

Washington University in St. Louis  
**Washington University Open Scholarship**

---

Engineering and Applied Science Theses &  
Dissertations

McKelvey School of Engineering

---


Spring 5-15-2016

# Revelation of Yin-Yang Balance in Microbial Cell Factories by Data Mining, Flux Modeling, and Metabolic Engineering

Gang Wu

*Washington University in St. Louis*

Follow this and additional works at: [https://openscholarship.wustl.edu/eng\\_etds](https://openscholarship.wustl.edu/eng_etds)

 Part of the [Biodiversity Commons](#), [Computer Sciences Commons](#), [Engineering Commons](#), and the [Evolution Commons](#)

---

## Recommended Citation

Wu, Gang, "Revelation of Yin-Yang Balance in Microbial Cell Factories by Data Mining, Flux Modeling, and Metabolic Engineering" (2016). *Engineering and Applied Science Theses & Dissertations*. 163.  
[https://openscholarship.wustl.edu/eng\\_etds/163](https://openscholarship.wustl.edu/eng_etds/163)

This Dissertation is brought to you for free and open access by the McKelvey School of Engineering at Washington University Open Scholarship. It has been accepted for inclusion in Engineering and Applied Science Theses & Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact [digital@wumail.wustl.edu](mailto:digital@wumail.wustl.edu).

WASHINGTON UNIVERSITY IN ST. LOUIS

School of Engineering and Applied Science

Department of Energy, Environmental & Chemical Engineering

Dissertation Examination Committee:

Yinjie J. Tang, Chair

Pratim Biswas

Zi Chen

Cynthia Lo

Tae Seok Moon

Ryan Senger

Revelation of 'Yin-Yang' Balance in Microbial Cell Factories by Data Mining, Flux Modeling,  
and Metabolic Engineering

by

Gang Wu

A dissertation presented to the  
Graduate School of Arts & Sciences  
of Washington University in  
partial fulfillment of the  
requirements for the degree  
of Doctor of Philosophy

May 2016  
Saint Louis, Missouri

© 2016, Gang Wu

## Table of Contents

<b>List of Figures .....</b>	<b>ix</b>
<b>List of Tables.....</b>	<b>xiv</b>
<b>Acknowledgement.....</b>	<b>iii</b>
<b>Abstract.....</b>	<b>xvi</b>

### **Chapter 1: Introduction of Fluxomics Studies and Metabolic Burden**

<b>Modeling.....</b>	<b>1</b>
1.1. Background of fluxomics .....	1
1.2. A brief overview of FBA.....	1
1.3. Introduction of <sup>13</sup> C-MFA.....	1
1.4. Overview of published <sup>13</sup> C-MFA papers on bacteria species .....	6
1.5. Modeling work related with metabolic burden.....	9
1.6. Outline of this disserttation .....	13
1.7. Reference.....	14

### **Chapter 2: Evaluating Physiological State of Engineered *E. coli* Strains by**

<b>Isotopomer Constrained Flux Balance Analysis .....</b>	<b>22</b>
2.1. Abstract.....	22
2.2. Introduction.....	23
2.3. Materials and Methods .....	26



2.3.1. Strains and plasmids .....	26
2.3.2. Medium and culture conditions .....	27
2.3.3. Quantification of biomass and extracellular metabolites .....	27
2.3.4. Mass isotopomer distribution of proteinogenic amino acids .....	28
2.3.5. Central metabolic flux determined by <sup>13</sup> C-MFA .....	29
2.3.6. Genome-scale model constrained by <sup>13</sup> C-MFA flux .....	30
2.4. Results and discussions .....	31
2.4.1. Physiological states of different Strains .....	31
2.4.2. Central metabolic flux determined by <sup>13</sup> C-MFA .....	32
2.4.3. Energy metabolism and evolutionary fitness analysis .....	33
2.4.4. The effects of P/O ratio, oxygen flux, and maintenance energy on isobutanol production.	34
2.5. Conclusions.....	36
2.6. References.....	36

**Chapter 3: Investigate Energy Metabolism of Microbial Cell Factories by Yin-Yang Theory Research .....60**

3.1. Abstract.....	60
3.2. Introduction.....	61
3.3. The energy losses in microbial cell factories .....	62
3.4. The tradeoff between product yield and energy fitness.....	64
3.5. Sensitivity analysis of the energy penalty on biofuel synthesis.....	66

3.6 Yin-Yang theory in metabolic engineering .....	68
3.7. Conclusions.....	73
3.8. References.....	73

## **Chapter 4: Enhance Energy State of Fatty Acid Producing *E. coli* Strains**

### **With *Vitreoscilla* Hemoglobin .....87**

4.1. Abstract.....	87
4.2. Introduction.....	87
4.3. Experimental and Methods .....	90
4.3.1. Chemicals and Strains .....	90
4.3.2. Plasmid construction .....	91
4.3.3. Medium and culture conditions .....	92
4.3.4. Fatty acid measurement .....	92
4.3.5. Simulate cellular physiology with flux balance model .....	93
4.4. Results and discussions .....	94
4.4.1. Growth kinetics and fatty acid production .....	99
4.4.2. Expression of VHb affects the degree of unsaturation of free fatty acid .....	95
4.4.3. Effect of oxygen and maintenance energy on fatty acid production .....	96
4.5. Conclusions.....	98
4.6. References.....	98

## **Chapter 5: ... Build Web-Based Platform for Fluxomics Studies: Microbesflux**

### **Rebuild and Website Development .....112**

5.1. Abstract.....	112
5.2. Introduction.....	113
5.3. Implementation .....	115
5.3.1. MicrobesFlux update .....	115
5.3.2. New features of reloaded MicrobesFlux.....	116
5.3.3. Development of websites for fluxomics studies .....	116
5.4. Results .....	116
5.5. Availability and requirements.....	117
5.6. References.....	118

## **Chapter 6: .Rapid Prediction of Bacterial Fluxomics Using Machine Learning**

### **and Constraint Programming .....125**

6.1. Abstract.....	125
6.2. Authors' Summary .....	126
6.3. Introduction .....	126
6.4. Methods .....	129
6.4.1. Data collection and preprocessing .....	129
6.4.2. Feature selection and scaling.....	130
6.4.3. Machine learning algorithm selection.....	131

6.4.4.	Error evaluation and cross validation.....	132
6.4.5.	Stoichiometric constraints and boundary .....	132
6.4.6.	Flux adjustment using stoichiometric constraints.....	134
6.4.7.	Constraint programming and input checking .....	136
6.4.8.	Overall system design .....	136
6.5.	Results and Discussion .....	137
6.5.1.	Pathway map and statistical analysis results.....	137
6.5.2.	Optimization of algorithm and parameters .....	138
6.5.3.	Flux correction by quadratic programming .....	139
6.5.4.	MFlux case studies .....	140
6.5.5.	Compare flux balance analysis with MFlux for <i>E. coli</i> metabolism.....	142
6.5.6.	Perspective of metabolic robustness and machine learning of fluxome patterns.....	143
6.5.7.	Limitations of machine learning .....	145
6.6.	Conclusion .....	146
6.7.	Supporting information.....	147
6.8.	References.....	147

## **Chapter 7: Enable Fast Literature Analysis by Text Mining And Big Data**

<b>Technology.....</b>	<b>167</b>
7.1. Abstract.....	167
7.2. Introduction.....	168

7.3. Methods .....	169
7.3.1. Database availability and record structure.....	169
7.3.2. Text mining methods.....	170
7.3.3. Fast search via BigQuery.....	171
7.3.4. Data analysis .....	172
7.4. Results and discussions .....	173
7.4.1. Most related words of ‘metabolic engineering’, ‘environmental engineering’, ‘synthetic biology’, ‘systems biology’, and ‘metabolic flux’ .....	173
7.4.2. Compare the difference and similarity between two different terms .....	174
7.4.3. Identify the developing trend of a specific term .....	175
7.4.4. Advantages and Limitations of Big Data workflow.....	177
7.5. Conclusions.....	178
7.6. References .....	178
<b>Chapter 8: Conclusions And Future Perspectives .....</b>	<b>192</b>
8.1. Conclusions .....	192
8.2. Future directions to solve intracellular energy crisis .....	194
8.3. Personal views on the future of microbial cell factories .....	198
8.4. References .....	199
<b>9. Appendix .....</b>	<b>204</b>
9.1. Appendix I .....	204

9.2. Appendix II .....	206
<b>10.Side Projects .....</b>	<b>276</b>
10.1. Phytotoxicity of metal oxide nanoparticle (NP) on plant seeds.....	276
10.2. NP electrospray facilitates seed germination.....	388
<b>Vita .....</b>	<b>322</b>

## List of Figures

Figure 1.1 Procedure of GSM reconstruction and FBA .....	3
Figure 1.2 A general procedure of $^{13}\text{C}$ -MFA.....	5
Figure 1.3 Percentage of $^{13}\text{C}$ -MFA paper on each bacteria species.....	7
Figure 1.4 Number of $^{13}\text{C}$ -MFA papers published on each journal.....	8
Figure 2.1 Diagram of $^{13}\text{C}$ flux as constraints for FBA .....	48
Figure 2.2 Central metabolic flux of each strain determined by $^{13}\text{C}$ -MFA .....	49
Figure 2.3a-b Energy analysis of four strains at different energy conditions .....	50
Figure 2.4 Evolutionary fitness of four strains in this study.....	54
Figure 2.5a-e: Influence of P/O ratio and maintenance energy on isobutanol production potential (growth rate for strain 1, JCL260), simulated by FBA .....	54
Figure 2.6a-e: Influence of oxygen uptake flux and maintenance energy on isobutanol production potential (growth rate for strain 1, JCL260), simulated by FBA.....	57
Figure 3.1 Cell carbon and energy metabolism illustrated by Yin-Yang Theory.....	80
Figure 3.2a-d: Genome-scale FBA models for microbial biofuel mole-carbon yields from glucose.....	81
Figure 3.3 Energy fitness and productivities in microbial cell factories. ....	84

Figure 3.3a: The trend of metabolic entropy changes (unit: ATP generation per glucose). In optimal metabolism, one mole of glucose generates 38 ATP for biosynthesis. Under constraints of P/O ratios and maintenance loss, less ATP can be generated (i.e., increase of metabolic entropy). .....	84
Figure 3.3b: The transition from carbon limitation to energy limitation with the increase of product yield. In many cases, the energy limitation prevents strains from achieving the yield and titer at break-even point.....	85
Figure 3.3c: Cascade of energy changes (Heat of combustion) during biofuel synthesis from glucose. ....	86
Figure 4.1 Structure of <i>Vitreoscilla</i> hemoglobin in active dimer form .....	106
Figure 4.2 Genetic manipulations to insert VHb into pA58c-TR.....	107
Figure 4.3 Optimization of IPTG concentrations for VHb50 expression .....	108
Figure 4.4 Growth curve for three fatty acid producing strains .....	108
Figure 4.5 Fatty acid productions after 24 hr of IPTG Induction .....	109
Figure 4.6 Free fatty acid production profiles for control strain and VHb strain after (a) 8 hr and (b) 24 hr of IPTG induction .....	110
Figure 4.7 Effect of oxygen flux and maintenance energy on fatty acid yield at (a) exponential phase (b) late exponential phase .....	111
Figure 5.1 <sup>13</sup> C-MFA protocol and sources of flux analysis variance.....	121



Figure 5.2 The webpage of our platform for comprehensive fluxomics studies (<http://fluxomics.net>).....122

Figure 5.3 The webpage of WUFlux (<http://13cmfa.org>), which can be accessed and freely download.....123

Figure 5.4 The webpage of Amazon server EC2, (a) all Amazon web services, (b) buckets of our websites .....124

Figure 6.1 A universal central metabolic pathway for bacteria: The central carbon metabolic pathway is simplified into 29 fluxes in MFlux.....157

Figure 6.2 Statistical analysis of central metabolic fluxes collected in our database. “Flux range” represents variations of each fluxes among <sup>13</sup>C-MFA database; “95% confidence interval” represents 95% of flux data were within a small range; “Average flux value” are the mean of flux values from <sup>13</sup>C-MFA database.....158

Figure 6.3 A flow chart of MFlux algorithm. This diagram is to illustrate the detailed procedures for our algorithm. ....159

Figure 6.4 A comparison of three different algorithms: SVM, kNN, and decision tree: The best cross-validation results on 29 fluxes are compared. All tests in this step were performed only on the WT database.....160

Figure 6.5 Best results by SVM for WT and WP databases. Both the linear and the RBF kernels are considered in a grid search, and the results from WP database is much better than from the WT database.....161

Figure 6.6 A comparison between the linear kernel and the RBF kernel for SVM. The results are quite similar.....162

Figure 6.7 Summary of root mean squared error (RMSE) from 20 case studies: averaged flux from <sup>13</sup>C-MFA database; machine learning, and MFlux. The average RMSE is 7.7 from machine learning alone and 5.6 from MFlux.....163

Figure 6.8 A comparison of <sup>13</sup>C-MFA, the flux predicted by ML, and the flux predicted by MFlux in case 8. *B. subtilis* was incubated in a shake flask (37 °C, 300 rpm, aerobic condition), and supplied with labeled succinate and glutamate as carbon sources in M9 minimal medium.....164

Figure 6.9 A comparison of <sup>13</sup>C-MFA, the flux predicted by ML flux, and the flux predicted by MFlux. *G. thermoglucosidasius* M10EXG was incubated in sealed bottles (micro-aerobic condition), supplied with glucose as a carbon source.  $RMSE_{ML} = 4.0$ ,  $RMSE_{MFlux} = 3.0$ .....165

Figure 6.10 A comparison of <sup>13</sup>C-MFA, MFlux and the flux predicted by FBA. FBA Analysis is simulated by *E. coli* iJO1366 model with defaulted boundary settings from the reference (Orth *et al.* 2011). (A) *E. coli* fluxome of glucose metabolism was precisely measured via parallel labeling experiments (a recent paper not in our database) (Crown *et al.* 2015).  $RMSE_{FBA} = 11.3$ ,  $RMSE_{MFlux} = 6.5$ . (B) *E. coli* fluxome of glycerol and glucose co-metabolism were measured by Dr. Yao and Dr. Shimizu (unpublished data). *E. coli* strain was cultured in chemostat fermentor with a working volume of 1 L (37°C). The dilution rates in the continuous culture were 0.35 h<sup>-1</sup>. [1-<sup>13</sup>C] glucose and [1, 3-<sup>13</sup>C] glycerol were used for tracer experiments. The flux calculation is

based on previous method (Fong *et al.* 2006; Peng *et al.* 2004).  $RMSE_{FBA} = 22.5$ ,  $RMSE_{MFlux} = 5.1$ .....166

Figure 7.1a Total number of journals in the database at different time.....185

Figure 7.1b Total number of papers in the database increase at different time.....185

Figure 7.2 Structure of nxml file.....186

Figure 7.3 Word cloud of ‘metabolic engineering’.....187

Figure 7.4 Word cloud of ‘environmental engineering’.....188

Figure 7.5 Word cloud of ‘synthetic biology’.....189

Figure 7.6 Word cloud of ‘systems biology’.....190

Figure 7.7 Word cloud of ‘metabolic flux’.....191

## List of Tables

Table 2.1 Detailed information of plasmids and strains used in this study.....	44
Table 2.2 Physiological information of strains used in this study.....	44
Table 2.3 Metabolic network for <sup>13</sup> C-MFA calculation.....	46
Table 2.4 Energy metabolism of different strains.....	47
Table 4.1 DNA sequence of all primers used in this work.....	105
Table 6.1 Summary of 20 cases of study.....	164
Table 7.1 Comparison of ‘metabolic engineering’ and ‘synthetic biology’, similarity 69.6%.....	182
Table 7.2 Development trend of ‘metabolic engineering’ during 2000 ~ 2009 and 2010 ~ 2015, similarity 81%.....	183
Table 7.3 Development trend of ‘biofuel’ during 2000 ~ 2009, 2010 ~ 2015, and 2015 similarity 72.4% and 88%.....	184

## Acknowledgements

During the past four years, I would like to convey my sincere gratitude to my adviser, Dr. Yinjie J. Tang, for bringing me into this PhD program, teaching me so much and providing tremendous supports on my PhD research work. I would also like to express my appreciation to all the members in my committee: Dr. Pratim Biswas, Dr. Cynthia Lo, Dr. Tae Seok Moon, Dr. Zi Chen, Dr. Ryan Senger, and Dr. Venkat Subramanian (previous committee member) for their helpful suggestions and guidance.

