the compositions of gut microbiota. Interestingly, three stable clusters (enterotypes) have been identified (*Arumugam et al.*, 2011), even though the causes of this clustering is still unclear.

Several factors that might shape the composition of gut microbiota have been identified and studied. The genotype of host is one of the factors that can explain some variations between individuals. For example, different genes involved in immune system of host can affect the gut microbiota through the host-microbiome interactions (*Ley et al.*, 2006). The composition of the initial colonizing microbial community is another important factor, which is evident from animal studies. For example, the composition of mouse gut microbiota can be controlled by maternal transmission (*Ley* *et al.*, 2005). Another dominant factor shaping the gut microbiota is the diet. The correlations between diet composition and microbiota composition were generated from model systems on mice (*Faith et al.*, 2011). Studies also indicate that metabolic activity of gut microbiota can also be affected by diet (*Martin et al.*, 2008). Other factors, such pH and antimicrobial compounds, are also important in determining gut microbiota composition.

Host-microbiome interaction is another interesting research area on this ecosystem. As mentioned, the host immune system affects the microbiota directly; however, the immune system of host matures with the gut microbiota. The microbiota shapes the development of the host immune system beginning at birth and the developing immune system also shapes the composition of the microbioata in return (*Nicholson et al.*, 2012). The chemical communications between host and gut microbiota are essential for these interactions. Regarding metabolism, organisms composing the microbiota alter their metabolic networks resulting in co-metabolism with the host. Many of the products identified in gut are synthesized through these chemical communication, including short-chain fatty acids (SCFAs), bile acids, and choline (*Peterson et al.*, 2008). Some of those products directly or indirectly involved in health disorders of host, such as obesity and inflammation (*Hooper et al.*, 2012; *Blumberg and Powrie*, 2012). Understanding the roles and metabolic functions of the gut microbiota is therefore crucial to reveal these profound interactions.

Culture-independent metagenomics provides us with not only the information about microbiota composition but also the genetic information of the organisms. Therefore, metagenomics became one of the most powerful tools to study the metabolism of microbial communities. Taking advantage of next-generation sequencing (NGS) techniques, more accurate and comprehensive metagenomic data have been collected. This has allowed researchers to answer the questions "what can the microorganisms do together?" after knowing "what is there?". For example, *Candela et al.* found the intestinal microbiota is enriched in several metabolism pathways, including carbohydrate metabolism, energy metabolism, biosynthesis of short-chain fatty acids, amino acid metabolism, biosynthesis of secondary metabolism and metabolism of cofactors and vitamins. As part of MetaHIT (Metagenimics of Human Intestinal Tract) project, the datasets collected by *Qin et al.* also indicated a functional group of genes, named as "minimal gut metagenome". According to the data, about 45% of the minimal gut metagenome is present in less than 10% of the sequenced bacterial genomes. More recently, *Greenblum et al.* studied the topological variations of the metabolic functions in human gut microbiome and linked them to community species composition and host state, e.g. obesity and inflammatory bowel disease. All these works demonstrate the potential of utilizing metabolic network and modeling in studying gut microbiota.

Currently, two fundamental challenges limit further studies in the metabolism and metabolic interactions of gut microbiota. The most direct one is the difficulty in generating high-quality metabolic networks and models at the community level. Even though there are effective tools and resources that can link the metagenomic data to functional analysis, e.g. KEGG database and MetaCyc database, accurate and complete metabolic networks of both individual organisms and whole microbiome are still missing, mainly due to the complexity of the ecosystem and scale of data. The second challenge is the lack of tools able to study the metabolism and metabolic interactions at the community level. Currently, statistical methods are the most commonly used tools, which normally cannot provide detailed mechanisms in molecular level.

In Chapter III, we reconstructed the community-wide metabolic networks for AMD biofilm for metagenomic datasets. To study the model gut microbiome, transcriptomic data are more frequently collected because the genomic sequences of the microorganisms in model microbial communities are pre-defined before they are inoculated. Therefore, we will present two case studies to demonstrate several strategies in reconstructing and modeling community-scale metabolic networks for model microbial communities by integrating genomic and transcriptomic data. There are three major objectives we will achieve.

- Community-wide metabolic model for a two-species model gut microbial community by integrating genomic data, growth test results, and genechip-based transcriptomic data.
- Community-wide metabolic model for a ten-species model gut microbial community by integrating genomic data and sequence-based transcriptomic data.
- Prediction of potential interactions between the organisms in model gut microbial communities.

4.2 Metabolic Network Modeling of a Two-species Model Microbial Community

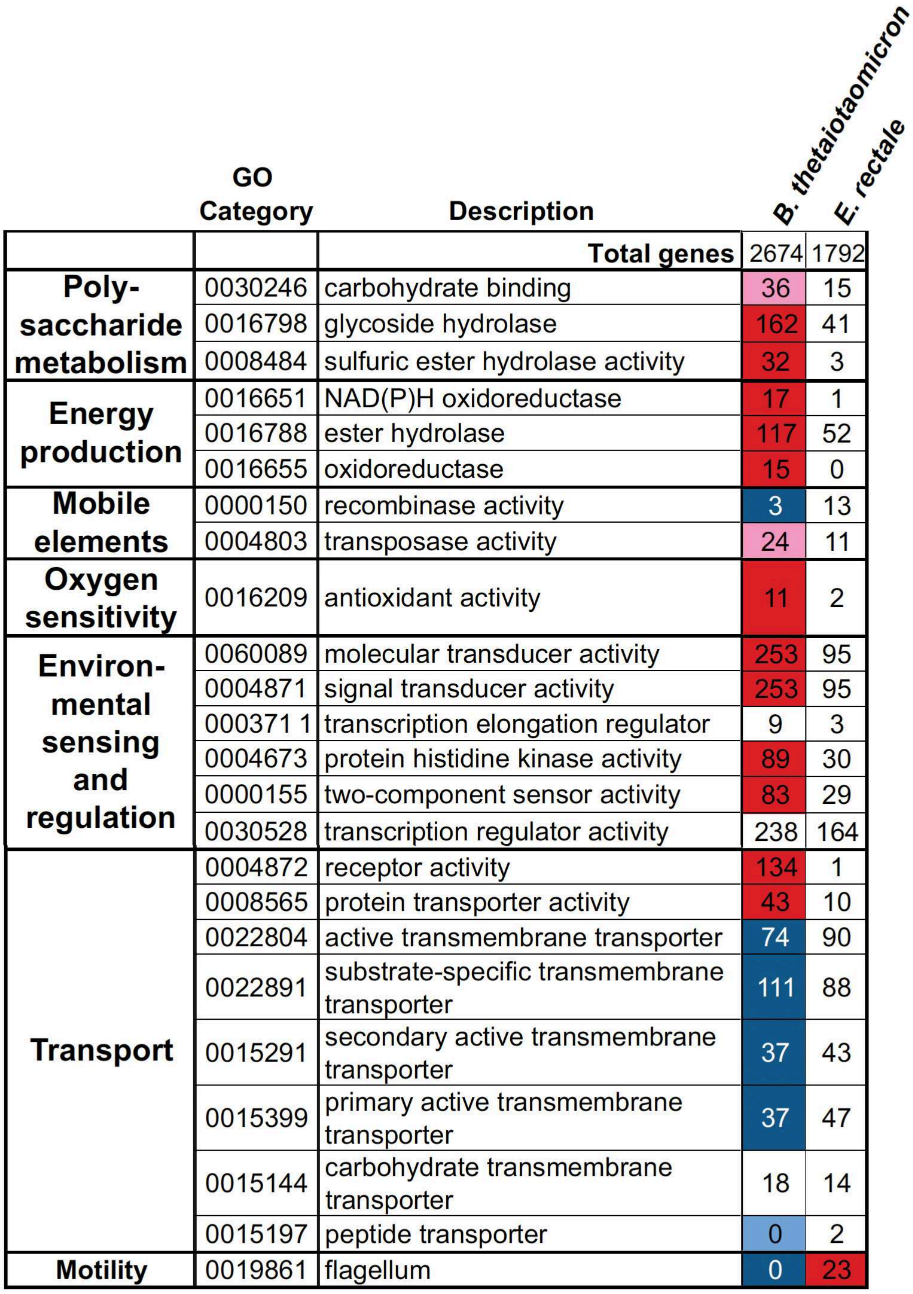
4.2.1 Background

The complexity of the natural gut microbial community is one of the major difficulties in studies of gut microbiota. To simplify the ecosystem, model microbial communities have been constructed to study specific questions. Gnotobiotic mice colonized with defined model microorganisms provide these simplified in vivo model systems. Because the microbiome is developed by inoculating know strains, complete genome sequences of all the organisms are available. Together with other detection methods, such as transcriptomics and proteomics, researchers are able to monitor both composition and metabolism of the microbiota. *Mahowald et al.*, developed a model system with two species and characterized metabolism and interactions genechip analysis and proteomic analysis.

To represent the human gut microbiota, this two-species model microbial community contains species from two bacterial phyla, Firmicutes and Bacteriodetes. These two phyla commonly dominate the gut microbial microbiota (*Turnbaugh et al.*, 2009). *Bacteroides thetaiotaomicron* VPI-5482 from Bacteroidetes and *E. rectale* ATCC 33656 from Clostridium are the two strains inoculated into the gnotobiotic mice. According to the genome of *B. thetaiotaomicron*, it contains a large repertoire of genes involved in polysaccharide acquisition and metabolism, including glycoside hydrolases (GHs) and polysaccharide lyases (PLs), myriad paralogs of SusC and SusD, and related environmental sensors and regulators (*Shipman et al.*, 2000; *Xu et al.*, 2007). Figure 4.1 demonstrates the numbers of genes in relative categories for both *B. thetaiotaomicron* and *E. rectale.* From the figure, it is clear that B. thetaiotaomicron is enriched in genes of polysaccharide related metabolism. This enrichment also indicates the potential interactions of monosaccharides and other carbon sources from B. thetaiotaomicron to E. rectale.

To study the in vivo interactions of the two organisms, both genechip based transcriptional analysis and tandem mass spectrometry based proteomic analysis were applied to both mono-inoculated samples and co-inoculated samples. Table 4.1 summarizes the results of the two sets of data. For both organisms, transcriptional data have much higher coverage than the proteomic data, and there are only very few genes that can only be observed in proteomic data. Therefore, expression data are more appropriate for genome-wide analysis.

The gnotobiotic mice were fed with different diets, low-fat and high plant polysaccharide chow, high fat and high-sugar Western-style chow, and corresponding control low fat and high-sugar chow. By comparing the expression data from the three diets, *Mahowald et al.* hypothesized that *B. thetaiotaomicron*



	GO Category	Description	B. thetaiotaomicron	E. rectale
		Total genes	2674	1792
Poly-saccharide metabolism	0030246	carbohydrate binding	36	15
	0016798	glycoside hydrolase	162	41
	0008484	sulfuric ester hydrolase activity	32	3
Energy production	0016651	NAD(P)H oxidoreductase	17	1
	0016788	ester hydrolase	117	52
	0016655	oxidoreductase	15	0
Mobile elements	0000150	recombinase activity	3	13
	0004803	transposase activity	24	11
Oxygen sensitivity	0016209	antioxidant activity	11	2
Environ-mental sensing and regulation	0060089	molecular transducer activity	253	95
	0004871	signal transducer activity	253	95
	000371 1	transcription elongation regulator	9	3
	0004673	protein histidine kinase activity	89	30
	0000155	two-component sensor activity	83	29
	0030528	transcription regulator activity	238	164
Transport	0004872	receptor activity	134	1
	0008565	protein transporter activity	43	10
	0022804	active transmembrane transporter	74	90
	0022891	substrate-specific transmembrane transporter	111	88
	0015291	secondary active transmembrane transporter	37	43
	0015399	primary active transmembrane transporter	37	47
	0015144	carbohydrate transmembrane transporter	18	14
	0015197	peptide transporter	0	2
Motility	0019861	flagellum	0	23

Figure 4.1: Genes involved in carbohydrate metabolism and energy production for *B. thetaiotaomicron* and *E. rectale* (adapted from *Mahowald et al.*, 2009). Red, enriched; blue, depleted; darker color $P \leq 0.001$ and light color $P \leq 0.05$ relative to the average of all Firmicute genomes.

with enriched PUL-associated GHs functions utilizes complex dietary plant polysaccharides and distributes carbon sources to *E. rectale*, which synthesizes a large amount of butyrate.

Table 4.1: Summary of proteins detected by mass spectrometry and GeneChip for B. thetaiotaomicron and E. rectale.

	E. rectale			*B. thetaiotaomicron*		
	Mono-inoculated	Bi-inoculated	Total	Mono-inoculated	Co-inoculated	Total
Detected by MS/MS	661	453	680	1608	1367	1687
Detected by GeneChip	2139	2010	2150	3798	3865	3995
GeneChip-/ MS/MS+ *	7	7	8	40	21	23
MS/MS- /GeneChip+ *	1608	1638	1603	2280	2569	2357

*: + means identified in any proteomic datasets or in $\geq 75\%$ of GeneChip datasets.

Provided with comprehensive expression data and annotated genomes, we believe this 2-species model microbial community is a good model system. With much simplified structure, we are able to reconstruct the genome-wide metabolic networks for all the organisms in the microbial community. Further, the expression data enable us to identify the active metabolic networks in various conditions, e.g. mono-inoculation or co-inoculation. By specifying these active metabolic networks, we can predict the potential metabolic interactions between the two organisms, which are not only determined by the metabolic capabilities but also the metabolic responses to each other. Therefore, we designed a three-step modeling procedure for the 2-species microbial community (Figure 4.2). In this model, we first reconstructed the metabolic networks for individual species using their annotated genomes and growth test results. Based on these individual metabolic networks, we build a two-species model, which considers both metabolic gaps and expression data. This two-species model allows exchange of all the metabolites between the two organisms. Therefore, in the third step, a two-species model with minimal exchange fluxes was built. Utilizing this three-step modeling procedure, we can reconstruct the community-scale metabolic model for this two-species model microbial community by integrating annotated genomes, growth test results, expression data, and transporter predictions. This model is able to predict not only metabolic capabilities, but also the metabolic interactions and responses of the two organisms in the gut environment with more detailed molecular mechanisms.

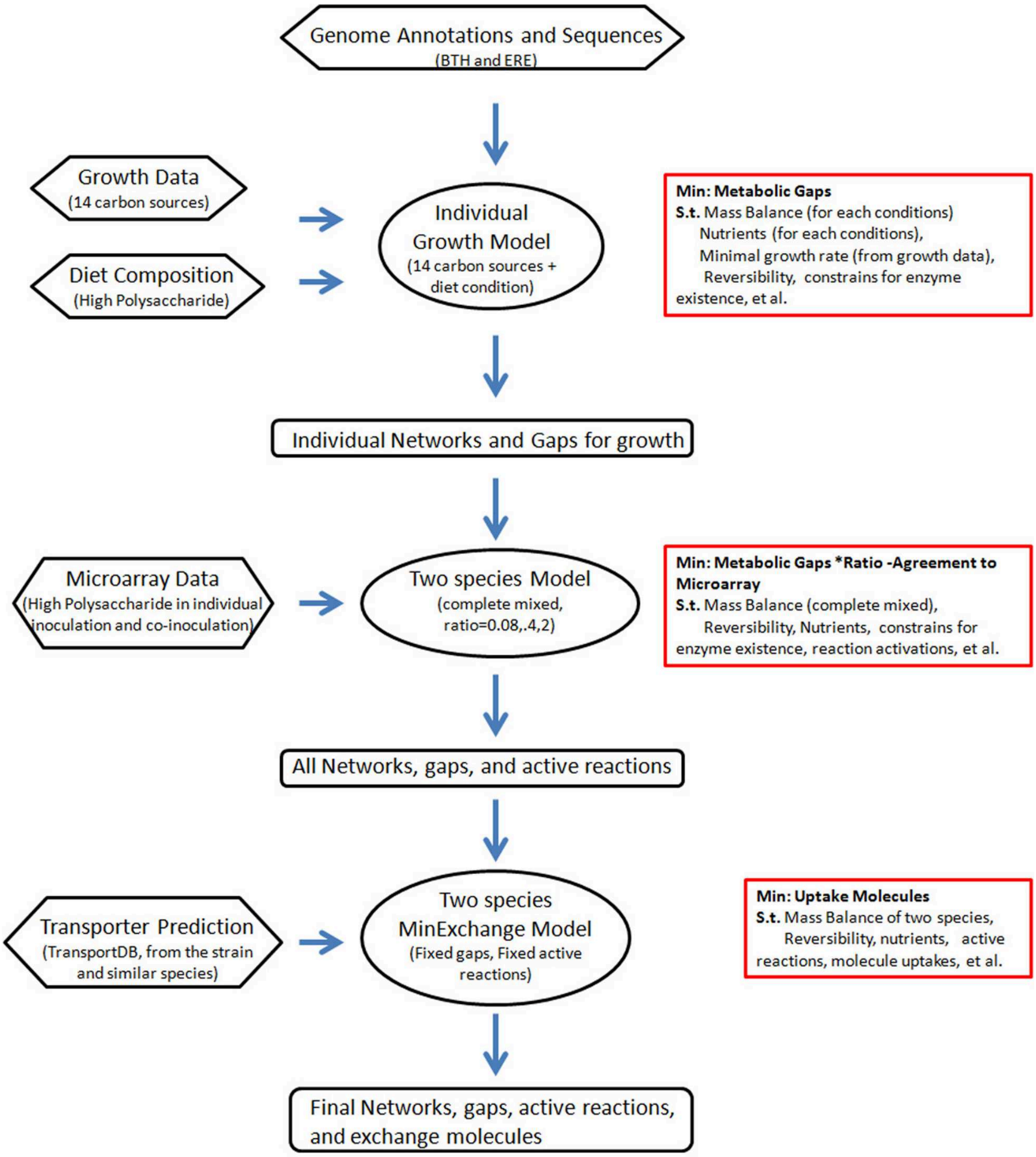

Genome Annotations and Sequences
(BTH and ERE)
Growth Data
(14 carbon sources)
Diet Composition
(High Polysaccharide)
Individual
Growth Model
(14 carbon sources +
diet condition)
Min: Metabolic Gaps
S.t. Mass Balance (for each conditions)
Nutrients (for each conditions),
Minimal growth rate (from growth data),
Reversibility, constrains for enzyme
existence, et al.
Individual Networks and Gaps for growth
Microarray Data
(High Polysaccharide in individual
inoculation and co-inoculation)
Two species Model
(complete mixed,
ratio=0.08,.4,2)
Min: Metabolic Gaps *Ratio -Agreement to
Microarray
S.t. Mass Balance (complete mixed),
Reversibility, Nutrients, constrains for
enzyme existence, reaction activations, et al.
All Networks, gaps, and active reactions
Transporter Prediction
(TransportDB, from the strain
and similar species)
Two species
MinExchange Model
(Fixed gaps, Fixed active
reactions)
Min: Uptake Molecules
S.t. Mass Balance of two species,
Reversibility, nutrients, active
reactions, molecule uptakes, et al.
Final Networks, gaps, active reactions,
and exchange molecules

Figure 4.2: Flow chart of three-step modeling procedure for the 2-species microbial community. The methods used in the three steps are briefly described. Detailed information is provided in Section 4.2.2.

4.2.2 Three-step Metabolic Network Reconstruction and Modeling Method for Two-Species Model

4.2.2.1 Individual Metabolic Networks of the Two Species

The first step in this three-step modeling procedure is reconstruction of individual metabolic networks of the two species. We applied a modified PEER to reconstruct the two metabolic networks. We made several modifications based on the algorithms described in Section 2.2. First, we collected the gene annotations of the two species from three resources, manual curated annotation (`http://gordonlab.wustl.edu/modeling_microbiota/`), RAST automated annotations tools (`http://rast.nmpdr.org/`), and annotations in KEGG database (T00122 and T00909, downloaded on Oct-26,2011). Unsurprising, there are disagreements among the three annotations. The RAST annotations generated its own gene identifications while the other two datasets use the same one as GenBank, which contain the locus tags. To compare the three annotations, we used sequence alignment method to match the genes identified by RAST with the genes identified in the other two annotations. We removed all the conflicts among the three annotations, meaning only the annotations that are the same in the three annotations will be accepted. The reason for this strategy is that PEER can fill the metabolic gaps caused by inconsistent annotations.

Another modification we made to PEER for this two-species ecosystem was that we assumed a minimal growth rate (0.05 1/h) on those carbon sources that can be utilized according to the growth test results. To achieve this, we used one compartment in the model for each growth condition, which shares the same metabolic gap filling but do not share the flux values and environmental conditions (Figure 4.3). Therefore, we were able to find the metabolic network for the organism which could agrees with the growth test results as accurately as possible. We collected 14 growth tests on different carbon sources together with the gut environmental conditions. We used a simplified biomass compositions to represent the growth function, which is the same as the biomass composition used in the metabolic network of *Prochlorococcus marinus* (Section 2.4).

$$\min_{v, b_{add}} \sum_{r \in R} b_{add}(r) * \text{Weight}(r) \qquad \text{(Find a best network under constraints)}$$

S.t.

$$\boxed{\begin{array}{ll} \sum_{r \in R} S(m,r) * v(r) = 0, \quad \forall m \in Metabolite & \text{(Mass balance of metabolites)} \\ v(r) \geq \epsilon, \quad \forall r \in \text{Product} & \text{(Products must be synthesized)} \\ F_{min} \leq v(r) \leq F_{max}, \quad \forall r \in \text{Exchange} & \text{(Exchange reactions with environment)} \\ v(r) \geq 0, \quad \forall r \in \text{Irreversible} & \text{(Reversibility of reactions)} \\ v(r) \leq F_{max} * b_{Active}(r), \quad \forall r \in \text{R} & \text{(Active reactions)} \\ b_{Active}(r) \leq b_{\text{Exist}}(r) + b_{Spon}(r) + b_{Add}(r), \quad \forall r \in \text{R} & \text{(Existence of reactions)} \\ \text{etc.} & \text{(Other constraints)} \end{array}} \quad \text{Condition 1}$$

$$\vdots$$

$$\boxed{\begin{array}{ll} \sum_{r \in R} S(m,r) * v(r) = 0, \quad \forall m \in Metabolite & \text{(Mass balance of metabolites)} \\ v(r) \geq \epsilon, \quad \forall r \in \text{Product} & \text{(Products must be synthesized)} \\ F_{min} \leq v(r) \leq F_{max}, \quad \forall r \in \text{Exchange} & \text{(Exchange reactions with environment)} \\ v(r) \geq 0, \quad \forall r \in \text{Irreversible} & \text{(Reversibility of reactions)} \\ v(r) \leq F_{max} * b_{Active}(r), \quad \forall r \in \text{R} & \text{(Active reactions)} \\ b_{Active}(r) \leq b_{\text{Exist}}(r) + b_{Spon}(r) + b_{Add}(r), \quad \forall r \in \text{R} & \text{(Existence of reactions)} \\ \text{etc.} & \text{(Other constraints)} \end{array}} \quad \text{Condition n}$$

$v(r)$: Flux of reaction r, with different values in all conditions

$b_{Add}(r)$: Whether r is putative reaction, binary variable, the same value in all conditions

$b_{Active}(r)$: Whether reaction r is active, binary variable, with different values in all conditions

Figure 4.3: Demonstration of the multi-condition MILP model for metabolic network gap filling. Red: binary variables, blue: continuous variables, black: parameters

The whole reconstruction process is the same as in PEER. A draft metabolic network was generated based on the combined annotations. All potential putative reactions were then predicted based on the multi-condition MILP model for metabolic network gap filling and evaluated with sequence alignment results. The final metabolic network was reconstructed using the same multi-condition MILP model for metabolic network gap filling together with the weight parameters calculated according to the sequence alignment results. Differently from what was used in PEER, we implemented these steps with the IBM ILOG CPLEX Optimizer (v12), which performs better than

XPRESS for large-scale MILP problems.

4.2.2.2 Community-Wide Model of the Two-Species Model Microbial Community

After the individual metabolic networks were reconstructed, we integrated them into a two-species model, which contains exchange fluxes between the two organisms. Furthermore, we utilized the gene expression data to indicate active and inactive metabolic reactions. Thus, we can not only capture the metabolic potentials of this model microbial community but also the metabolic responses between the two organisms in the gut environment. The genechip-based gene expression data were carried out in both co-inoculation and mono-inoculation conditions. The absolute values of gene expression level from genechip-based data for a specific condition cannot accurately identify the activity of gene transcription due to the varied background signals and non-specific bindings. Therefore, we utilized the changes of gene expression levels between the co-inoculation and the mono-inoculation conditions rather than the absolute values in the two conditions. The active and inactive genes in the two conditions were identified according to these changes.

We assumed if the gene expression change was more than one fold, then the gene in the condition with higher expression would be marked as active and the one with lower gene expression would be marked as inactive. To eliminate the genes with highly fluctuating expression levels among the biologic replicates, the standard deviations of fold changes of gene expression were calculate using Equation 4.1 (E_{mono} represents the expression value in mono-inoculation condition and E_{co} represents the expression value in co-inoculation condition).

$$\begin{aligned}\sigma_{log(E_{co}/E_{mono})} &= \sigma_{log(E_{co})-log(E_{co})} \\ &= \sqrt{(\sigma_{log(E_{co})})^2 + \sigma_{log(E_{mono})})^2} \end{aligned} \tag{4.1}$$

If the calculated standard deviations $\sigma_{log(E_{co}/E_{mono})} > 0.608 * log(E_{co}/E_{mono})$, we removed the corresponding gene from the list of differentially expressed gene and the activity of this gene was not considered due to the low quality of the gene expression data. The metabolic reactions associated with these genes were marked as the same activities as the gene.

To consider the active and inactive reactions suggested by genechip data, we developed a two-species model that considered the co- and mono-inoculation conditions simultaneous. In this two-species model, the activity of metabolic reaction was determined by the flux value. We assumed the active reactions carry flux more than 1 $mmol/(gDCW \cdot h)$. To work with MILP framework, the relationship was linearized with two binary variables as shown in Equation 4.2 (v_r is the reaction flux, $P(r)$ and $N(r)$ are the binary variables representing whether v_r is positive or negative).

$$\begin{aligned} v_r &\geq P(r) - 1000N(r) \\ v_r &\leq (-1)N(r) + 1000P(r) \\ 1 &\geq P(r) + N(r) \end{aligned} \tag{4.2}$$

We assumed the minimal flux for active reactions is 1 mmol/h/g DCW in this two-species model. Therefore, if reaction r is active, $P(r) + N(r) = 1$; otherwise, $P(r) + N(r) = 0$. In order to utilize the reaction activities predicted from genechip data, the objective function in this two-species model was divided into two parts. The first part was the metabolic gaps, which is same as the objective function in the individual metabolic network reconstruction model. The second part was the agreement of the reaction activities predicted by model versus calculated from expression data. Equation 4.3 is the formula of this two-part objective function, which contains a ratio parameter to balance the two parts of the objective function. The parameters $I_{Active}(r)$ and $I_{Inactive}(r)$ indicate whether the reaction is predicted to be active or inactive according to the expression data. For reactions lacking activity data, the two parameters were set to zero.

$$\begin{aligned} obj : \min_{v,b_{add},P,N} \quad & ratio \cdot \textstyle\sum_{r \in R} b_{add}(r) \cdot Weight(r) \\ & - \textstyle\sum_{r \in R} \left[(P(r) + N(r)) I_{Active}(r) - (P(r) + N(r)) I_{Inactive}(r) \right] \end{aligned} \tag{4.3}$$

The ratio parameter in the equation balances the two parts of the objective function. For lower values of this parameter, the model will emphasize more the expression data compared to the metabolic gaps. When the ratio parameter is set to higher level, the model fills less metabolic gaps and the agreement between expression data and model prediction is reduced. Three different level of the ratio parameters were tested and combined. The discovery rate of each putative reaction under one ratio setting was calculated based on the suboptimal solutions within 5% optimality gaps. Only the putative reactions discovery rates with higher than 50% in at least two ratio settings were considered. The active and inactive reactions for both mono- and co-inoculation conditions were calculated using the model by fixing these selected putative reactions.

4.2.2.3 Prediction of Potential Interactions in Microbial Community

In addition to the community-wide metabolic network reconstruction and reaction activity prediction, we applied a two-species model with minimal uptake fluxes to predict potential interactions. In this two-species model, the metabolic gaps were fixed according to the results generated by the two-species model that balances expression data and putative reactions. The reaction activities were also derived from the co-inoculation condition results. The objective function of the new two-species model is minimization of uptake fluxes as shown in Equation 4.4, in which I_{uptake} is the binary variable indicating whether the corresponding uptake reaction is active and $weight_{uptake}$ is the weight for the uptake reaction.

$$obj : \min_{v,ve,I_{uptake}} \sum_{re \in Re} I_{uptake}(re) * weight_{uptake} \tag{4.4}$$

We tried to identify the interactions that were necessary due to growth requirement or expression data. The objective function (Equation 4.4) can eliminate unnecessary uptake reactions, which enables us to predict the most desired interactions. The $weight_{uptake}$ used in the objective function reflects the probability of the uptake reaction, which is predicted based on transporter predictions. We utilized the data in TransportDB (`http://www.membranetransport.org/`) to identify transporter. Only the transporter prediction of *B. thetaiotaomicron* was identified in the TransportDB. For metabolites associated with identified transporters, the $weight_{uptake}$ was set to 0.1. For metabolites associated with transporters not identified in *B. thetaiotaomicron*, the $weight_{uptake}$ was set to 0.9, while if no transporters are associated, the $weight_{uptake}$ was set to 1. Therefore, the parameter $weight_{uptake}$ can help the model identify necessary interactions that are associated with transporter predictions.

4.2.3 Results

4.2.3.1 Individual Metabolic Network Reconstruction

We used the multi-step method to reconstruct the metabolic network of this two-species gut microbiome utilizing the genome sequences; three annotations from KEGG database, RAST automated annotation tool and authors' manual annotations; growth phenotypes on carbon sources; expression data of gut microbiome in individual and mixed inoculation conditions. By integrating these data and information, we were able to reconstruct the most plausible metabolic networks for *Bacteroides thetaiotaomicron* and *Eubacterium rectale*, together with their active metabolic reactions in the gut environment. Table 4.2 provides a overview of the reconstructed metabolic networks.

Table 4.2: Summary of the reconstructed metabolic networks of B. thetaiotaomicron and E. rectale.

Species	Genome size (Gene)	Identified metabolic reactions *	Putative metabolic reactions	Total metabolic reactions
B. thetaiotaomicron	6.3Mbp (4917)	681	69	750
E. rectale	3.4Mbp (3693)	538	74	612

∗: Based on the annotations that are consistent in all the three annotations.

From the results, we found *B. thetaiotaomicron* has a larger genome than *E. rectale* and the metabolic network of *B. thetaiotaomicron* is also larger, this translates into a higher number of identified metabolic enzymes in *B. thetaiotaomicron.* Furthermore, *E. rectale* has a higher percentage of putative reaction (12%) than *B. thetaiotaomicron* (9%). This finding indicates that either the annotation of *E. rectale* genome is of poorer quality than the annotation of *B. thetaiotaomicron*, or the metabolism of *E. rectale* is different from that of known organisms. Because we only used the annotations that are consistent in all of the three annotations, we might be able to evaluate the quality of the annotations by comparing the annotations from different sources. Table 4.3 lists the number of annotated genes in the three sets of annotation and the comparison between them. It is clear that author's annotation contains more annotated genes and both KEGG and RAST annotate fewer genes. Furthermore, the annotated genes in RAST annotations of *E. rectale* is much less than that of the other two annotations, which might explain the higher number of metabolic gaps in the metabolic network of *E. rectale*.

Table 4.3: Comparison of the gene annotations from author's annotation, KEGG annotation and RAST annotation.

Species	Author's annotation	KEGG annotation	RAST annotation	Consistent annotation in 3 annotations
B. thetaiotaomicron	2290	1094	804	424
E. rectale	1818	1055	571	333

We classified the metabolic reactions according to the pathways they belong to, as shown in Figure 4.4. Purine and pyrimidine metabolism pathways are the most abundant pathways, which is not surprising. The third abundant pathway is the galactose metabolism pathway, which is relevant to polysaccharide and monosaccharide utilization. This observation is in agreement with the functional genome analysis of the two species, which indicates that both of the two organisms have many genes involved in acquisition and utilization of poly- and monosaccharide (*Mahowald et al.*, 2009). We also found that *E. rectale* has more putative reactions (7 reactions) in glycolysis/gluconeogenesis pathway than *B. thetaiotaomicron* (2 reactions), some of which are essential for certain carbon sources utilization. *B. thetaiotaomicron* has much more metabolic reactions in pentose and glucuronate interconversions pathway. Many of these reactions are involved in pectin utilization. In human nutrition, pectin is one of the most important sources of dietary fiber which is almost completely utilized by the gut microflora (*Titgemeyer et al.*, 1991; *Dongowski et al.*, 2000). *B. thetaiotaomicron* is one of the organisms that can utilize pectin (*Dongowski et al.*, 2000) while *E. rectale* cannot (*Mahowald et al.*, 2009), which can explain why *B. thetaiotaomicron* contains more metabolic reactions in pentose and glucuronate interconversions pathway. Interestingly, *E. rectale* has more metabolic reactions in porphyrin and chlorophyll metabolism, even though the reason is still unknown.

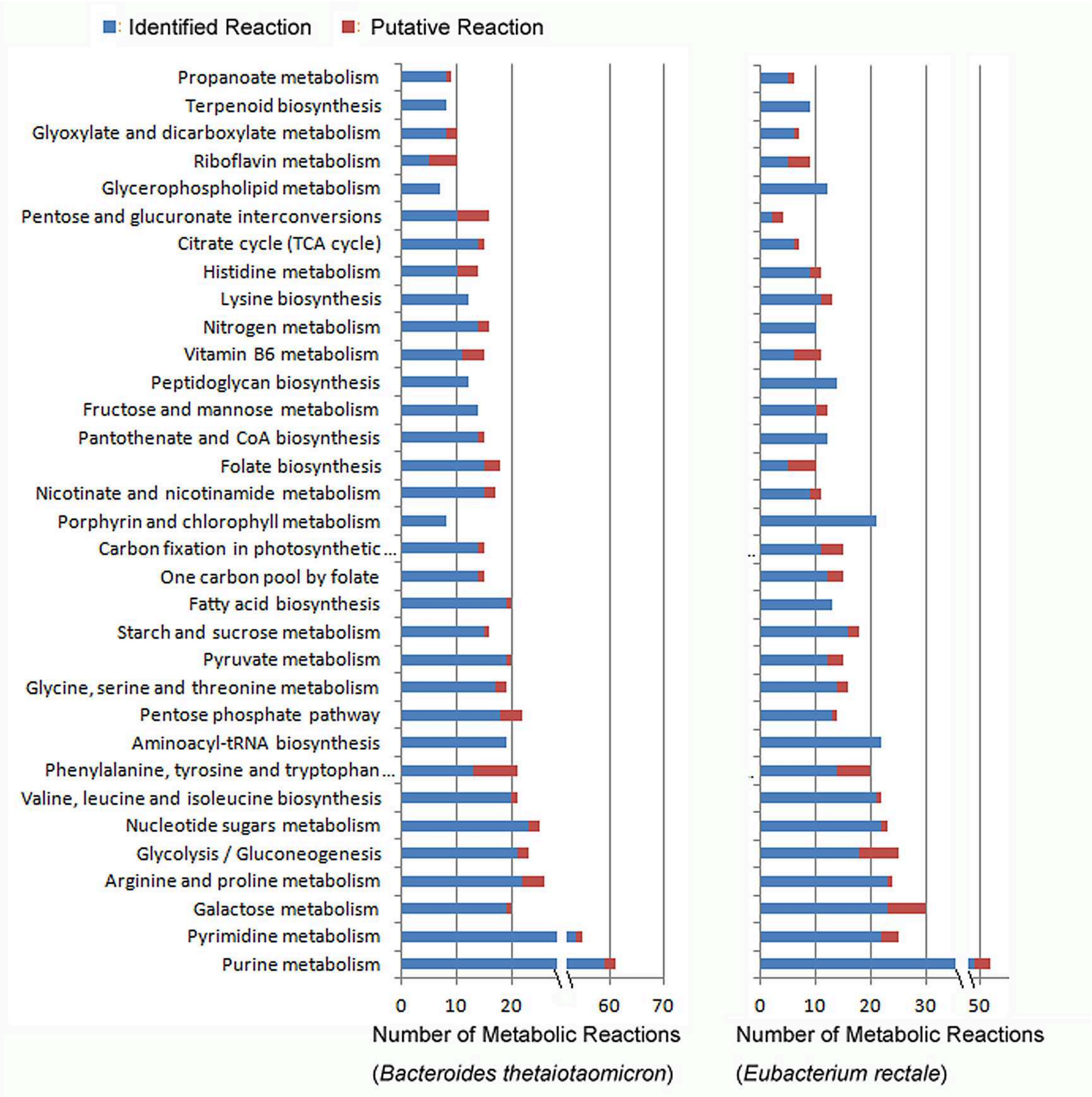
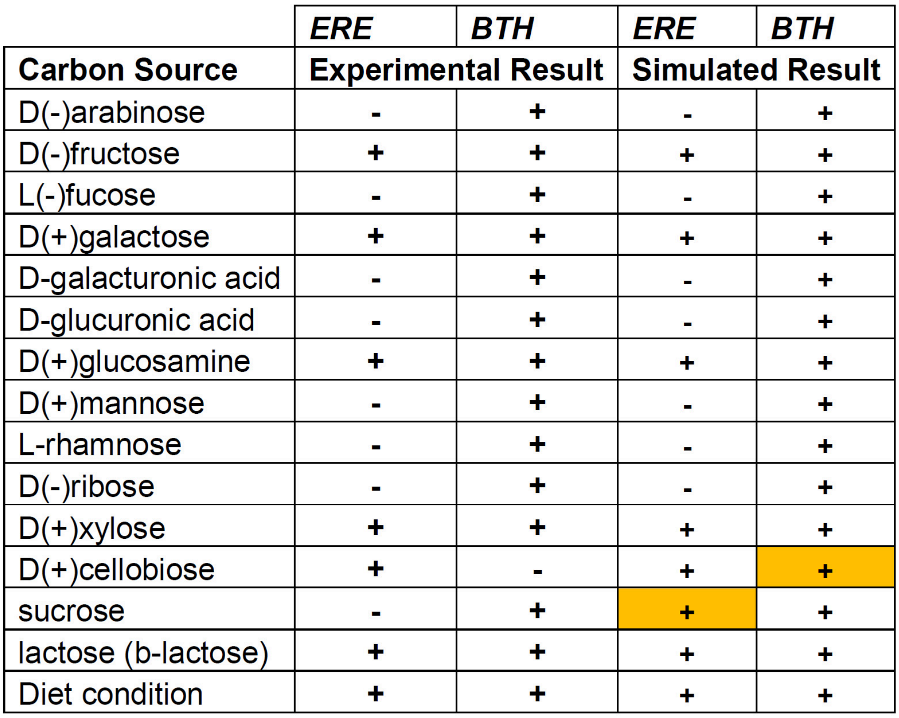


Figure 4.4: Pathway analysis of the reconstructed metabolic network of *Bacteroides thetaiotaomicron* and *Eubacterium rectale*. Ordered by the abundance of the pathways.

B. thetaiotaomicron and *E. rectale* have different capabilities in terms of carbon source utilization, as studied by *Mahowald et al.*. Table 2 lists the experimental growth phenotypes together with the model predictions based on the reconstructed metabolic networks. According to the table, there are only two false positive predictions and no false negative predictions in the simulated results (accuracy is 93%). However, according to the work of *Tannock*, *B. thetaiotaomicron* is capable of utilizing cellobiose, which is consistent with our model prediction.



	ERE	*BTH*	*ERE*	*BTH*
Carbon Source	**Experimental Result**		**Simulated Result**	
D(-)arabinose	-	+	-	+
D(-)fructose	+	+	+	+
L(-)fucose	-	+	-	+
D(+)galactose	+	+	+	+
D-galacturonic acid	-	+	-	+
D-glucuronic acid	-	+	-	+
D(+)glucosamine	+	+	+	+
D(+)mannose	-	+	-	+
L-rhamnose	-	+	-	+
D(-)ribose	-	+	-	+
D(+)xylose	+	+	+	+
D(+)cellobiose	+	-	+	+
sucrose	-	+	+	+
lactose (b-lactose)	+	+	+	+
Diet condition	+	+	+	+

Figure 4.5: Growth phenotypes of *B. thetaiotaomicron* and *E. rectale* on different carbon sources. BTH: *Bacteroides thetaiotaomicron*, ERE: *Eubacterium rectale*. Two false positive predictions are highlighted.

4.2.3.2 Metabolic Network Modeling with Expression Data

According to the expression data, we are able to predict the active metabolic networks for *B. thetaiotaomicron* and *E. rectale* in both mono- and co- inoculation conditions. As described in the methods, we set the threshold to 1 fold change to distinguish the active and inactive genes. By balancing the agreement of expression data and penalty of putative reactions we are able to determine the active and inactive parts of the two metabolic networks. Table 4.4 summarizes the agreement of expression data of the model results. The accuracy of the model prediction ranges from 63% to 72%, the false positive error and false negative error contribute equally to the overall accuracy, indicating there is no bias to either of the two errors in the model.

Table 4.4: Summary of the agreement of model prediction with expression data. BTH: *B. thetaiotaomicron* , ERE: *Eubacterium rectale.* Mixed: co-inoculation, mono: mono-inoculation.

Inoculated Species	Active Reaction	True Positive	False Positive	True Negative	False Negative	Accuracy
Mixed (BTH)	350	38	3	13	17	0.72
Mixed (ERE)	282	31	18	31	16	0.65
Mono (BTH)	297	7	17	38	9	0.63
Mono (ERE)	242	31	16	31	18	0.65

We found the accuracy of the model prediction is not sensitive to the penalty parameter, indicating most of the disagreements cannot be fixed by simply adding a few putative reactions. We mapped the two types of errors into pathways and the most abundant pathways are shown in the Figure 4.6. Seven metabolic reactions in arginine and proline metabolism pathway are predicted to be active in mono-inoculated condition but suggested to be inactive according to expression data. Most of these reactions are involved in urea cycle, which may be explained by the nitrogen cycling between gut microbiota and the host that is not considered in our model. According to the figure, the model predicts the metabolism of *E. rectale* less accurately than that of *B. thetaiotaomicron*, which is reasonable because the metabolism of *B. thetaiotaomicron* is better understood.

Besides these metabolic reactions with significant changes of corresponding gene expressions, the active fraction of each pathway was also calculated (shown in Figure 4.7). It is interesting that the pathway active states are clustered according to inoculation status rather than species. This clustering indicates that the environmental conditions of the gut microbiota shape the metabolism of the

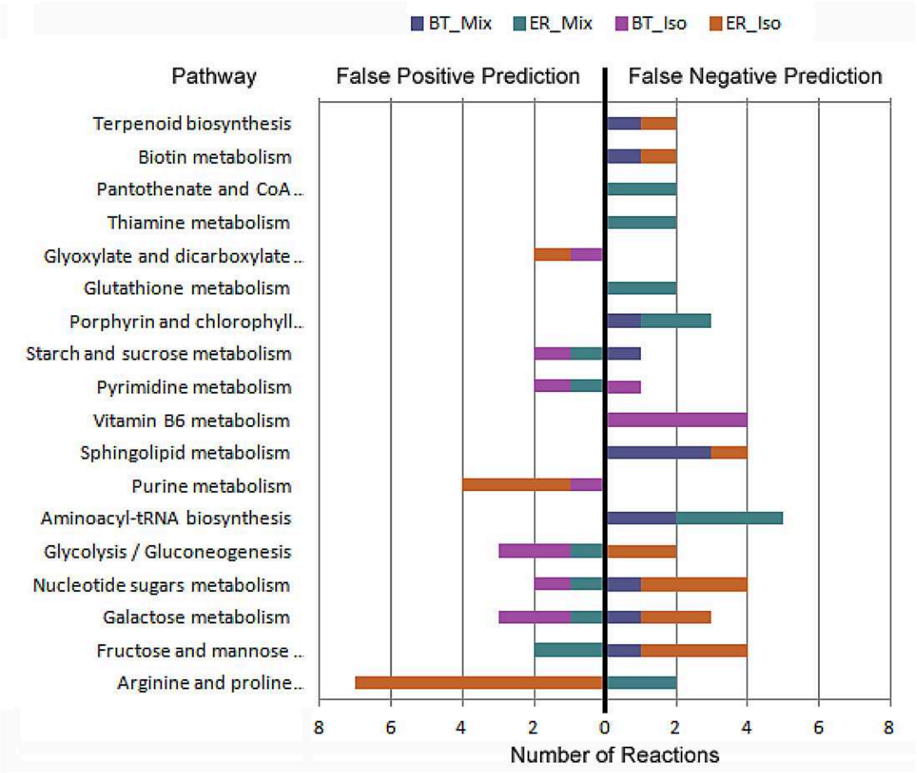


Figure 4.6: False positive and negative predictions of the model. BT: *Bacteroides thetaiotaomicron*; ER: *Eubacterium rectale*; Mix: co-inoculated; Iso: mono-inoculated.

microbes in the microbiome. For riboflavin metabolism and phenylalanine, tyrosine and tryptophan biosynthesis pathways, both organisms have lower active fractions in co-inoculated condition compared to mono-inoculated condition. This phenomenon can be explained by the interactions between the two organisms, which allow the two to cooperate. In contrast, certain pathways were up-regulated in the co-inoculated condition, including folate biosynthesis, Nicotinate and nicotinamide metabolism and fatty acid metabolism in *Eubacterium rectale*. These changes reflect the response of *E. rectale* to *B. thetaiotaomicron*. For example, the biosynthesis of short chain fatty acid (SCFA) is up-regulated in *E. rectale* when *B. thetaiotaomicron* presents, such results have been observed by *Mahowald et al.*(2009).

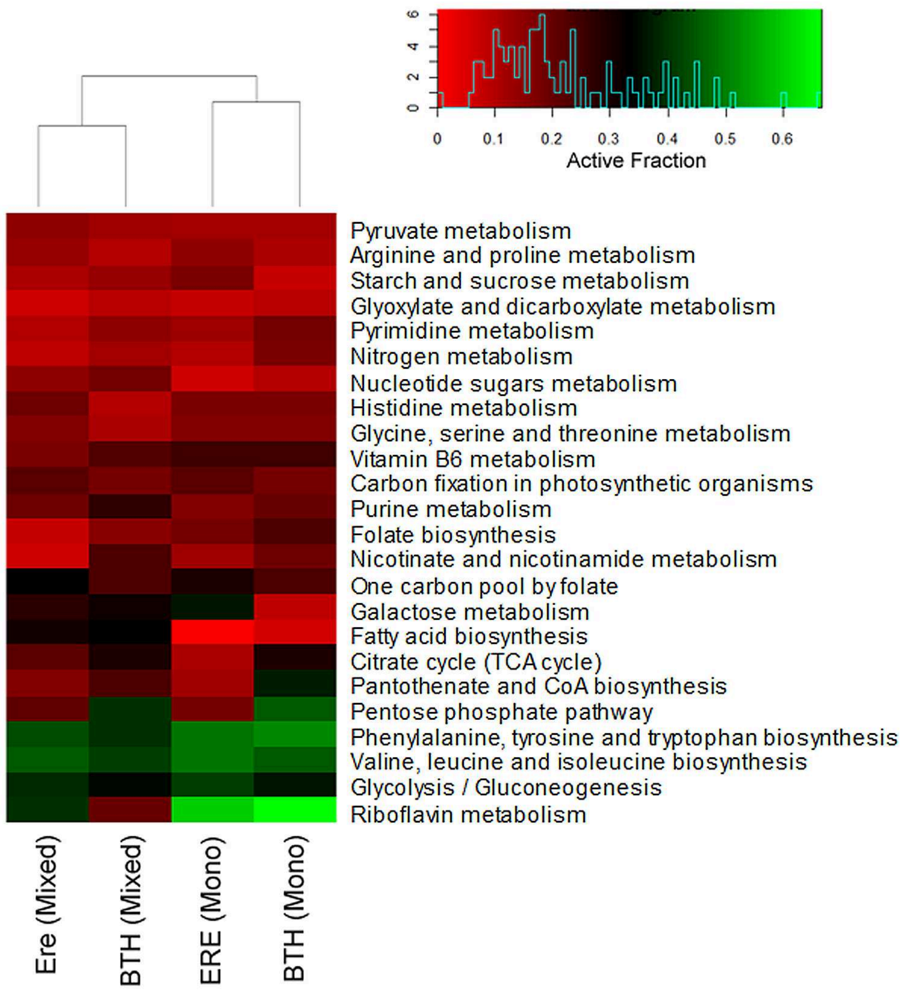


Figure 4.7: Pathway activities in both mono- and co- inoculated conditions. BTH: *Bacteroides thetaiotaomicron*; ERE: *Eubacterium rectale*; Mixed: co-inoculated; Iso: mono-inoculated.

4.2.3.3 Interaction Predictions

Utilizing reconstructed community-wide metabolic network of this gut microbiome, we were able to predict the active metabolic reactions according to the transcriptome data. The metabolic networks and the active metabolic reactions were further used to predict potential necessary interactions between the two species. These interactions were predicted based on not only the biosynthetic capability of the microbiome (represented by metabolic networks) but also their cellular regulations (represented by transcriptome). Therefore, these predicted interactions reflect either the metabolic synergy between the two bacteria or the metabolic responses to each other. Figure 4.8 demonstrates the interactions predicted by our model, the darkness of color indicates the probability of the interaction.

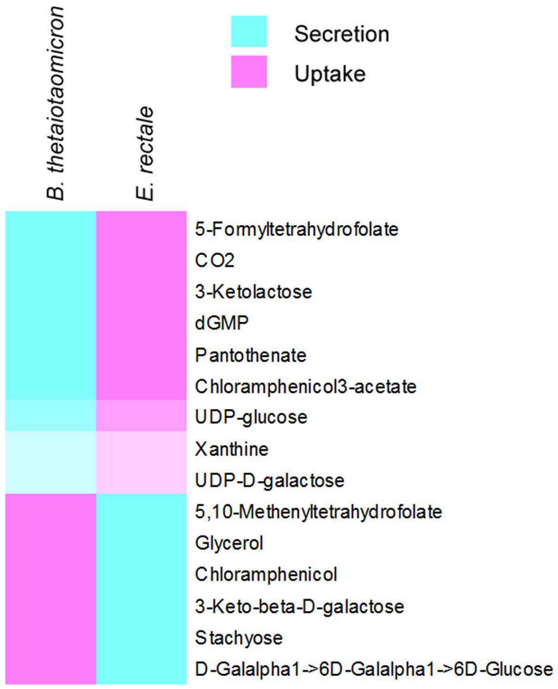


Figure 4.8: Interactions between *E. rectale* and *B. thetaiotaomicron* predicted by 2-species model. The grayscale indicates the probability of the predictions.

One of the predicted interaction between the two species is *E. rectale* providing stachyose to *B. thetaiotaomicron*. According to the work of *Salyers and Pajeau*, stachyose can function as growth enhancer for *B. thetaiotaomicron*. The microarray data indicates the expression of stachyose synthase(GH36, EUBREC_0489 and EUBREC_3387), which is involved in the synthesis of stachyose in *Eubacterium rectale*. The direct enzyme utilizing stachyose in *B. thetaiotaomicron* is alpha-galactosidase (BT_3065 and BT_3133), which can covert stachyose to raffinose and D-galactose. These genes are also expressed according to the expression data. All these experimental data support the hypothesis that *E. rectale* synthesize stachyose for *B. thetaiotaomicron*.

Another interaction predicted by the model is *B. thetaiotaomicron* providing pantothenate to *E. rectale*. Pantothenate is one of the vitamins used in cultivation of anaerobic faecal flora (*Wensinck et al.*, 1981; *Van de Merwe et al.*; *Faith et al.*, 2011). According to expression data and reconstructed metabolic networks, enzyme involved in pantothenate synthesis, pantothenate synthetase (BT_4308) and 2-oxopantoate reductase (BT_3117) in *B. thetaiotaomicron* are expressed. Pantothenate kinase (EUBREC_0060) and phosphopantothenoylcysteine synthetase(EUBREC_0828), the enzymes involved in utilizing pantothenate in *Eubacterium rectale*, are also found to be expressed in co-cultivation conditions. Therefore it is reasonable to hypothesize the *B. thetaiotaomicron* provides pantothenate to *Eubacterium rectale*.

Carbon dioxide or bicarbonate was predicted to be transferred from *B. thetaiotaomicron* to *Eubacterium rectale*. Research indicate that *E. rectale* has the capability of fixing carbon dioxide to produce propionic acid through dicarboxylic pathway (*Purwani et al.*, 2012). According to the metabolic networks, the enzymes involved in fixing CO2 are phosphoenolpyruvate carboxykinase (EUBREC_2002) and acetyl-CoA carboxylase (EUBREC_3141). Both of the two were expressed in the co-inoculation conditions. Therefore, carbon dioxide is another potential interaction between the two organisms in the gut environment, which was also predicted by *Mahowald et al.*.

Both false positive and negative prediction of interactions are possible due to the inaccuracy of either reconstructed metabolic networks or expression data. Therefore, the predicted interactions by the two-species model provide researchers with a list of candidates of metabolic interactions and should be reviewed in further research. The interactions predicted by the two-species model provide possible mechanisms to explain the observations in the experiments, especially for those expressed enzymes that catalyze reactions requiring metabolites that cannot be synthesized by the organism in the experimental condition. Metabolic response to other organisms is another reason that can cause these interactions, e.g. the inactivation of certain enzymes in the expression data. Thus, this two-species model can be used as a framework to interpret these gene expression data.

4.3 Metabolic Network Modeling of a Ten-species Model Microbial Community

4.3.1 Background

To further study the relationships between diets and the structure of gut microbial communities, a ten-species model community has been established (Faith et al, 2011). By inoculating 10 sequenced human gut bacteria into gnotobiotic mice and tracing the changes in response to the diets, the authors were able to develop a statistical model that can predict over 60% of the variations in the population structure changes. The 10 sequenced human gut bacteria can be divided into four classes: i) bacteria that can utilize complex dietary polysaccharides, including *B. thethaiotaomicron*, B. ovatus and B. caccae; ii) bacteria consuming oligosaccharides and simple sugar, including *E. rectale*, M. formatexigens, C. aerofaciens, and E. coli; iii) bacteria that can ferment amino acids, including E. coli and C. symbiosum; iv) H_2-consuming bacteria, including D. piger and B. hydrogenotrophica.

To study the changes in structure of this ten-species model community, sequence-based transcriptome was applied to measure the gene expression levels. Table 4.5 lists the coverage of the transcriptome for these ten species. Besides *E. rectale*, the coverage of the transcriptomes for the remaining nine organisms were higher than 60%. The inoculated strain names are also listed in the table. According to these sequence-based transcriptomes, the expressed genes were predicted by the number of reads for all the genes. Both unique reads and non-unique reads were counted.

However, the non-unique reads only contribute to the gene expression levels partly, which is proportional to the number of unique reads of the corresponding genes. After normalization, if more than 64 sequences have been identified for one gene, this gene is considered as expressed. Compared to the genechip based expression data, sequence-based expression is more reliable in terms of absolute value of individual gene expressions. Therefore, the sequence-based transcriptomes can be applied to predict the gene activity for individual conditions and datasets.

Table 4.5: Strains inoculated in the 10-species model community and the transcriptome coverages. The gene with more than 64 sequencing reads were considered as expressed.

Species	Gene identified in transcriptome	Total genes in genome	Transcriptome coverage
B. thethaiotaomicron VPI-5482	3696	4498	77%
B. ovatus ATCC 8483	3785	5536	68%
B. caccae ATCC 43185	3375	3855	88%
E. rectale ATCC 33656	453	3621	13%
M. formatexigens DSM 14469	3173	4896	65%
C. aerofaciens ATCC 25986	1779	2367	75%
E. coli K-12 MG1655	2969	4132	72%
C. symbiosum ATCC 14940	3141	5128	61%
D. piger GOR1	1660	2487	67%
B. hydrogenotrophica DSM 10507	2612	3869	68%

Utilizing a three-step modeling procedure that was similar to the one used in the two-species model microbial community, we could reconstruct the community-wide metabolic network for this ten-species model microbial community (Figure 4.9). We made three major modifications to fit the data of ten-species model microbial community. First, there were no comprehensive growth test results for the ten species inoculated. Therefore, in the first step, the growth model was only applied in the diet environment. Second, the sequence-based expression data can provide accurate prediction of reaction activity and there were no data considering mono-inoculated cases. To make use of these expression data, only the co-inoculated condition was considered. Third, in the last step, the ten-species MinExchange model did not fix the reaction activities predicted in the second step. In opposite, the reaction activity was still a variable which was considered in the objective function together with potential exchanges.