Many thanks go to lab colleagues in our group, including Lian He, Whitney Hollinshead, Mary Abernathy, Tolutola Oyetunde, alumni Dr. Le You, Dr. Arul Varman, Dr. Yi Xiao, and visiting Professors, Dr. Yi Yu, Dr. Haifeng Hang, Dr. Lifeng Peng, for their countless assistances in research and friendship. Thanks also extend to Patty Wurm for laboratory supports, Tim McHugh for helping with remote computing and video. I also would like to thank all staffs in EECE department, who make the department a wonderful place to work and study. Further, I would like to thank James Ballard at writing center of engineering school for help in my writings.

Besides, I would like to thank my colleague Di Liu, Wen Jiang for helping the VHB experiments, Dr. Forrest Sheng Bao at University of Akron for numerous discussions and helps for ThunderFlux project, and Dr. Eric You Xu for many insightful suggestions for MicrobesFlux reconstruction. My sincere thanks go to Dr. Shi Chen, Dr. Cristina Lanzas, Dr. Alison Buchan, and Dr. Louis Gross at the University of Tennessee for their encouragements during my visit at NIMBioS, and to Dr. Yi Sun, Dr. Yang Liu, Dr. Qunyu Zhang, and Dr. Fei Zhong for their helps during my summer intern at Sigma-Aldrich.

Also, I want to express my deepest gratitude to my parents and my wife, for their unending supports, encouragements, and love during my PhD study. Last but most important, nothing can be realized without the grace of the ONE, who saved me from the Death Valley, changed me through nine years of graduate life. All glory goes to Him.

Gang Wu

Nov 2015

## ABSTRACT OF THE DISSERTATION

Decipher ‘Yin-Yang’ Balance of Microbial Cell Factories by Data Mining, Flux Modeling, and  
Metabolic Engineering

By

Gang Wu

Doctor of Philosophy in Energy, Environmental, and Chemical Engineering

Washington University in St. Louis, 2015

Professor Yinjie Tang, Chair

The long-held assumption of never-ending rapid growth in biotechnology and especially in synthetic biology has been recently questioned, due to lack of substantial return of investment. One of the main reasons for failures in synthetic biology and metabolic engineering is the metabolic burdens that result in resource losses. Metabolic burden is defined as the portion of a host cell’s resources — either energy molecules (e.g., NADH, NADPH and ATP) or carbon building blocks (e.g., amino acids) — that is used to maintain the engineered components (e.g., pathways). As a result, the effectiveness of synthetic biology tools heavily depends on cell capability to carry on the metabolic burden. Although genetic modifications can effectively engineer cells and redirect carbon fluxes toward diverse products, insufficient cell ATP powerhouse is limited to support diverse microbial activities including product synthesis. Here, I employ an ancient Chinese philosophy (Yin-Yang) to describe two contrary forces that are interconnected and interdependent, where Yin represents energy metabolism in the form of ATP, and Yang represents carbon metabolism. To decipher “Yin-Yang” balance and its implication to microbial cell factories, this dissertation applied metabolic engineering, flux analysis, data

mining tools to reveal cell physiological responses under different genetic and environmental conditions.

Firstly, a combined approach of FBA and  $^{13}\text{C}$ -MFA was employed to investigate several engineered isobutanol-producing strains and examine their carbon and energy metabolism. The result indicated isobutanol overproduction strongly competed for biomass building blocks and thus the addition of nutrients (yeast extract) to support cell growth is essential for high yield of isobutanol. Based on the analysis of isobutanol production, 'Yin-Yang' theory has been proposed to illustrate the importance of carbon and energy balance in engineered strains. The effects of metabolic burden and respiration efficiency (P/O ratio) on biofuel product were determined by FBA simulation. The discovery of 'energy cliff' explained failures in bioprocess scale-ups. The simulation also predicted that fatty acid production is more sensitive to P/O ratio change than alcohol production. Based on that prediction, fatty acid producing strains have been engineered with the insertion of *Vitreoscilla* hemoglobin (VHb), to overcome the intracellular energy limitation by improving its oxygen uptake and respiration efficiency. The result confirmed our hypothesis and different level of trade-off between the burden and the benefit from various introduced genetic components. On the other side, a series of computational tools have been developed to accelerate the application of fluxomics research. Microbesflux has been rebuilt, upgraded, and moved to a commercial server. A platform for fluxomics study as well as an open source  $^{13}\text{C}$ -MFA tool (WUFlux) has been developed. Further, a computational platform that integrates machine learning, logic programming, and constrained programming together has been developed. This platform gives fast predictions of microbial central metabolism with decent accuracy. Lastly, a framework has been built to integrate Big Data technology and text mining to interpret concepts and technology trends based on the literature survey. Case studies have been



performed, and informative results have been obtained through this Big Data framework within five minutes.

In summary,  $^{13}\text{C}$ -MFA and flux balance analysis are only tools to quantify cell energy and carbon metabolism (i.e., Yin-Yang Balance), leading to the rational design of robust high-producing microbial cell factories. Developing advanced computational tools will facilitate the application of fluxomics research and literature analysis.

# CHAPTER ONE

## INTRODUCTION OF FLUXOMICS STUDIES AND METABOLIC BURDEN MODELING

### 1.1. Background of fluxomics

Systems biology reveals intricate cellular metabolic and regulatory activities by a series of high-throughput methods. Development and application of those high-throughput methods raise up their respective research field defined as ‘omics’, including genomics (sequencing and annotation of genomic DNA), transcriptomics (determination of global gene transcriptional level), proteomics (characterization of structure and function of individual protein), metabolomics (assay of metabolite profile), and fluxomics (infer rate of each single biochemical reaction within metabolic network) (Tang *et al.* 2009a).

In the realm of fluxomics,  $^{13}\text{C}$  Metabolic Flux Analysis ( $^{13}\text{C}$ -MFA) and Flux Balance Analysis (FBA) are basic quantitative approaches to unveil activities of metabolic reactions. Both approaches are built upon many assumptions, for instance, steady state or quasi-steady state assumption (which indicates there is no net flux of those intermediates, and the sum of input fluxes equals to sum of output fluxes), homogenous cell culture assumption (local environment and metabolic state of each cell is considered to be equally same). During the exponential growth phase of a well cultivated single cell organism culture, these assumptions are reasonable for most times and modeling calculations based on them seem relatively accurate.

### 1.2. A brief overview of FBA

FBA is a bio-mathematical approach to calculate metabolic flux profiles and has been extensively developed during recent two decades. By building up genome-scale metabolic network model, FBA is able to calculate large scale models with more than 2000 reactions which include over 30% of genes of the whole genome. (Orth *et al.* 2011; Monk *et al.* 2013) As a powerful mathematical modeling tool, FBA is able to make predictions on growth rate, product yield, nutrient requirements, physiology of knockout phenotype, track extreme pathways, and guide rational metabolic engineering based on only a few inputs. (Kauffman *et al.* 2003; Edwards *et al.* 2002; Orth *et al.* 2010a; Becker *et al.* 2007; Bordbar *et al.* 2014) Therefore, it became popular and widely accepted by researchers of diverse fields (Tang *et al.* 2009a). Until now, Genome scale model (GSM) reconstruction has been performed for more than one hundred species (<http://systemsbiology.ucsd.edu/InSilicoOrganisms/OtherOrganisms>) including bacteria, eukaryotic, and archaeal species and this number is still increasing now (Orth *et al.* 2010a). Meanwhile, many efforts have been made to integrate FBA model with other omics data (e.g., transcriptomics, proteomics, and metabolomics) (Gowen and Fong 2010; Schellenberger *et al.* 2011; Åkesson *et al.* 2004; Coquin *et al.* 2008; Winter and Krömer 2013).

A general procedure for GSM reconstruction and FBA is depicted in Figure 1.1.: first collecting information from pathway database and identify existing genes and pathways, then performing gap filling and patching the gaps among metabolic network of GSM. The stoichiometry matrix of metabolic reactions and list of metabolites will be extracted out subsequently. Constrained linear programming is carried out based on the guidance of objective function and flux profile will be obtained. Maximization of biomass growth is the most common objective function and has been proven to be reasonable in many case studies, especially for those microbes in exponential growth phase (Orth *et al.* 2010a; Khannapho *et al.* 2008).

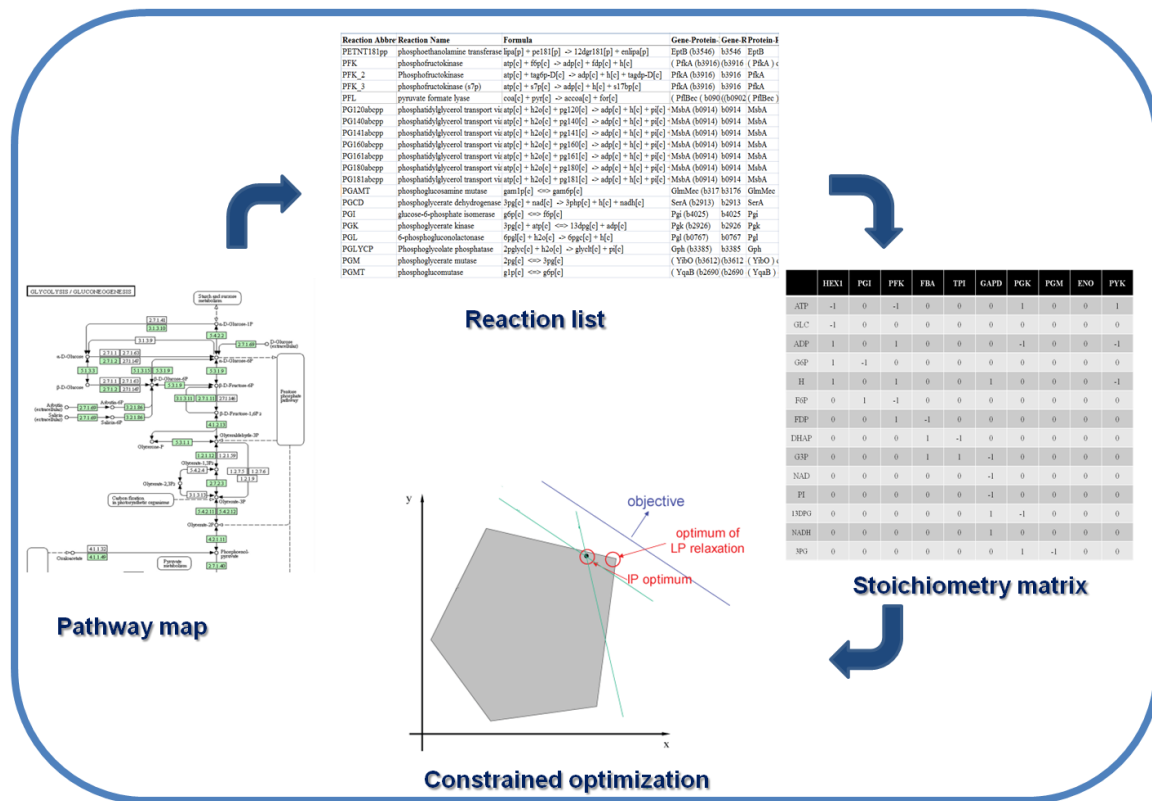


Figure 1.1 Procedure of GSM reconstruction and FBA

Calculation of FBA is a LP (Linear Programming) process. In general, there are much more fluxes than metabolites in FBA, resulting in large freedom of solution space and underdetermined system (Vallino and Stephanopoulos 1993). Even with a defined objective function, a series of constraints (e.g., thermodynamic properties, exchange/overflow fluxes, pH balance etc) and other information (e.g., biomass composition) are essential for meaningful simulation results of cellular metabolism. Among those constraints, overflow measurements are regularly used to define FBA boundaries. Precise measurements of some overflows require researchers developing professional skills in instruments such as HPLC. Similarly, biomass composition determination is expensive and labor intensive (Tang *et al.* 2009a). Recently, lots of efforts have been made to develop effective constraints to relieve the tiresome work of measurement, including addition of flux ratio on critical nodes (FBrAtio) (Yen *et al.* 2013;

McAnulty *et al.* 2012), linking with other omics data such as RNAseq or microarray data (as mentioned above), specific proton flux (SPF) (Senger and Papoutsakis 2008), enzyme activity assay, or even the flux values calculated from  $^{13}\text{C}$ -MFA results (Blank *et al.* 2005).

### 1.3. Introduction of $^{13}\text{C}$ -MFA

As an alternative tool for metabolic flux analysis,  $^{13}\text{C}$ -MFA adopts the information from labeling experiments as the constraints for modeling calculation and takes a different procedure to obtain the flux profile (Shown in Figure 1.2.).  $^{13}\text{C}$  labeling experiment is an important part in  $^{13}\text{C}$ -MFA. Labeled substrates are added into minimal medium to feed cells, and biomass is harvested when a metabolic & isotopic steady state is reached. Harvested biomass is pretreated and derivatized for further analysis (NMR or GC-MS). Proteinogenic amino acids can be analyzed via GC-MS measurement following a well-developed protocol (Fischer and Sauer 2003; Zamboni *et al.* 2009). The advantages of proteinogenic amino acids approach lie in its easy manipulation, fast process, relatively high accuracy, high robustness and relatively low requirements on instruments. Therefore, it can be potentially developed into an automatic process (e.g., use robot). Using labeling information from central metabolites requires different derivatization methods (You *et al.* 2014). Some unstable metabolites requires fast quenching method, and further analysis of many metabolites needs LC MS-MS (Young *et al.* 2011) , which provides more analytic power yet costs much more than GC-MS.