4.3.2 Three-step Metabolic Network Reconstruction and Modeling Method

4.3.2.1 Metabolic Reconstruction for Growth Model of the Ten Species

The first step in this three-step modeling procedure is the reconstruction of metabolic networks for the growth model of the ten species. First, we collected the gene annotations of the ten species from three resources, manual annotations (`http://gordonlab.wustl.edu/modeling_microbiota/`), RAST automated annotations tools (`http://rast.nmpdr.org/`), and annotations in KEGG database (`http://www.genome.jp/kegg/`). Only the annotations that were consistent in all three datasets were selected, which was the same as in two-species model. Because there were no extra growth test results, we applied the growth requirement only to the diet conditions. There were numbers of similar diets with different amounts of nutrients used in the ten-species model microbial community; however, the compositions of these diets were not changed. Therefore, we can identify the potential nuterients from the controlled diet.

We assumed all metabolites were freely exchanged in the growth model, which will be refined in the third step. We made this assumption because there were no individual inoculation studies in the work of Faith et al. All the other assumptions and methods were exactly the same as in the two-species model (Section 4.2.2).

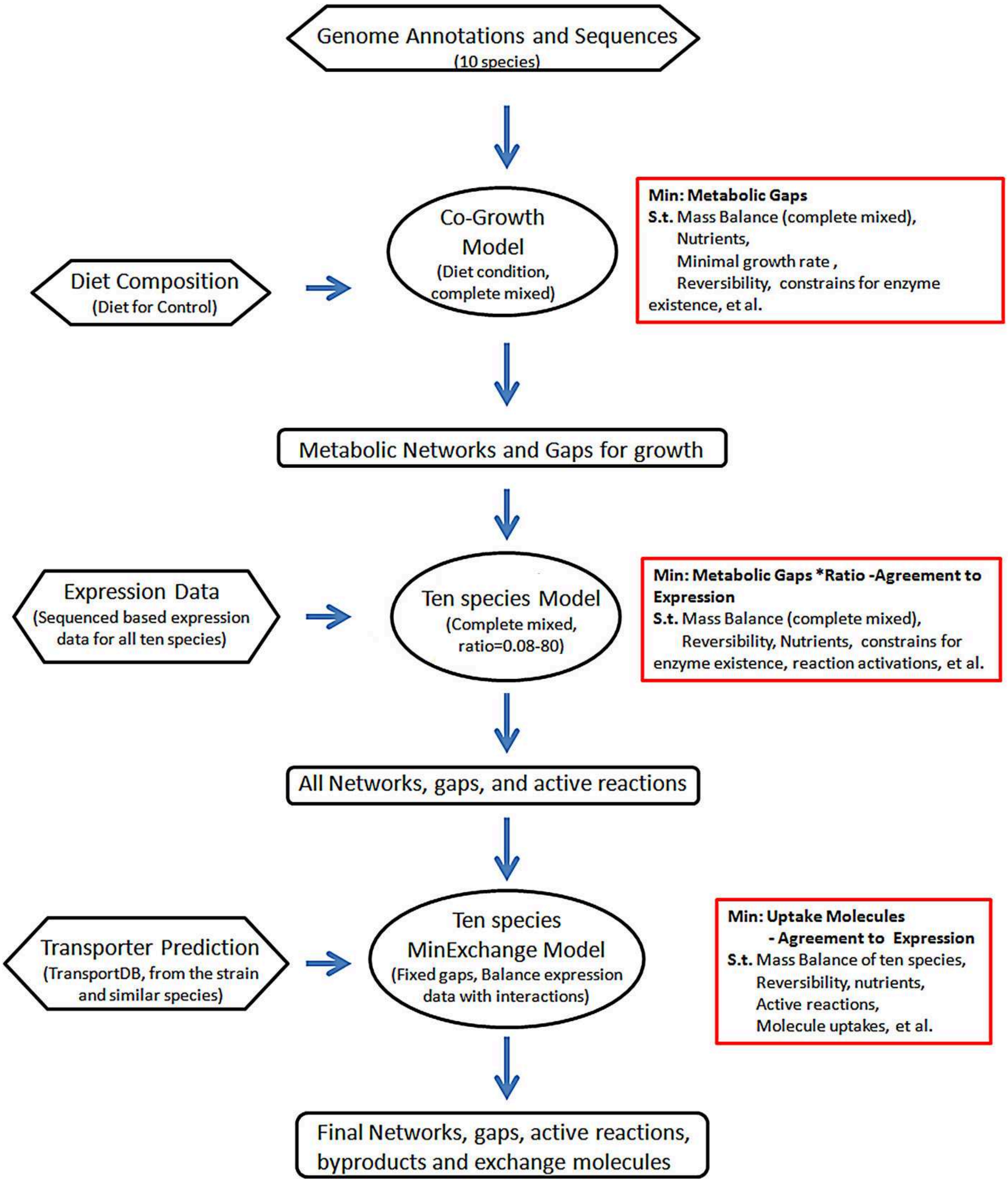

Genome Annotations and Sequences
(10 species)
Co-Growth Model
(Diet condition, complete mixed)
Min: Metabolic Gaps
S.t. Mass Balance (complete mixed),
Nutrients,
Minimal growth rate ,
Reversibility, constrains for enzyme existence, et al.
Diet Composition
(Diet for Control)
Metabolic Networks and Gaps for growth
Expression Data
(Sequenced based expression data for all ten species)
Ten species Model
(Complete mixed, ratio=0.08-80)
Min: Metabolic Gaps *Ratio -Agreement to Expression
S.t. Mass Balance (complete mixed),
Reversibility, Nutrients, constrains for enzyme existence, reaction activations, et al.
All Networks, gaps, and active reactions
Transporter Prediction
(TransportDB, from the strain and similar species)
Ten species MinExchange Model
(Fixed gaps, Balance expression data with interactions)
Min: Uptake Molecules
- Agreement to Expression
S.t. Mass Balance of ten species,
Reversibility, nutrients,
Active reactions,
Molecule uptakes, et al.
Final Networks, gaps, active reactions, byproducts and exchange molecules

Figure 4.9: Flow chart of three-step modeling procedure for the ten-species microbial community. The methods used in the three steps are briefly described. More details are in Section 4.3.2

4.3.2.2 Ten-Species Model Considering Sequence-based Expression Data

In the second step, the sequenced-based expression data was used to further refine the metabolic networks of the ten species. Equation 4.2 demonstrates the constraints that determine the activity of reactions. Only co-inoculation condition was considered, the same as in the first step. All the other constraints used in this step were the same as in two-species model(Section 4.2.2.2).

We made an assumption that if a reaction with higher expression level, it is more likely to be active in the tested condition. Therefore, a different objective function was developed to fit the new expression data. Equation 4.5 is the objective function. In the objective function, $P(r)$ and $N(r)$ are the binary variables representing whether v_r is positive or negative; $b_{add}(r)$ indicates whether reaction r is a putative reaction. $Weight(r)$ is the penalty level of putative reactions as defined in Equation 2.1.

$$obj : \min_{v,b_{add},P,N} ratio \cdot \sum_{r \in R} b_{add}(r) \cdot Weight(r) - \sum_{r \in R} (P(r) + N(r)) \cdot D(r) \tag{4.5}$$

In this objective function, D(r) is a new parameter that is correlated to the expression level. Due to the large range of expression level, D(r) must be normalized. Here we chose a stable definition, 10th-quantiles (deciles) of the expression value (Equation 4.6). Therefore, if the gene catalyzing reaction r is among the highest 10% average expression level, D(r) is set to 1. If the reaction r cannot be identified in the expression data, D(r) is set to 0. The ratio parameter in the objective function balances the two parts in it. With a higher ratio value, the model will emphasize the metabolic gap filling, and oppositely, the model will emphasize the expression data for lower ratio value. In this work, four values of the ratio parameter were chosen (0.08, 0.8, 8, and 8). The discovery rate of each putative reaction under one ratio setting was calculated based on the suboptimal solutions within 5% optimality gaps. Only the putative reactions with higher than 50% discovery rates in at least two ratio settings were considered. The active and inactive reactions for both mono- and co-inoculation conditions were calculated using the model by fixing these selected putative reactions.

$$D(r) = deciles(Expression(r)) \tag{4.6}$$

4.3.2.3 Prediction of Potential Interactions in Microbial Community

To predict potential interactions, we assumed all the metabolites can be exchanged across the species and if the exchanged molecules are required either because of the expression data or the growth requirement, we considered them as minimal required exchanges. The MinExchange model used in the two-species model was modified and applied to the ten-species model microbial community. The major modification from the 2-species model was MinExchange does not fix the active reactions that were predicted in the second step. One reason for this modification was the interactions in the ten-species model were much more complex than in the two-species model. To capture the most significant interactions, we relaxed the constraints on reaction activity and balanced this relaxation with the number of interactions. Therefore, the objective function for this MinExchage model contains two parts (Equation 4.7). In this objective function, I_{uptake} is a binary variable indicating whether re is an uptake flux. $weight_{uptake}(re)$ represents the probability of the uptake flux, which was calculated based on transporter predictions.

$$obj : \min_{v, b_{add}, P, N, I_{uptake}} ratio \cdot \sum_{re \in Re} I_{uptake}(re) \cdot Weight_{uptake}(re) - \sum_{r \in R} (P(r) + N(r)) \cdot D(r) \tag{4.7}$$

We assumed that the interactions with more specific transporter prediction in the genome or the genome of similar organisms in the phylogenetic tree will be more likely to take place. To predict this probability, we used TransportDB (`http://www.membranetransport.org/`) for transporter predictions. However, only part of the ten species has been analyzed by TransportDB. According to the assumption, to predict the transporters for the strains not in TransportDB, the $Weight_{uptake}$ parameters were calculated based on some strains in TransportDB close to the target strains in the phylogenetic tree. Table 4.6 lists the strains that were chosen to calculate the weight parameters. The minimal byproducts can be predicted using the same methods, which gives penalty to secretion rather than uptake.

Table 4.6: Strains in TransportDB used for $Weight_{uptake}$ calculation in the ten-species model. If more than one strains was chosen, the $Weight_{uptake}$ was calculated according to all the transporters identified in the reference strains.

Species in ten-species model	Strains for $Weight_{uptake}$ calculation
B. caccae ATCC 43185	*G. forsetii* KT0803 and *P. vibrioformis* DSM 265
B. ovatus ATCC 8483	*G. forsetii* KT0803 and *P. vibrioformis* DSM 265
B. thetaiotaomicron VPI-5482	*B. thetaiotaomicron* VPI-5482
B. hydrogenotrophica DSM 10507	*E. faecalis* V583 and *C. acetobutylicum* ATCC 824
M. formatexigens DSM 14469	*E. faecalis* V583 and *C. acetobutylicum* ATCC 824
C. symbiosum ATCC 14940	*E. faecalis* V583 and *C. acetobutylicum* ATCC 824
C. aerofaciens ATCC 25986	*B. longum* NCC2705
E. coli str. K-12 substr. MG1655	*E. coli* str. K-12 substr. MG1655
E. rectale ATCC 33656	*E. faecalis* V583 and *C. acetobutylicum* ATCC 824
D. piger GOR1	*D. piger* GOR1

4.3.3 Results

4.3.3.1 Individual Metabolic Networks of Microbial Community

We used the three-step modeling procedure to reconstruct the metabolic networks of all the ten species simultaneously. Two types of automated curations were carried out in the first two steps, including growth requirements and expression data curations. In the second step, four ratio parameters were utilized, from a low penalty level to a high penalty level. We integrated the solutions under different level of penalty and generated the final metabolic networks for all the ten species. Table 4.7 contains the information of the final metabolic network reconstructions.

Table 4.7: Summary of the reconstructed metabolic networks of the ten species in the model microbial community.

Species	Number of genes	Identified reaction	Putative reaction
B. thethaiotaomicron VPI-5482	4498	880	75 *
B. ovatus ATCC 8483	5536	901	19
B. caccae ATCC 43185	3855	861	19
E. rectale ATCC 33656	3621	688	76 *
M. formatexigens DSM 14469	4896	778	26
C. aerofaciens ATCC 25986	2367	665	25
E. coli K-12 MG1655	4132	1162	17
C. symbiosum ATCC 14940	5128	847	20
D. piger GOR1	2487	687	26
B. hydrogenotrophica DSM 10507	3869	775	26

*: Metabolic gaps identified in 2-species model were included

Besides the two species studied in the two-species model, the number of putative reactions for one species is around 3%. However, this number can simply increase to 10% when decreasing the penalty level. These results indicate that part of the reactions identified in the expression data are isolated from the rest, or belong to incomplete pathways. Thus, more putative reactions are needed to activate these reactions. In this work, these isolated reactions were not considered because the metabolic networks integrated from four penalty levels were very close to the solutions with high penalty level. In all of the putative reactions, 19.5% of them can be found in one of the annotation. Therefore, these metabolic gaps are caused by inconsistent gene annotation. Another 33.4% of the putative reactions can be assigned to gene candidates with P value less than e^{-30}. Figure 4.10 demonstrates the most abundant pathways of the metabolic networks for the ten species, including putative reactions. Purine metabolism, pyrimidine metabolism, and nucleotide sugars metabolism are the most abundant pathways for all of the ten species. Across the species, the differences in pathway level are almost neglectable. This similarity in abundant pathways indicates that these functions are all essential for these organisms.

4.3.3.2 Metabolic Network Modeling with Expression Data

The sequence based expression data provide direct information about the activity of individual genes and associated reactions. In the work of Faith et al (2011), a threshold (64=2^6) was chosen to identify active genes. This threshold was decided arbitrarily and may not be accurate for all the genes and organisms in the microbiome. To avoid the arbitrarily determined threshold, the 10th-quantiles were used to indicate the activity of the genes. Four values of the ratio parameter, from lower penalty to high penalty of metabolic gaps, were tested. As shown in Figure 4.11, the fractions of active reactions in the reactions predicted to be active by the expression data decreased when the ratio parameter was increased. This is because the model emphasizes more metabolic gaps than gene expression data. After integration of the four sets of solutions, the final prediction is between the solutions from ratio equal to 0.8 and 8. Furthermore, we found that the active fractions from the top 10% to 80% of the highly expressed reactions did not decrease significantly, but the active fractions for the 10% lowest expressed reactions was significantly lower than the rest. These results indicate that there was no bias in the model for the reactions with expression above one threshold. This finding was consistent with the assumption Faith et al made. Compared to the results obtained without curation of gene expression data, all the four sets of results have much higher active fraction for all active reactions. Therefore, the curation of gene expression was essential in this community-wide metabolic model.

Based on the model prediction after curation of gene expression data, we can calculate the activities of pathways for all the ten species in the gut environment

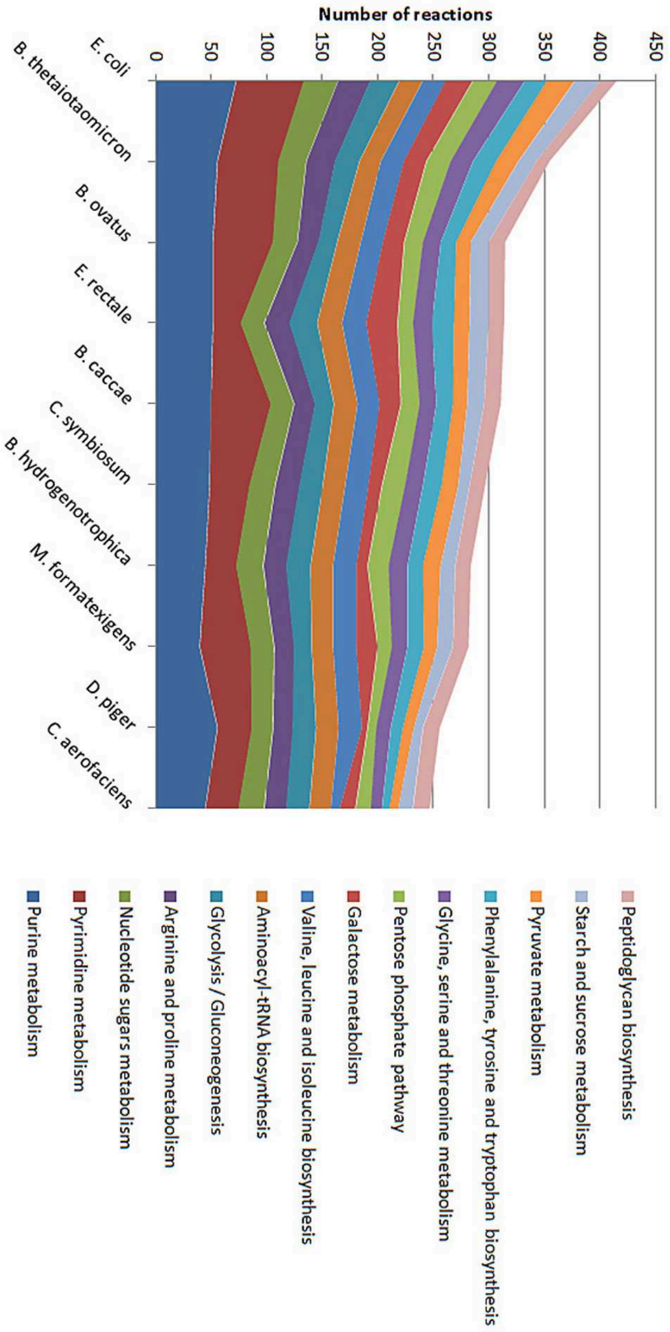


Figure 4.10: The most abundant metabolic pathways in the metabolic networks of the ten species. Putative reactions (metabolic gaps) are also included. The species are ordered by the total number of reactions in these pathways.

(Figure 4.12). From the results, we found that the purine and pyrimidine metabolism pathways were the two pathways with the highest number of active

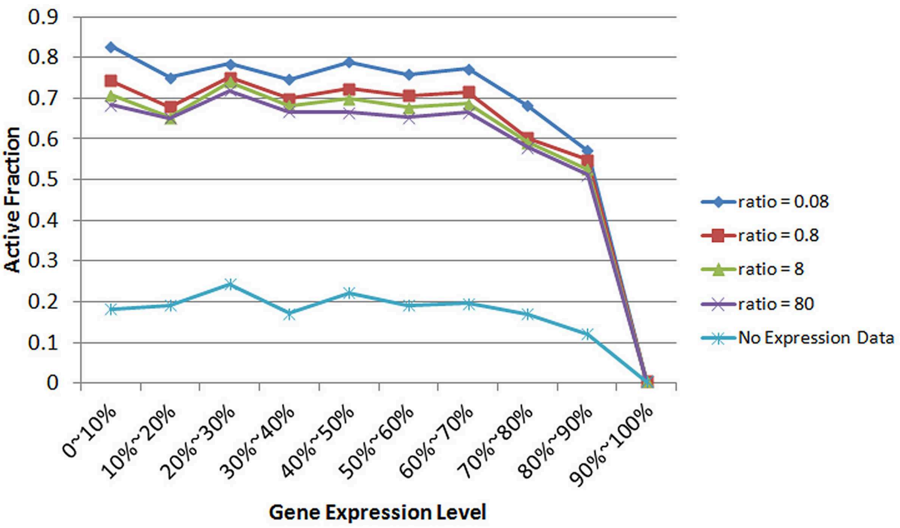


Figure 4.11: Fractions of predicted active reaction in the total active reactions accroding to expression data. The results predicted under four ratio parameters were shown. The results without gene expression data are provided for comparison.

reactions, which was consistent with the total number of reactions identified in the microbial community. However, the average active fraction for these two pathways were significant lower than some other pathways, such as glycolysis, phenylalanine tyrosine tryptophan biosynthesis, and valine, leucine and isoleucine biosynthesis pathways. these results indicate the most active pathway in the gut microbiota are those related to carbon utilization and amino acid metabolism. Compared to the pathway distributions of all the metabolic reactions, including both active and inactive, fatty acid biosynthesis, folate biosynthesis and pantothenate and CoA biosynthesis biosynthesis were new in this active pathway list. Therefore, these pathways play important roles in the gut microbial community. For the four pathways with an active fraction higher than 70%, two of them were carbon utilization related. These results indicate that the carbon utilization is the most important function for certain organisms, which can be explained by the host-microbiome interactions.

From the figure, we found that the activities of certain pathways vary

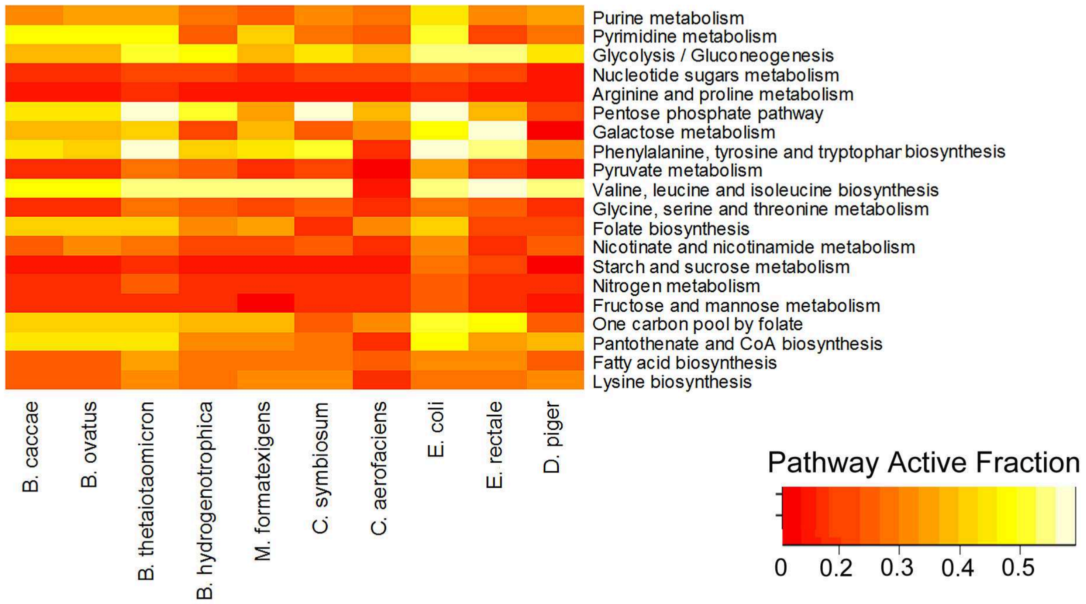


Figure 4.12: Active metabolic pathways in the metabolic networks of the ten species. The pathway active fraction is the fraction of active reactions in all the reaction associated with the pathway. The 20 pathways with most active reactions were listed.

significantly across the ten species. For example, folate biosynthesis and pantothenate and CoA biosynthesis biosynthesis both have higher active fractions in all the three Bacteroidetes strains but lower in all the four Firmicutes strains. For valine, leucine and isoleucine biosynthesis pathway, nine of the ten species have significantly high active fraction, but not in C. aerofaciens. This indicates that C. aerofaciens may not synthesize these amino acids by itself. Arginine and proline metabolism pathways have low active fractions in all the ten species, indicating the microbiome have extrageneous sources for them, such as host diet.

4.3.3.3 Metabolic Interactions

The metabolic gaps and activities of reactions have been predicted by the ten-species metabolic network reconstruction with gene expression data curation. However, we assumed that all the metabolites can exchange across the species, which is not accurate. In Section 4.3.2.3 we described the methods we used to predict potential interactions among the ten species. Similar to the reaction activity prediction, one parameter was introduced to balance the two parts of objective function: potential interactions and curation of gene expression data. With a lower penalty level, the model requires less interactions but with less active fractions according to gene expression data. As shown in Figure 4.13, the average active fraction decreased from 55% to 47% when the ratio parameter was increased from 0.1 to 10. We calculated the error bars based on all the suboptimal solutions within a range for which the objective values increased slowly. We observed a significant decrease of predicted interactions when the ration parameter increased from 1 to 10 but the average active fraction of reaction were almost the same for the two set of solutions when considering the error bars. Therefore, the range of ratio parameter from 0.1 to 1 includes most of the variations of active fractions and we focused on this range.

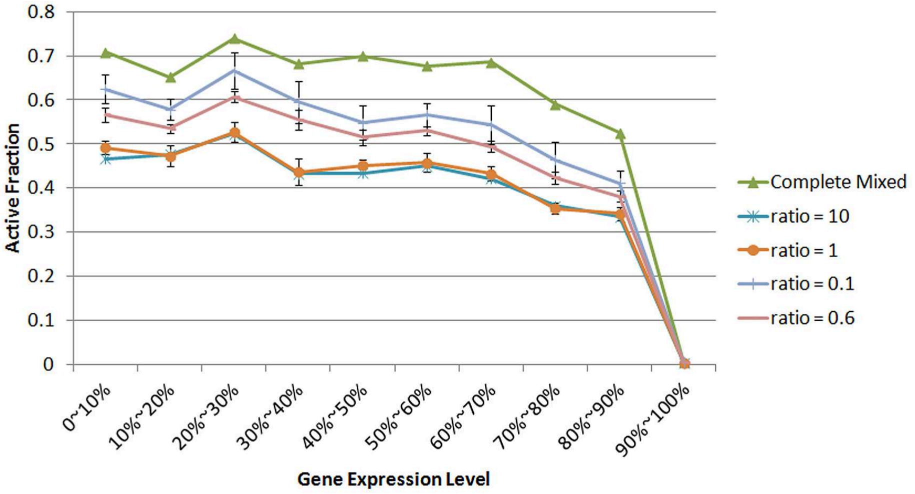


Figure 4.13: Fractions of active reaction in the reaction predicted to be active accroding to expression using MinExchange method under four ratio parameters and compared with complete mixed condition. Error bars were calculated from the selected suboptimal solutions.

Different numbers of interactions were predicted for the four settings of ratio parameter (Figure 4.14). When ratio parameter was set to 10, four molecules were predicted to be exchanged between the organisms. One of the molecule was citrate and there were only two strains, *M. formatexigens* and *C. symbiosum*, utilizing citrate. *Antranikian et al.* tested 44 species in Clostridium and found *C. symbiosum* was able to utilize citrate. This experiment supports our prediction. According to Figure 4.14, the uptake fluxes that can be confirmed by metabolomic data were consistent for all of the three set of solutions. As discussed, the ratio parameter from 0.1 to 1 contributes most to the variations in reaction active fractions. Therefore, the results from these three sets of solution were most representative of this microbial community. Agmatine is an important intermediate for polyamine synthesis. Research indicate that *C. aerofaciens* lacks polyamine biosynthesis function and was a polyamine auxotroph. This conclusion is consistent with our prediction of *C. aerofaciens* requiring agmatine uptake. Urea was another molecule found to be utilized by some of the strains, including *B. thetaiotaomicron*, *E. rectale* and *C. aerofaciens.* Kinetics data from human and pig indicate that the gut microbiome utilizes almost 25% of urea synthesized by the host. The predicted interactions in this 10-species metabolic model provide three candidates that may contribution to the urea utilization in gut.

The minimal byproducts that were produced by the microbial community were predicted in addition to uptake fluxes (Figure 4.15). For example, citrate must be secreted by *B. hydrogenotrophica.* Citrate was produced by the highly expressed citrate synthase (RUMHYD00774 and RUMHYD00862), but isocitrate dehydrogenase, an enzyme utilizing isocitrate, was not identified in this strain. Therefore either citrate or isocitrate was secreted by *B. hydrogenotrophica.* Because *C. symbiosum* needs citrate uptakes, the model predicts citrate to be the byproduct of *B. hydrogenotrophica.* Glutathione was predicted to be another byproduct produced in this model microbial community. *Escherichia coli* was predicted to produce glutathione because both gamma-glutamylcysteine synthetase (b2688) and glutathione synthase (b2947) have been identified in the transcriptome of

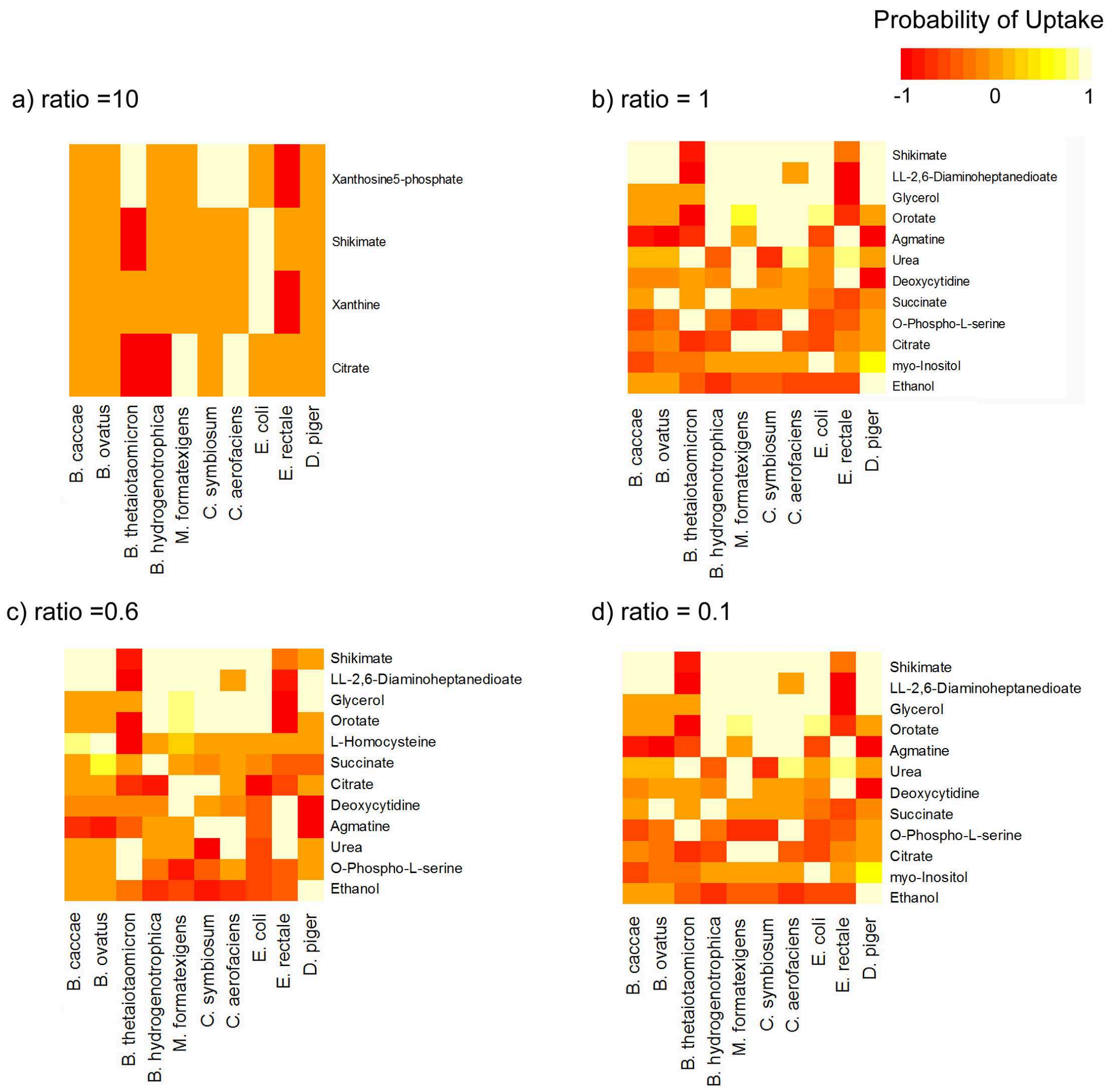


Figure 4.14: Predicted uptake fluxes under four settings of ratio parameter in 10-species model. The complete predictions are reported for ratio = 10, and for the rest three settings, only molecules found in gut environment were reported, which are almost constant under all three settings.

Escherichia coli. Therefore, *Escherichia coli* produces glutathione from cysteine and glutamate. Gamma-glutamyltransferase (b3447) and glutathione peroxidase (b1710), the enzymes utilize glutathione, was predicted to be inactive due to the low expression level. This is another reason for predicted glutathione secretion. Metabolomic studies also indicate glutathione play roles in response to oxidative stress through the host-microbiome interactions (*Musso et al.*, 2011; *Matsumoto* *et al.*, 2012). As predicted by the model, urea is produced by *Escherichia coli* and secreted. The agmatine amidinohydrolase (b2937) was identified in the transcriptome of *Escherichia coli*, which converts agmatine to putrescine and urea in the arginine metabolism pathway. However, urease (EC:3. 5. 1. 5) was not found in *Escherichia coli*, which forced *Escherichia coli* to secrete urea. *Morris and Koffron* observed *Escherichia coli* secreted urea when it converted arginine to putrescine using isotopic labeling experiments. This experimental result confirmed this secretion of urea in *Escherichia coli*.

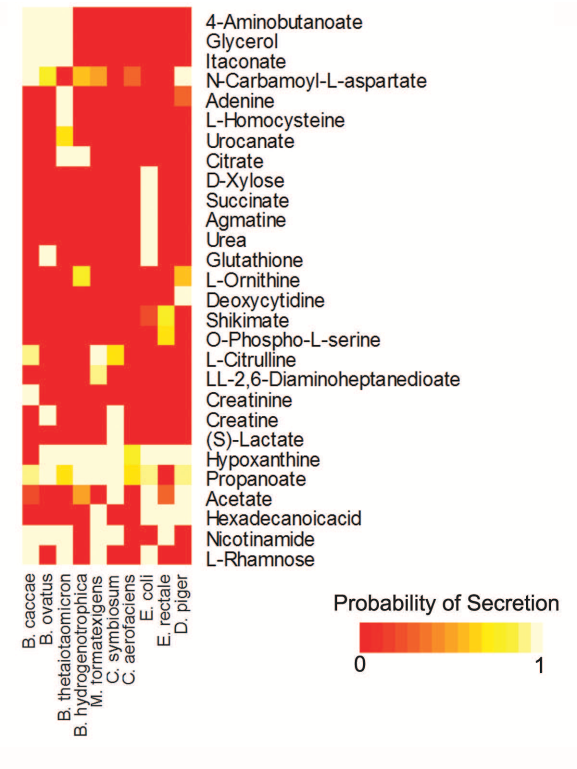


Figure 4.15: Predicted minimal byproducts in the ten-species microbial community. The uptake fluxes predicted under 0.1 ratio setting were reported.

4.4 Discussion and Conclusions

In this chapter, we introduced two cases studies applying metabolic network reconstruction in community-wide modeling of model gut microbial communities.

For the two-species model microbial community, comprehensive growth tests and genechip based transcriptomes were included in the study of *Mahowald et al.*. These data enabled us to reconstruct high-quality metabolic networks for both individual organisms and the whole microbial community following the multi-step modeling procedures.

The three-step metabolic network modeling procedure for two-species microbial community took advantage of these data and refined the metabolic networks sequentially. There were several reasons for this sequential procedure. First, the accuracies of these datasets were different. The growth tests provided accurate indications to determine the capability of utilizing different carbon sources. In contrast, genechip based expression data provide quantitative measurements about the changes of gene expression levels under mono- and co- inoculated conditions. A threshold was chosen to predict the active and inactive metabolic reactions based on the level of changes. Therefore, the predicted activities of genes and reactions were not as accurate as the growth phenotypic data. Non-specific binding and background noises were some other factors that decreased the accuracy of gene expression data. To represent the differences in data quality and accuracy, we first curated the metabolic networks with growth test results and then expression data. The transporter predictions completely rely on computational calculations from genomic sequences, which was less inaccurate than the remaining two datasets. Thus, interaction predictions, which considered the exchange fluxes between the two organisms, were applied in the last step.

Similar to the two-species microbial community, the three-step metabolic network modeling procedure used in ten-species microbial community followed the same sequences. The draft metabolic networks for all the ten species were reconstructed and refined with growth model. However, there were no comprehensive growth tests for the ten-species microbial community. Thus, we applied a community-wide growth model to enforce the growth requirements. In the second step, sequence-based gene expression data were used to curate the metabolic networks. The sequence-based gene expression data provided more accurate measurements about gene expression levels than genechip based methods. Therefore, the absolute values of the gene expression data were utilized rather than the changes between conditions used in two-species model. The model predictions indicated nearly constant active fractions for the top 80% expressed genes and reactions. These results suggested the absolute values of sequence-based gene expression levels did not affect the gene activity predictions once the absolute values were above a threshold. This observation confirmed the existence of such a threshold, which was used in many works.

We observed that the active fractions of metabolic reactions predicted to be active by expression data was in the range of 60% to 80% for different organisms and parameter settings. Even at the highest level, there were still 20% reactions not connected to the active metabolic networks. These inactive reactions were caused mainly by two reasons. First, metabolic gaps in the metabolic networks were not identified completely. In our models, the agreement between model predictions and expression data was balanced by ratio parameters. Therefore, the false negative predictions of metabolic gaps can be controlled but not eliminated under the framework we developed. Another reason for those inactive reactions was the incomplete metabolic pathways in reaction database. The knowledge on cellular metabolism is increasing rapidly. Therefore, the inconsistency caused by incomplete metabolic pathways will decrease as the knowledge grows.

The interactions, for both uptake and secretion, between the ten species were predicted based on the reconstructed metabolic networks and activity predictions. These interactions were predicted to be the minimal requirements to meet certain level of agreement with gene expression data, which was governed by the ratio parameter.

These minimal predictions might not cover some interactions that could be carried out by more than one species but was not necessary for any of these species. Therefore, to predict the interaction more comprehensively, we applied both minimal uptake model and minimal secretion model to the ten-species microbial community. The situation was different for the two-species model. Because there were only two species, the interactions were symmetric and only one of the two directions was needed.

Part of the predicted interactions in two-species model and ten-species model can be supported by experimental data in literature. However, the number of these confirmed interactions is still small compared to the total number of predicted interactions. This phenomenon reflects some of the challenges we are facing today in studying microbial community. The culture-independent methods provide researcher with a great amount of data about the population structure and metabolism of the microbial community. However, characterizations of the organisms in microbial communities still largely rely on cultivation of the microorganisms. Community-wide models provide a possible path to resolve this issue. As demonstrated by our results, the metabolic network models can not only predict the known properties of the microbial community but also provide hypotheses about unknown metabolic functions and interactions. The power of model prediction might become more important as more culture-independent data are collected, and our work provides a promising direction to model microbial community by integrating -omics data.

In this work, the complexities of the two model gut microbial communities were relatively low compared to natural gut microbiota. Directly modeling the natural microbial community definitely will bring more comprehensive understanding of the metabolic functions and interactions of these microbial communities. However, the complexity and scale of the data will bring new challenges in both metabolic network reconstructions and computation requirements. In the future, in order to apply the community-wide metabolic models to natural microbial communities, we need more simplifications during the modeling process. We will discuss some of the possible simplifications and assumptions in Chapter VI.

CHAPTER V

Metabolic and Regulatory Network Design for Biochemical Production

5.1 Introduction

5.1.1 Metabolic and Regulatory Network Design

Metabolic engineering has emerged as an effective method to improve biosynthesis of different products, including biofuels. Localized modifications of metabolic pathways, such as eliminating competing branches and over-expressing enzymes of rate-limiting steps, have been commonly applied and provide significant improvements. However, these modifications can also cause changes on the cell metabolism in those pathways which are not directly relevant. Another limitation of these modifications is that they may fail to re-engineer cell metabolism as expected due to the complex regulation of cell metabolism. As a result, systematic designs of modifications of metabolic networks are desired, which may resolve or partly resolve the issues mentioned above.

Researchers have been developing computational tools for systematic designs of modifications of metabolic networks, and many works have demonstrated the advantages of applying such an optimization process. Some of the early algorithms required detailed enzyme kinetics (*Fell, et al.* 1996) that are largely missing for most systems. Metabolic control analysis (MCA) (*Domach et al.*, 2000), based on flux control coefficients, provides a different way to predict cell metabolism changes based on experiment-based data, and can include both effects of enzyme kinetics and cellular regulation. However, MCA type models require a large amount of experimental data, which is still not available for most organisms and pathways.

To avoid the requirements of either enzyme kinetics or experimental data, constraint-based models that can help researcher for strain design have been developed and widely used. Alper et al. used flux balance analysis to screen single, double and triple gene knockouts that can increase the lycopene biosynthesis in *Escherichia coli* (*Alper et al.*, 2005). Jin Hwan Park et al. used a similar strategy to verify their designed *Escherichia coli* strain for L-valine production by carrying out *in silico* gene knockout simulation before they applied the design in the lab (*Park et al.*, 2007). Besides these models, more systematic design models have been developed for better performance. For example, Kiran Patil used one evolutionary programming method to design *Saccharomyces cerevisiae* strains for acid, glycerol and vanillin production (*Patil et al.*, 2005). Using a different strategy, OptKnock (*Burgard et al.*, 2003) is able to predict gene knockouts for better biochemical production after adaptive evolution by solving a bi-level Mixed Integer Linear Programming (MILP) problem, which can provide optimized designs for different purposes. In the OptKnock algorithm, the whole problem has two levels. In the inner level, cell metabolism is modeled by flux balance analysis (FBA) model, and genetic manipulations (gene knockouts) are included in the model by forcing the corresponding reactions to carry zero flux. In the outer lever, the best choice of genetic manipulations is searched according to desired functions, e.g. maximum of production. The bi-level problem is converted into single level MILP according to duality theory and then solved. After the OptKnock was developed, a number of models utilizing this bi-level optimization framework were introduced. For example, OptStrain (*Pharkya et al.*, 2004), allows both gene knockouts and gene knock-ins, which represents expressing novel heterogeneous enzymes into the designed strains. By doing so, functions that do not exist in wildtype strains can be achieved in the designed strain, which enlarges the design space. Along this line, OptReg was developed by *Pharkya and Maranas* (2006), which considers up- and down-regulation of metabolic reactions in addition to gene knockouts. Using this method, researchers want to further improve the performance of the strains which cannot be achieved by simply removing or adding reactions.

Cellular regulations always play important roles in cell metabolism, which should not be neglected when designing strains. Regulatory flux balance analysis (rFBA)(*Covert et al.*, 2004) was developed to model the effects of regulatory network on metabolic network. In the rFBA model, logic rules that represent regulatory networks are used to predict the inhibited metabolic reactions under different conditions. Along the same line, integrated flux balance analysis (iFBA) *Covert et al.* (2008) was developed to model effects of cellular regulation by utilizing ordinary differential equations (ODEs) to represent signal pathways and then integrating these ODEs into the constraints of FBA framework. Differently, Chandrasekaran and Price introduced a probabilistic regulation of metabolism (PROM) model (*Chandrasekaran and Price*, 2010), using the state of regulatory factors to predict the expression level of the genes and then to predict the fractions of activity of the corresponding reactions. The expression levels of transcriptional factors are estimated based on transcriptome data from various conditions.

Researchers are not satisfied by merely predicting the effects of regulatory network on cellular metabolism. Joonhoon Kim and Jennifer Reed developed a computational framework, OptORF(*Kim and Reed*, 2010), to design both metabolic and regulatory perturbations under bi-level MILP framework by embedding the linear regulatory constraints within the design model. In this model, linear regulatory constraints are used to represent the regulatory network, which will identify inhibited reactions. Thus, besides gene knockouts, removal of inhibition is another type of perturbation that the model allows. Because the model considers cellular regulation, it is able to search for strains that can retrieve certain functions which are not available normally in designed conditions. A different type of model, OptForce (*Ranganathan et al.*, 2010), was developed to identify sets of genes/reactions that should be down/up regulated, or eliminated by predicting desired metabolic flux values in optimal conditions. Even though this algorithm does not suggest the elimination of transcriptional factors to remove inhibition, similar suggestions can be derived when reviewing the reactions that require a increase of flux from zero.

5.1.2 Fatty Acids As Potential Biofuel Precursors

Biodiesel, methyl esters of fatty acids, is one of the major types of biofuels that are commercially available. Currently, biodiesel is exclusively produced from plant oils, which are mainly composed of triacylglycerols. Therefore, glycerol, an undesired byproduct during the production of biodiesel from triacylglycerols, presents a major problem in biodiesel manufacture. To avoid glycerol, direct production of fatty acid ethyl esters or free fatty acids followed by conversion into biofuel molecules is an alternative choice. As presented in Schirmer et al., C13 to C17 mixtures of alkanes and alkenes can be synthesized in *Escherichia coli* by introducing an alkane biosynthesis pathway (*Schirmer et al.*, 2010). Similarly, Steen et al., engineered *Escherichia coli* to produce fatty acid ethyl ester leading to a yield of 674 mg/L (*Steen et al.*, 2010). All these works demonstrate the potential of biosynthesis of fatty acids for biofuel production.

Fatty acids are important precursors for the biosynthesis of cell envelopes, and most bacteria have the function of biosynthesis of fatty acids. However, challenges remain for practical production of fatty acids or derived biofuels. One of these challenges is the efficiency and yield of free fatty acid production. One explanation for the low efficiency and yield is that the anabolic and catabolic processes involved in fatty acid metabolism are strongly regulated in bacteria transcriptionally and post-transcriptionally (*Magnuson et al.*, 1993). Through these regulated pathways, cells are able to produce and consume these molecules precisely, indicating significant re-programming of these functions is required to produce and accumulate free fatty acids in cells. To achieve this, mechanisms for fatty acid metabolism and their regulations are required. Fatty acid metabolism in bacteria has been extensively studied. In *Escherichia coli*, for instance, fatty acids synthesis starts exclusively from acetyl-CoA, which is the same as in plants. Acetyl-CoA and bicarbonate are first converted into malonyl-CoA by Acetyl-CoA carboxylase(ACC) with the requirement of ATP, which is believed to be the rate-limiting step for fatty acid synthesis. Malonyl-CoA is then converted into fatty acyl-ACPs (acyl carrier proteins) by fatty acid synthase (FAS). These acyl-ACPs can be either used to synthesize phospholipids by glycerol-3-phosphate acyl transferase and other transferases or converted to free fatty acids by acyl-ACP thioesterases which can be further degraded to acetyl-CoA through β oxidation pathway. The biosynthesis of fatty acids or phospholipids requires both ATP and NADPH(NADH), which is a energy consuming process and is well regulated.

Researchers have tried different ways to enhance the biosynthesis of fatty acids and redirect fatty acid metabolism to improve the production of free fatty acids (*Lu et al.*, 2008). These localized optimizations for fatty acid production reach titers of 2.5g/L with 4.8% carbon efficiency, suggesting both opportunities and challenges in synthesizing fatty acids for biofuel production. One possible reason for the low conversion is that engineered cell metabolism is still far from the optimal state for producing free fatty acids, even though the biosynthesis pathway is enhanced.

Therefore, we hypothesize if cell metabolism and regulation can be further optimized globally, the production of fatty acids can be more efficient. To achieve this goal, we need to develop a method that enables us to design a metabolic network and regulatory network at the same time.

The reconstructed metabolic networks contain comprehensive information about cellular metabolism. One application for these metabolic networks is strain optimization for specific purpose, e.g. biofuel molecules production. In this chapter, we will utilize both metabolic network and regulatory network in a bi-level optimization framework to design strains for biofuel molecules production. Two specific objectives should be achieved in this chapter.

- Develope a bi-level optimization framework for metabolic network and regulatory network design.
- Design mutated *E. coli* strains for overproduction of fatty acids and some other biofuel molecules utilizing the bi-level optimization framework.

5.2 Bi-level Optimization Framework for Metabolic and Regulatory Network Design

To design a metabolic network and corresponding regulatory network, we developed a model framework that considers the modification of both metabolic and regulatory networks. Here, we first choose regulatory flux balance analysis (rFBA) to model the metabolic network with logic rules representing transcriptional regulations. Thus we can derive a bi-level optimization framework to design the metabolic network and corresponding regulatory network. This algorithm is similar to OptORF described above; the major difference is that instead of using embedded linear regulatory constraints to represent the regulation, logic rules indicating the bi-states of gene expressions are employed, which can be non-linear. Therefore, our model can avoid inaccurate predictions due to the non-linear structure of regulatory network. More details will be discussed in the following two sections.

5.2.1 Assumptions and Simplifications

Regulation of cell metabolism is complex, and both transcriptional regulation and post transcriptional regulation have strong effects. Developing accurate and reliable mathematic models for the these regulations is still a challenge in system biology. Boolean network and some other methods have been used to describe cellular regulation. For simplification, we use the discrete model, Boolean network, to represent complex interactions in cellular regulation. One basic assumption made in Boolean networks for cellular regulation is the states of all the components in the network, such as gene expression, enzyme activity, compound concentration and other extracellular signals, can be described by binary numbers (0 or 1). One reason for this discretization is ultrasensitivity of these systems(*Huang*, 2001). Under this assumption, gene expression is classified into two states, ON and OFF, as well as the existence of nutrients, extracellular signals and some other components in the network. Another assumption made in this Boolean network model is that interactions between components in the network can be described by logic relations, such as AND, OR, and NOT. By doing so, enzyme activity or gene expression can be predicted according to the states of extracellular signals, other genes, and enzymes through the corresponding logic rules. Figure 5.1 demonstrates how this Boolean network works.

To estimate the effects of cellular regulation on metabolism, predicted states of enzyme activities must be converted to regulatory constraints under the rFBA framework. As there are only two states of enzyme activity or gene expression, two states of a metabolic reaction can be modeled by the Boolean network, ON or OFF. If a metabolic reaction is indicated as ON by the regulatory network, it is able to

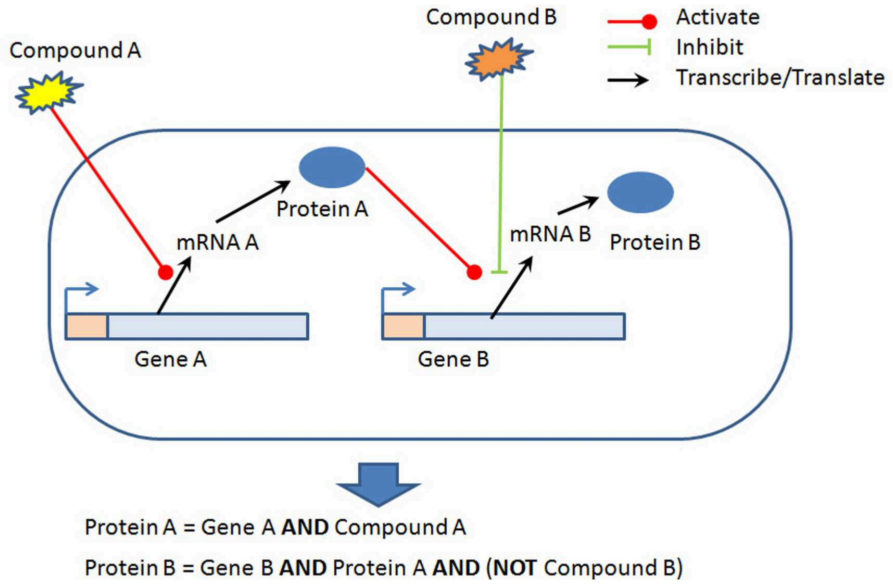

Compound A
Compound B
Activate
Inhibit
Transcribe/Translate
Protein A
mRNA A
mRNA B
Protein B
Gene A
Gene B
Protein A = Gene A **AND** Compound A
Protein B = Gene B **AND** Protein A **AND** (**NOT** Compound B)

Figure 5.1: Example of boolean network and logic rules

carry flux; and if indicated as OFF, which means the reaction is inhibited under this condition and its flux is set to zero. Therefore, a set of inhibited reactions predicted by the Boolean network can further constrain the solution space in flux balance analysis. However, certain intracellular signals (e.g. the flux or direction of one metabolic reaction) in the Boolean network cannot be provided until the later flux balance analysis is done. To solve this issue, iterations are employed in this framework. Thus, the states of these intracellular signals are initially set to hypothetical values. Then, after the Boolean network and flux balance analysis are solved, updated values based on flux predictions will be used. The iteration is stopped when there is no difference between the hypothetical values and updated values. This process is described in Figure 5.2.

The rFBA framework discussed above is able to provide a solid prediction of metabolism under cellular regulation, and has been applied to *Escherichia coli* to improve the phenotype predictions. Based on this we can further develop the algorithm for design of metabolic and regulatory networks. As mentioned before, the inhibited reaction set together with extracellular signals (nutrient information)

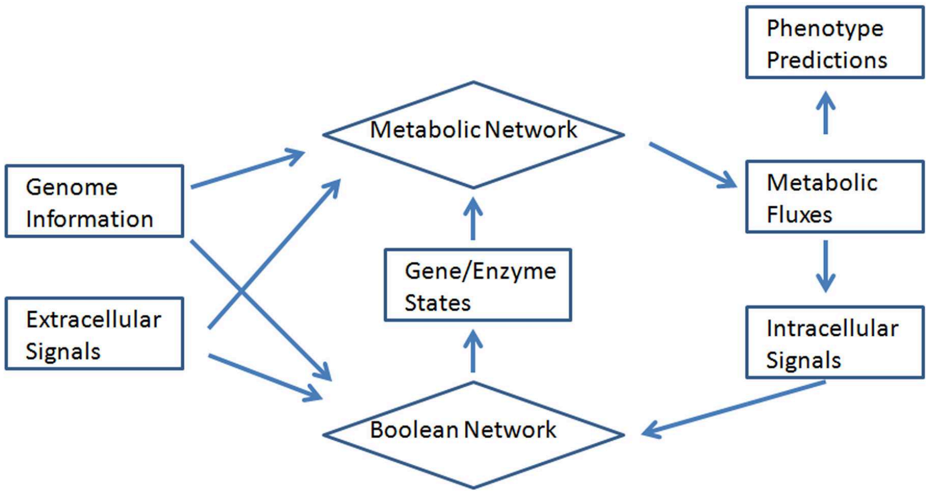


Figure 5.2: Flow chart of regulatory Flux Balance Analysis(rFBA)

is needed when predicting metabolic fluxes in rFBA framework. Thus, two types of modifications can be modeled under this framework, adding reactions into the non-active list and removing reactions from the non-active list. The biological implications for these two modifications are very clear. Adding reactions into the non-active list can be achieved by eliminating corresponding metabolic genes. Removing reactions from the non-active list can be achieved by either eliminating the transcriptional factors inhibiting these genes or changing the promoters of these genes so they are no longer inhibited in the desired condition. There is another assumption implied in this framework, that is, there are certain objective functions that the wildtype strain or modified strain are trying to accomplish. This assumption is necessary to make prediction of metabolic flux. Different objective functions have been used, for example maximization of growth (FBA) or ATP production, minimization of flux value, and minimization of metabolic adjustment (MOMA). In this framework, maximization of growth is selected as the objective function, which is commonly used for both wildtype strains and mutant strains.

5.2.2 Bi-level optimization Framework

We developed a framework for bi-level metabolic network and regulatory network design based on the assumptions and simplifications described in Section 5.2.1. There are two levels in the framework, which is demonstrated in Figure 5.3. The outer level of the model is optimization for desired property, which is maximization of product synthesis. The inner level of the model is the same as rFBA model, which is trying to determine the metabolic flux by assuming the cellular metabolism is optimized for growth, which is represented by the inner level objective function. There are several constraints that will define the solution spaces for all possible values of metabolic fluxes. The first one,

$$\sum_{r \in R} S(m,r) \cdot v(r) = 0 \tag{5.1}$$

is the mass balance constraint, which indicates no net changes in metabolites at steady state and is applied to all metabolites. The second type of constraints,

$$v(r) \leq F_{max} \cdot I_{reactive}(r) \tag{5.2}$$

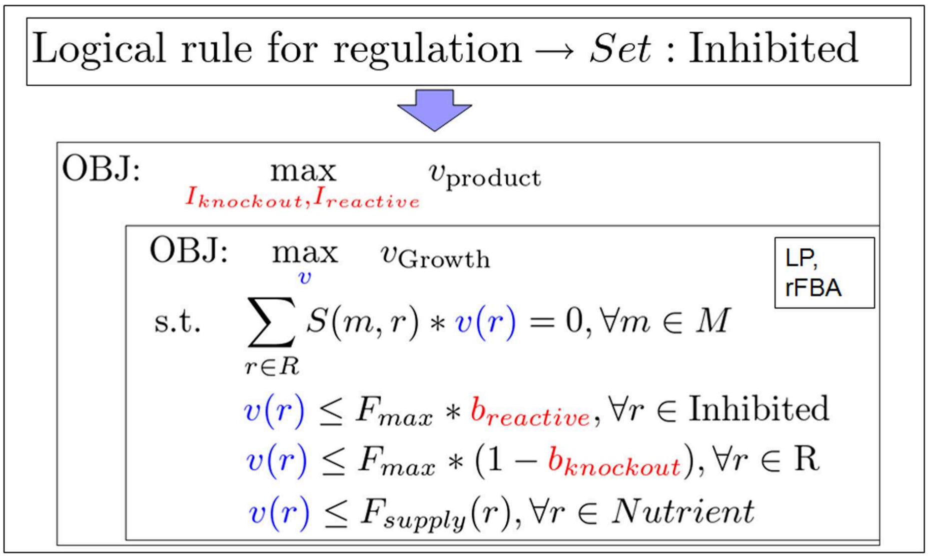
indicates inhibited reactions must carry zero fluxes unless they are re-activated, which will be reflected by $I_{reactive}(r) = 1$. Similarly, the third constraint,

$$v(r) \leq F_{max} \cdot (1 - I_{knockout}(r)) \tag{5.3}$$

indicates the reactions that are eliminated must carry zero fluxes, which will be reflected by $I_{knockout} = 0$. The last constraint,

$$v(r) \leq F_{supply}(r) \tag{5.4}$$

indicates all nutrients are supplied with certain amounts, which are limited by F_{supply}. There are a set of constraints that are not shown in Figure 5.3. One of them is constraint of reversibility, which forces irreversible reactions to carry positive flux ($v(r) \geq 0, \quad \forall r \in Irreversible$). Another constraint not shown in Figure 5.3 is the constraint for ATP maintenance, which sets the lower bound of ATP maintenance.