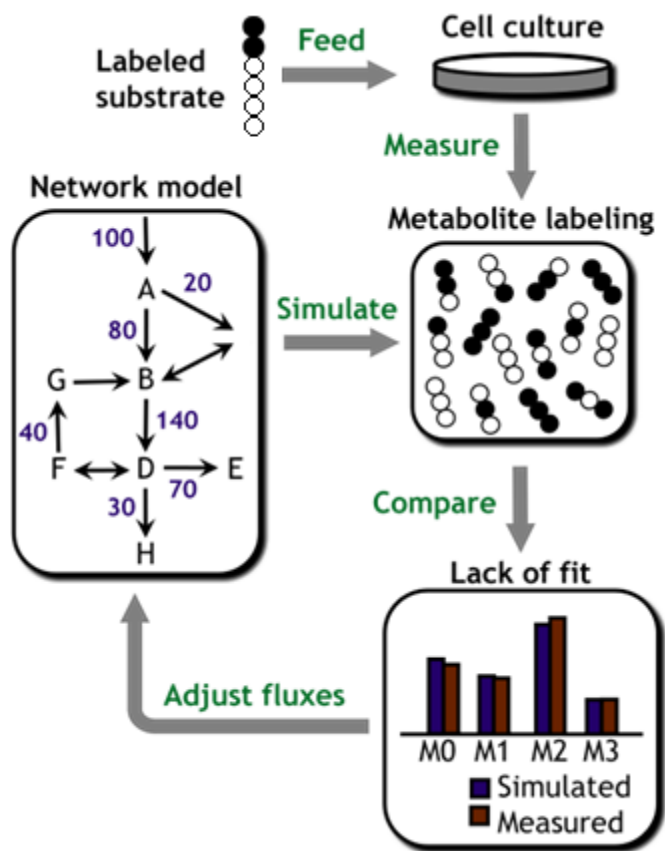


Figure 1.2. A general procedure of  $^{13}\text{C}$ -MFA

Normally mass isotopomer distribution data from 16 amino acids can be used for flux profile calculation (Tang *et al.* 2009a). The objective function of  $^{13}\text{C}$ -MFA is to minimize error between simulated and experimental determined mass isotopomer distribution of amino acids, which can be expressed as equation 1. The computational methods for  $^{13}\text{C}$ -MFA include the cumomer method (Möllney *et al.* 1999), isotopomer path tracing method (Forbes *et al.* 2001), and most recently, the Elementary Metabolic Unit (EMU) method (Antoniewicz *et al.* 2007a). Among them, EMU uses the minimum set of information to track atom transition among central metabolism, and is proven to be very efficient (Young *et al.* 2008).

$^{13}\text{C}$ -MFA has proven its competence in finding new pathway (Tang *et al.* 2007a), validation of gene functions (Tang *et al.* 2007a), medium design (Zhuang *et al.* 2011a), and profiling metabolism of engineered strain (He *et al.* 2014; Becker *et al.* 2011). However, there are only a few cases with the aid of  $^{13}\text{C}$ -MFA can improve yield of desired product, (Tang *et al.* 2012), and  $^{13}\text{C}$ -MFA is employed as a powerful tool of validation rather than prediction. Furthermore, it is still difficult to quickly determine flux profiles of strain library (hundreds of strains) to know the variances among different phenotypes at the flux level.

A series of computational tools have been developed in the field of fluxomics. Among them, COBRA toolbox developed by Palsson's group at UCSD is the most famous software (Becker *et al.* 2007; Schellenberger *et al.* 2011). COBRA is capable in many functions, including both FBA and  $^{13}\text{C}$ -MFA. In the field of  $^{13}\text{C}$ -MFA, published modeling framework includes 13C-FLUX and 13C-FLUX2 by Wiechert group (Wiechert *et al.* 2001; Weitzel *et al.* 2013), FiatFlux by Zamboni group (Zamboni *et al.* 2005a), Metran by Antoniewicz group, and INCA by Young group (Young 2014). All those software tools have greatly promote researches in the fluxomics field.

#### **1.4. Overview of published $^{13}\text{C}$ -MFA papers on prokaryotic species**

We collected most  $^{13}\text{C}$ -MFA papers on bacteria species published during the past twenty years (by Dec 2014). Through a brief survey, we found some important facts about  $^{13}\text{C}$ -MFA research:

(1) Fact of microbial species that  $^{13}\text{C}$ -MFA papers worked on: Most  $^{13}\text{C}$ -MFA papers are focusing on three model species: *E. coli*, *B. subtilis*, and *C. glutamicum*, which cover nearly 70% of our paper collections (shown in Figure 1.3.). This can be explained by that genetic

manipulations are so mature for those three model species, that many mutants and engineered strains have been created by researchers around the world. Also, there are lots of reports on metabolic network and biomass composition for those model species, researchers don't need to spend time or money on those experiments. For the rest 30% papers, they are on pathogenic species, environmental essential species, and chemical or fuel potential producers.

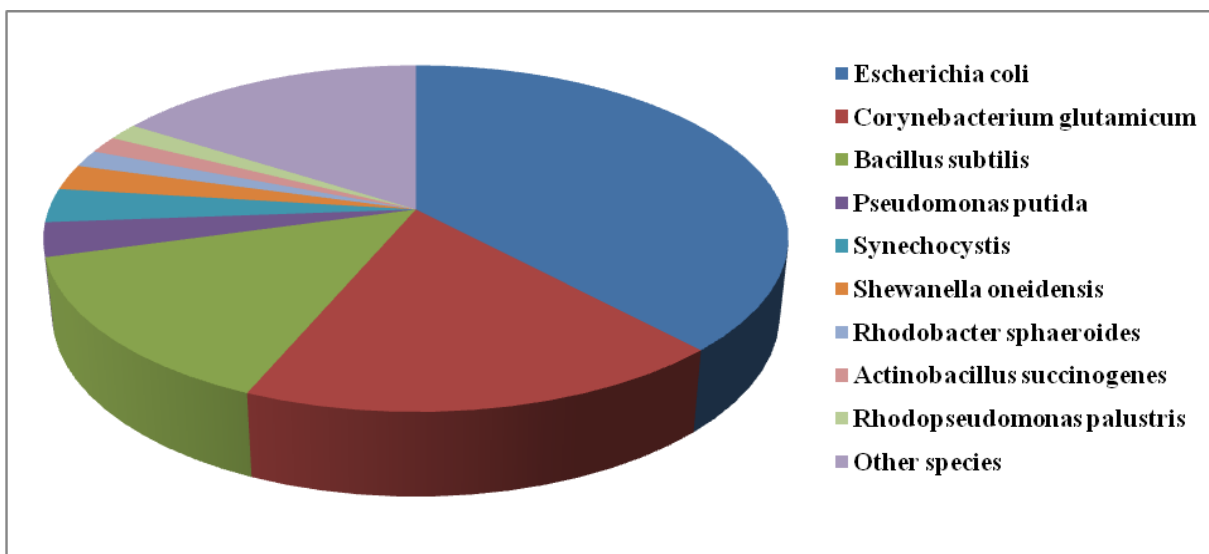


Figure 1.3 Percentage of  $^{13}\text{C}$ -MFA paper on each bacteria species

(2) Fact of scientific journals  $^{13}\text{C}$ -MFA papers published on: the top three journals that  $^{13}\text{C}$ -MFA published are 'Metabolic Engineering', 'Applied and Environmental Microbiology', and 'Biotechnology and Bioengineering' (as shown in Figure 1.4.). This result is very informative: people employed  $^{13}\text{C}$ -MFA as a complimentary tool of metabolic engineering and  $^{13}\text{C}$ -MFA do provide quantitative information of central carbon metabolism.



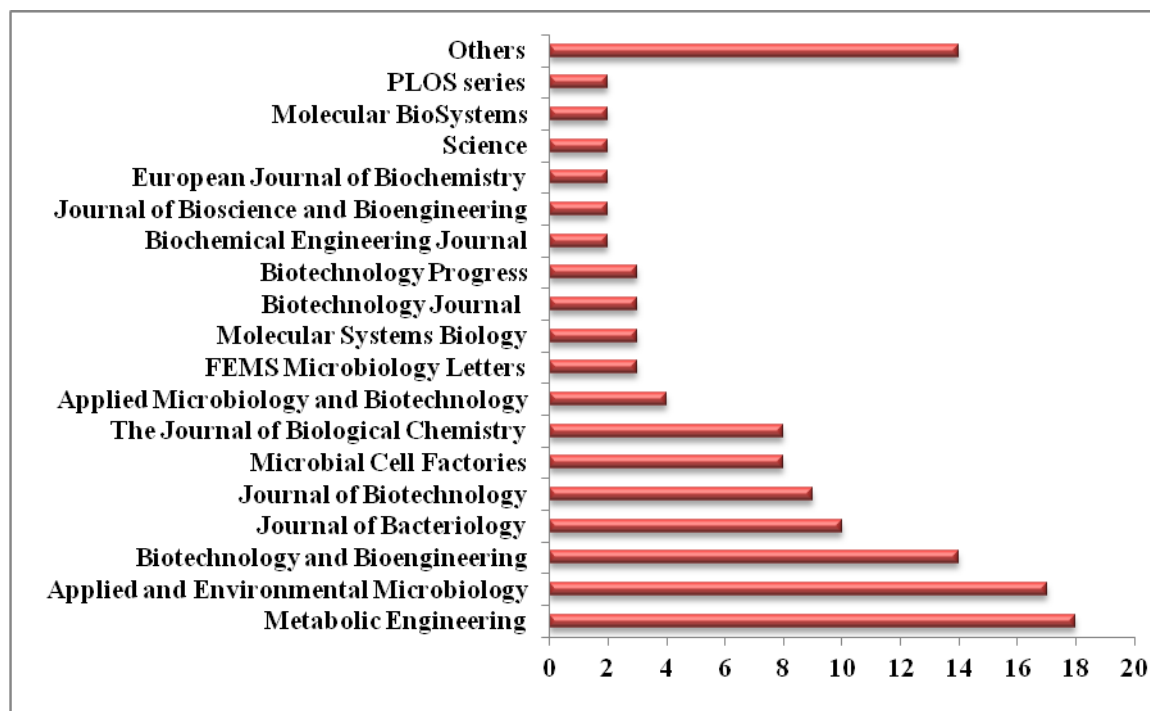


Figure 1.4 Number of  $^{13}\text{C}$ -MFA papers published on each journal

(3) Fact of researchers active in  $^{13}\text{C}$ -MFA field: the top three researchers published most  $^{13}\text{C}$ -MFA papers are Dr. Uwe Sauer (mainly work on *Bacillus subtilis* and *Escherichia coli*), Dr. Christoph Wittmann (mainly work on *Corynebacterium glutamicum*), and Dr. Kazuyuki Shimizu (mainly work on *Escherichia coli*).

### 1.5. Modeling work related with metabolic burden

Metabolic burden was first defined as ‘expression of foreign proteins utilize a significant amount of the host cell’s resources, removing those resources away from host cell metabolism and placing a metabolic load (or burden) on the host’ by Glick (Glick 1995). With product scope of genetic modifications extended to various chemicals; we also extend the concept of metabolic

burden to all cellular energy not used for biomass synthesis or product formation. There are six major sources of metabolic burden:

- (a) Defense of internal stress (e.g., imbalanced NADH/NADPH or NAD/NADH ratio, excess proton). A famous example is that isobutanol production leads to imbalance of cofactor utilization in *E.coli* (Bastian *et al.* 2011).
- (b) Defense of environmental stress (e.g., toxic compounds, O<sub>2</sub> stress). A good instance is that the fluxome of *Shewanella oneidensis* show robustness under salt stress (Tang *et al.* 2009b).
- (c) Cost of protein overexpression and plasmid maintenance (including the cost for turnover and protein incorrect folding). For instance, protein overexpression significantly boosted fluxes of the TCA cycle and acetate overflow (Heyland *et al.* 2011).
- (d) Defense of specific stress (toxicity) from metabolites or enzymes. A typical example is that overproduction of fatty acid will cause severe stress on cell membrane (Lennen *et al.* 2011).
- (e) Energy spilling reactions (e.g., futile cycle). Cells just waste the energy when the energy source is sufficient (Hoehler and Jorgensen 2013a; Russell and Cook 1995).
- (f) Energy for cell mobility, which only costs for 2% of total energy (Russell and Cook 1995).

Burden from protein expression has been noticed even over fifty years ago: Induced *E. coli* cells demonstrated a significant decrease in growth rate compared with uninduced cells (Novick and Weiner 1957). With the advance of genetic manipulation technology (e.g., restriction enzyme, DNA sequencing method) and successful commercialization of recombinant protein (e.g., insulin by Genentech) during late 1970s and early 1980s, overexpression of various proteins in order to get desired products has been attempted. The negative effects of protein and

plasmid burden were realized by a broad range of researchers (Schaaff *et al.* 1989; Jensen *et al.* 1993; Birnbaum and Bailey 1991). To better understand and simulate the impacts of metabolic burdens, a series of models were proposed or began to include the effect of burden into consideration. Ollis and Chang took the factor of plasmid instability into unstructured kinetic models; and their model is able to predict the effect of different inoculation ratio on final product formation (Ollis and Chang 1982). Lee and Bailey developed a model include the mathematical description of plasmid replication. And their model simulation results well matched the experimental observations that increased plasmid content leads to decrease growth rate. Further, the model also predicted the maximal intracellular product accumulation with respect to growth rate, which was simultaneously verified by experimental data.(Lee and Bailey 1984) In another work authored by Bailey, an empirical parabolic relationship was adopted to integrate with Monod equation to investigate the effects of different plasmid copies and various medium composition on beta-lactamase specific activity (Seo and Bailey 1985). In a later work, Bailey employed a structure model mathematically describing the competition between chromosomal- and plasmid-based expression system for cellular resource (e.g., transcription and translation machinery). This model well predicted the influences of different promoters and RBS strength on growth rate. In addition, the simulation result also revealed that the capability of intracellular transcriptional machinery was a limiting factor of heterogeneous gene expression (Peretti and Bailey 1987).

Other models rather than kinetic or structured model were also employed to simulate the effects of metabolic burden. For instance, Snoep *et al.* took the factor of protein burden into the metabolic control model by assigning a coefficient. Their model was able to appropriately explain the effects of glycolytic enzymes overexpression in *Zymomonas mobilis* (Snoep *et al.*

1995). An empirical model was also proposed to interpret protein heterologous expression in *E. coli*. This model was experimentally verified and the authors identified ribosome allocation as the limiting factor of growth during this process (Carrera *et al.* 2011; Somerville *et al.* 1994). A recent paper proposed a mechanistic model which considered three trade-offs on cellular resource (i.e., total protein, free ribosomes, and cellular energy) (Weiße *et al.* 2015). This model predicted well on cellular behaviors and the interaction between a synthetic circuit and its host. All those models provide reasonable quantifications of the effects by various factors related with the metabolic burden on cellular metabolism. However, it is still a challenge to apply those models (with their respective parameter set) directly into another system without extensive efforts on experiments and data fitting.

Flux balance model potentially provides an alternative strategy to quantify the metabolic burden. Weber *et al.* first employed FBA to investigate the effects of recombinant protein production on flux distribution and growth.(Weber *et al.* 2002) In their work, the amino acid composition of human basic fibroblast growth factor (hFGF-2) was considered and calculated as a specific flux. With maximization of biomass growth as the objective function, FBA successfully predicted that increased energy demand was satisfied by up-regulation of fluxes within EMP pathway and TCA cycle, as well as activated transhydrogenase flux. Later, Ozkan *et al.* included the effect of heterogeneous plasmid maintenance and antibiotics marker protein expression into FBA model to simulate the heterogeneous protein expression induced by IPTG in a minimal medium supplied with amino acids (Özkan *et al.* 2005). The authors claimed that the predicted relative flux distribution was in the same trend as reported expression profiles; however, there are some apparent variations between prediction and real values (e.g., P/O ratio).

Besides, the researchers didn't set any boundary of amino acid uptake fluxes, which might lead to further variations.

<sup>13</sup>C-MFA was also employed to interpret the metabolic burden of plasmid maintenance. (Wang *et al.* 2006a) The authors found that the strain hosting high copy plasmid had significantly lower relative flux in TCA cycle and higher flux in acetate secretion and ATP maintenance. The energy metabolism of engineered strains producing biofuel was also investigated by using <sup>13</sup>C-MFA. And the researchers found that the maintenance energy of fatty acid producing strain is two-fold of maintenance energy in the control strain (He *et al.* 2014). And this result partially explained numerous failures of biofuel production scale-up projects: high maintenance requirement leads to strain instability. To further explore the metabolism of strains under industrial fermentation (normally use complex medium rather than minimal medium), dynamic flux analysis and fast quenching method are required to resolve the metabolic flux and energy metabolism (Antoniewicz *et al.* 2007b; Zamboni 2011).

## **1.6. Outline of this dissertation**

In Chapter two, we integrated FBA and <sup>13</sup>C-MFA to investigate the metabolism of several isobutanol-producing strains. In particular, the energy metabolism, and several related factors such as the P/O ratio, the oxygen condition, and the maintenance energy were investigated.

In Chapter three, after reviewing recent successes and failures of metabolic engineering projects, we proposed the Yin-Yang theory of metabolic engineering: carbon metabolism and energy metabolic within microbial cell factories should be balanced. We employed FBA to predict the effects of maintenance energy (metabolic burden) and P/O ratio on biofuel yields in *E. coli*. We also provided several strategies to solve the energy bottleneck in engineered strains.

In Chapter four, we attempted to solve the energy bottleneck in fatty acid producing strain by insertion of *vhb* gene, which facilitates oxygen uptake of the host cell. We tested three VHB variants with different oxygen transfer capabilities. Compared with control strain, engineered strain with wild-type VHB only showed a decreased growth as well as reduced fatty acid production because genetic modification brought more metabolic burden than benefits; while strain with VHB50 demonstrated higher cell density, as well as increased fatty acid accumulation.

In Chapter five, we developed a series of modeling tools for fluxomics studies. First, we rebuilt MicrobesFlux on a commercial server (Amazon AWS) to make the systems more usable. Second, we also developed an open source  $^{13}\text{C}$ -MFA tool (WUFlux) in MATLAB. Third, we designed and developed a web-based platform, to make all our fluxomics tools freely accessed and downloaded through the Internet.

In Chapter six, we collected  $^{13}\text{C}$ -MFA data from published literature. Based on that information, we developed a web-based computational platform (MFlux) that directly predicts bacterial central metabolism via machine learning, constraint programming, and quadratic programming. We performed cases studies with our platform and compared with FBA predictions. The results indicated that MFlux can yield decent results close to  $^{13}\text{C}$ -MFA values, and better than FBA predictions.

In Chapter seven, we developed a platform providing fast literature analysis by using text mining and Big Data technology. We performed several case studies to demonstrate its functionality: (a) display word cloud of a specific term; (b) compare difference between different terms; (c) show the developing trend and current status of a specific term.

In Chapter eight, we summarized all projects in this dissertation and provided personal suggestions for the future directions.

## 1.7. Reference

Tang, Y. J., Martin, H. G., Myers, S., Rodriguez, S., Baidoo, E. E. K. and Keasling, J. D. (2009). Advances in analysis of microbial metabolic fluxes via  $^{13}\text{C}$  isotopic labeling. *Mass. Spectrom. Rev.* 28:362-375.

Orth, J., Conrad, T., Na, J., Lerman, J., Nam, H., Feist, A. and Palsson, B. (2011). A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism-2011. *Mol. Syst. Biol.* 7:535.

Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M. and Palsson, B. Ø. (2013). Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proc. Natl. Acad. Sci.* 110:20338-20343.

Kauffman, K. J., Prakash, P. and Edwards, J. S. (2003). Advances in flux balance analysis. *Curr. Opin. Biotech.* 14:491-496.

Edwards, J., Covert, M. and Palsson, B. (2002). Metabolic modeling of microbes: the flux-balance approach. *Environ. Microbiol.* 4:133 - 140.

Orth, J., Thiele, I. and Palsson, B. (2010). What is flux balance analysis? *Nat. Biotechnol.* 28:245 - 248.

- Becker, S., Feist, A., Mo, M., Hannum, G., Palsson, B. and Herrgard, M. (2007). Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nat. Protoc.* 2:727 - 738.
- Bordbar, A., Monk, J. M., King, Z. A. and Palsson, B. O. (2014). Constraint-based models predict metabolic and associated cellular functions. *Nat. Rev. Genet.* 15:107-120.
- Gowen, C. M. and Fong, S. S. (2010). Genome-scale metabolic model integrated with RNAseq data to identify metabolic states of *Clostridium thermocellum*. *Biotechnol. J.* 5:759-767.
- Schellenberger, J., Que, R., Fleming, R., Thiele, I., Orth, J., Feist, A., Zielinski, D., Bordbar, A., Lewis, N. and Rahmanian, S. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.* 6:1290 - 1307.
- Åkesson, M., Förster, J. and Nielsen, J. (2004). Integration of gene expression data into genome-scale metabolic models. *Metab. Eng.* 6:285-293.
- Coquin, L., Feala, J. D., McCulloch, A. D. and Paternostro, G. (2008). Metabolomic and flux - balance analysis of age-related decline of hypoxia tolerance in *Drosophila* muscle tissue. *Mol. Syst. Biol.* 4.
- Winter, G. and Krömer, J. O. (2013). Fluxomics – connecting ‘omics analysis and phenotypes. *Environ. Microbiol.* 15:1901-1916.
- Khannapho, C., Zhao, H., Bonde, B. K., Kierzek, A. M., Avignone-Rossa, C. A. and Bushell, M. E. (2008). Selection of objective function in genome scale flux balance analysis for process feed development in antibiotic production. *Metab. Eng.* 10:227-233.
- Vallino, J. J. and Stephanopoulos, G. (1993). Metabolic flux distributions in *Corynebacterium glutamicum* during growth and lysine overproduction. *Biotechnol. Bioeng.* 41:633-646.



Yen, J. Y., Nazem-Bokaei, H., Freedman, B. G., Athamneh, A. I. M. and Senger, R. S. (2013). Deriving metabolic engineering strategies from genome-scale modeling with flux ratio constraints. *Biotechnol. J.* 8:581-594.

McAnulty, M., Yen, J., Freedman, B. and Senger, R. (2012). Genome-scale modeling using flux ratio constraints to enable metabolic engineering of clostridial metabolism *in silico*. *BMC Syst. Biol.* 6:42.

Senger, R. S. and Papoutsakis, E. T. (2008). Genome-scale model for *Clostridium acetobutylicum*: Part II. Development of specific proton flux states and numerically determined sub-systems. *Biotechnol. Bioeng.* 101:1053-1071.

Blank, L., Kuepfer, L. and Sauer, U. (2005). Large-scale  $^{13}\text{C}$ -flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biol.* 6:R49.

Fischer, E. and Sauer, U. (2003). Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using GC-MS. *Eur. J. Biochem.* 270:880 - 891f.

Zamboni, N., Fendt, S., Ruhl, M. and Sauer, U. (2009).  $^{13}\text{C}$ -based metabolic flux analysis. *Nat. Protoc.* 4:878 - 892.

You, L., Berla, B., He, L., Pakrasi, H. B. and Tang, Y. J. (2014).  $^{13}\text{C}$ -MFA delineates the photomixotrophic metabolism of *Synechocystis sp. PCC 6803* under light- and carbon-sufficient conditions. *Biotechnol. J.* 9:684-692.

Young, J. D., Shastri, A. A., Stephanopoulos, G. and Morgan, J. A. (2011). Mapping photoautotrophic metabolism with isotopically nonstationary  $^{13}\text{C}$  flux analysis. *Metab. Eng.* 13:656-665.

- Möllney, M., Wiechert, W., Kownatzki, D. and de Graaf, A. A. (1999). Bidirectional reaction steps in metabolic networks: IV. Optimal design of isotopomer labeling experiments. *Biotechnol. Bioeng.* 66:86-103.
- Forbes, N. S., Clark, D. S. and Blanch, H. W. (2001). Using isotopomer path tracing to quantify metabolic fluxes in pathway models containing reversible reactions. *Biotechnol. Bioeng.* 74:196-211.
- Antoniewicz, M. R., Kelleher, J. K. and Stephanopoulos, G. (2007a). Elementary metabolite units (EMU): A novel framework for modeling isotopic distributions. *Metab. Eng.* 9:68-86.
- Young, J. D., Walther, J. L., Antoniewicz, M. R., Yoo, H. and Stephanopoulos, G. (2008). An elementary metabolite unit (EMU) based method of isotopically nonstationary flux analysis. *Biotechnol. Bioeng.* 99:686-699.
- Tang, Y. J., Chakraborty, R., Martín, H. G., Chu, J., Hazen, T. C. and Keasling, J. D. (2007). Flux Analysis of Central Metabolic Pathways in *Geobacter metallireducens* during Reduction of Soluble Fe(III)-Nitrilotriacetic Acid. *Appl. Environ. Microb.* 73:3859-3864.
- Zhuang, W.-Q., Yi, S., Feng, X., Zinder, S. H., Tang, Y. J. and Alvarez-Cohen, L. (2011). Selective Utilization of Exogenous Amino Acids by *Dehalococcoides ethenogenes* strain 195 and the Effects on Growth and Dechlorination Activity. *Appl. Environ. Microb.*:AEM.05676-05611.
- He, L., Xiao, Y., Gebreselassie, N., Zhang, F., Antoniewicz, M. R., Tang, Y. J. and Peng, L. (2014). Central metabolic responses to the overproduction of fatty acids in *Escherichia coli* based on <sup>13</sup>C-metabolic flux analysis. *Biotechnol. Bioeng.* 111:575-585.

Becker, J., Zelder, O., Häfner, S., Schröder, H. and Wittmann, C. (2011). From zero to hero—Design-based systems metabolic engineering of *Corynebacterium glutamicum* for l-lysine production. *Metab. Eng.* 13:159-168.

Tang, J. K.-H., You, L., Blankenship, R. E. and Tang, Y. J. (2012). Recent advances in mapping environmental microbial metabolisms through  $^{13}\text{C}$  isotopic fingerprints. *J. Royal Soc. Interface* 9:2767-2780.

Wiechert, W., Mollney, M., Petersen, S. and de Graaf, A. (2001). A universal framework for  $^{13}\text{C}$  metabolic flux analysis. *Metab. Eng.* 3:265 - 283.

Weitzel, M., Nöh, K., Dalman, T., Niedenfür, S., Stute, B. and Wiechert, W. (2013). 13CFLUX2—high-performance software suite for  $^{13}\text{C}$ -metabolic flux analysis. *Bioinformatics* 29:143-145.

Zamboni, N., Fischer, E. and Sauer, U. (2005). FiatFlux - a software for metabolic flux analysis from  $^{13}\text{C}$ -glucose experiments. *BMC Bioinformatics* 6:209.

Young, J. D. (2014). INCA: a computational platform for isotopically non-stationary metabolic flux analysis. *Bioinformatics* 30:1333-1335.

Glick, B. R. (1995). Metabolic load and heterologous gene expression. *Biotechnol. Adv.* 13:247-261.

Bastian, S., Liu, X., Meyerowitz, J. T., Snow, C. D., Chen, M. M. Y. and Arnold, F. H. (2011). Engineered ketol-acid reductoisomerase and alcohol dehydrogenase enable anaerobic 2-methylpropan-1-ol production at theoretical yield in *Escherichia coli*. *Metab. Eng.* 13:345-352.

Tang, Y. J., Martin, H. G., Deutschbauer, A., Feng, X., Huang, R., Llorca, X., Arkin, A. and Keasling, J. D. (2009b). Invariability of central metabolic flux distribution in *Shewanella oneidensis* MR-1 under environmental or genetic perturbations. *Biotechnol. Progr.* 25:1254-1259.

Heyland, J., Blank, L. M. and Schmid, A. (2011). Quantification of metabolic limitations during recombinant protein production in *Escherichia coli*. *J. Biotechnol.* 155:178-184.

Lennen, R. M., Kruziki, M. A., Kumar, K., Zinkel, R. A., Burnum, K. E., Lipton, M. S., Hoover, S. W., Ranatunga, D. R., Wittkopp, T. M., Marnier, W. D. and Pflieger, B. F. (2011). Membrane Stresses Induced by Overproduction of Free Fatty Acids in *Escherichia coli*. *Appl Environ Microb* 77:8114-8128.

Hoehler, T. M. and Jorgensen, B. B. (2013). Microbial life under extreme energy limitation. *Nat Rev Micro* 11:83-94.

Russell, J. and Cook, G. (1995). Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiol Rev* 59:48 - 62.

Novick, A. and Weiner, M. (1957). Enzyme induction as an all-or-none phenomenon. *Proc. Natl. Acad. Sci.* 43:553-566.

Schaaff, I., Heinisch, J. and Zimmermann, F. K. (1989). Overproduction of glycolytic enzymes in yeast. *Yeast* 5:285-290.

Jensen, P. R., Westerhoff, H. V. and Michelsen, O. (1993). Excess capacity of H(+)-ATPase and inverse respiratory control in *Escherichia coli*. *The EMBO journal* 12:1277-1282.

Birnbaum, S. and Bailey, J. E. (1991). Plasmid presence changes the relative levels of many host cell proteins and ribosome components in recombinant *Escherichia coli*. *Biotechnol Bioeng* 37:736-745.

Ollis, D. F. and Chang, H.-T. (1982). Batch fermentation kinetics with (unstable) recombinant cultures. *Biotechnol Bioeng* 24:2583-2586.

Lee, S. B. and Bailey, J. E. (1984). Analysis of growth rate effects on productivity of recombinant *Escherichia coli* populations using molecular mechanism models. *Biotechnol Bioeng* 26:66-73.

Seo, J.-H. and Bailey, J. E. (1985). Effects of recombinant plasmid content on growth properties and cloned gene product formation in *Escherichia coli*. *Biotechnol Bioeng* 27:1668-1674.

Peretti, S. W. and Bailey, J. E. (1987). Simulations of host–plasmid interactions in *Escherichia coli*: Copy number, promoter strength, and ribosome binding site strength effects on metabolic activity and plasmid gene expression. *Biotechnol Bioeng* 29:316-328.

Snoep, J. L., Yomano, L. P., Westerhoff, H. V. and Ingram, L. O. (1995). Protein burden in *Zymomonas mobilis*: negative flux and growth control due to overproduction of glycolytic enzymes. *Microbiology* 141:2329-2337.

Carrera, J., Rodrigo, G., Singh, V., Kirov, B. and Jaramillo, A. (2011). Empirical model and *in vivo* characterization of the bacterial response to synthetic gene expression show that ribosome allocation limits growth rate. *Biotechnol. J.* 6:773-783.