Logical rule for regulation → Set : Inhibited

OBJ: $\max\limits_{I_{knockout}, I_{reactive}} \quad v_{\text{product}}$

OBJ: $\max\limits_{v} \quad v_{\text{Growth}}$

$$\text{s.t.} \quad \sum_{r\in R} S(m,r) * v(r) = 0, \forall m \in M$$

$$v(r) \leq F_{max} * b_{reactive}, \forall r \in \text{Inhibited}$$

$$v(r) \leq F_{max} * (1 - b_{knockout}), \forall r \in \text{R}$$

$$v(r) \leq F_{supply}(r), \forall r \in Nutrient$$

Figure 5.3: Structure of bi-level metabolic network and regulatory network design model. R: all reactions, M: all metabolites, Nutrient: nutrient uptake reactions, Inhibited: inhibited reactions, v:variable of flux, v_{Growth}: growth rate, $v_{product}$: rate of product synthesis, $I_{reactive}$: binary variables of re-activate reactions, $I_{knockout}$: eliminated reactions, S: stoichiometric matrix, F_{max}: maximum flux value, F_{supply}: nutrient uptake rates.

To solve this bi-level mixed integer linear programming (MILP) problem, it is converted into a single level MILP problem based on duality theory. According to duality theory, if there is a feasible solution, the inner level linear programming problem (binary variables are treated as parameters in the inner level) can be replaced by the primal LP problem and its dual problem with another constraint that forces the two problems to have the same objective function values. There is one remaining issue to address before converting the inner level into a set of linear equations and linear inequalities which can be solved in MILP framework. When writing the dual problem for the primal problem, the binary variables $I_{knockout}$ and $I_{reactive}$ are treated as parameters, which is definitely true for the inner level. However, these binary variables must be solved in the outer level when the inner level is converted into primal and dual problems. Then, the products between variables in dual problem and these binary variables from outer level can make the whole problem non-linear. To solve this problem, the constraints used in the inner level must be rewritten. For the constraints for eliminating reactions (Equation 5.3), can be reformed into

$$v(r) = 0, \quad \forall r \in \{r | I_{knockout}(r) = 1\} \tag{5.5}$$

and similarly, constraints 5.2 can be reformed into

$$v(r) = 0, \quad \forall r \in \mathit{Inhibited} \cap \{r | I_{reaction}(r) = 0\} \tag{5.6}$$

After these transformations, the dual problem for the transformed primal problem can be written as

$$\sum_{m \in M} S(m,r) \cdot u_1(m) + A_{Inhibited}(r) \cdot u_2(r) + u_3(r) + A_{Nutrient}(r) \cdot u_4(r) \geq 0, \qquad \forall r \in \mathit{Irreversible}; \tag{5.7}$$

$$\sum_{m \in M} S(m,r) \cdot u_1(m) + A_{Inhibited}(r) \cdot u_2(r) + u_3(r) + A_{Nutrient}(r) \cdot u_4(r) = 0, \qquad \forall r \in \mathit{Reversible}; \tag{5.8}$$

$$u_2(r) \leq u_{max} \cdot (1 - I_{reactive}(r)), \quad \forall r \in Inhibited; \tag{5.9}$$

$$u_3(r) \leq u_{max} \cdot I_{knockout}(r), \quad \forall r \in R; \tag{5.10}$$

$$u_4(r) \geq 0, \quad \forall r \in Nutrient; \tag{5.11}$$

in which $Reversible$ is the set for reversible reactions, $Irreversible$ is the set for irreversible reactions, $u_1(m)$ is the dual variable for these mass balance constraints, $u_2(r)$ is the dual variable for these constraints of inhibited reactions, $u_3(r)$ is the dual variable for these constraints of eliminated reactions, and $u_4(r)$ is the dual variable for these constraints of nutrient uptake limits. Two arrays are introduced, $A_{Inhibited}$ and $A_{Inhibited}$. They are defined as below,

$$A_{Inhibited}(r) = \begin{cases} 0 & \forall r \notin Inhibited \\ 1 & \forall r \in Inhibited \end{cases} \tag{5.12}$$

and

$$A_{Nutrient}(r) = \begin{cases} 0 & \forall r \notin Nutrient \\ 1 & \forall r \in Nutrient \end{cases} \tag{5.13}$$

To enforce values of the two objective functions to be the same, a final constraint,

$$v_{Growth} = \sum_{r \in Nutrient} F_{supply}(r) * u_4(r) \tag{5.14}$$

is added also.

By replacing the inner level LP problem to these linear equations and linear inequalities, not including the transformed constraints for primal problem (Equation 5.5 and 5.6), the whole problem is now a common MILP problem and can be solved with conventional algorithm. In practice, two more constraints to limit total number of modifications can be added to avoid solutions with more genetic manipulations than expected, and these maximum numbers of modifications should be determined by both the scale of the two networks and difficulty of the problem.

5.3 Strain Network Design for Fatty Acids Derived Hydrocarbons

The bi-level framework for metabolic network and regulatory network design described in the last section enable us to optimize metabolic network and regulatory network for maximization of product rate. To apply this framework, metabolic network reconstruction and logic rules representing regulatory network are required. We choose *Escherichia coli* as the organism to over-produce fatty acid for bio-hydrocarbon production.

5.3.1 Metabolic and Regulatory Networks of *Escherichia coli*

There are several versions of metabolic network reconstructions for *Escherichia coli*, and some of them also include logic rules to represent cellular regulations. In this work, we start with a metabolic network reconstruction (*Covert and Palsson*, 2002) focused on central carbon metabolism. The logic rules of cellular regulation for this metabolic network also were generated and tested. However, fatty acid metabolism is not included in this metabolic network reconstruction, which is essential for our design. We added relevant metabolic reactions and metabolites (*Raetz*, 1978) into the metabolic network reconstruction. Further, we add the uptake of xylose to enable the prediction with xylose as carbon source. The final metabolic network is shown in Figure 5.4. Finally, there are 148 metabolic reactions and 25 exchange reactions involved in the metabolic network, and 57 logic rules in the regulatory network. The details about the metabolic network and logic rules used in this design model can be found in Appendix II.

From the Figure 5.4, we can find the metabolic network reconstruction is not

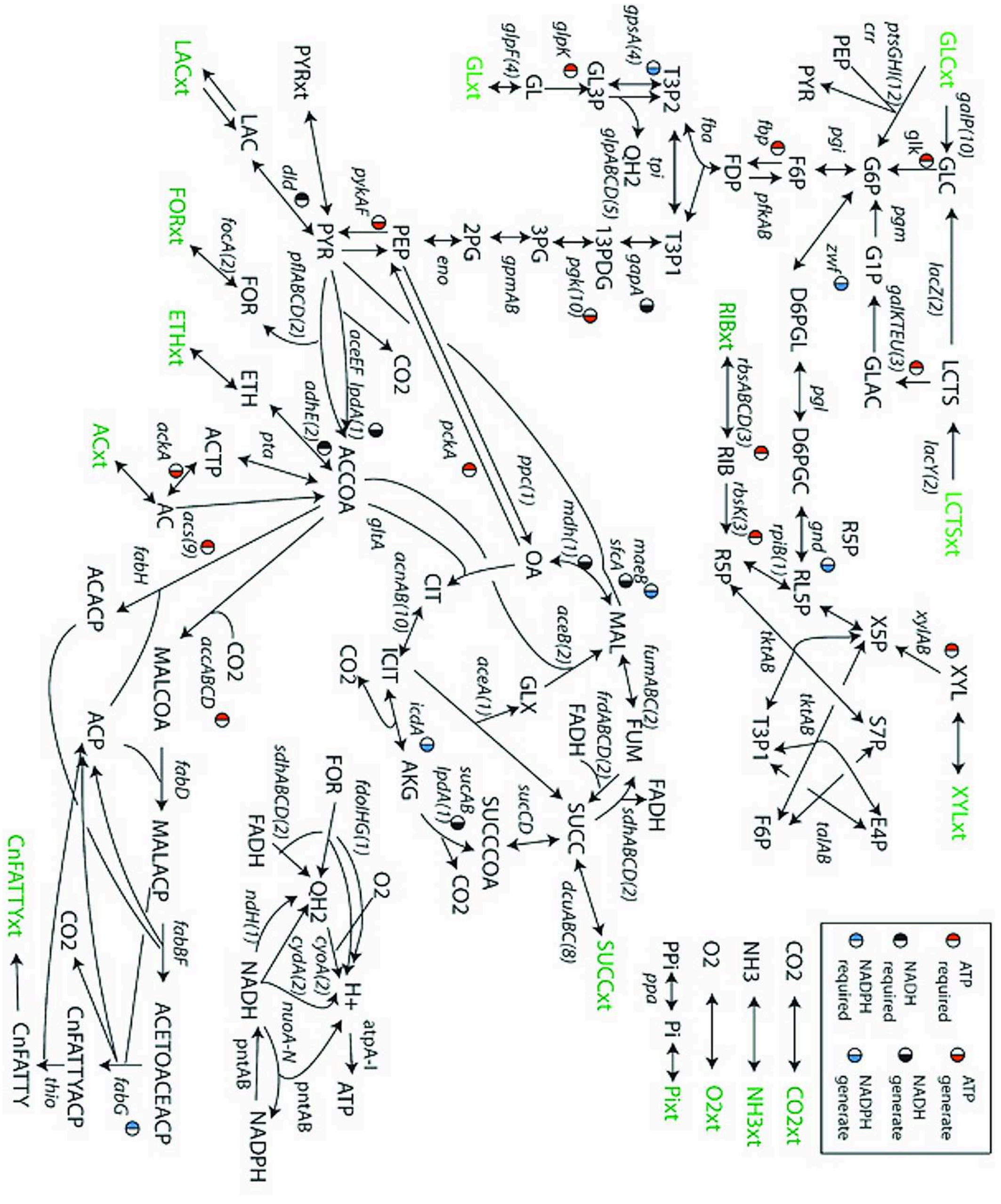


Figure 5.4: Metabolic network of *Escherichia coli* for the design model. The number behind the gene name is the number of transcription factors or signals regulating that reaction.

a genome scale, and there are some other genome scale models for *Escherichia coli* that also consider both metabolic network and regulatory network. One reason for us to choose this simplified model is the cellular regulation for carbon metabolism has been extensively studied compared to some other pathways. Thus, we are more confident to apply this design model on this simplified model to avoid errors from the raw data, so we can focus more on the methods and results. Because this simplified network does not include some essential pathways, e.g. amino acid metabolism, we need to pay more attention to the designs suggested by the model, and verify them experimentally if possible.

5.3.1.1 Optimal Design of Metabolic Network and Regulatory Network for Fatty Acids

We applied the design model of *Escherichia coli* to find best design for biofuel production. First, we looked for best design for fatty acids or triacylglycerols (TAGs) from D-glucose, which is a common nutrient used as carbon source. Second, we looked into the possibility of producing fatty acid from more economically feasible feedstock, e.g. glycerol, and conditions for mixed carbon sources. Finally we also applied the model to design strains for other valuable products, such as succinate and ethanol. To further study these designed strains, Dr. Fengming Lin [Lin et al., to be submitted] has been working on implement the designed strain for fatty acid/TAG production from D-glucose.

We first study the conditions with only D-glucose as carbon source in anaerobic and aerobic conditions. For comparison, we also tried to design different strains allowing only gene knockouts, only gene re-activation and both two types of modifications. We also constrained the possible modifications to avoid problematic designs. For example, the modifications of fatty acid synthesis cycles are excluded, as they are essential for cellular metabolism. Table 5.1 summarizes these results.

From the table, there are two major conclusions. First, applying modifications to both metabolic network and regulatory network can bring better performance than applying only one of them with an improvement of more than 25%, which is the reason we want to design a model enabling both changes. Second, aerobic conditions are preferred for fatty acids over production. This might be explained by the high energy requirement during fatty acid synthesis, and anaerobic condition is not favorable for energy production. We also found if not applying gene knockouts, but only re-activation of genes/reactions, no improvement can be made, which suggests re-activating genes/reactions only plays an auxiliary role. The growth rates for the designed strains are greatly reduced, which means cells put more resources into fatty acid metabolism that can be used for growth. The optimal products of fatty acids are different in aerobic and anaerobic conditions. In aerobic condition, octanoic acid (C8) is preferred product but in anaerobic condition stearic acid (C18) is chosen.

We are also interested in converting different feedstock into biofuels, for example glycerol which is much cheaper than glucose or other sugars. In addition, we want to explore the possibility of using more than one carbon source, and whether the mixed carbon sources can bring better efficiency for fatty acid production. Another reason for us to investigate mixed carbon source conditions is that hexose and pentose sugars always co-exist in lignocellulosic biomass, which is an attractive carbon source for biofuel production. Here we choose D-xylose to represent pentose sugars and D-glucose to represent hexose sugars. So it is interesting to investigate the fatty acid production with both D-xylose and D-glucose. Table5.2 lists the results of using glycerol, D-xylose and D-glucose or their combinations as carbon source for fatty acids production.

From the results, we found the improvement of introducing regulatory network modifications is more significant in mixed carbon sources conditions. This is different from what we observed in single carbon source supply conditions, and it seems in

Table 5.1: Summary of optimal designs for fatty acid production from D-glucose. The numbers of carbons in optimal fatty acid products are also listed. The results for wildtype strain are highlighted. Production rate is calculated based on 10mM Glucose/hr/g DCW.

Methods	Growth Condition	Objectives	Growth Rate (1/hr)	Fatty Acid Production (mM)*	Carbon Efficiency
Wildtype	Aerobic	Growth	0.95	0	0
Re-active Only	Aerobic	Fatty Acid	NA	NA	NA
Knockout Only	Aerobic	Fatty Acid	0.27	3.1 (C8)	41%
Re-active & Knockout	Aerobic	Fatty Acid	0.16	3.90 (C8)	52%
Wildtype	Anaerobic	Growth	0.33	0	0
Re-active Only	Anaerobic	Fatty Acid	NA	NA	NA
Knockout Only	Anaerobic	Fatty Acid	0.198	0.16 (C18)	4.7%
Re-active & Knockout	Anaerobic	Fatty Acid	0.135	0.57 (C18)	17%

*:Extra free fatty acid, and counted based on 10 mM Glucose.

mixed carbon sources supply conditions, re-design of regulatory network is necessary and important. These results can be explained by the effects of catabolic repression. In wildtype strain, catabolic repression will force the cell to utilize its favorite carbon source first even when there are other sources available. Before removing catabolite repression, the cell can only utilize one carbon source, Glucose in our cases, and the other carbon sources are wasted. However, after modifying regulatory network there is no catabolite repression, and the cell can make use of all the carbon sources provided. Interestingly, when glucose and glycerol are mixed and provided to the modified cell, the carbon efficiency is higher than any of the two if provided separately, which is unexpected. The details about the modifications suggested by the model for mixed glycerol and glucose condition is shown in Figure 5.5.

Table 5.2: Summary of optimal designs for fatty acid production from glycerol, glucose, xylose and their combinations in aerobic condition. The numbers of carbons in optimal fatty acid products are also listed. The results for wildtype strain are highlighted. Production rate is calculated based on 10mM Glucose/hr/g DCW.

Methods	Carbon Source *	Objective	Growth Rate (1/hr)	Fatty Acid Production (mM)	Carbon Efficiency
Wildtype	Glycerol	Growth	1.1	0	0
Knockout Only	Glycerol	Fatty Acids	0.375	3.25 (C8)	43%
Re-active & Knockout	Glycerol	Fatty Acids	0.375	3.25 (C8)	43%
Wildtype	Glycerol & Glucose	Growth	0.47	0	0
Knockout Only	Glycerol & Glucose	Fatty Acids	0.13	1.56 (C8)	21%
Re-active & Knockout	Glycerol & Glucose	Fatty Acids	0.122	4.95 (C8)	66%
Wildtype	Xylose & Glucose	Growth	0.47	0	0
Knockout Only	Xylose & Glucose	Fatty Acids	0.180	1.56 (C8)	21%
Re-active & Knockout	Glycerol & Glucose	Fatty Acids	0.109	2.37 (C8)	34%

* All carbon sources are supplied with equal carbon amount as 60 mM/hr/gDCW. When mixed together, two carbon sources have the same carbon amount.

5.3.1.2 Optimal Design of Metabolic and Regulatory Network for Other Products

We applied the same model to *Escherichia coli* to design metabolic network and regulatory network for some other products, including ethanol, succinate, lactate and pyruvate. Table 5.3 summarizes the optimal production rates for these products from glucose in both anaerobic and aerobic conditions.

As shown in Table 5.3, ethanol, succinate and lactate can be synthesized by designed strains in both aerobic and anaerobic conditions but with higher yield in anaerobic condition, while pyruvate can be produced better in the aerobic condition. Further, by modifying regulatory network and metabolic network together, products

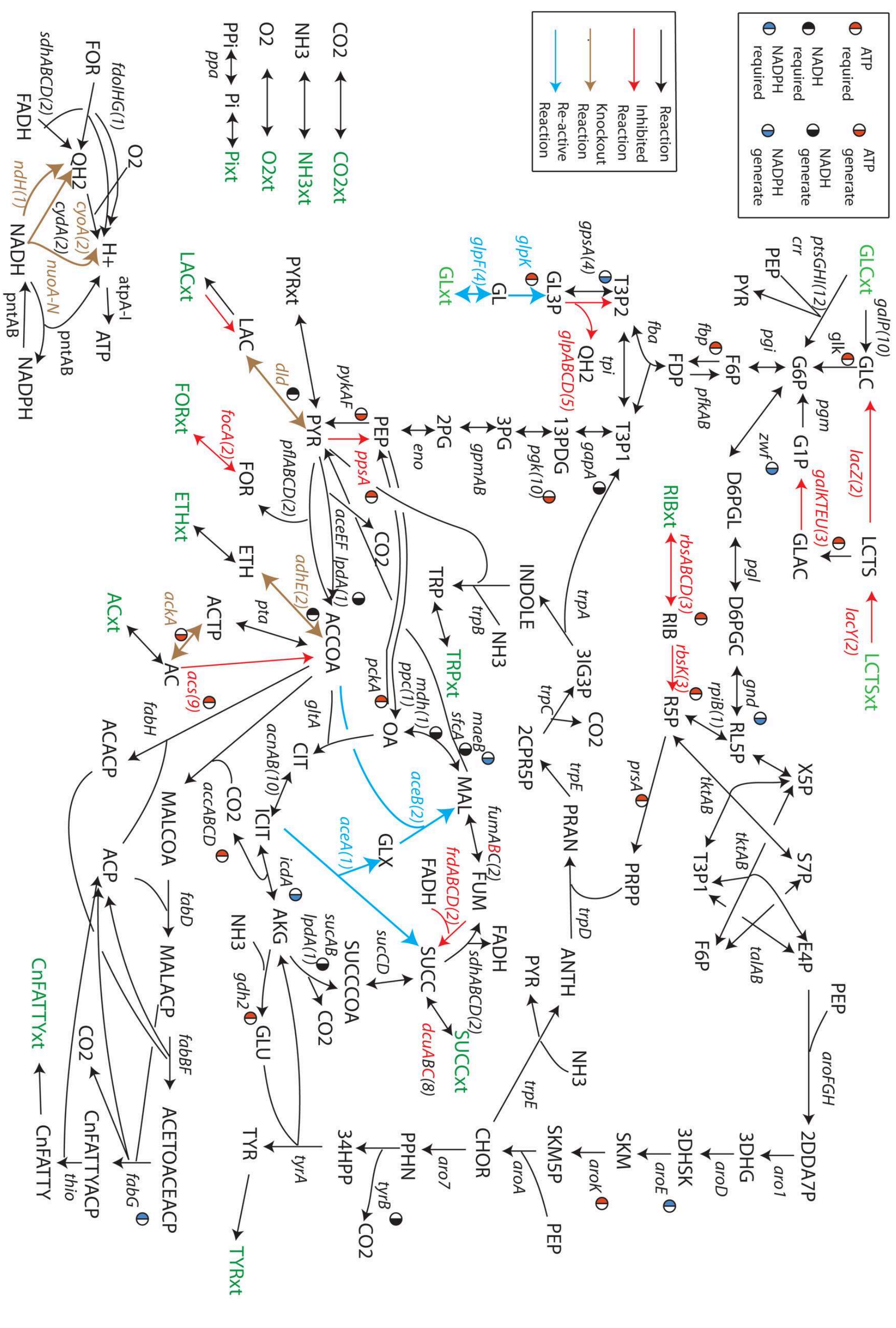


Figure 5.5: Designed strain for fatty acids production from glucose and glycerol mixed supply.

that cannot be produced in certain conditions by only applying gene knockouts can be synthesized, even though the conditions are not ideal for them.

Researchers have tried to produce succinate in aerobic conditions. Henry Lin

Product	Condition	Knockout Only	Re-active Only	Re-active & Knockout
Ethanol	Aerobic	0	0.2	8.81
Ethanol	Anaerobic	9.76	7.80	16.80
Succinate	Aerobic	0	3.03	5.88
Succinate	Anaerobic	0	0	7.37
Lactate	Aerobic	7.13	0.95	9.13
Lactate	Anaerobic	11.54	0	16.66
Pyruvate	Aerobic	12.44	0	12.44
Pyruvate	Anaerobic	5.73	0	8.32

Table 5.3: Optimal production rate for designed strain from D-glucose. Production rate is calculated based on 10mM Glucose/hr/g DCW.

and his group members applied a very similar genetic manipulation suggested by our model to *Escherichia coli* and successfully produced succinate with 45% to 70% of theoretical yield. Compared to our model, the suggested optimal design has a yield of 58% of theoretical yield. We also calculated the yield for the strain used in Lin's work, which is 52% of theoretical yield. The details about these two designs can be found in Figure 5.6. It is clear that our model prediction is in agreement with the published experimental data. Thus we also expect the optimal design suggested by the model to have similar or even better performance.

5.3.2 Experimental Verification of Designed Strain for Fatty Acid Production

In Section 5.3.1.1 we described several designed *Escherichia coli* strains for fatty acid production in different conditions utilizing different feedstock. We are interested in experimentally implementing some of these designs and testing their performance. Dr. Fengming Lin in our lab was working to develop an *Escherichia coli* strain to produce triacylglycerols(TAGs), which also requires over-production of fatty acids to enhance the TAGs production. Therefore, the strain designed for production of fatty acids from D-glucose on aerobic conditions is selected. To derive the designed strain, there are 6 metabolic reactions need to be eliminated and 2

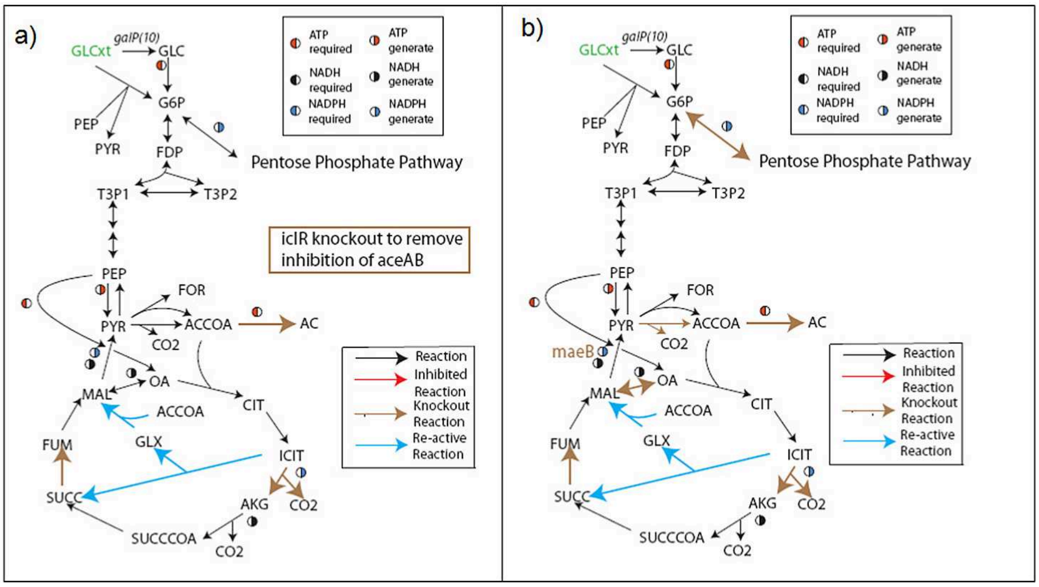


Figure 5.6: Model suggested Strain and Published Strain for Aerobic Succinate Production. a) Left is published strain for aerobic succinate production with 45% to 70% of theoretical yield. Our model predicts this strain with 52% of theoretical yield. b) Right is the optimal strain suggested by out model, with 58% of theoretical yield.

reactions to be re-activated from the wildtype strain. To eliminate the 6 metabolic reactions, cyoA(subunit II of the cytochrome bo terminal oxidase complex encoded by cyoABCDE); nuoA(part of the inner membrane component of NADH dehydrogenase I); ndh(NADH dehydrogenase II); adhE(alcohol dehydronase); pta(Phosphate acetyltransferase); dld and ldh(D-lactate dehydrogenase) need to be removed from the chromosome. To re-activate the two glyoxylate bypass reactions, aceA and aceB, genes iclR and icdA should be removed to stop their inhibition on glyoxylate bypass operon (aceBAK) (*Lee et al.*, 2009; *Lin et al.*, 2005).

We tried to understand the mechanism and effects of these modifications suggested by the computational model. Four of them are simply removal of competing pathways, including eliminating genes of pta, adhE, dld and adhE. The purpose of this type of modification is straightforward, that is to drive most of the carbon and energy to fatty acid production. The second type of modifications in the designed strain is reaction re-activation, including iclR and icdA gene knockouts that can promote glyoxylate bypass pathways. When glyoxylate bypass pathways is active, less NADH(NADPH) but more malate and succinate are produced through tricarboxylic acid (TCA) cycle. Research indicates these effects will lead to a increase of organic acid production (*Meijer and Otero* 2009). Another type of modification suggested by the design model is change of aerobic respiration, including gene knockouts of cyoA, nuoA and ndh. The purpose of these modifications might be to change the ratio between NADPH and NADH production and ATP generation as well. Fatty acid synthesis pathway requires both NADPH and NADH, and is the major consumer of NADPH in cell metabolism. Increase of NADPH production level may be beneficial for fatty acid production.

The design model in Section 5.3.1 provides us all these modifications on metabolic network, and we also can study their effects in different combinations. In addition, we are able to determine the best sequence of manipulations, in which the genetic manipulation with largest effect is chosen at each step. The predicted sequence of manipulations is ΔcyoA, Δadh, ΔnuoA, Δndh, Δdld and Δldh, Δpta, and ΔiclR and ΔicdA. Using two different algorithms, flux balance analysis (FBA) and minimization of metabolic adjustment (MOMA), we predict the phenotypes after each step of modification, so we can compare the model predictions along this path with the experimental results. Figure 5.7 demonstrates the predictions along the optimal sequence.

From the figure, we found the two algorithms, MOMA and FBA, give different predictions for the first four modifications. FBA predicts no extra fatty acid production before the fifth manipulation is introduced, while MOMA predicts the fatty acid production increases gradually. Further, because of the mechanism of MOMA, no improvement can be predicted for re-activated reactions.

Dr. Fengming Lin carried out most of the designed manipulations suggested by the model, along the sequence of ΔcyoA, Δadh, ΔnuoA, Δndh, Δpta, Δdld and

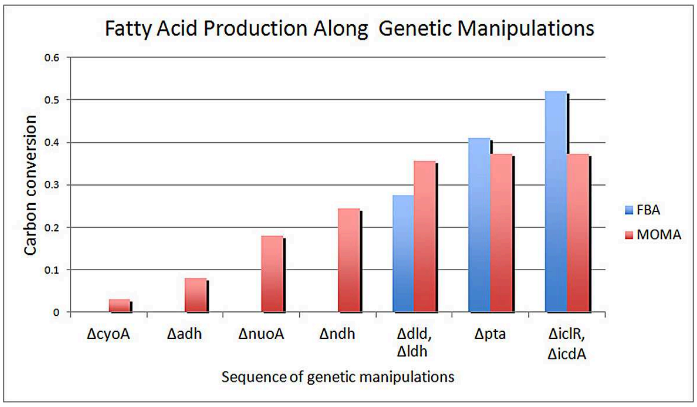


Figure 5.7: Model predictions of fatty acid production along optimal sequence of genetic manipulations. The results predicted by FBA and MOMA methods are shown in the optimal sequences calculated by the models.