Somerville, J. E., Jr., Goshorn, S. C., Fell, H. P. and Darveau, R. P. (1994). Bacterial aspects associated with the expression of a single-chain antibody fragment in *Escherichia coli*. *Appl Microbiol Biot* 42:595-603.

Weiß, A. Y., Oyarzún, D. A., Danos, V. and Swain, P. S. (2015). Mechanistic links between cellular trade-offs, gene expression, and growth. *Proc. Natl. Acad. Sci.* 112:E1038-E1047.

Weber, J., Hoffmann, F. and Rinas, U. (2002). Metabolic adaptation of *Escherichia coli* during temperature-induced recombinant protein production: 2. Redirection of metabolic fluxes. *Biotechnol Bioeng* 80:320-330.

Özkan, P., Sariyar, B., Ütkür, F. Ö., Akman, U. and Hortaçsu, A. (2005). Metabolic flux analysis of recombinant protein overproduction in *Escherichia coli*. *Biochem Eng J* 22:167-195.

Wang, Z., Xiang, L., Shao, J., Wegrzyn, A. and Wegrzyn, G. (2006). Effects of the presence of ColE1 plasmid DNA in *Escherichia coli* on the host cell metabolism. *Microb. Cell Fact.* 5:34.

Antoniewicz, M. R., Kraynie, D. F., Laffend, L. A., González-Lergier, J., Kelleher, J. K. and Stephanopoulos, G. (2007b). Metabolic flux analysis in a nonstationary system: Fed-batch fermentation of a high yielding strain of *E. coli* producing 1,3-propanediol. *Metab. Eng.* 9:277-292.

Zamboni, N. (2011). <sup>13</sup>C metabolic flux analysis in complex systems. *Curr Opin Biotech* 22:103-108.

# CHAPTER TWO

## EVALUATING PHYSIOLOGICAL STATE OF ENGINEERED *E. COLI* STRAINS BY ISOTOPOMER CONSTRAINED FLUX BALANCE ANALYSIS

### 2.1. Abstract

Metabolic engineering, especially the introduction of exogenous plasmids into the cell, imposes considerable burdens on cell physiology. For example, plasmid replication, protein overexpression and metabolite accumulation significantly affect the growth rate, expression of native proteins, energetic metabolism, and cell composition of the host cell. Furthermore, biosynthesis of products may result in severe metabolic stresses and cause deleterious impact on both the cell membrane and the energy metabolism. This study aims to understand the metabolic shifts in engineered microbial hosts. Specifically, we have integrated  $^{13}\text{C}$ -MFA and genome-scale FBA (flux balance analysis) to investigate the physiologies of engineered *E.coli* strains for isobutanol production.

On the experimental side, we performed labeling experiments on several engineered *E. coli* strains (high performance JCL260 strain from James Liao Lab, low performance BW25113 wild type strain). Under aerobic growth conditions, we measured both the strain's growth and isotopomer data of their key proteinogenic amino acids. Subsequently, we used a  $^{13}\text{C}$ -MFA ( $^{13}\text{C}$ -metabolic flux analysis) model to profile the central metabolism based on the isotopomer data.  $^{13}\text{C}$ -MFA could only determine scratchy ranges of fluxes in central metabolism. To obtain a broad metabolic solution, we built a large-scale flux balance analysis (FBA) model, which is constrained by  $^{13}\text{C}$ -MFA results. The integrated FBA model relied on objective functions to

evaluate flux distributions. In addition, we tested the sensitivity of the model prediction towards changes in the energy metabolism (ATP maintenance and P/O ratios) and biomass composition equations. By extensively comparing the fluxomics results between the engineered strains, we discovered several metabolic features of the high performance JCL260. First, the JCL260 strain could up-regulate its NADPH production pathways and minimize its overflow metabolism. Second, P/O ratios have relatively a small impact on its optimal isobutanol yield. Third, isobutanol overproduction strongly competes for biomass building blocks and thus addition of nutrients (yeast extract) to support cell growth is essential for high yield of isobutanol. Finally, model sensitivity analysis also implied that isobutanol production is very sensitive to the metabolic burden. Furthermore, isobutanol production pathway is less susceptible to oxygen limitation therefore more likely to achieve high yield compared with biodiesel.

**Key words:** isobutanol,  $^{13}\text{C}$ -MFA, P/O ratio, maintenance energy, production yield

## 2.2. Introduction

Metabolic engineering aims to get desired products through reshaping metabolic network with the aid of DNA recombination technology (Bailey 1991). In general, exogenous plasmids are introduced into host strains over-expressing enzymes to pull flux into related pathways; while chromosomal genes encoding bypass fluxes are knocked out to reduce flux competition. Genetic modifications, especially the presence of plasmids inside cell, impose a considerable influence on cell physiology. Metabolic burdens contributed by protein expression and plasmid replication significantly affect growth rate, expression of native proteins, energetic metabolism, and cell composition of host cells (Birnbaum and Bailey 1991; Özkan *et al.* 2005; Rozkov *et al.* 2004). Furthermore, generation of certain metabolites such as alcohol or fatty acid results in severe



oxidative stress *in vivo*, damages on cell membrane as well as on transport system and energy metabolism (Lennen *et al.* 2011; Nicolaou *et al.* 2010; 2009). All those factors enforce host cells to adjust their metabolism in response to burdens and stress (He *et al.* 2014). Quantitative understandings over such shifts would be beneficial for a deeper view of cellular regulation, and may shed light on strain rational design for the sake of metabolic engineering and bioprocess scale-up (Sauer *et al.* 1998; Garcia-Ochoa and Gomez 2009).

Fluxomics, the functional systems biology tool, have been employed to investigate the physiological alternations of engineered strains quantitatively (Vallino and Stephanopoulos 1993; Antoniewicz *et al.* 2007b). As the two basic approaches in the realm of fluxomics,  $^{13}\text{C}$ -MFA ( $^{13}\text{C}$  metabolic flux analysis) measures *in vivo* flux information mainly based on mass isotopomer distribution of amino acids (isotopic fingerprint), while FBA (Flux Balance Analysis) predicts flux distribution in genome-wide metabolic network under presumed objective functions, which describe the “possible” metabolic potential of microbial hosts (Orth *et al.* 2010a; Stephanopoulos 1999).  $^{13}\text{C}$ -MFA is able to precisely determine flux profile, but limited its scope to central carbon metabolic pathways (Chen *et al.* 2011). FBA has been widely employed to interpret metabolism of a variety of species at the genome-scale, to identify gene essentiality, and to reveal the trend of adaptive evolution (Feist *et al.* 2009; Ibarra *et al.* 2002; Famili *et al.* 2003). Accurate quantification of cellular metabolism is highly dependent on the selection of appropriate objective function for FBA. Maximizing biomass growth is the most common objective function and it works well for cells in exponential growth phase (Schuetz *et al.* 2007; Knorr *et al.* 2007). However, in many cases, cell does not behave in the manner of any optimal strategies, which raise a challenge for choosing an appropriate objective function (Schuetz *et al.* 2007; Schuetz *et al.* 2012). On the other hand, introduction of a series of constraints in the process of linear

programming would greatly reduce the solution space; therefore enable more accurate predictions on flux profile. Constraints such as overflow flux, energy balance, membrane site occupancy, and proton gradient represent the laws of thermodynamics, physics and physiology working on cellular metabolism (Senger and Papoutsakis 2008; Beard *et al.* 2002; Zhuang *et al.* 2011b). The idea to combine the advantages of both approaches together as synergistic tools to overcome shortcomings of individual has been attempted in distinctive ways by various groups (Blank *et al.* 2005; Chen *et al.* 2011).

To decipher physiological conditions of engineered strains by fluxomics tools, several challenges have to be seriously circumvented. The first challenge is regarding the ATP generation by oxidative phosphorylation: Theoretically, the ideal maximal P/O ratio is 2 for *B. subtilis* and 3 for *E. coli* under aerobic condition; however, realistic values are always lower and vary from case to case in different strains (Özkan *et al.* 2005; Sauer and Bailey 1999), precise determination of P/O ratio for a specific engineered strain requires labor-intensive measurements. The second challenge is the maintenance energy, which was found to be not only including the heat generation, but also related with cell mobility, protein and DNA turnover, osmoregulation, and pathway shift (Hoehler and Jorgensen 2013a). Traditionally, maintenance energy can be determined through employing chemostat culture at different dilution rates (Pirt 1965), which is not considered as a good reflection of maintenance energy now (Hoehler and Jorgensen 2013a). Due to the intricate and dynamic essence of maintenance energy and various types of genetic modifications, setting an appropriate value of maintenance energy for FBA also requires experimental assistant.

In order to address problems mentioned above, we propose an integrated platform of <sup>13</sup>C-MFA and FBA in this work as isotopomer constrained flux balance analysis (icFBA, shown in

Figure 1). This method is utilized to explore the physiological status of several engineered *E. coli* strains that produce isobutanol (Baez *et al.* 2011a). Isobutanol, a promising biofuel potentially to replace ethanol, has attracted enormous attentions from both biochemical engineers and the public since 2008 (Atsumi *et al.* 2008a). However, there are still obstacles on the way to scale-up laboratory high yield isobutanol producing strains into industrial production. Several approaches have been attempted to work out the concerns such as product toxicity (Atsumi *et al.* 2010). A thorough examination on isobutanol producing strains by this fluxomics platform might help us to gain more insights into their physiology and provide some hints for industrial-scale fermentation. Furthermore, the knowledge obtained from this study can potentially applied in the design of other engineered strains.

We performed  $^{13}\text{C}$  labeling experiments on several isobutanol engineered *E. coli* strains. The growth characteristics and amino acid labeling patterns were used for fluxomics studies. Consequently, we obtained the flux distribution of central carbon metabolism. Subsequently, flux values of central metabolic pathways determined by  $^{13}\text{C}$ -MFA are taken as constraints for genome scale FBA calculation (iJO1366 is employed here) (Orth *et al.* 2011). Different values of maintenance energy, oxygen flux, as well as P/O ratio are also employed in the simulation of FBA.

## **2. 3. Materials and Methods**

### **2.3.1. Strains and plasmids**

BW25113 (CGSC# 7636) was the wild-type strain used in this study (Atsumi *et al.* 2008b) and it was purchased from the *Escherichia coli* Genetic Stock Center (<http://cgsc.biology.yale.edu/>). The knockout strain JCL260 ( $\Delta\text{adhE}$ ,  $\Delta\text{ldhA}$ ,  $\Delta\text{frdBC}$ ,  $\Delta\text{fnr}$ ,  $\Delta\text{pta}$ ,  $\Delta\text{pflB}$ ) and the plasmids pSA65 and pSA69 (Baez *et al.* 2011b; Atsumi *et al.* 2008c) were provided by Prof. James Liao

(University of California, Los Angeles). The plasmid pSA65 contains the genes *kivd* and *adhA* from *Lactococcus lactis*. The plasmid pSA69 contains the genes *alsS* of *Bacillus subtilis* and *ilvCD* of *E. coli*. The detailed information of strains and plasmids are also listed in Table 2.1.

### **2.3.2. Medium and culture conditions**

A minimal medium containing 0.5% [1, 2-<sup>13</sup>C] glucose, M9 salts (Difco), and 10 mg/L vitamin B1 was used for the labeling experiments. The antibiotics, Ampicillin (100 µg/mL) and Kanamycin (50 µg/mL) were added as appropriate. For the pre-culture, a single colony of cells from a fresh plate was used to inoculate a 5 mL Luria-Bertani (LB) media. The pre-culture was grown overnight at 37 °C on a rotary shaker at 225 rpm. The pre-culture was used to inoculate (0.2%, v/v) a main culture of 10 mL minimal medium, grown on a rotary shaker in 250 mL shake flasks at 30 °C and 225 rpm. The cultures were grown in duplicates and kept airtight by closing the flasks with rubber stoppers to prevent any loss of isobutanol due to evaporation. IPTG was added to a final concentration of 0.1 mM for induction. Samples were collected before and after (both at mid-log phase and late-log phase) the addition of IPTG. The liquid cultures (~10 mL) from each flask were centrifuged and the supernatant was separated from the biomass. Both the biomass and the supernatant samples were stored in -20 °C prior to analysis.

### **2.3.3. Quantification of biomass and extracellular metabolites**

Growth of the *E. coli* cells was monitored by measuring the optical density of the cultures at 600 nm (OD<sub>600</sub>) using an Agilent Cary 60 UV-Vis spectrophotometer. The dry biomass concentration in gram per liter was determined based on the correlation of, OD<sub>600</sub> - 1 is equivalent to 0.338 g dry weight/L (Xiao *et al.* 2013). The concentration of acetic acid and glucose were measured by using the corresponding enzymatic kit (R-Biopharm) and by following the manufacturer's protocol. The enzymatic reactions were conducted at room temperature in a 96-well plate reader

(Infinite 200 PRO microplate photometer, TECAN). Isobutanol analysis was done using a gas chromatograph (GC) (Hewlett Packard model 7890A [Agilent Technologies] equipped with a DB5-MS column [J&W Scientific]) and a mass spectrometer (MS) (5975C, Agilent Technologies) (Xiao *et al.* 2012). Isobutanol was extracted from 800  $\mu\text{L}$  of the supernatant using 400  $\mu\text{L}$  of toluene as the extractant. Both the supernatant and the toluene were vortexed together for 2 minutes, followed by centrifugation at  $10,000 \times g$  for 5 min to separate the aqueous and organic phase. 1  $\mu\text{L}$  of the organic layer was injected into the GC column with helium as the carrier gas. The GC oven was held at 70  $^{\circ}\text{C}$  for 2 minutes and then raised to 200  $^{\circ}\text{C}$  with a temperature ramp of 30  $^{\circ}\text{C min}^{-1}$ , and the post run was set at 300  $^{\circ}\text{C}$  for 6 minutes. The MS scan mode was set from  $m/z$  of 20 to  $m/z$  of 200. The quantification of isobutanol was done based on a calibration curve prepared with known concentrations of isobutanol ranging from 25 mg/L to 200 mg/L. Methanol was used as an internal standard for all the samples.

#### **2.3.4. Mass isotopomer distribution of proteinogenic amino acids**

The mass isotopomer distribution (MID) of proteinogenic amino acids were performed as described elsewhere (You *et al.* 2012). In short, the biomass pellets were hydrolyzed with 6 M hydrochloric acid and dried under a stream of air. The hydrolysates were dissolved in tetrahydrofuran and derivatized with N-Methyl-N-[tert-butyldimethyl-silyl] trifluoroacetamide (Sigma-Aldrich, MO) at 70  $^{\circ}\text{C}$  for 1 hour. The amino acid analysis was also performed on the GC-MS equipped with a DB5-MS column. The sample injection volume was 1  $\mu\text{L}$  and a 1:10 split ratio was utilized with helium as the carrier gas (1.2 mL/min). The GC oven was held at 150  $^{\circ}\text{C}$  for 2 minutes and then raised to 280  $^{\circ}\text{C}$  with a temperature ramp of 3  $^{\circ}\text{C min}^{-1}$ , followed by another temperature ramp of 20  $^{\circ}\text{C min}^{-1}$  to a final temperature of 300  $^{\circ}\text{C}$  and was held at 300  $^{\circ}\text{C}$  for 5 minutes. The mass spectra were acquired from the MS with an  $m/z$  range of 60 to

500. The mass isotopomer distributions of the amino acids were corrected for the presence of naturally abundant isotopes [ $^{13}\text{C}$  - 1.13%,  $^{18}\text{O}$  - 0.20%,  $^{29}\text{Si}$  - 4.70%, and  $^{30}\text{Si}$  - 3.09%] using published algorithm (Tang *et al.* 2009a; Wahl *et al.* 2004).