ΔiclR. This sequence is slightly different from the optimal sequence predict by the model but introduces Δpta before Δdld and excludes Δldh and ΔicdA. Figure 5.8 demonstrates the experimental results after applying each manipulation in M9 growth medium.

From these results, we found the production of fatty acid increases gradually, but the most significant increase takes place in the fifth step, indicating a state of metabolism between MOMA prediction and FBA prediction. Because the strains in Dr. Fengming Lin's work are different from the model suggested after the fifth steps, we cannot make further conclusion about the sequential predictions. In the designed fifth step, competing pathway of lactate synthesis is removed by Δdld and Δldh. In the experiments, ldh is still left, which leads to a high production rate of lactate and decrease in fatty acid production.

Cell growth rate is also measured during the process, which is reduced significantly as the model predicted. The compositions of fatty acids are shown in the Figure 5.8, which do not change significantly along the modifications. Dr. Fengming Lin also cultivated the strains in another two growth mediums, LB 1-2 and LB 5-10. The results of these experiments are shown in Figure5.9. Comparing the results with only reaction eliminations to the results with both reactions elimination and re-

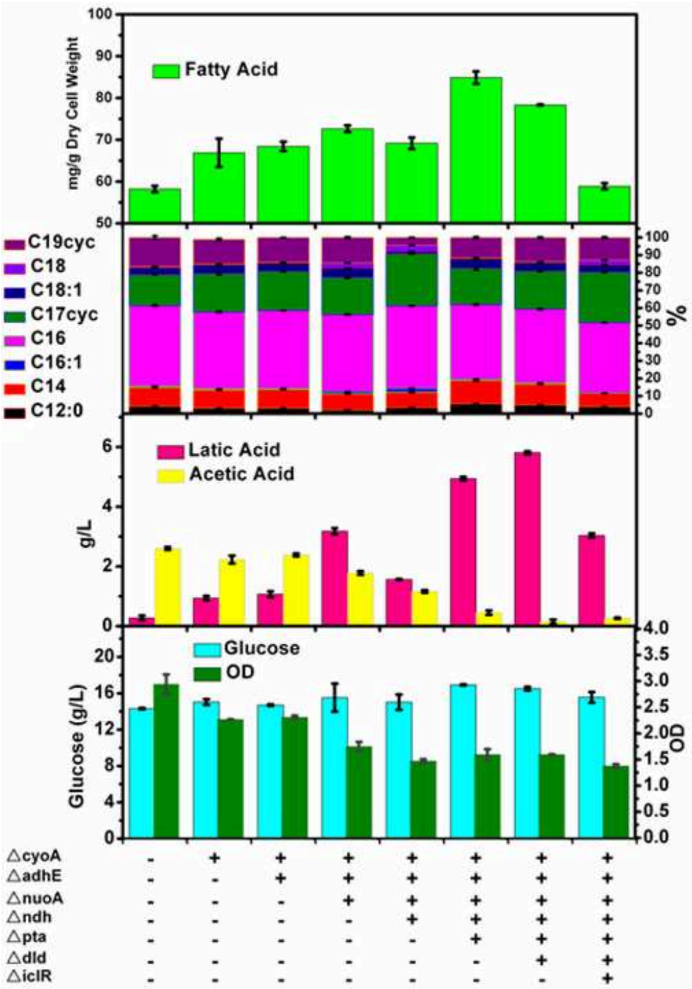


Figure 5.8: Profiling of gene modified strains based on the model prediction. The strains were cultured in M9 minimal medium with 2% glucose for 48 hrs. The total amount of fatty acids was quantified by GC-FID as well as the fatty acid composition was identified. The other fermentation parameters were determined, including the final concentration of by-products lactate and acetate, the growth (OD), and the final glucose concentration.

activations, we found the reaction re-activation has strong positive effects on fatty acid production only when cultivated in LB 5-10. In contrast, re-activating the glyoxylate bypass reactions will decrease the fatty acid production in LB 1-2 medium. The mechanism causing these ambiguous results is still unclear. One hypothesis is that the designed strain does not grow healthily compared to wildtype strain due to the gene modifications, which is indicated by the slower growth. LB 5-10 medium has a higher glucose concentration (5%) than LB 1-2(1%) medium, so the sick cells may prefer a richer medium.

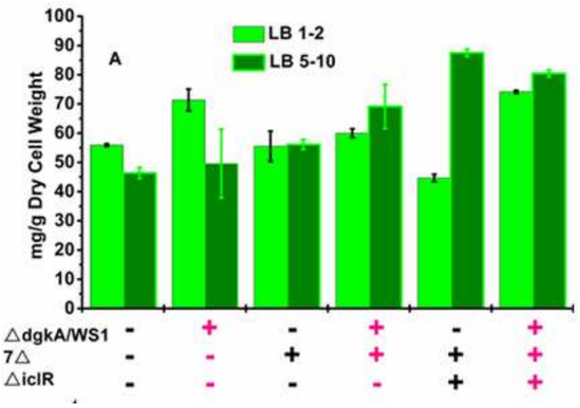


Figure 5.9: Total amount of fatty acid of strains suggested by the optimization model. Two culture mediums LB 1-2 and LB 5-10 were tested. All strains were cultured at 30?? for 48 hrs. 7Δ represents strain with ΔcyoA, Δadh, ΔnuoA, Δndh, Δpta, Δpta.

5.4 Discussion and Conclusions

As mentioned before, the metabolic network and regulatory network used in this work is not a genome-wide model, because the quality of regulatory networks in genome scale might be not as good as those in central carbon metabolism. However, researchers are getting more and more data to generate more accurate genome-wide models for cellular regulation. We expect there should be genome-wide metabolic and regulatory model for model organisms with desired quality in the near future. Then all prediction models and design models relying on the genome-wide models can be further improved.

Different to some existing design methods, such as OptORF, our framework does not linearized the cellular regulation but use the logic rules to represent these complex interactions. Due to the non-linearity of some logic rules, the regulatory network is no longer able to be solved within mixed integer linear programming (MILP) framework. Therefore, this solving process is done iteratively in our model. The regulatory network is solved separately before the optimization model for strain design, and then this process repeats. Once the iteration converges to a stable state, we can make a reliable approximation about the effects of regulatory network on the metabolic network. One limitation for our model and all other models based on MILP framework is the objective function must also be linear. This indicates the model in the inner level of the bi-level model can only using linear functions as its objective function. Thus, some model, such as Minimization of Metabolic Adjustment (MOMA), cannot be utilized in this framework. Currently, there are only limited objective functions for constraint based metabolic models that are widely use, and finding a proper objective function for a specific system is still a challenge. In this work, we choose the commonly used objective function, maximization for growth, which can be used for both wildtype strains and mutated strains.

According to the predicted results, we found re-activated metabolic reactions will have strong effects only in those conditions that cellular regulation controls cellular metabolism strongly, which is the same as we expected. This means our framework for strain design will give more powerful predictions in those cases that either the desired products are unfavorable in the conditions or the cells are cultivated in a condition different from its common growth environments. For example, producing succinate aerobically belongs to the first class and cultivation of *Escherichia coli* using mixed glucose and glycerol belongs to the latter one.

In summary, we discussed the algorithm to design metabolic networks together with regulatory networks and implement this method in *Escherichia coli* for fatty acid production. We first introduced a bi-level optimization computational framework enabling the design for both two networks. Then we applied this framework to a metabolic and regulatory network of *Escherichia coli* which mainly considers carbon metabolism. We used this model to design strains for fatty acids production in both aerobic and anaerobic conditions. Different carbon sources were examined as well as their combinations. Results indicate that modifications of regulatory network can bring further improvement on fatty acid production, especially in mixed carbon sources conditions. The possible reason for this significant increase of fatty acid production is caused by the removal of catabolite repression, which regulates cellular carbon utilization. We also found when mixing glucose and glycerol, a high yield of fatty acid can be achieved, which is even higher than any of the two cases they are provided individually. This result indicates that there are synergetic interactions for fatty acid production from glucose and glycerol, so when cells utilize both of them, an extra benefit besides of removal of catabolite repression can be achieved. We also designed strains for some other products, including succinate and ethanol. The designed strain for aerobic succinate production is very similar to one of the designs in literature, and the predicted yield is within the experimental range. Dr. Fengming Lin carried out part of the manipulations for fatty acid production designed by this method. By comparing the results along the manipulations, we found the experimental results stand between the prediction of FBA method and MOMA method, which are calculated along the optimal sequence of manipulations. The fatty acid production in M9 medium is partly in agreement with model predictions. An improvement of fatty acid production by re-activating two glyoxylate bypass reactions was observed in the experiments using LB 5-10 as growth medium but not in LB 1-2. In future, the remaining modifications would be carried out and we expect the final strain can produce more fatty acids than current engineered strains.

CHAPTER VI

Concluding Remarks and Future Directions

6.1 Concluding Remarks

In this dissertation, we demonstrated strategies for automatically reconstructing metabolic network for microorganisms by integrating their genomes, transcriptomes and proteomes. We further introduced multiple-organism models to utilize these metabolic networks for community-wide metabolic network modeling. These community-wide metabolic network models were able to describe the cellular metabolism, biosynthesis potentials, as well as interspecies interactions for microbial communities. Besides metabolic modeling, a bi-level strain design model was developed to optimize metabolic networks and regulatory networks of microorganisms for production of biofuel molecules.

In Chapter II, we first developed a bioinformatic pipeline for genome-wide metabolic network reconstruction (PEER), which can automatically reconstruct high-quality genome-wide metabolic networks from annotated genomes. This bioinformatic pipeline was tested with a model organism *E. coli.* By comparing the metabolic network reconstructed by this automated bioinformatic pipeline with the manual curated reference metabolic networks, we demonstrated PEER can provide high-quality and complete genome-wide metabolic networks. The PEER was further applied to twelve strains of *P. marinus* to generate pan and core metabolic networks of the species. We found the metabolic networks of the twelve strains can be classified into two groups, which are consistent with the two ecotypes of the strains. Therefore, we can define the metabolic functions and biosynthesis capabilities that are essential for all the strains of *P. marinus*, as well as the diversities of metabolism across the species. By mapping the variable part of metabolic networks of the twelve strains, we identified several factors that differentiate the metabolism of strains. The dominant factor, light, shaped the metabolic networks of low-light-adapted strains, which are enriched with two reactions involved in citric-acid cycle (CAC) converting 2-oxoglutarate to succinyl-CoA. This observation suggest that the low-light-adapted *P. marinus* are not obligate autotrophs, which agrees with experimental data. Sulfur source is another factor identified in our work that may differentiate the metabolic networks of the two ecotypes. In addition, we found phosphorus source and nitrogen source did not affect the metabolic networks of the two ecotypes, even though different types phosphorus sources and nitrogen sources were available in the two environments. All these results demonstrate the extra benefits of reconstructing pan and core metabolic networks for one species, which can only be achieved by high-throughput automated metabolic network reconstruction tools.

After the bioinformatic pipeline for genome-wide metabolic network reconstruction (PEER) was developed, we could start to reconstruct metabolic networks for all the organisms identified in microbial communities. The PEER was applied to Acid Mine Drainage (AMD) biofilm in Chapter III to model the AMD biofilm and identify essential interactions associated with the biofilm formation. The genome-scale metabolic networks of the five major organisms were reconstructed, providing mechanisms for certain essential metabolisms in the biofilm, such as nitrogen fixation, carbon fixation and biomass synthesis. In addition to the individual metabolic networks, we developed community-wide metabolic networks using multi-organism models. These models considered both intracellular metabolism and interspecies interactions simultaneously. According to the prediction, several essential interactions were identified, including cross-feeding like interactions of amino acids and ammonia. These interactions provided us potential treatments for AMD pollution by blocking these interactions that are essential for the biofilm formation. We incorporated the proteomic datasets with the reconstructed metabolic network to further refine the metabolic networks. 88.4% and 75.7% active reactions for the two dominant organisms were identified in the proteome, which were significantly higher than the proteome coverage (p-value $< 10^{-200}$). This enrichment of active reactions not only verified the reconstructed metabolic networks but also indicated that growth is the major objective for this biofilm.

Gut microbiomes are important host-related microbial communities that directly relevant to health issues. We were interested in reconstructing community-wide metabolic networks and identifying interspecies interactions to reveal the mechanisms for the metabolism and relationship of the species inside the gut microbiome. In Chapter IV, we applied two three-step metabolic modeling procedures to two model gut microbial communities. By integrating growth phenotypic data, genechip-based transcriptomes, and annotated genomes, we reconstructed the metabolic networks for the two species in the two-species microbiome. For the two-species model gut microbiome, the reconstructed metabolic networks can predict the growth test results with 93% accuracy and agree with 63% to 73% of the transcriptome. The community-wide model predicted *B. thetaiotaomicron* provided pantothenate to *E. rectale* and *E. rectale* produced stachyose for *B. thetaiotaomicron* which can be verified by experimental data. We reconstructed the metabolic networks for a more comprehensive model gut microbial community which includes ten species. Ten-species model was developed by a different three-step modeling procedure that can utilize sequence-based transcriptomes. More than 60% of metabolic reactions identified in the transcriptome was predicted to be active in the model. By modeling the interactions, the model predicted the organisms that required extrageneous sources of certain metabolites as well as the organisms synthesized them. Urea, citrate and agmatine were found in the list of these interacted metabolites, and experimental evidences were found to support these prediction.

In addition to reconstructing and modeling metabolic networks, we developed a bi-level optimization based framework for strain design, which considering both modifications of cellular metabolism and gene regulation. This method was used to predict gene manipulations for some biofuel products including fatty acids, succinate, and ethanol. From the model predictions, we confirmed that optimizing both metabolic network and regulatory network can provide higher productivities than only revising one of them, and the improvements were more significant for the cases that either the products were not the favorite products of the cell, or the growth conditions were far from the environments the organism evolved with. Part of the design for fatty acid over-produced *E. coli* has already been implemented in experiment by Dr. Fengming Lin.

From these results, we demonstrated the possibility of automatically reconstructing genome-scale and community-wide metabolic network. We generated the community-wide metabolic networks of three model microbial communities and explored the potential of utilizing metabolic modeling methods to predict not only cellular metabolism but also inter-species interactions. Despite of challenges, we believe the bioinformatic pipeline for automated metabolic network reconstruction we developed is a powerful tool for generating (meta)genome-scale metabolic networks. These complex metabolic networks contain comprehensive information about both cellular metabolism and the interactions with environments. Through the studies of two model gut microbiota, we introduced strategies of integrating growth phenotypic data and gene expression data with metabolic models. This integration of information from multiple levels provides us more comprehensive knowledge about these complex biological systems. Therefore, community-wide metabolic modeling is a promising method to study the metabolic functions and ecological roles of microorganisms in microbial communities, and a powerful tool to analyze and integrate large-scale culture-independent data.

6.2 Future Directions

In this dissertation, we demonstrated that reconstructing metabolic networks from metagenomes or data collected by other cultivation-independent methods is a promising tool for studying the intracellular metabolism and interspecies interactions of microbial communities. However, there are several challenges during this reconstruction process. First, the metagenomic sequences are less comprehensive than genomic sequences collected from single organisms. *Mavromatis et al.*(2007) predicted that about 20% of all the genes in a dominant microorganism co-existing with others in a community could not be identified from metagenomic sequences. The missing of these genes causes more metabolic gaps in the reconstructed metabolic networks. To fill these gaps, more putative reactions are needed, which increases the uncertainty of the final metabolic network. This increase of putative reactions can be observed by comparing the metabolic networks of AMD biofilm (Chapter IV) with the metabolic networks for *P. marinus* (Chapter II). Therefore, to reconstruct high-quality community-wide metabolic networks, accurate metabolic gap filling and gene candidate identification methods are required.

The inter-species interactions are important properties in microbial communities. However, these interactions have not been comprehensively characterized for most microbial communities. To elucidate these interactions, we reconstructed the metabolic networks together with inter-species interactions during our community-wide metabolic modeling. This method can generate informative predictions about these interactions, but largely relies on metabolic gap filling and transporter prediction. Therefore, to accurately predict the inter-species interactions in microbial communities, transporter predictions must be improved, especially for the specificity of the exchange reactions.

The complexity of microbial communities further increases the difficulty of community-wide metabolic networks modeling. Because of the inter-species interactions, the metabolic networks of member organisms must be reconstructed simultaneously, which can be demonstrated by comparing the metabolic networks of individual organisms with the five-species metabolic network for the AMD biofilm (Chapter III). As a result, the computational requirement will grow exponentially with the number of organisms in the microbial community. Therefore, simplifications that can reduce the complexity of the system but will not overly reduce the accuracy and predictive power are required for these large-scale microbial systems.

In this section, I will discuss several potential strategies for solving some of these challenges, including utilizing more gene annotation/classification methods to improve gap filling, utilizing multiple culture-independent data to improve the community-wide model, and reconstructing metabolic networks for phylogenetic groups to reduce the complexity. By introducing these methods, we expect to further extend our metabolic modeling framework and its application to a wider range of microbial communities. For example, one day we might be able to create personalized gut microbe signatures using these methods, which could either diagnose related diseases or guide our behavior in everyday life.

6.2.1 Future Directions for Metabolic Network Reconstruction and Modeling

6.2.1.1 Incorpration of Gene Annotation Methods

Including PEER framework, there are only several tools developed to automatically reconstruct (meta)genome-wide metabolic networks. The desires of high-quality metabolic networks drive us to keep improving these tools to overcome several limitations. One of the challenges is to predict and annotate gene accurately. The electronic annotation methods have been extensively studied and different mapping algorithms were developed. Currently, BLAST against COG or KEGG datasets is the most commonly used method for metagenomic data. In PEER, BLAST against KEGG datasets was the only method used to identify gene candidates for metabolic gaps. This process can be further improved by introducing other gene annotation methods. Table 6.1 lists several potential methods or databases that may work with PEER.

Gene Classification Systems and Databases	Classification Methods	Reference
HAMAP families	HAMAP−Scan	*Lima et al.*, 2009
Pfam protein domains	HMMER3	*Punta et al.*, 2012
InterPro protein families, domains and functional sites	InterProScan	*Hunter et al.*, 2009
Clusters of Orthologous Groups (COG)	BLAST	*Tatusov et al.*, 2003

Table 6.1: Potential Gene Classification Methods and Datasets for PEER.

The listed annotation and classification methods utilize different algorithms from different classification aspects to provide systematic gene annotation and classification. By mapping these classification systems and databases to the metabolic reactions we collected, we will be able to predict the gene-reaction associations more accurately than merely considering KEGG datasets. Therefore, we can improve the metabolic network reconstruction by incorporating these methods when searching for gene candidates for metabolic gaps.

6.2.1.2 Incorpration of More -Omics data

In the Chapter III and IV, we described methods that incorporate either transcriptomic or proteomic datasets during reconstruction of metabolic networks of Acid Mine Drainage (AMD) biofilm and model gut microbial communities. Those results indicate the advantage of integrating either gene expression or protein identification with metabolic network reconstruction process based on genome annotation and sequence. Therefore, considering both transcriptomic and proteomic data might bring us to a better understanding of the metabolism of cells. Furthermore, metabolomic data are becoming more and more available for many biological systems, which are another potential source of information about these microorganisms.

There are several challenges in integrating different -omics data. First, these -omics data always have different coverages. Therefore, only part of the metabolic network can be connected with all of these data. We need to develop methods to avoid the bias caused by the different coverages of these -omics datasets. Another challenge of integrating -omics data is the conflicts between different -omics data. Specific strategies are needed to resolve these conflicts according to the properties of these experimental data. For example, genechip based expression data normally have higher coverages than sequenced-based expression data but less accurate. Therefore, different strategies are needed to resolve the conflicts caused by the two types of gene expression data. If we can overcome these challenges and integrate all the -omics data in reconstructing process, we will generate much more complete and accurate metabolic networks, which can bring us to a better understanding of cellular metabolism and interactions.

6.2.2 Metabolic Network Modeling of Real Human Gut Microbiota

In Chapter IV, we demonstrated two case studies on model gut microbial communities utilizing metabolic reconstruction methods. These model microbial communities were designed to study some specific questions about the natural gut microbiota, for example, the microbial structure changes in response to diet. However, the structure of natural gut microbiota is much more complex. Due to this complexity, assembling the metagenomic sequences is still challenge. Therefore, to associate the gene sequences with corresponding organisms, binning methods are more frequently applied to these metagenomic datasets. Because of the limitation of the binning methods, only incomplete genomes for organisms can be retrieved. The incomplete genomes may not provide comprehensive information to reconstruct the genome-scale metabolic network of all the individual organisms. To overcome this shortage, we can reconstruct metabolic networks either for the dominant organisms or for phylogenetic groups. However, due to variations and complexities of the data, the required reconstruction strategies will be different from those used for single organism or model communities, and we will discuss the possibility to utilize metabolic network reconstruction methods in studying natural gut microbial communities.

6.2.2.1 Preliminary Results

There are a number of metagenomic datasets for natural human gut microbial communities. Here we introduce one collected by *Qin et al.* (2010). As part of MetaHIT (Metagenimics of Human Intestinal Tract) project, this 576.7 Gb dataset contains DNA sequences from 124 European adults. 3,299,822 non-redundant genes were mapped to 89 frequent reference microbial genomes to generate a gene catalogue of the human gut microbiome. According to the works, only 18 bacteria have been found in all the 124 human faecal samples, while 57 were found in 90% of the human faecal samples and 155 were found in at least 1 human fecal sample. Therefore, sample-wise variation is significant across the dataset, which cannot be neglected when reconstructing metabolic networks for core human gut microbiome. Even though the 124 individuals can be classified into healthy people, people with ulcerative colitis (UC) and people with Crohn's disease (CD), we did not observe clear separation of the gut microbiomes from the three groups, which can be explained partly by the sample-wise variations.

Provided with this comprehensive metagenomic dataset, we are capable of predicting the core and pan metabolic networks of human gut microbiome from the metabolic networks reconstructed from 124 individuals. Utilizing the same method described in Section 2.2.1, we reconstructed the draft metabolic networks from the metagenomes of 124 individuals. In total, 3261 metabolic reactions were identified for the gut microbiomes from 124 people. To identify the core metabolic network for human gut microbiome, we set the cutoff value of 60 individuals to select core metabolic reactions. There were 1777 metabolic reactions found in more than 60 human faecal samples, which were considered as core metabolic reactions.

Figure 6.1a lists the products according to the model predictions with different levels of penalty for metabolic gaps. Some common fermentation products can be produced without any gap filling, including acetate, ethanol and propanoate. Several secondary metabolites also belong to this category, such as nicotinate and indole. Riboflavin, ascorbate, and taurine require significantly more metabolic gaps if they are produced.

The active reactions for synthesizing these products were listed in Figure 6.1b. From the results, the more products lead to more active reactions, as well as metabolic gaps. This result also indicates that the last several products in the product list are mainly synthesized by putative reactions. Therefore, we shall not keep decreasing the penalty levels for metabolic gaps, which will make the

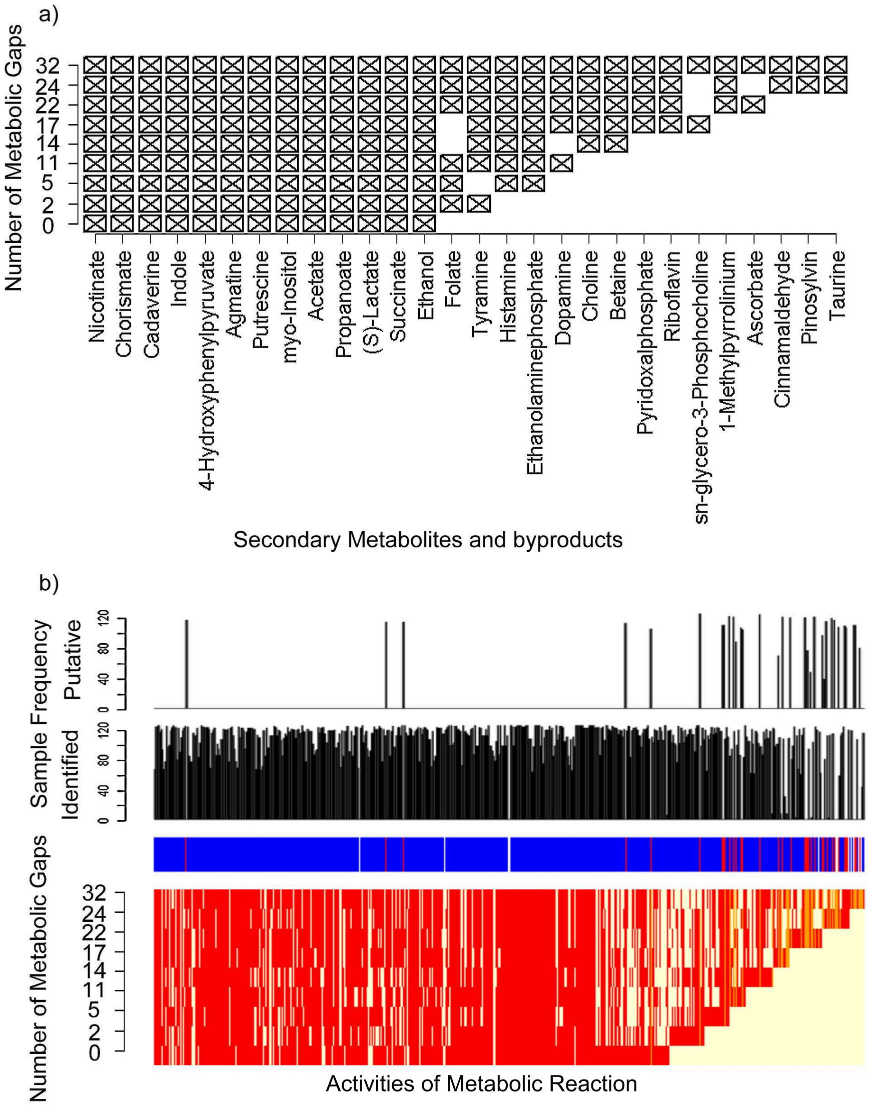


Figure 6.1: a) Secondary metabolites and byproducts synthesized by the core metabolic network of human gut microbiome. The numbers of metabolic gaps that are needed to fill are listed on the left. The compounds in the biomass composition were not counted; b) the corresponding active reactions for synthesis of the products (red bars). The frequencies of these reactions identified in samples are also listed. Putative: metabolic gaps filled by reaction not in pan metabolic network; Identified: metabolic gaps filled by reaction in pan metabolic network. The side bar indicates the sources of metabolic reactions. Blue: core metabolic networks; red: metabolic gaps.

predictions less reliable. The sample frequencies of both pan metabolic reactions and putative reactions demonstrate that almost all the metabolic reactions involved can be found in more than half of the populations. This result ensures the predictions are representative and the derived core metabolic network is capable of producing these metabolites with limited metabolic gap filling.

We reconstructed the metabolic network of several phylogenetic groups identified in this human gut metagenomic dataset. The sequences were classified into different groups from phylum to genus. Even though we can generate metabolic networks for all these groups, the discovery rates of genes in all the samples prevent us reconstructing high-quality metabolic networks for some lower-level groups in the phylogenetic tree. To generate representative metabolic networks, we select those reactions widely identified not only in the whole gut microbiome but also in specific classified groups.

After reconstructing the metabolic networks of the three major phyla, we were able to generate community-wide metabolic networks of Firmicutes, Bacteroidetes and Proteobacteria using the same methods described in Section 4.3.2. Figure 6.2 are the predicted interactions among the three phyla under three levels of penalty of interaction. Some interactions that were suggested by the prediction of biosynthetic capabilities were found in this result. For example, Firmicutes was the major phylum corresponding to butyrate production. There are only a few interactions predicted under the highest penalty level, including riboflavin (Vitamin B2), ascorbate(Vitamin C), sn-glycero-3 phosphocholine, and pyridoxalphosphate(Vitamin B6). Interestingly, most of these interacted molecules are vitamins. According to the results in medium penalty level, 11, 12 and 18 putative reactions were filled as metabolic gaps for Firmicutes, Bacteroidetes and Proteobacteria respectively. These metabolic gaps were about 3% of the total active metabolic networks. The low fraction of metabolic gaps can be explained by the methods we used to identify metabolic reactions from the metagenomic dataset. We assigned all the metabolic reactions in lower levels to the corresponding higher level. Therefore, metabolic reactions from multiple organisms were including for one phylum and the chance for missing an important metabolic reaction was significantly reduced, compared to assignment metabolic reactions to genus or species level.