### 2.3.5. Central metabolic flux determined by $^{13}\text{C}$ -MFA

$^{13}\text{C}$ -MFA was carried out based on the MID information from the proteinogenic amino acids. The metabolic network of *E. coli* strains includes major pathways such as the glycolysis, the pentose phosphate pathway, the Entner–Doudoroff (ED) pathway, the tricarboxylic acid (TCA) cycle, the glyoxylate shunt, and the anaplerotic pathways. For biomass flux, we adopted the same equation previously reported (He *et al.* 2014). In engineered strains that produce isobutanol, a simplified reaction was employed to describe the isobutanol flux ( $2*\text{PYR} + \text{NADH} + \text{NADPH} \rightarrow \text{IB} + 2*\text{CO}_2$ ). The detailed information of metabolic network is listed in Table 2.3. The carbon substrate uptake rate was defined as 100, while other metabolic fluxes were normalized to a scale of 100. The energy metabolism was not included in  $^{13}\text{C}$ -MFA calculation, because different P/O ratio values affect the central carbon flux profiles seriously (data not shown). The EMU (elementary metabolite units) method was adopted for  $^{13}\text{C}$ -MFA. The MATLAB-based software WUFlux, developed by Tang lab, was employed for  $^{13}\text{C}$ -MFA calculation (available in [13cmfa.org](http://13cmfa.org)). 100 randomly generated initial values were used and the solution set with minimal objective function value (best fit) was selected as the final fluxes. The relative flux profile in each strain was calculated by minimizing the difference between predicted and measured isotopomer patterns. The measured production rates of acetic acid and biomass are used as the constraints for  $^{13}\text{C}$ -MFA, while isobutanol flux is predicted by the flux model. Since the precise measurement of volatile extracellular metabolites (i.e., acetate) is difficult, calculation of  $^{13}\text{C}$ -MFA did not place tight constraints on overflow metabolite fluxes. Due to its high volatility,

isobutanol is difficult to be determined precisely thus the measured value will only be used as reference.

### 2.3.6. Genome-scale model constrained by <sup>13</sup>C-MFA flux

The *E. coli* iJO1366 genome scale model was employed for FBA studies with minor modifications (Orth *et al.* 2011): the isobutanol flux (3mob[c] + nadh[c] + h[c] --> isobutanol + nad[c] + co2[c]) was added for engineered strains; a few specific fluxes were knockout for JCL260 strain and BW25113 strain (Monk *et al.* 2013). The <sup>13</sup>C-MFA flux  $v^{mfa}$  was used as the boundaries to constrain central metabolic flux of genome scale model using the following rules (Blank *et al.* 2005):

$$s.t. S \cdot v' = 0$$

$$v_i^{mfa} \cdot (1 - \delta) \leq v_i' \leq v_i^{mfa} \cdot (1 + \delta)$$

$$\delta = 20\%$$

The objective function is to maximize isobutanol production (for engineered strains, set growth rates to be experimental values) or to maximize growth rate (for control strain). Default boundary conditions of iJO1366 were employed here, except for the fluxes of glucose uptake, oxygen uptake, and maintenance energy. We test the sensitivities of oxygen uptake rate, the maintenance energy, and the P/O ratio on the isobutanol production flux. Considering of oxygen supply conditions in our <sup>13</sup>C-MFA experiments, we set the oxygen flux to be 16 mmol/gCDW h (Varma *et al.* 1993). For isobutanol production by strain 5 (JCL260/pSA65+69, see Table 2.1.) in the bioreactor, we employed a modified biomass equation based on the <sup>13</sup>C labeling information -- non-labeled biomass derived from yeast extract was deducted (Xiao *et al.* 2012), as well as a reduced oxygen flux (average 13 mmol/gCDW h) (Xu *et al.* 1999).

To access strain stability, we also employ evolutionary fitness to quantify each strain: measured growth rate relative to the predicted maximal growth rate by FBA under the same conditions (Steinmetz *et al.* 2002).

## **2.4. Results and discussions**

### **2.4.1. Physiological states of different Strains**

We performed labeling experiments by feeding *E. coli* strains with 1, 2-<sup>13</sup>C labeled glucose in minimal medium under aerobic conditions. The physiological states of different strains are recorded in Table 2.2.: As expected, the control strain (JCL260, strain 1) has a much higher specific growth rate compared with the other three isobutanol-producing (strains 2 – 4). Significant lower biomass yield was observed in engineered strains compared with control strain. JCL260 has higher efficiency for biomass yield while JCL260/pSA65 displays the lowest biomass yield compared with BW25113 engineered strain. Acetic acid production was found in cases of two BW25113 host strains, but was not detected for either JCL260 strain in the minimal medium. Strain JCL260/(pSA65+69) demonstrated poor or no growth repeatedly in minimal medium even without IPTG induction. Thereby, we used previous experimental data of JCL260/(pSA65+69) growth in presence of yeast extract, which was published by our lab three years ago (strain 5, in Table 2.2.) (Xiao *et al.* 2012). Apparently, strain 5 exhibited a significantly higher yield of isobutanol in the presence of yeast extract, compared with other strains growth in a minimal medium. Previous reports have proved that yeast extract mainly contributed to the synthesis of biomass building blocks while most of isobutanol was converted from glucose (Xiao *et al.* 2012). However, yeast extract is indispensable for this process: Isobutanol production pathway poses considerable metabolic burdens on host strains, thereby, cellular energy generated from glucose catabolism is insufficient to power both biomass



synthesis and isobutanol production in high-yield strain JCL260/(pSA65+69). Biomass synthesis, especially the biosynthesis of amino acids consumes lots of ATP and NADH (Russell and Cook 1995; Akashi and Gojobori 2002), while yeast extract relieves such burden by supplying most amino acids (Xiao *et al.* 2012; Selvarasu *et al.* 2009). Notably, isobutanol itself will disrupt cell membrane integrity and reduce the efficiency of oxidative phosphorylation, leading to decreased pH and ATP supply (Atsumi *et al.* 2010; Wu *et al.* 2015). Supplement of yeast extract relieves the burden from energy intensive biomass synthesis, thereby well solves the energy crisis inside isobutanol producing strains.

#### **2.4.2. Central metabolic flux determined by $^{13}\text{C}$ -MFA**

To quantify the intracellular fluxes of the central metabolic pathway,  $^{13}\text{C}$ -MFA was employed to solve the flux profile based on the mass isotopic distribution (MID) of proteinogenic amino acid. The results of the central flux map are illustrated in Figure 2.2a-d. Central metabolic flux profile always undergoes apparent adjustments in response to metabolic burdens and cell stress within engineered strains (He *et al.* 2014; Antoniewicz *et al.* 2007c). In this work, significant changes in central metabolism were observed in response to the genetic alternations and isobutanol production for engineered strain 2 (JCL260/pSA65): The pentose phosphate (PP) pathway is up-regulated in engineered strains to balance NADPH demanding (compared with JCL260, strain 1); the glyoxylate shunt increases to reduce the carbon loss as  $\text{CO}_2$ ; while the anaplerotic flux reduced significantly, enable more flux from glucose rerouting into pyruvate that is a precursor for isobutanol synthesis. The glucose-inhibited glyoxylate shunt is recovered by a decreased glucose uptake flux (inhibited by isobutanol) (Atsumi *et al.* 2010; Bowles and Ellefson 1985), while knocking out *ppc* gene leads to increase activity of glyoxylate shunt (Fong *et al.* 2006).

Compared with BW25113 strains (strain 3 and 4), JCL260 strains (1, 2) only produced little amount of acetate in the minimal medium, which was even out of detection limit (Table 2.2). Knocking out acetate related genes not only leads to the loss of acetate production capability in JCL260, but also reroutes more carbon fluxes into isobutanol production pathway. However, significant acetate production was still observed for strain JCL260/(pSA65+69) in the presence of yeast extract. A possible source of acetate is the biosynthesis and degradation of amino acids (e.g.,  $SER + AceCoA + 3 ATP + 4 NADPH == CYS + Ac$ ), especially supplied with abundant exogenous amino acids (i.e., yeast extract).

### 2.4.3. Energy metabolism and evolutionary fitness analysis

To illustrate the energy status of different strains under genetic modifications (e.g., gene knockout, exogenous plasmids) and internal stress (e.g., protein overexpression, isobutanol toxicity), we performed a simple analysis by assuming ideal energy metabolism (i.e., P/O ratio = 3) and the results are shown in Table 2.4 and Figure 2.3. In ideal condition (Figure 2.3a), the control strain (JCL260) has highest excess energy while strain 2 (JCL260/pSA65) is of lowest excess energy; this situation overturned when inefficient energy metabolisms (P/O ratio = 1, shown in Figure 2.3b) apply for energy analysis. P/O ratio of 1 is closer to the real situations of  $^{13}C$ -MFA experiments, based on literature reports on *E. coli* strains (Özkan *et al.* 2005; Noguchi *et al.* 2004). The ‘excess energy’ actually is used as cellular maintenance cost, whereas the control strain (strain 1) has the lowest maintenance energy, while strain 2 (JCL260/pSA65) shows the highest maintenance energy cost. Further, strain 4 (BW/pSA65+69) has a higher maintenance requirement compared with strain 3 (BW/pSA65). Such differences in maintenance cost are caused by additional energy expense used for extra plasmid maintenance, as well as heterogeneous protein expression (Glick 1995).

Strain stability is a crucial factor for industrial-scale fermentation. Evolutionary fitness is used to estimate the genetic stability of engineered strains, based on the assumption that cells with fastest growth at given conditions has the largest fitness of survival (Blank *et al.* 2005; Steinmetz *et al.* 2002). Considering of complicated genetic modifications in this study, evolutionary fitness is used here rather than physiological fitness (Blank *et al.* 2005). From the results shown in Figure 2.4, strain 1 (JCL260) has the highest evolutionary fitness while strain 4 (BW/pSA65+69) demonstrates the lowest degree of evolutionary fitness. Besides, strain 5 may have even lower evolutionary fitness in the minimal medium because it contains dual plasmids in addition to a series of gene knockouts. It is well known that maintenance of exogenous plasmids brings considerable metabolic burdens that lead to increased instability of genotype (Silva *et al.* 2012). Researchers have proposed a series of approaches to reduce the negative impacts from plasmids, such as employment of low copy plasmid (Jones *et al.* 2000) or insertion of genes into chromosome (Tyo *et al.* 2009). With advancements of novel genome editing tools (e.g., CRISPR, TALEN), genetic modifications at chromosome level will result in strains with improved genetic stability in the future (Luo *et al.* 2015).

#### **2.4.4. The effects of P/O ratio, oxygen flux, and maintenance energy on isobutanol production**

One major bottleneck for metabolic engineering is that laboratory high-yield strains fail to make good performance during the scale-up process. Quantification of several factors involving this process (e.g., environmental factors, such as oxygen, nutrient availability, and strain physiological factors such as P/O ratio, maintenance energy) may provide novel insights into this problem. In this work, we employed FBA to investigate the effects of oxygen flux, nutrient availability, and maintenance energy on product yield. In particular, central metabolic fluxes determined by  $^{13}\text{C}$ -MFA are utilized to constrain flux boundary of FBA to ensure the final flux

profile to be in a reasonable range (Blank *et al.* 2005). Meanwhile, we also determine the effect of yeast extract on isobutanol production based on the  $^{13}\text{C}$  labeling data of strain 5. The details of simulations are described in section 2.3.5, and the results are presented in Figure 2.5 and 2.6.

The maximal yield of isobutanol differs from strain to strain: strain 5 > strain 2 > strain 4 > strain 3. This prediction exactly matches the experimental observations of isobutanol yield recorded in Table 2.2. Meanwhile, isobutanol production is more sensitive to the variation in maintenance energy and oxygen flux, rather than P/O ratio. Isobutanol synthesis requires NADH/NADPH rather than ATP in its synthetic pathway, which explains its robustness to P/O ratio change (Wu *et al.* 2015). Oxygen is not required for isobutanol synthesis; however, oxygen-involved oxidative phosphorylation contributes most energy for biomass synthesis at growth phase. The competition between biomass growth/protein expression and isobutanol production on intracellular energy makes ATP-rich oxidative respiration pathway very favorable, leading to increased sensitivity to oxygen concentration. During industrial-scale fermentations, environmental conditions (e.g., pH, oxygen concentration) at different locations of the bioreactor are changing with the time (Garcia-Ochoa and Gomez 2009; Zou *et al.* 2012). Further, massive consumption of oxygen and insufficient mixing during exponential growth leads to heterogeneous oxygen distribution within the reactor. The conflict between increasing oxygen demand for cell growth and decreasing oxygen supply capability in many regions of bioreactor always leads to poor performance of engineered strains (fall off the cliff of isobutanol production as shown in Figure 2.5 and 2.6). In many cases, supply of rich nutrients will not only lesson the cost of carbon source, and energy/reducing power (ATP, NADH, and NADPH) on amino acids synthesis, but also alleviate the requirement on activation of corresponding synthesis pathways, that's why strain 5 have the best performance of all test conditions in this study (results shown in

Table 2.2, Figure 2.5e, and Figure 2.6e). In practice, induction of protein overexpression at mid or late growth phase also can relieve the resource crisis mentioned above (Xu *et al.* 2014). Alcohol production process has an apparent advantage: Even cell membrane and oxidative phosphorylation system are destructed by aldehyde and alcohol, the synthesis process of isobutanol is still going on.

## 2.5. Conclusions

In this work, we have resolved the central carbon metabolism of isobutanol-producing *E. coli* strains by <sup>13</sup>C-MFA. The results indicated that genetically modified strains can make adjustments over their flux profile in response to the requirements of isobutanol production. Also, extensive genetic modifications will lead to decreased evolutionary fitness as well as increased maintenance cost. Further, isobutanol production is very sensitive to the increase of cellular maintenance energy while rich nutrients (e.g., yeast extract) can relieve the stress caused by metabolic burdens. Lastly, genome editing will bring less metabolic burden and more evolutionary fitness compared with plasmid-based modification, thus is suitable for bioprocess scale-up in the future.

## 2.6. References

Bailey, J. (1991). Toward a science of metabolic engineering. *Science* 252:1668-1675.

Birnbaum, S. and Bailey, J. E. (1991). Plasmid presence changes the relative levels of many host cell proteins and ribosome components in recombinant *Escherichia coli*. *Biotechnol Bioeng* 37:736-745.

Rozkov, A., Avignone-Rossa, C. A., Ertl, P. F., Jones, P., O'Kennedy, R. D., Smith, J. J., Dale, J. W. and Bushell, M. E. (2004). Characterization of the metabolic burden on *Escherichia coli* DH1 cells imposed by the presence of a plasmid containing a gene therapy sequence. *Biotechnol Bioeng* 88:909-915.

Özkan, P., Sariyar, B., Ütkür, F. Ö., Akman, U. and Hortaçsu, A. (2005). Metabolic flux analysis of recombinant protein overproduction in *Escherichia coli*. *Biochem Eng J* 22:167-195.

Nicolaou, S. A., Gaida, S. M. and Papoutsakis, E. T. (2010). A comparative view of metabolite and substrate stress and tolerance in microbial bioprocessing: From biofuels and chemicals, to biocatalysis and bioremediation. *Metab. Eng.* 12:307-331.

Lennen, R. M., Kruziki, M. A., Kumar, K., Zinkel, R. A., Burnum, K. E., Lipton, M. S., Hoover, S. W., Ranatunga, D. R., Wittkopp, T. M., Marner, W. D. and Pfleger, B. F. (2011). Membrane Stresses Induced by Overproduction of Free Fatty Acids in *Escherichia coli*. *Appl Environ Microb* 77:8114-8128.

(2009). *An integrated network approach identifies the isobutanol response network of Escherichia coli.*

He, L., Xiao, Y., Gebreselassie, N., Zhang, F., Antoniewicz, M. R., Tang, Y. J. and Peng, L. (2014). Central metabolic responses to the overproduction of fatty acids in *Escherichia coli* based on <sup>13</sup>C-metabolic flux analysis. *Biotechnol Bioeng* 111:575-585.

Sauer, U., Cameron, D. C. and Bailey, J. E. (1998). Metabolic capacity of *Bacillus subtilis* for the production of purine nucleosides, riboflavin, and folic acid. *Biotechnol Bioeng* 59:227-238.

Garcia-Ochoa, F. and Gomez, E. (2009). Bioreactor scale-up and oxygen transfer rate in microbial processes: An overview. *Biotechnol Adv* 27:153-176.

Antoniewicz, M. R., Kraynie, D. F., Laffend, L. A., González-Lergier, J., Kelleher, J. K. and Stephanopoulos, G. (2007a). Metabolic flux analysis in a nonstationary system: Fed-batch fermentation of a high yielding strain of *E. coli* producing 1,3-propanediol. *Metab. Eng.* 9:277-292.

Vallino, J. J. and Stephanopoulos, G. (1993). Metabolic flux distributions in *Corynebacterium glutamicum* during growth and lysine overproduction. *Biotechnol Bioeng* 41:633-646.

Stephanopoulos, G. (1999). Metabolic Fluxes and Metabolic Engineering. *Metab. Eng.* 1:1-11.

Orth, J., Thiele, I. and Palsson, B. (2010). What is flux balance analysis? *Nat Biotechnol* 28:245 - 248.

Chen, X., Alonso, A. P., Allen, D. K., Reed, J. L. and Shachar-Hill, Y. (2011). Synergy between <sup>13</sup>C-metabolic flux analysis and flux balance analysis for understanding metabolic adaptation to anaerobiosis in *E. coli*. *Metab. Eng.* 13:38-48.

Feist, A. M., Herrgard, M. J., Thiele, I., Reed, J. L. and Palsson, B. O. (2009). Reconstruction of biochemical networks in microorganisms. *Nat Rev Micro* 7:129-143.

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

Famili, I., Förster, J., Nielsen, J. and Palsson, B. O. (2003). *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc. Natl. Acad. Sci.* 100:13134-13139.

Schuetz, R., Kuepfer, L. and Sauer, U. (2007). Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol. Syst. Biol.* 3:119.

Knorr, A. L., Jain, R. and Srivastava, R. (2007). Bayesian-based selection of metabolic objective functions. *Bioinformatics* 23:351-357.

Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M. and Sauer, U. (2012). Multidimensional Optimality of Microbial Metabolism. *Science* 336:601-604.

Beard, D. A., Liang, S.-d. and Qian, H. (2002). Energy Balance for Analysis of Complex Metabolic Networks. *Biophys J* 83:79-86.

Senger, R. S. and Papoutsakis, E. T. (2008). Genome-scale model for *Clostridium acetobutylicum*: Part II. Development of specific proton flux states and numerically determined sub-systems. *Biotechnol Bioeng* 101:1053-1071.

Zhuang, K., Vemuri, G. N. and Mahadevan, R. (2011). Economics of membrane occupancy and respiro-fermentation. *Mol. Syst. Biol.* 7:500.

Blank, L., Kuepfer, L. and Sauer, U. (2005). Large-scale <sup>13</sup>C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biol.* 6:R49.

Sauer, U. and Bailey, J. E. (1999). Estimation of P-to-O ratio in *Bacillus subtilis* and its influence on maximum riboflavin yield. *Biotechnol Bioeng* 64:750-754.

Hoehler, T. M. and Jorgensen, B. B. (2013). Microbial life under extreme energy limitation. *Nat Rev Micro* 11:83-94.

Pirt, S. (1965). The maintenance energy of bacteria in growing cultures. *Proc. R. Soc. Lond. B Biol. Sci.* 163:224 - 231.

Baez, A., Cho, K.-M. and Liao, J. (2011a). High-flux isobutanol production using engineered *Escherichia coli*: a bioreactor study with in situ product removal. *Appl Microbiol Biot* 90:1681-1690.



Atsumi, S., Hanai, T. and Liao, J. C. (2008a). Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature* 451:86-89.

Atsumi, S., Wu, T. Y., Machado, I. M. P., Huang, W. C., Chen, P. Y., Pellegrini, M. and Liao, J. C. (2010). Evolution, genomic analysis, and reconstruction of isobutanol tolerance in *Escherichia coli*. *Mol. Syst. Biol.* 6: 449.

Orth, J., Conrad, T., Na, J., Lerman, J., Nam, H., Feist, A. and Palsson, B. (2011). A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism-2011. *Mol. Syst. Biol.* 7:535.

Atsumi, S., Cann, A. F., Connor, M. R., Shen, C. R., Smith, K. M., Brynildsen, M. P., Chou, K. J., Hanai, T. and Liao, J. C. (2008b). Metabolic engineering of *Escherichia coli* for 1-butanol production. *Metabolic engineering* 10:305-311.

Baez, A., Cho, K.-M. and Liao, J. C. (2011b). High-flux isobutanol production using engineered *Escherichia coli*: a bioreactor study with in situ product removal. *Applied Microbiology and Biotechnology* 90:1681-1690.

Atsumi, S., Hanai, T. and Liao, J. C. (2008c). Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature* 451:86-89.

Xiao, Y., Ruan, Z. H., Liu, Z. G., Wu, S. G., Varman, A. M., Liu, Y. and Tang, Y. J. (2013). Engineering *Escherichia coli* to convert acetic acid to free fatty acids. *Biochemical Engineering Journal* 76:60-69.

Xiao, Y., Feng, X., Varman, A. M., He, L., Yu, H. and Tang, Y. J. (2012). Kinetic Modeling and Isotopic Investigation of Isobutanol Fermentation by Two Engineered *Escherichia coli* Strains. *Ind Eng Chem Res* 51:15855-15863.

You, L., Page, L., Feng, X., Berla, B., Pakrasi, H. B. and Tang, Y. J. (2012). Metabolic pathway confirmation and discovery through  $^{13}\text{C}$ -labeling of proteinogenic amino acids. *J Vis Exp*:e3583.

Tang, Y. J., Martin, H. G., Myers, S., Rodriguez, S., Baidoo, E. E. K. and Keasling, J. D. (2009). Advances in Analysis of Microbial Metabolic Fluxes Via  $^{13}\text{C}$  Isotopic Labeling. *Mass Spectrom Rev* 28:362-375.

Wahl, S. A., Dauner, M. and Wiechert, W. (2004). New tools for mass isotopomer data evaluation in  $^{13}\text{C}$  flux analysis: Mass isotope correction, data consistency checking, and precursor relationships. *Biotechnology and Bioengineering* 85:259-268.

Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M. and Palsson, B. Ø. (2013). Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proc. Natl. Acad. Sci.* 110:20338-20343.

Varma, A., Boesch, B. and Palsson, B. (1993). Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. *Appl. Environ. Microbiol.* 59:2465 - 2473.

Xu, B., Jahic, M. and Enfors, S.-O. (1999). Modeling of Overflow Metabolism in Batch and Fed-Batch Cultures of *Escherichia coli*. *Biotechnol Progr* 15:81-90.

Steinmetz, L. M., Scharfe, C., Deutschbauer, A. M., Mokranjac, D., Herman, Z. S., Jones, T., Chu, A. M., Giaever, G., Prokisch, H., Oefner, P. J. and Davis, R. W. (2002). Systematic screen for human disease genes in yeast. *Nat Genet* 31:400-404.

Russell, J. and Cook, G. (1995). Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiol Rev* 59:48 - 62.

Akashi, H. and Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc. Natl. Acad. Sci.* 99:3695-3700.

Selvarasu, S., Ow, D. S.-W., Lee, S. Y., Lee, M. M., Oh, S. K.-W., Karimi, I. A. and Lee, D.-Y. (2009). Characterizing *Escherichia coli* DH5 $\alpha$  growth and metabolism in a complex medium using genome-scale flux analysis. *Biotechnol Bioeng* 102:923-934.

Wu, S., He, L., Wang, Q. and Tang, Y. (2015). An ancient Chinese wisdom for metabolic engineering: Yin-Yang. *Microb. Cell Fact.* 14:39.

Antoniewicz, M., Kraynie, D., Laffend, L., Gonzalez-Lergier, J., Kelleher, J. and Stephanopoulos, G. (2007b). Metabolic flux analysis in a nonstationary system: fed-batch fermentation of a high yielding strain of *E. coli* producing 1,3-propanediol. *Metab. Eng.* 9:277 - 292.

Bowles, L. K. and Ellefson, W. L. (1985). Effects of butanol on *Clostridium acetobutylicum*. *Appl Environ Microb* 50:1165-1170.

Fong, S. S., Nanchen, A., Palsson, B. O. and Sauer, U. (2006). Latent Pathway Activation and Increased Pathway Capacity Enable *Escherichia coli* Adaptation to Loss of Key Metabolic Enzymes. *J Biol Chem* 281:8024-8033.

Noguchi, Y., Nakai, Y., Shimba, N., Toyosaki, H., Kawahara, Y., Sugimoto, S. and Suzuki, E.-i. (2004). The Energetic Conversion Competence of *Escherichia coli* during Aerobic Respiration Studied by <sup>31</sup>P NMR Using a Circulating Fermentation System. *J Biochem* 136:509-515.

Glick, B. R. (1995). Metabolic load and heterologous gene expression. *Biotechnol Adv* 13:247-261.

Silva, F., Queiroz, J. A. and Domingues, F. C. (2012). Evaluating metabolic stress and plasmid stability in plasmid DNA production by *Escherichia coli*. *Biotechnol Adv* 30:691-708.

- Jones, K. L., Kim, S.-W. and Keasling, J. D. (2000). Low-Copy Plasmids can Perform as Well as or Better Than High-Copy Plasmids for Metabolic Engineering of Bacteria. *Metab. Eng.* 2:328-338.
- Tyo, K. E. J., Ajikumar, P. K. and Stephanopoulos, G. (2009). Stabilized gene duplication enables long-term selection-free heterologous pathway expression. *Nat Biotech* 27:760-765.
- Luo, Y., Li, B.-Z., Liu, D., Zhang, L., Chen, Y., Jia, B., Zeng, B.-X., Zhao, H. and Yuan, Y.-J. (2015). Engineered biosynthesis of natural products in heterologous hosts. *Chem. Soc. Rev.* 44:5265-5290.
- Zou, X., Xia, J.-y., Chu, J., Zhuang, Y.-p. and Zhang, S.-l. (2012). Real-time fluid dynamics investigation and physiological response for erythromycin fermentation scale-up from 50 L to 132 m<sup>3</sup> fermenter. *Bioprocess Biosyst. Eng.* 35:789-800.
- Xu, P., Li, L., Zhang, F., Stephanopoulos, G. and Koffas, M. (2014). Improving fatty acids production by engineering dynamic pathway regulation and metabolic control. *Proc. Natl. Acad. Sci.* 111:11299-11304.

Strain/ Plasmid	Genetic Information
JCL260 (Strain)	BW25113/F'[traD36, proAB+ ,lacIq ZΔM15] Δ <i>adhE</i> , Δ <i>frdBC</i> , Δ <i>fnr</i> , Δ <i>ldhA</i> , Δ <i>pta</i> , Δ <i>pf1B</i>
pSA65 (Plasmid)	ColE1ori;AmpR; PLlacO1: <i>kivd-adhA</i> ( <i>Lactococcus lactis</i> )
pSA69 (Plasmid)	P15ori;KanR; PLlacO1: <i>alsS</i> ( <i>Bacillus subtilis</i> )- <i>ilvCD</i>

Table 2.1 Detailed information of plasmids and strains used in this study

NO	Host Strain/ Plasmid	Growth rate (h <sup>-1</sup> )	Biomass yield (g/g glucose)	Acetate yield (g/g glucose)	Isobutanol yield (g/g glucose)	Glucose uptake rate (mmol/gDW*h)
1	JCL260	0.345	0.432	~0	0	3.921
2	JCL260/pSA65	0.126	0.240	~0	0.143	3.968
3	BW25113/pSA65	0.095	0.353	0.065	0.098	1.943
4	BW25113/ (pSA65+pSA69)	0.091	0.354	0.114	0.097	3.249
5	JCL260/ (pSA65+pSA69)*	0.112	0.390	0.350	0.360	7.589

\* Data adopted from previous experiments published by Tang lab, with 5 g/L yeast extract in bioreactor (Xiao *et al.* 2012).

Physiological information of strain 1-4 came from <sup>13</sup>C-MFA experiments in this study, while physiological data of strain 5 was from a paper published by our lab.