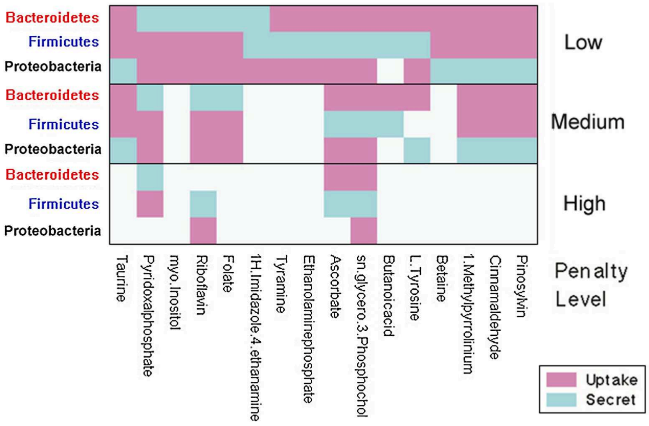


Figure 6.2: Interaction results on Three-Phylum-Model of human gut microbiome. Higher penalty leads to more conservative prediction.

6.2.2.2 Future Work

The above preliminary work on the MetaHIT dataset demonstrated that we can capture and predict metabolic phenotypes as well as inter-groups interactions by applying metabolic network reconstruction to those phylogenetic groups. One advantage of this method is that we can avoid the uncertainty caused by inaccurate sequence binning. In addition, this method integrated the organisms with relatively low discovery rates into several phylogenetic groups, which contain more complete genomic sequences and higher discovery rates than individual organisms. We further found that the sequence depth and coverage have significant effects on the accuracy of model predictions. Therefore, applying similar methods to more comprehensive datasets lead to more accurate and complete understanding about these microbial communities.

Human Microbiome Project (HMP) is such a project aiming to provide comprehensive structural and genetic information of human microbiomes in various body sites including the gastrointestinal tract. One of the recently released datasets were collected from 242 screened and phenotyped healthy adults (*Human Microbiome Project Consortium*, 2012). This dataset contains more samples than the MetaHIT dataset and the information of host phenotype contains more details. Another advantage of the new dataset is that only samples from healthy adults were included. Therefore, sample-wise variations caused by differences in host phenotypes were minimized and the metabolic network reconstructed according to this dataset might be more accurate. The metabolic pathways were found to be consistent across the stool samples (*Human Microbiome Project Consortium*, 2012), which suggests the consistency of the dataset.

Provided with more comprehensive and consistent Human Microbiome Project (HMP) datasets, we will be able to generate high-quality metabolic networks for natural human gut microbiomes. The detailed phenotype records of the host are another type of informative data. By associating the metabolic networks with the phenotypes, e.g. Body Mass Index (BMI), we might be able to generate new hypotheses about the host-microbiome interaction. With these comprehensive sequencing data, multiple-group metabolic models can be applied at lower levels, e.g. at the species level. Results from such models with lower-level groups could provide more accurate predictions about inter-species interactions by reducing the number of organisms in each group.

APPENDICES

APPENDIX A

Biomass Compositions Assumed for *E. coli*

Table A.1: Biomass synthesis function assumed for *E. coli*. This composition was modified from *i*AF1260 model. Negative value represents requirment.

Metabolites	Coefficients (mmol/g DCW)	Metabolites	Coefficients (mmol/g DCW)
L-alanine	-0.5137	L-arginine	-0.2958
L-asparagine	-0.2411	L-aspartate	-0.2411
L-cysteine	-0.09158	L-glutamate	-0.2632
L-glutamine	-0.2632	L-glycine	-0.6126
L-histidine	-0.09474	L-iso-leucine	-0.2905
L-leucine	-0.4505	L-lysine	-0.3432
L-methionine	-0.1537	L-phenylalanine	-0.1759
L-proline	-0.2211	L-serine	-0.2158
L-threonine	-0.2537	L-tryptophan	-0.05684
L-tyrosine	-0.1379	L-valine	-0.4232
dATP	-0.02617	dGTP	-0.02702
dCTP	-0.02702	dTTP	-0.02617

Continued on next page

Metabolites	Coefficients (mmol/g DCW)	Metabolites	Coefficients (mmol/g DCW)
GTP	-0.2151	CTP	-0.1335
UTP	-0.1441	NAD	-0.001831
NADP	0.000447	COA	-0.000576
FAD	-0.000223	ATP	-59.984
10-Formyltetrahydrofolate	-0.000223	2-Oxo-3-hydroxy-4-phosphobutanoate	-0.000223
S-Adenosyl-L-methionine	-0.000223	5,10-Methylenetetrahydrofolate	-0.000223
Pyridoxal 5'-phosphate	-0.000223	Riboflavin	-0.000223
5,6,7,8-Tetrahydrofolate	-0.000223	ADP	59.810000

APPENDIX B

Automated Curation Model in PEER

- Sets

  R, all the metabolic reactions;

  R_e, all the transporter reactions;

  R_{envi}, environmental conditions;

  M, all the metablites;

  S, all the species (subdivisions) in the model;

  Ire, irreservable reactions;

  In_{envi}, all the incomming environmental exchange fluxes;

  Out_{envi}, all the outgoing environmental exchange fluxes;

  In, all the uptake fluxes;

  Out, all the secretion fluxes;

- Parameters

  $S(r, m)$, stoichiometric coefficient of metabolite m in metabolic reaction r;

$S_e(r_e, m)$, stoichiometric coefficient of metabolite m in transport reaction r_e;

$S_{envi}(r_{envi}, m)$, stoichiometric coefficient of metabolite m of environmental exchange flux r_{envi};

$weight(r, s)$, the scaled risk of added reaction $r$in species s;

$Exs(r, s)$, existent of reaction r in specise s (0 or 1);

$Elim$, the upbound of the reaction flux;

$spon(r)$, parameter indicates whether r is spontaneous reaction;

$Nutrient(r_{envi})$, the maximum nutrient supply of the incomming environmental exchange fluxes;

$Growth(s)$, the minimal growth rate requirement of species s in this environment.

- Continuous variables

  $v(r, s)$, flux of reaction $r$in species s;

  $v_e(r_e, s)$, flux of reaction r_e in species s;

  $v_{envi}(r_{envi})$, flux of environmental exchange reaction r_{envi};

- Binary variables

  $b_{act}(r, s)$, activity of reaction r in species s (0 or 1);

  $b_{add}(r, s)$, Add reaction r into species s (0 or 1);

- Objective function

$$\min_{v, v_e, b_{add}} \sum_{s \in S, r \in R} weight(r, s) * b_{add}(r, s)$$

This objective function is trying to minimize the all the putative reactions in a metabolic network. *weight* parameters give weights to all these putative reactions based on the sequence alignment. Therefore, the model will fill the metabolic gaps with those reactions associated with good gene candidates if possible.

- Constrains
  - Mass balance constraints of each species.

    $$\sum_{r \in R} S(r,m) * v(r,s) + \sum_{r_e \in R_e} S_e(r_e,m) * v_e(r_e,s) = 0$$

    , $\forall m \in M, s \in S$. These constraints will force all the metabolites in each species without net changes. The effects of exchange fluxes have been included in these mass balance constraints.

  - Mass balance constraints of the whole compartment

    $$\sum_{r \in R} \sum_{s \in S} S_e(r_e,m) * v_e(r_e,s) + \sum_{r_{envi} \in R_{envi}} S_{envi}(r_{envi},m) * v_{envi}(r_{envi}) = 0$$

    , $\forall m \in M$. These constraints requires mass balance of all the exchange fluxes. All the net changes must be balanced by the exchange fluxes with environments. These mass balance constraints can be omitted if there is only one organism or compartment containted in the model.

  - Reversibility

    $$v(r,s) \geq 0 \quad \forall r \in R, s \in S$$

    These constraints requires all the irreversible reaction with positive flux.

  - Active reaction constraints

    $$v(r,s) \leq Elim * b_{act}(r,s) \quad \forall r \in R, s \in S$$

$$v(r,s) \geq -Elim * b_{act}(r,s) \quad \forall r \in R, s \in S$$

These two set of constriants will only allow those reactions predicted to be active to carry fluxes.

- Existance of reaction

$$b_{act}(r,s) \leq Exs(r,s) + b_{add}(r,s) + spon(r) \quad \forall r \in R, s \in S$$

The active reactions can be spontaneous reactions, putative reactions, or reactions identified in the genome annotation.

- Nutrient uptake and product secret

$$Nutrient(r_{envi}) \geq v_{envi}(r_{envi}) \geq 0 \quad \forall r_{envi} \in In_{envi}$$

Environmental incomming fluxes, the upbounds of the incomming fluxes are determined by the nutrient supply.

$$v_{envi}(r_{envi}) \leq 0 \quad \forall r_{envi} \in Out_{envi}$$

Environmental outgoing fluxes, there is no limit for these outgoing fluxes.

$$v_e(r_e, s) \geq 0 \quad \forall r_e \in In, s \in S$$

Uptake fluxes for each organism;

$$v_e(r_e, s) \leq 0 \quad \forall r_e \in Out, s \in S$$

Secret fluxes for each organism.

– Growth Requirement

$$v(r\_Growth, s) \leq Growth(s) \quad \forall s \in S$$

r_Growth is the biomass synthesis reaction. It must larger than the minimal growth rate defined by $Growth(s)$.

APPENDIX C

Byproducts in Model Gut Microbiota

The minimal products were predicted by the multi-organisms model by minimizing the number of byproducts. Table C.1 lists all the necessary byproducts predicted in two-species model. Most of the byproducts are synthesized by *E. rectale* except 3-beta-D-Galactosyl-sn-glycerol.

Table C.1: Minimal byproducts in the two-species microbial community.

Species	Minimal Byproducts
B .thetaiotaomicron	3-beta-D-Galactosyl-sn-glycerol
E. rectale	Hydroxylamine
E. rectale	2,3-Bisphospho-D-glycerate
E. rectale	Deoxyguanosine
E. rectale	3-Hydroxy-3-methyl-2-oxobutanoicacid
E. rectale	alpha-D-Galactose

Figure 4.15 provides the list of byproducts that have been identified in relevant metabolomic studies. The complete list of the byproducts are listed in Table C.2. There are only less than 15% of predicted byproducts were identified in relevent metabolomic studies. This low coverage was caused by either the incomplete metabolomic datasets or false positive predictions by the ten speceis model.

Table C.2: Minimal byproducts in the ten-species microbial community. Only the byproducts predicted by all the solutions were listed.

Metabolite	Metabolite	Metabolite	Metabolite
CoA	D-Fructose1,6-bisphosphate	Hexadecanoicacid	Shikimate3-phosphate
Glyoxylate	D-Glucosamine6-phosphate	1,2-Diacyl-sn-glycerol	L-Rhamnose
Hydrogen	Sucrose6-phosphate	Glycolaldehyde	L-Rhamnulose
Oxidizedferredoxin	D-Mannose6-phosphate	D-Xylose	D-Mannonate
3-Phospho-D-glycerate	Sugarphosphate	D-Xylulose	(R)-Pantoate
Maltose	Sugar	D-Fructuronate	N-Acetyl-L-glutamate5-phosphate
Thiosulfate	Anthranilate	D-Arabinose5-phosphate	Deoxyribose
Raffinose	Sorbitol6-phosphate	2-Dehydro-3-deoxy-D-gluconate	4,6-Dideoxy-4-oxo-dTDP-D-glucose
Inosine	Glycerol	D-Altronate	dTDP-4-dehydro-6-deoxy-L-mannose
5-Dehydro-4-deoxy-D-glucuronate	D-Fructose1-phosphate	2-Dehydro-3-deoxy-6-phospho-D-gluconate	D-O-Phosphoserine
Glutathione	D-Sorbitol	(4S)-4,6-Dihydroxy-2,5-dioxohexanoate	Melibiitol

Continued on next page

Metabolite	Metabolite	Metabolite	Metabolite
Nicotinamide	D-Mannose	beta-D-Glucose1-phosphate	D-Tagatose6-phosphate
Hydrogensulfide	Propanoyl-CoA	Creatine	3-Hydroxy-3-methyl-2-oxobutanoicacid
Sulfite	Propanoate	Thymidine	Chloramphenicol
Triphosphate	10-Formyltetrahydrofolate	3-Dehydroshikimate	Chloramphenicol3-acetate
Adenine	Uridine	Deoxycytidine	2-Amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine

APPENDIX D

Metabolites Identified in Metabolomic Studies

We collect metabolites that have been identified in the gut environments from literatures (*Bjerrum et al.*, 2010; *Chuang et al.*, 2012; *Jansson et al.*, 2009; *Le Gall et al.*, 2011; *Li et al.*, 2011; *Martin et al.*, 2010; *Wikoff et al.*, 2009; *Wu et al.*, 2010; *Zheng et al.*, 2011). This list of metabolite was used as a filter to simplify the prediction of uptake and secretion in Chapter IV. Table D.1 is the complete list of these metabolites.

Table D.1: Complete list of metabolites identified in relevant metabolomic studies.

Metabolite	Metabolite	Metabolite	Metabolite
4-Imidazolone-5-propanoate	L-Glutamine	Ascorbate	3-Hydroxykynurenine
L-Citrulline	2-Oxoglutarate	Choline	O-(4-Hydroxy-3,5-diidophenyl)-3,5-diiodo-L-tyrosine
Urea	beta-D-Glucose	(R)-3-Hydroxybutanoate	4-(2-Aminoethyl)-1,2-benzenediol
Adenine	alpha-D-Glucose	(S)-Lactate	Homovanillate

Continued on next page

Metabolite	Metabolite	Metabolite	Metabolite
D-Ribose5-phosphate	L-Tyrosine	Acetone	L-Normetanephrine
L-Homocysteine	Raffinose	Dimethylamine	Octanoicacid
Acetate	Formate	Creatine	Tetradecanoicacid
4-Aminobutanoate	L-Serine	Creatinine	Hexadecanoicacid
Citrate	L-Alanine	Allantoin	L-Arabitol
Succinate	Isocitrate	3-Methylguanine	D-Arabinose
Cadaverine	Glycine	Thymine	Xylitol
N-Carbamoyl-L-aspartate	D-Alanine	Deoxycytidine	D-Xylose
L-Ornithine	Fumarate	Cytosine	6-Deoxy-D-galactose
O-Phospho-L-serine	L-Lysine	Malonate	L-Rhamnose
L-Proline	L-Asparagine	Cytidine	D-Gluconicacid
Indole	2-Oxobutanoate	Imidazole-4-acetaldehyde	Citramalate
Ethanol	L-Tryptophan	Imidazole-4-acetate	Aminomalonate
Glycerol	L-Phenylalanine	Methylimidazoleaceticacid	2'-Hydroxydihydrodaidzein
Hypoxanthine	Phenylpyruvate	Hypotaurine	I-Urobilin
L-Histidine	beta-D-Fructose	5-Oxoproline	N,N-Dimethylformamide
Urocanate	Chorismate	L-Methionine	Salicyluricacid
myo-Inositol	L-Leucine	N6-(L-1,3-Dicarboxypropyl)-L-lysine	S-Succinyldihydrolipoamide
Orotate	Nicotinate	L-Pipecolate	L-Carnitine

Continued on next page

Metabolite	Metabolite	Metabolite	Metabolite
Shikimate	Xanthine	Sarcosine	Agmatine
LL-2,6-Diaminoheptanedioate	L-Isoleucine	Putrescine	Tyramine
CO2	alpha-D-Galactose	Itaconate	1H-Imidazole-4-ethanamine
Pyruvate	Glutathione	Tryptamine	Skatole
D-Glucose	sn-glycero-3-Phosphocholine	Nicotinamide	Propanoate
Glyoxylate	Betaine	N-Methyltryptamine	Pentanoate
L-Glutamate	Taurine	4-Hydroxy-2-quinolinecarboxylicacid	Ethanolaminephosphate

BIBLIOGRAPHY

BIBLIOGRAPHY

Albertsen, M., L. B. S. Hansen, A. M. Saunders, P. H. Nielsen, and K. L. Nielsen (2012), A metagenome of a full-scale microbial community carrying out enhanced biological phosphorus removal, *ISME J*, *6*(6), 1094–106, doi:10.1038/ismej.2011.176.

Allen, E. E., and J. F. Banfield (2005), Community genomics in microbial ecology and evolution., *Nat Rev Microbiol*, *3*, 489–498, doi:10.1038/nrmicro1157.

Almaas, E., B. Kovács, T. Vicsek, Z. N. Oltvai, and A.-L. Barabási (2004), Global organization of metabolic fluxes in the bacterium *Escherichia coli*, *Nature*, *427*, 839–43, doi:10.1038/nature02289.

Alper, H., K. Miyaoku, and G. Stephanopoulos (2005), Construction of lycopene-overproducing e. coli strains by combining systematic and combinatorial gene knockout targets, *Nat Biotechnol*, *23*(5), 612–6, doi:10.1038/nbt1083.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (1990), Basic local alignment search tool, *J Mol Biol*, *215*, 403–410, doi:10.1006/jmbi.1990.9999.

Antranikian, G., C. Friese, A. Quentmeier, H. Hippe, and G. Gottschalk (1984), Distribution of the ability for citrate utilization amongst clostridia, *Archives of Microbiology*, *138*(3), 179–182, doi:10.1007/BF00402115.

Arumugam, M., et al. (2011), Enterotypes of the human gut microbiome, *Nature*, *473*(7346), 174–80, doi:10.1038/nature09944.

Aziz, R., D. Bartels, A. Best, M. DeJongh, T. Disz, R. Edwards, and et al (2008), The rast server: rapid annotations using subsystems technology, *BMC Genomics*.

Barrett, C. L., C. D. Herring, J. L. Reed, and B. O. Palsson (2005), The global transcriptional regulatory network for metabolism in *Escherichia coli* exhibits few dominant functional states, *Proc Natl Acad Sci U S A*, *102*, 19,103–8, doi:10.1073/pnas.0505231102.

Bategeli, V., and A. Mrvar (2003), *Graph Drawing Software*, chap. Pajek-Analysis and Visualization of Larget Networkd, pp. 77–103, 1 ed., Springer.

Baumler, D. J., K.-C. Jeong, B. G. Fox, J. F. Banfield, and C. W. Kaspar (2005), Sulfate requirement for heterotrophic growth of "*Ferroplasma acidarmanus*" strain fer1., *Res Microbiol*, *156*, 492–498, doi:10.1016/j.resmic.2004.12.007.

Becker, S. A., N. D. Price, and B. Ø. Palsson (2006), Metabolite coupling in genome-scale metabolic networks, *BMC Bioinformatics*, *7*, 111, doi:10.1186/1471-2105-7-111.

Becker, S. A., A. M. Feist, M. L. Mo, G. Hannum, B. Ø. Palsson, and M. J. Herrgard (2007), Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox, *Nat Protoc*, *2*, 727–38, doi:10.1038/nprot.2007.99.

Biddle, J. F., S. Fitz-Gibbon, S. C. Schuster, J. E. Brenchley, and C. H. House (2008), Metagenomic signatures of the peru margin subseafloor biosphere show a genetically distinct environment, *Proc Natl Acad Sci U S A*, *105*(30), 10,583–8, doi:10.1073/pnas.0709942105.

Bizukojc, M., D. Dietz, J. Sun, and A.-P. Zeng (2010), Metabolic modelling of syntrophic-like growth of a 1,3-propanediol producer, clostridium butyricum, and a methanogenic archeon, methanosarcina mazei, under anaerobic conditions, *Bioprocess Biosyst Eng*, *33*(4), 507–23, doi:10.1007/s00449-009-0359-0.

Bjerrum, J. T., O. H. Nielsen, F. Hao, H. Tang, J. K. Nicholson, Y. Wang, and J. Olsen (2010), Metabonomics in ulcerative colitis: diagnostics, biomarker identification, and insight into the pathophysiology, *J Proteome Res*, *9*(2), 954–62, doi:10.1021/pr9008223.

Blumberg, R., and F. Powrie (2012), Microbiota, disease, and back to health: a metastable journey, *Sci Transl Med*, *4*(137), 137rv7, doi:10.1126/scitranslmed.3004184.

Brüls, T., and J. Weissenbach (2011), The human metagenome: our other genome?, *Hum Mol Genet*, *20*(R2), R142–8, doi:10.1093/hmg/ddr353.

Burgard, A. P., P. Pharkya, and C. D. Maranas (2003), Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization, *Biotechnol Bioeng*, *84*, 647–657, doi:10.1002/bit.10803.

Burgard, A. P., E. V. Nikolaev, C. H. Schilling, and C. D. Maranas (2004), Flux coupling analysis of genome-scale metabolic network reconstructions, *Genome Res*, *14*(2), 301–12, doi:10.1101/gr.1926504.

Candela, M., S. Maccaferri, S. Turroni, P. Carnevali, and P. Brigidi (2010), Functional intestinal microbiome, new frontiers in prebiotic design, *Int J Food Microbiol*, *140*(2-3), 93–101, doi:10.1016/j.ijfoodmicro.2010.04.017.

Chance, B., and G. R. Williams (1955), Respiratory enzymes in oxidative phosphorylation. iii. the steady state, *J Biol Chem*, *217*(1), 409–27.

Chandrasekaran, S., and N. D. Price (2010), Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in escherichia coli and mycobacterium tuberculosis, *Proc Natl Acad Sci U S A*, *107*(41), 17,845–50, doi:10.1073/pnas.1005139107.

Chuang, H.-L., Y.-T. Huang, C.-C. Chiu, C.-D. Liao, F.-L. Hsu, C.-C. Huang, and C.-C. Hou (2012), Metabolomics characterization of energy metabolism reveals glycogen accumulation in gut-microbiota-lacking mice, *J Nutr Biochem*, *23*(7), 752–8, doi:10.1016/j.jnutbio.2011.03.019.

Cogen, A. L., V. Nizet, and R. L. Gallo (2008), Skin microbiota: a source of disease or defence?, *Br J Dermatol*, *158*(3), 442–55, doi:10.1111/j.1365-2133.2008.08437.x.

Covert, M. W., and B. Ø. Palsson (2002), Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*, *J Biol Chem*, *277*, 28,058–64, doi:10.1074/jbc.M201691200.

Covert, M. W., E. M. Knight, J. L. Reed, M. J. Herrgard, and B. O. Palsson (2004), Integrating high-throughput and computational data elucidates bacterial networks, *Nature*, *429*, 92–6, doi:10.1038/nature02456.

Covert, M. W., N. Xiao, T. J. Chen, and J. R. Karr (2008), Integrating metabolic, transcriptional regulatory and signal transduction models in escherichia coli, *Bioinformatics*, *24*(18), 2044–50, doi:10.1093/bioinformatics/btn352.

Cvijovic, M., R. Olivares-Hernández, R. Agren, N. Dahr, W. Vongsangnak, I. Nookaew, K. R. Patil, and J. Nielsen (2010), Biomet toolbox: genome-wide analysis of metabolism, *Nucleic Acids Res*, *38*(Web Server issue), W144–9, doi:10.1093/nar/gkq404.

DeLong, E. F., et al. (2006), Community genomics among stratified microbial assemblages in the ocean's interior, *Science*, *311*(5760), 496–503, doi:10.1126/science.1120250.

Dewhirst, F. E., T. Chen, J. Izard, B. J. Paster, A. C. R. Tanner, W.-H. Yu, A. Lakshmanan, and W. G. Wade (2010), The human oral microbiome, *J Bacteriol*, *192*(19), 5002–17, doi:10.1128/JB.00542-10.

Divakaruni, A. S., and M. D. Brand (2011), The regulation and physiology of mitochondrial proton leak, *Physiology (Bethesda)*, *26*(3), 192–205, doi:10.1152/physiol.00046.2010.

Domach, M. M., S. K. Leung, R. E. Cahn, G. G. Cocks, and M. L. Shuler (2000), Computer model for glucose-limited growth of a single cell of escherichia coli b/r-a. reprinted from biotechnology and bioengineering, vol. 26, issue 3, pp 203-216 (1984), *Biotechnol Bioeng*, *67*(6), 827–40.

Dongowski, G., A. Lorenz, and H. Anger (2000), Degradation of pectins with different degrees of esterification by bacteroides thetaiotaomicron isolated from human gut flora, *Appl Environ Microbiol*, *66*(4), 1321–7.

Duarte, N. C., M. J. Herrgård, and B. Ø. Palsson (2004), Reconstruction and validation of saccharomyces cerevisiae ind750, a fully compartmentalized genome-scale metabolic model, *Genome Res*, *14*(7), 1298–309, doi:10.1101/gr.2250904.

Duarte, N. C., S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. Ø. Palsson (2007), Global reconstruction of the human metabolic network based on genomic and bibliomic data, *Proc Natl Acad Sci U S A*, *104*, 1777–82, doi:10.1073/pnas.0610772104.

Eckburg, P. B., E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson, and D. A. Relman (2005), Diversity of the human intestinal microbial flora, *Science*, *308*(5728), 1635–8, doi:10.1126/science.1110591.

Edwards, J. S., and B. O. Palsson (1999), Systems properties of the *Haemophilus influenzae* rd metabolic genotype, *J Biol Chem*, *274*, 17,410–17,416.

Edwards, J. S., and B. O. Palsson (2000a), The *Escherichia coli* mg1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities, *Proc Natl Acad Sci U S A*, *97*, 5528–5533.

Edwards, J. S., and B. O. Palsson (2000b), Robustness analysis of the escherichia coli metabolic network, *Biotechnol Prog*, *16*(6), 927–39, doi:10.1021/bp0000712.

Edwards, J. S., R. U. Ibarra, and B. O. Palsson (2001), In silico predictions of escherichia coli metabolic capabilities are consistent with experimental data, *Nat Biotechnol*, *19*(2), 125–30, doi:10.1038/84379.

Faith, J. J., N. P. McNulty, F. E. Rey, and J. I. Gordon (2011), Predicting a human gut microbiota's response to diet in gnotobiotic mice, *Science*, *333*(6038), 101–4, doi:10.1126/science.1206025.

Feist, A. M., C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. Ø. Palsson (2007), A genome-scale metabolic reconstruction for *Escherichia coli* k-12 mg1655 that accounts for 1260 orfs and thermodynamic information, *Mol Syst Biol*, *3*, 121, doi:10.1038/msb4100155.

Floudas, C. (1995), *Nonlinear And Mixed-Integer Optimization: Fundamentals And Applications*, 0195100565, Oxford University Press, USA.

Förster, J., I. Famili, P. Fu, B. Ø. Palsson, and J. Nielsen (2003), Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network, *Genome Res*, *13*, 244–53, doi:10.1101/gr.234503.

Forth, T., G. McConkey, and D. Westhead (2010), Metnetmaker: A free and open-source tool for the creation of, *Bioinformatics*.

Freilich, S., R. Zarecki, O. Eilam, E. S. Segal, C. S. Henry, M. Kupiec, U. Gophna, R. Sharan, and E. Ruppin (2011), Competitive and cooperative metabolic interactions in bacterial communities, *Nat Commun*, *2*, 589, doi:10.1038/ncomms1597.

Frey, E. (2010), Evolutionary game theory: Theoretical concepts and applications to microbial communities, *Physica A-Statistical Mechanics and Its Applications*, *389*(20), 4265–4298, doi:DOI10.1016/j.physa.2010.02.047.

Garcia Martin, H., et al. (2006), Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities., *Nat Biotechnol*, *24*, 1263–1269, doi:10.1038/nbt1247.

Ghim, C.-M., K.-I. Goh, and B. Kahng (2005), Lethality and synthetic lethality in the genome-wide metabolic network of escherichia coli, *J Theor Biol*, *237*(4), 401–11, doi:10.1016/j.jtbi.2005.04.025.

Gill, S. R., et al. (2006a), Metagenomic analysis of the human distal gut microbiome, *Science*, *312*(5778), 1355–9, doi:10.1126/science.1124234.

Gill, S. R., et al. (2006b), Metagenomic analysis of the human distal gut microbiome, *Science*, *312*(5778), 1355–9, doi:10.1126/science.1124234.

Gille, C., S. Hoffmann, and H.-G. Holzhütter (2007), Metannogen: compiling features of biochemical reactions needed for the reconstruction of metabolic networks, *BMC Syst Biol*, *1*, 5, doi:10.1186/1752-0509-1-5.

Giovannoni, S. J., M. S. Rappé, K. L. Vergin, and N. L. Adair (1996), 16s rrna genes reveal stratified open ocean bacterioplankton populations related to the green non-sulfur bacteria, *Proc Natl Acad Sci U S A*, *93*(15), 7979–84.

Goltsman, D. S. A., et al. (2009), Community genomic and proteomic analyses of chemoautotrophic iron-oxidizing "*Leptospirillum rubarum*" (group ii) and "*Leptospirillum ferrodiazotrophum*" (group iii) bacteria in acid mine drainage biofilms., *Appl Environ Microbiol*, *75*, 4599–4615, doi:10.1128/AEM.02943-08.

Goto, S., T. Nishioka, and M. Kanehisa (1999), LIGAND database for enzymes, compounds and reactions., *Nucleic Acids Res*, *27*, 377–379.

Goto, S., T. Nishioka, and M. Kanehisa (2000), LIGAND: chemical database of enzyme reactions., *Nucleic Acids Res*, *28*, 380–382.

Goto, S., Y. Okuno, M. Hattori, T. Nishioka, and M. Kanehisa (2002), LIGAND: database of chemical compounds and reactions in biological pathways., *Nucleic Acids Res*, *30*, 402–404.

Green, M. L., and P. D. Karp (2004), A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases, *BMC Bioinformatics*, *5*, 76–76, doi:10.1186/1471-2105-5-76.

Greenblum, S., P. J. Turnbaugh, and E. Borenstein (2012), Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease, *Proc Natl Acad Sci U S A*, *109*(2), 594–9, doi:10.1073/pnas.1116053109.

Grice, E. A., et al. (2009), Topographical and temporal diversity of the human skin microbiome, *Science*, *324*(5931), 1190–2, doi:10.1126/science.1171700.

Handorf, T., O. Ebenhöh, and R. Heinrich (2005), Expanding metabolic networks: scopes of compounds, robustness, and evolution, *J Mol Evol*, *61*(4), 498–512, doi:10.1007/s00239-005-0027-1.

Handorf, T., N. Christian, O. Ebenhöh, and D. Kahn (2008), An environmental perspective on metabolism, *J Theor Biol*, *252*, 530–7, doi:10.1016/j.jtbi.2007.10.036.

Hao, T., H.-W. Ma, X.-M. Zhao, and I. Goryanin (2012), The reconstruction and analysis of tissue specific human metabolic networks, *Mol Biosyst*, *8*(2), 663–70, doi:10.1039/c1mb05369h.

Henry, C., M. DeJongh, A. Best, P. Frybarger, B. Linsay, and R. Stevens (2010), High-throughput generation, optimization and analysis of genome-scale metabolic models,, *Nat Biotechnol*.

Hooper, L. V., J. Xu, P. G. Falk, T. Midtvedt, and J. I. Gordon (1999), A molecular sensor that allows a gut commensal to control its nutrient foundation in a competitive ecosystem, *Proc Natl Acad Sci U S A*, *96*(17), 9833–8.

Hooper, L. V., D. R. Littman, and A. J. Macpherson (2012), Interactions between the microbiota and the immune system, *Science*, *336*(6086), 1268–73, doi:10.1126/science.1223490.

Hosokawa, T., Y. Kikuchi, N. Nikoh, M. Shimada, and T. Fukatsu (2006), Strict host-symbiont cospeciation and reductive genome evolution in insect gut bacteria, *PLoS Biol*, *4*(10), e337, doi:10.1371/journal.pbio.0040337.

Huang, S. (2001), Genomics, complexity and drug discovery: insights from boolean network models of cellular regulation, *Pharmacogenomics*, *2*(3), 203–22, doi:10.1517/14622416.2.3.203.

Human Microbiome Project Consortium (2012), Structure, function and diversity of the healthy human microbiome, *Nature*, *486*(7402), 207–14, doi:10.1038/nature11234.

Hunter, S., et al. (2009), Interpro: the integrative protein signature database, *Nucleic Acids Res*, *37*(Database issue), D211–5, doi:10.1093/nar/gkn785.

Huynen, M., T. Dandekar, and P. Bork (1999), Variation and evolution of citric-acid cycle: a genomic perspective, *Trends in Microbiology*.

Ibarra, R. U., J. S. Edwards, and B. O. Palsson (2002), *Escherichia coli* k-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth, *Nature*, *420*, 186–189, doi:10.1038/nature01149.

Imielinski, M., C. Belta, H. Rubin, and A. Halász (2006), Systematic analysis of conservation relations in escherichia coli genome-scale metabolic network reveals novel growth media, *Biophys J*, *90*(8), 2659–72, doi:10.1529/biophysj.105.069278.

Jansson, J., B. Willing, M. Lucio, A. Fekete, J. Dicksved, J. Halfvarson, C. Tysk, and P. Schmitt-Kopplin (2009), Metabolomics reveals metabolic biomarkers of crohn's disease, *PLoS One*, *4*(7), e6386, doi:10.1371/journal.pone.0006386.

Kennedy, J., B. Flemer, S. A. Jackson, D. P. H. Lejon, J. P. Morrissey, F. O'Gara, and A. D. W. Dobson (2010), Marine metagenomics: new tools for the study and exploitation of marine microbial metabolism, *Mar Drugs*, *8*(3), 608–28, doi:10.3390/md8030608.

Kettler, G. C., et al. (2007), Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*, *PLoS Genet*, *3*, e231, doi:10.1371/journal.pgen.0030231.

Kharchenko, P., D. Vitkup, and G. M. Church (2004), Filling gaps in a metabolic network using expression information, *Bioinformatics*, *20 Suppl 1*, 178–185, doi:10.1093/bioinformatics/bth930.

Kim, J., and J. L. Reed (2010), Optorf: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains, *BMC Syst Biol*, *4*, 53, doi:10.1186/1752-0509-4-53.

Klitgord, N., and D. Segrè (2010), Environments that induce synthetic microbial ecosystems, *PLoS Comput Biol*, *6*(11), e1001,002, doi:10.1371/journal.pcbi.1001002.

Kumar, V., M. Dasika, and C. Maranas (2007), Optimization based automated curation of metabolic reconstructions, *BMC Bioinformatics*.

Kumar, V. S., and C. D. Maranas (2009), Growmatch: an automated method for reconciling in silico/in vivo growth predictions, *PLoS Comput Biol*, *5*(3), e1000,308, doi:10.1371/journal.pcbi.1000308.

Lay, C., et al. (2005), Colonic microbiota signatures across five northern european countries, *Appl Environ Microbiol*, *71*(7), 4153–5, doi:10.1128/AEM.71.7.4153-4155.2005.

Le Gall, G., S. O. Noor, K. Ridgway, L. Scovell, C. Jamieson, I. T. Johnson, I. J. Colquhoun, E. K. Kemsley, and A. Narbad (2011), Metabolomics of fecal extracts detects altered metabolic activity of gut microbiota in ulcerative colitis and irritable bowel syndrome, *J Proteome Res*, *10*(9), 4208–18, doi:10.1021/pr2003598.

Lee, D.-Y., H. Yun, S. Park, and S. Y. Lee (2003), MetaFluxNet: the management of metabolic reaction information and quantitative metabolic flux analysis, *Bioinformatics*, *19*, 2144–6.

Lespinet, O., and B. Labedan (2005), Orphan enzymes?, *Science*, *307*, 42, doi:10.1126/science.307.5706.42a.

Ley, R. E., F. Bäckhed, P. Turnbaugh, C. A. Lozupone, R. D. Knight, and J. I. Gordon (2005), Obesity alters gut microbial ecology, *Proc Natl Acad Sci U S A*, *102*(31), 11,070–5, doi:10.1073/pnas.0504978102.

Ley, R. E., P. J. Turnbaugh, S. Klein, and J. I. Gordon (2006), Microbial ecology: human gut microbes associated with obesity, *Nature*, *444*(7122), 1022–3, doi:10.1038/4441022a.

Li, J., G. Wijffels, Y. Yu, L. K. Nielsen, D. O. Niemeyer, A. D. Fisher, D. M. Ferguson, and H. J. Schirra (2011), Altered fatty acid metabolism in long duration road transport: An nmr-based metabonomics study in sheep, *J Proteome Res*, *10*(3), 1073–87, doi:10.1021/pr100862t.

Liao, Y.-C., J. Chen, M.-H. Tsai, Y.-H. Tang, F.-C. Chen, and C. Hsiung (2011), Mrbac: A web server for draft metabolic network reconstructions for bacteria, *Bioengineered Bugs*.

Liao, Y.-C., M.-H. Tsai, F.-C. Chen, and C. A. Hsiung (2012), Gemsirv: A software platform for genome-scale metabolic model simulation, reconstruction and visualization, *Bioinformatics*, doi:10.1093/bioinformatics/bts267.

Likens, G., and G. Likens (1981), Some perspectives of the major biogeochemical cycles.

Lima, T., et al. (2009), Hamap: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in uniprotkb/swiss-prot, *Nucleic Acids Res*, *37*(Database issue), D471–8, doi:10.1093/nar/gkn661.

Lo, I., et al. (2007), Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria., *Nature*, *446*, 537–541, doi:10.1038/nature05624.

Lu, X., H. Vora, and C. Khosla (2008), Overproduction of free fatty acids in e. coli: implications for biodiesel production, *Metab Eng*, *10*(6), 333–9, doi:10.1016/j.ymben.2008.08.006.

Magnuson, K., S. Jackowski, C. O. Rock, and J. E. Cronan, Jr (1993), Regulation of fatty acid biosynthesis in escherichia coli, *Microbiol Rev*, *57*(3), 522–42.

Mahadevan, R., J. S. Edwards, and F. J. Doyle, 3rd (2002), Dynamic flux balance analysis of diauxic growth in escherichia coli, *Biophys J*, *83*(3), 1331–40, doi:10.1016/S0006-3495(02)73903-9.

Mahowald, M. A., et al. (2009), Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla, *Proc Natl Acad Sci U S A*, *106*(14), 5859–64, doi:10.1073/pnas.0901529106.

Majewski, R. A., and M. M. Domach (1990), Simple constrained-optimization view of acetate overflow in e. coli, *Biotechnol Bioeng*, *35*(7), 732–8, doi:10.1002/bit.260350711.

Martin, F.-P. J., N. Sprenger, I. Montoliu, S. Rezzi, S. Kochhar, and J. K. Nicholson (2010), Dietary modulation of gut functional ecology studied by fecal metabonomics, *J Proteome Res*, *9*(10), 5284–95, doi:10.1021/pr100554m.

Martin, F.-P. J., et al. (2008), Top-down systems biology integration of conditional prebiotic modulated transgenomic interactions in a humanized microbiome mouse model, *Mol Syst Biol*, *4*, 205, doi:10.1038/msb.2008.40.

Martiny, A., S. Kathuria, and P. Berube (2009), Widespread metabolic potential for nitrite and nitrate assimilation among prochlorococcus ecotypes., *Proc Natl Acad Sci U S A.*

Matsumoto, M., R. Kibe, T. Ooga, Y. Aiba, S. Kurihara, E. Sawaki, Y. Koga, and Y. Benno (2012), Impact of intestinal microbiota on intestinal luminal metabolome, *Sci Rep*, *2*, 233, doi:10.1038/srep00233.

Mavromatis, K., et al. (2007), Use of simulated data sets to evaluate the fidelity of metagenomic processing methods, *Nat Methods*, *4*, 495–500, doi:10.1038/nmeth1043.

Morris, D. R., and K. L. Koffron (1967), Urea production and putrescine biosynthesis by escherichia coli, *J Bacteriol*, *94*(5), 1516–9.

Musso, G., R. Gambino, and M. Cassader (2011), Interactions between gut microbiota and host metabolism predisposing to obesity and diabetes, *Annu Rev Med*, *62*, 361–80, doi:10.1146/annurev-med-012510-175505.

Nicholson, J. K., E. Holmes, J. Kinross, R. Burcelin, G. Gibson, W. Jia, and S. Pettersson (2012), Host-gut microbiota metabolic interactions, *Science*, *336*(6086), 1262–7, doi:10.1126/science.1223813.

Oh, Y.-K., B. O. Palsson, S. M. Park, C. H. Schilling, and R. Mahadevan (2007), Genome-scale reconstruction of metabolic network in bacillus subtilis based on high-throughput phenotyping and gene essentiality data, *J Biol Chem*, *282*(39), 28,791–9, doi:10.1074/jbc.M703759200.

Okuyama, T., and J. N. Holland (2008), Network structural properties mediate the stability of mutualistic communities., *Ecol Lett*, *11*, 208–216, doi:10.1111/j.1461-0248.2007.01137.x.

Orth, J. D., T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and B. Ø. Palsson (2011), A comprehensive genome-scale reconstruction of escherichia coli metabolism–2011, *Mol Syst Biol*, *7*, 535, doi:10.1038/msb.2011.65.

Pál, C., B. Papp, M. J. Lercher, P. Csermely, S. G. Oliver, and L. D. Hurst (2006), Chance and necessity in the evolution of minimal metabolic networks, *Nature*, *440*, 667–70, doi:10.1038/nature04568.

Palsson, B. (2006), *Systems biology: properties of reconstructed networks*, Cambridge University Press, Cambridge.

Papoutsakis, E. T. (1984), Equations and calculations for fermentations of butyric acid bacteria, *Biotechnol Bioeng*, *26*(2), 174–87, doi:10.1002/bit.260260210.

Papoutsakis, E. T., and C. L. Meyer (1985a), Fermentation equations for propionic-acid bacteria and production of assorted oxychemicals from various sugars, *Biotechnol Bioeng*, *27*(1), 67–80, doi:10.1002/bit.260270109.

Papoutsakis, E. T., and C. L. Meyer (1985b), Equations and calculations of product yields and preferred pathways for butanediol and mixed-acid fermentations, *Biotechnol Bioeng*, *27*(1), 50–66, doi:10.1002/bit.260270108.

Park, E.-J., J. Chun, C.-J. Cha, W.-S. Park, C. O. Jeon, and J.-W. Bae (2012), Bacterial community analysis during fermentation of ten representative kinds of kimchi with barcoded pyrosequencing, *Food Microbiol*, *30*(1), 197–204, doi:10.1016/j.fm.2011.10.011.

Park, J. H., K. H. Lee, T. Y. Kim, and S. Y. Lee (2007), Metabolic engineering of escherichia coli for the production of l-valine based on transcriptome analysis and in silico gene knockout simulation, *Proc Natl Acad Sci U S A*, *104*(19), 7797–802, doi:10.1073/pnas.0702609104.

Partensky, F., W. R. Hess, and D. Vaulot (1999), *Prochlorococcus*, a marine photosynthetic prokaryote of global significance, *Microbiol Mol Biol Rev*, *63*, 106–27.

Patil, K. R., I. Rocha, J. Förster, and J. Nielsen (2005), Evolutionary programming as a platform for in silico metabolic engineering, *BMC Bioinformatics*, *6*, 308, doi:10.1186/1471-2105-6-308.

Peterson, D. A., D. N. Frank, N. R. Pace, and J. I. Gordon (2008), Metagenomic approaches for defining the pathogenesis of inflammatory bowel diseases, *Cell Host Microbe*, *3*(6), 417–27, doi:10.1016/j.chom.2008.05.001.

Petrosino, J. F., S. Highlander, R. A. Luna, R. A. Gibbs, and J. Versalovic (2009), Metagenomic pyrosequencing and microbial identification, *Clin Chem*, *55*(5), 856–66, doi:10.1373/clinchem.2008.107565.

Pharkya, P., and C. D. Maranas (2006), An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems, *Metab Eng*, *8*, 1–13, doi:10.1016/j.ymben.2005.08.003.

Pharkya, P., A. P. Burgard, and C. D. Maranas (2004), Optstrain: a computational framework for redesign of microbial production systems, *Genome Res*, *14*(11), 2367–76, doi:10.1101/gr.2872004.

Price, N. D., J. L. Reed, J. A. Papin, S. J. Wiback, and B. O. Palsson (2003), Network-based analysis of metabolic regulation in the human red blood cell, *J Theor Biol*, *225*(2), 185–94.

Puchałka, J., M. A. Oberhardt, M. Godinho, A. Bielecka, D. Regenhardt, K. N. Timmis, J. A. Papin, and V. A. P. Martins dos Santos (2008), Genome-scale reconstruction and analysis of the pseudomonas putida kt2440 metabolic network facilitates applications in biotechnology, *PLoS Comput Biol*, *4*(10), e1000,210, doi:10.1371/journal.pcbi.1000210.

Punta, M., et al. (2012), The pfam protein families database, *Nucleic Acids Res*, *40*(Database issue), D290–301, doi:10.1093/nar/gkr1065.

Purwani, E. Y., T. Purwadaria, and M. T. Suhartono (2012), Fermentation rs3 derived from sago and rice starch with clostridium butyricum bcc b2571 or eubacterium rectale dsm 17629, *Anaerobe*, *18*(1), 55–61, doi:10.1016/j.anaerobe.2011.09.007.

Qin, J., et al. (2010), A human gut microbial gene catalogue established by metagenomic sequencing, *Nature*, *464*(7285), 59–65, doi:10.1038/nature08821.

Raetz, C. R. (1978), Enzymology, genetics, and regulation of membrane phospholipid synthesis in escherichia coli, *Microbiol Rev*, *42*(3), 614–59.

Ram, R. J., N. C. Verberkmoes, M. P. Thelen, G. W. Tyson, B. J. Baker, R. C. n. Blake, M. Shah, R. L. Hettich, and J. F. Banfield (2005), Community proteomics of a natural microbial biofilm., *Science*, *308*, 1915–1920, doi:10.1126/science.1109070.

Ranganathan, S., P. F. Suthers, and C. D. Maranas (2010), Optforce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions, *PLoS Comput Biol*, *6*(4), e1000,744, doi:10.1371/journal.pcbi.1000744.

Reed, J. L., and B. Ø. Palsson (2004), Genome-scale in silico models of e. coli have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states, *Genome Res*, *14*(9), 1797–805, doi:10.1101/gr.2546004.

Ren, Q., K. Chen, and I. T. Paulsen (2007), TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels., *Nucleic Acids Res*, *35*, D274–9, doi:10.1093/nar/gkl925.

Rocap, C., F. Larimer, J. Lamerdin, S. Malfatti, P. Chain, N. Ahlgren, and et al (2003), Genome divergence in two prochlorococcus ecotypes reflects oceanic niche differentiation, *NATURE*.

Rocha, I., et al. (2010), Optflux: an open-source software platform for in silico metabolic engineering, *BMC Syst Biol*, *4*, 45, doi:10.1186/1752-0509-4-45.

Rusch, D. B., et al. (2007), The sorcerer ii global ocean sampling expedition: northwest atlantic through eastern tropical pacific, *PLoS Biol*, *5*(3), e77, doi:10.1371/journal.pbio.0050077.

Salyers, A. A., and M. Pajeau (1989), Competitiveness of different polysaccharide utilization mutants of bacteroides thetaiotaomicron in the intestinal tracts of germfree mice, *Appl Environ Microbiol*, *55*(10), 2572–8.

Satish Kumar, V., M. S. Dasika, and C. D. Maranas (2007), Optimization based automated curation of metabolic reconstructions, *BMC Bioinformatics*, *8*, 212, doi:10.1186/1471-2105-8-212.

Schellenberger, J., et al. (2011), Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox v2.0, *Nat Protoc*, *6*(9), 1290–307, doi:10.1038/nprot.2011.308.

Schilling, C. H., M. W. Covert, I. Famili, G. M. Church, J. S. Edwards, and B. O. Palsson (2002), Genome-scale metabolic model of helicobacter pylori 26695, *J Bacteriol*, *184*(16), 4582–93.

Schirmer, A., M. A. Rude, X. Li, E. Popova, and S. B. del Cardayre (2010), Microbial biosynthesis of alkanes, *Science*, *329*(5991), 559–62, doi:10.1126/science.1187936.

Schwarz, R., et al. (2007), Integrated network reconstruction, visualization and analysis using yanasquare, *BMC Bioinformatics*, *8*, 313, doi:10.1186/1471-2105-8-313.

Segrè, D., D. Vitkup, and G. M. Church (2002), Analysis of optimality in natural and perturbed metabolic networks, *Proc Natl Acad Sci U S A*, *99*, 15,112–15,117, doi:10.1073/pnas.232349399.

Shipman, J. A., J. E. Berleman, and A. A. Salyers (2000), Characterization of four outer membrane proteins involved in binding starch to the cell surface of bacteroides thetaiotaomicron, *J Bacteriol*, *182*(19), 5365–72.

Shlomi, T., Y. Eisenberg, R. Sharan, and E. Ruppin (2007), A genome-scale computational study of the interplay between transcriptional regulation and metabolism, *Mol Syst Biol*, *3*, 101, doi:10.1038/msb4100141.

Shlomi, T., M. N. Cabili, M. J. Herrgård, B. Ø. Palsson, and E. Ruppin (2008), Network-based prediction of human tissue-specific metabolism, *Nat Biotechnol*, *26*(9), 1003–10, doi:10.1038/nbt.1487.

Singer, P. C., and W. Stumm (1970), Acidic mine drainage: The rate-determining step, *Science*, *167*, 1121–1123, doi:10.1126/science.167.3921.1121.

Sogin, M. L., H. G. Morrison, J. A. Huber, D. Mark Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl (2006), Microbial diversity in the deep sea and the underexplored "rare biosphere", *Proc Natl Acad Sci U S A*, *103*(32), 12,115–20, doi:10.1073/pnas.0605127103.

Sroka, J., et al. (2011), Acorn: a grid computing system for constraint based modeling and visualization of the genome scale metabolic reaction networks via a web interface, *BMC Bioinformatics*, *12*, 196, doi:10.1186/1471-2105-12-196.

Steen, E. J., Y. Kang, G. Bokinsky, Z. Hu, A. Schirmer, A. McClure, S. B. Del Cardayre, and J. D. Keasling (2010), Microbial production of fatty-acid-derived fuels and chemicals from plant biomass, *Nature*, *463*(7280), 559–62, doi:10.1038/nature08721.

Stolyar, S., S. Van Dien, K. L. Hillesland, N. Pinel, T. J. Lie, J. A. Leigh, and D. A. Stahl (2007), Metabolic modeling of a mutualistic microbial community, *Mol Syst Biol*, *3*, 92, doi:10.1038/msb4100131.

Taffs, R., et al. (2009), In silico approaches to study mass and energy flows in microbial consortia: a syntrophic case study, *BMC Syst Biol*, *3*, 114, doi:10.1186/1752-0509-3-114.

Tannock, G. W. (1977), Characteristics of bacteroides isolates from the cecum of conventional mice, *Appl Environ Microbiol*, *33*(4), 745–50.

Tatusov, R. L., et al. (2003), The cog database: an updated version includes eukaryotes, *BMC Bioinformatics*, *4*, 41, doi:10.1186/1471-2105-4-41.

Thiele, I., and B. Ø. Palsson (2010), A protocol for generating a high-quality genome-scale metabolic reconstruction, *Nat Protoc*, *5*, 93–121, doi:10.1038/nprot.2009.203.

Thiele, I., N. D. Price, T. D. Vo, and B. Ø. Palsson (2005), Candidate metabolic network states in human mitochondria. impact of diabetes, ischemia, and diet, *J Biol Chem*, *280*, 11,683–95, doi:10.1074/jbc.M409072200.

Thorleifsson, S., and I. Thiele (2011), rbionet: A cobra toolbox extension for reconstructing high-qualuty biochemical networks, *Bioinformatics*.

Titgemeyer, E. C., L. D. Bourquin, G. C. Fahey, Jr, and K. A. Garleb (1991), Fermentability of various fiber sources by human fecal bacteria in vitro, *Am J Clin Nutr*, *53*(6), 1418–24.

Tollis, E. T. P. P. I. G., and M. Reczko (2009), Computational identification of bacterial communities, *Int J Biol Life Sci*, *1*, 185–191.

Turnbaugh, P. J., R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon (2007), The human microbiome project, *Nature*, *449*(7164), 804–10, doi: 10.1038/nature06244.

Turnbaugh, P. J., et al. (2009), A core gut microbiome in obese and lean twins, *Nature*, *457*(7228), 480–4, doi:10.1038/nature07540.

Turroni, F., A. Ribbera, E. Foroni, D. van Sinderen, and M. Ventura (2008), Human gut microbiota and bifidobacteria: from composition to functionality, *Antonie Van Leeuwenhoek*, *94*(1), 35–50, doi:10.1007/s10482-008-9232-4.

Tyson, G. W., I. Lo, B. J. Baker, E. E. Allen, P. Hugenholtz, and J. F. Banfield (2005), Genome-directed isolation of the key nitrogen fixer *Leptospirillum ferrodiazotrophum* sp. nov. from an acidophilic microbial community., *Appl Environ Microbiol*, *71*, 6319–6324, doi:10.1128/AEM.71.10.6319-6324.2005.

Tyson, G. W., et al. (2004), Community structure and metabolism through reconstruction of microbial genomes from the environment, *Nature*, *428*, 37–43, doi:10.1038/nature02340.

Tzamali, E., P. Poirazi, I. G. Tollis, and M. Reczko (2011), A computational exploration of bacterial metabolic diversity identifying metabolic interactions and growth-efficient strain communities, *BMC Syst Biol*, *5*, 167, doi:10.1186/1752-0509-5-167.

Van de Merwe, J. P., A. M. Schröder, F. Wensinck, and M. P. Hazenberg (), The obligate anaerobic faecal flora of patients with crohn's disease and their first-degree relatives, *Scand J Gastroenterol*, *23*(9), 1125–31.

Varma, A., B. W. Boesch, and B. O. Palsson (1993a), Stoichiometric interpretation of escherichia coli glucose catabolism under various oxygenation rates, *Appl Environ Microbiol*, *59*(8), 2465–73.

Varma, A., B. W. Boesch, and B. O. Palsson (1993b), Biochemical production capabilities of escherichia coli, *Biotechnol Bioeng*, *42*(1), 59–73, doi:10.1002/bit.260420109.

Venter, J. C., et al. (2004), Environmental genome shotgun sequencing of the Sargasso Sea, *Science*, *304*, 66–74, doi:10.1126/science.1093857.

Ventura, M., F. Turroni, C. Canchaya, E. E. Vaughan, P. W. O'Toole, and D. van Sinderen (2009), Microbial diversity in the human intestine and novel insights from metagenomics, *Front Biosci*, *14*, 3214–21.

Wensinck, F., C. van Lieshout, P. A. Poppelaars-Kustermans, and A. M. Schröder (1981), The faecal flora of patients with crohn's disease, *J Hyg (Lond)*, *87*(1), 1–12.

Wiechert, W. (2001), 13c metabolic flux analysis, *Metab Eng*, *3*(3), 195–206, doi:10.1006/mben.2001.0187.

Wikoff, W. R., A. T. Anfora, J. Liu, P. G. Schultz, S. A. Lesley, E. C. Peters, and G. Siuzdak (2009), Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites, *Proc Natl Acad Sci U S A*, *106*(10), 3698–703, doi:10.1073/pnas.0812874106.

Wintermute, E. H., and P. A. Silver (2010), Emergent cooperation in microbial metabolism, *Mol Syst Biol*, *6*, 407, doi:10.1038/msb.2010.66.

Wright, J., and A. Wagner (2008), The systems biology research tool: evolvable open-source software, *BMC Syst Biol*, *2*, 55, doi:10.1186/1752-0509-2-55.

Wu, J., Y. An, J. Yao, Y. Wang, and H. Tang (2010), An optimised sample preparation method for nmr-based faecal metabonomic analysis, *Analyst*, *135*(5), 1023–30, doi:10.1039/b927543f.

Wunderlich, Z., and L. A. Mirny (2006), Using the topology of metabolic networks to predict viability of mutant strains, *Biophys J*, *91*(6), 2304–11, doi:10.1529/biophysj.105.080572.

Xu, J., et al. (2007), Evolution of symbiotic bacteria in the distal human intestine, *PLoS Biol*, *5*(7), e156, doi:10.1371/journal.pbio.0050156.

Zheng, X., et al. (2011), The footprints of gut microbial-mammalian co-metabolism, *J Proteome Res*, *10*(12), 5512–22, doi:10.1021/pr2007945.

Zhuang, K., M. Izallalen, P. Mouser, H. Richter, C. Risso, R. Mahadevan, and D. R. Lovley (2011), Genome-scale dynamic modeling of the competition between rhodoferax and geobacter in anoxic subsurface environments, *ISME J*, *5*(2), 305–16, doi:10.1038/ismej.2010.117.

Zomorrodi, A. R., and C. D. Maranas (2012), Optcom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities, *PLoS Comput Biol*, *8*(2), e1002,363, doi:10.1371/journal.pcbi.1002363.

Zubkov, M., G. Tarran, and B. Fuchs (2004), Depth related amino acid uptake by prochlorococcus cyanobacteria in the southern atlantic tropical gyre, *FEMS Microbiology Ecology*.