Table 2.2 Physiological information of strains used in this study

'Glucose + ATP == G6P'
'G6P == F6P'
'F6P + ATP == FBP'
'FBP == F6P'
'FBP == DHAP + GAP'
'DHAP == GAP'
'GAP == G3P + ATP + NADH'
'G3P == PEP'
'PEP == PYR + ATP'
'PYR + 2*ATP == PEP'
'PYR == AceCoA + CO2 + NADH'
'AceCoA + OAA == CIT'
'CIT == ICIT'
'ICIT == AKG + CO2 + NADPH'
'AKG == SucCoA + CO2 + NADH'
'SucCoA == SUC + ATP'
'SUC == FUM + FADH2'
'FUM == MAL'
'MAL == OAA + NADH'
'MAL == PYR + CO2 + NADH'
'MAL == PYR + CO2 + NADPH'
'PEP + CO2 == OAA'
'OAA + ATP == PEP + CO2'
'ICIT == GLX + SUC'
'GLX + AceCoA == MAL'
'G6P == PG6 + NADPH'
'PG6 == CO2 + Ru5P + NADPH'
'Ru5P == X5P'
'Ru5P == R5P'
'X5P + R5P == GAP + S7P'
'GAP + S7P == E4P + F6P'
'X5P + E4P == GAP + F6P'
'PG6 == PYR + GAP'
'AceCoA == Ac + ATP'
'AKG + NADPH == GLU'
'GLU + ATP == GLN'
'GLU + ATP + 2*NADPH == PRO'
'GLU + GLN + CO2 + ASP + AceCoA + 5*ATP + NADPH == ARG + AKG + FUM + Ac'
'OAA + GLU == ASP + AKG'
'ASP + 2*ATP == ASN'
'PYR + GLU == ALA + AKG'
'G3P + GLU == SER + AKG + NADH'

'SER == GLY + Methylene_THF'
'GLY == Methylene_THF + CO2 + NADH'
'Methylene_THF + NADH == Methyl_THF'
'Methylene_THF == Formyl_THF + NADPH'
'ASP + 2*ATP + 2*NADPH == THR'
'THR == GLY + AceCoA + NADH'
'SER + AceCoA + 3*ATP + 4*NADPH == CYS + Ac'
'ASP + PYR + GLU + SucCoA + ATP + 2*NADPH == LYS + CO2 + AKG + SUC'
'ASP + Methyl_THF + CYS + SucCoA + ATP + 2*NADPH == MET + PYR + SUC'
'GLU + NADPH + 2*PYR == VAL + AKG + CO2'
'AceCoA + 2*PYR + GLU + NADPH == LEU + AKG + NADH + 2*CO2'
'THR + PYR + GLU + NADPH == ILE + AKG + CO2'
'E4P + 2*PEP + GLU + ATP + NADPH == PHE + AKG + CO2'
'E4P + 2*PEP + GLU + ATP + NADPH == TYR + AKG + NADH + CO2'
'SER + R5P + 2*PEP + E4P + GLN + 3*ATP + NADPH == TRP + GAP + PYR + GLU + CO2'
'R5P + Formyl_THF + GLN + ASP + 5*ATP == HIS + AKG + FUM + 2*NADH'
'NADH == NADPH'
'NADH == 3*ATP'
'FADH2 == 2*ATP'
'ATP == 0*Ex'
'Ac == Ace_measure'
'CO2 == 0*Ex'
'CO2_air + CO2 == CO2_cell + CO2'
0.488*ALA + 0.281*ARG + 0.229*ASN + 0.229*ASP + 0.087*CYS + 0.250*GLU + 0.250*GLN + 0.582*GLY + 0.090*HIS + 0.276*ILE + 0.428*LEU + 0.326*LYS + 0.146*MET + 0.176*PHE + 0.210*PRO + 0.205*SER + 0.241*THR + 0.054*TRP + 0.131*TYR + 0.402*VAL + 0.205*G6P + 0.071*F6P + 0.754*R5P + 0.129*GAP + 0.619*G3P + 0.051*PEP + 0.083*PYR + 2.510*AceCoA + 0.087*AKG + 0.340*OAA + 0.443*Methylene_THF + 33.247*ATP + 5.363*NADPH == 39.68*Biomass + 1.455*NADH
'2*PYR + NADH + NADPH == IB + 2*CO2'

Table 2.3 Metabolic network for <sup>13</sup>C-MFA calculation

	NADH				NADPH				ATP				FADH <sub>2</sub>			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Glycolysis	169	175	173	175	0	0	0	0	109	152	142	154	0	0	0	0
PP pathway	0	0	0	0	44	62	44	39	0	0	0	0	0	0	0	0
TCA cycle	214	119	172	162	52	6	15	8	38	-2	4	-2	55	32	36	25
Amino acid synthesis	25	9	16	12	-123	-57	-87	-76	-68	-29	-46	-38	0	0	0	0
Acetic acid formation	0	0	0	0	0	0	0	0	0	0	14	30	0	0	0	0
Biomass formation	13	7	10	9	-48	-26	-37	-35	-299	-163	-227	-216	0	0	0	0
Isobutanol production	0	-43	-20	-20	0	-43	-20	-20	0	0	0	0	0	0	0	0
One-carbon metabolism	-1	-1	-1	-1	1	0	1	1	0	0	0	0	0	0	0	0
Anaplerotic Pathway*	7	15	10	0	0	0	0	0	0	0	0	-5	0	0	0	0
Net flux from central carbon metabolism	426	281	359	339	-74	-58	-86	-82	-219	-41	-114	-78	55	32	36	25

All flux values are normalized to a glucose uptake rate of 100 mol/h in each strain.

a. The anaplerotic pathway is assumed to only produce NADH and consume minimal amount of ATP.

b. The excessive NADH & FADH<sub>2</sub> are assumed to be converted to ATP via oxidative phosphorylation at maximal P/O ratio (NADH → 3 ATP, FADH<sub>2</sub> → 2 ATP)

Table 2.4 Energy metabolism of different strains



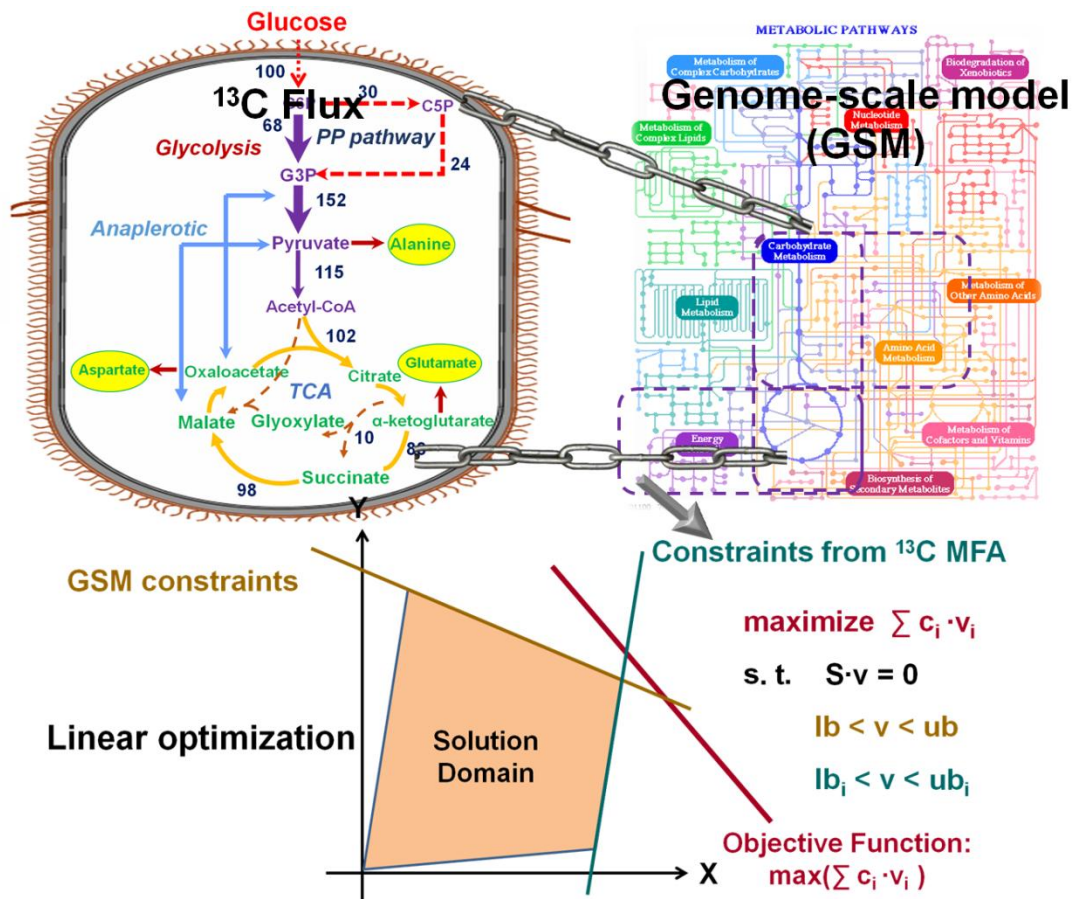
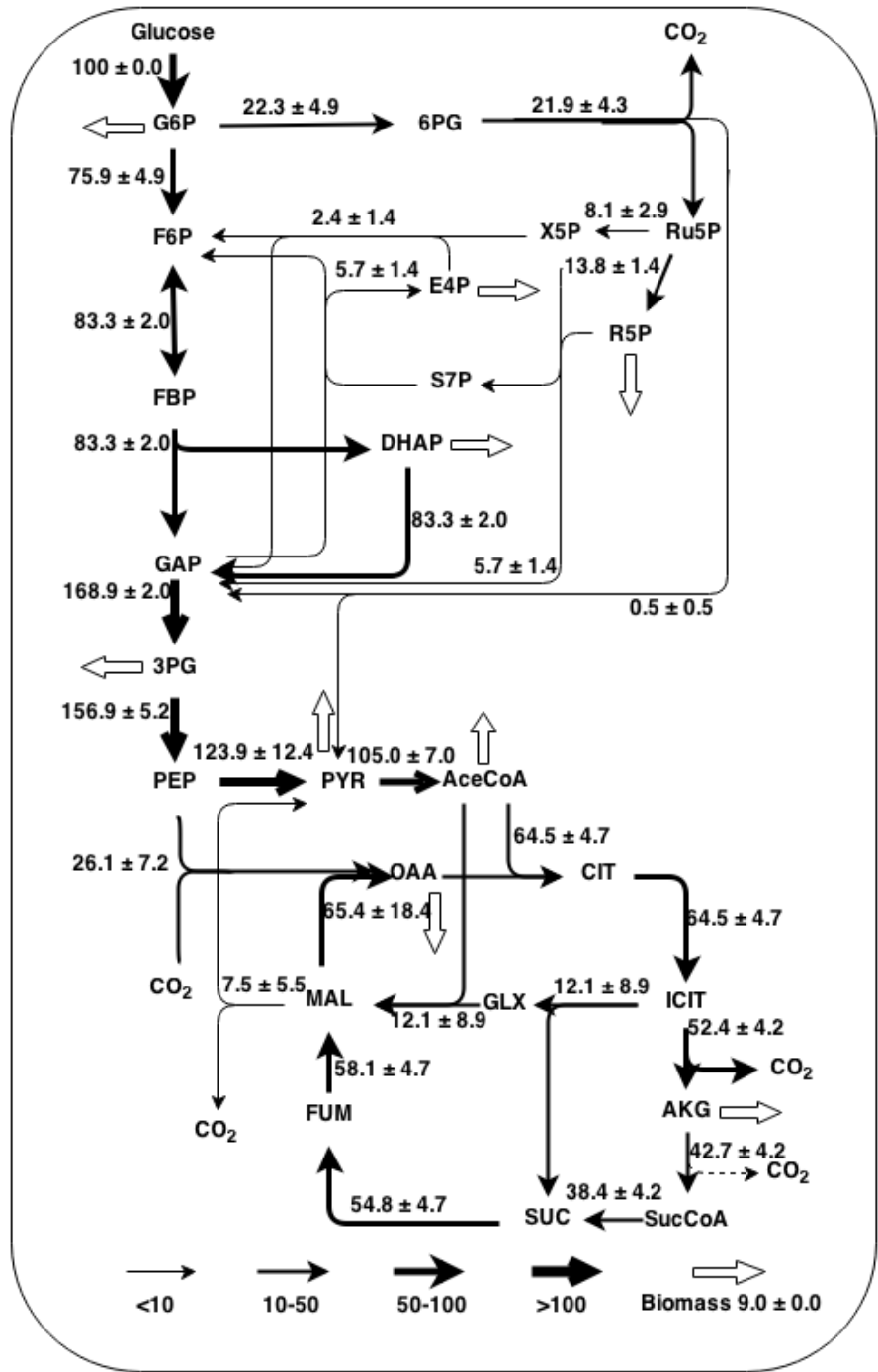
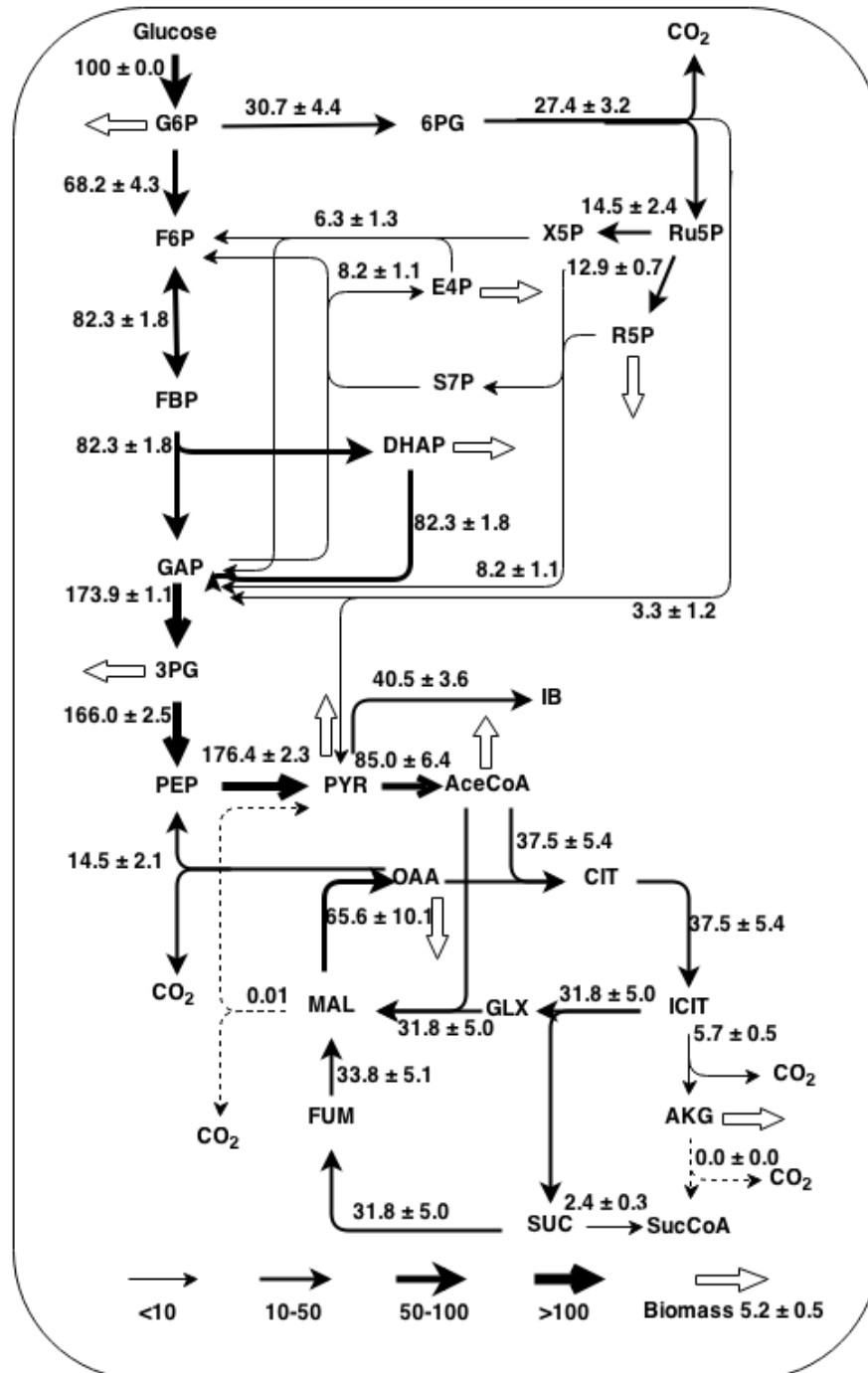


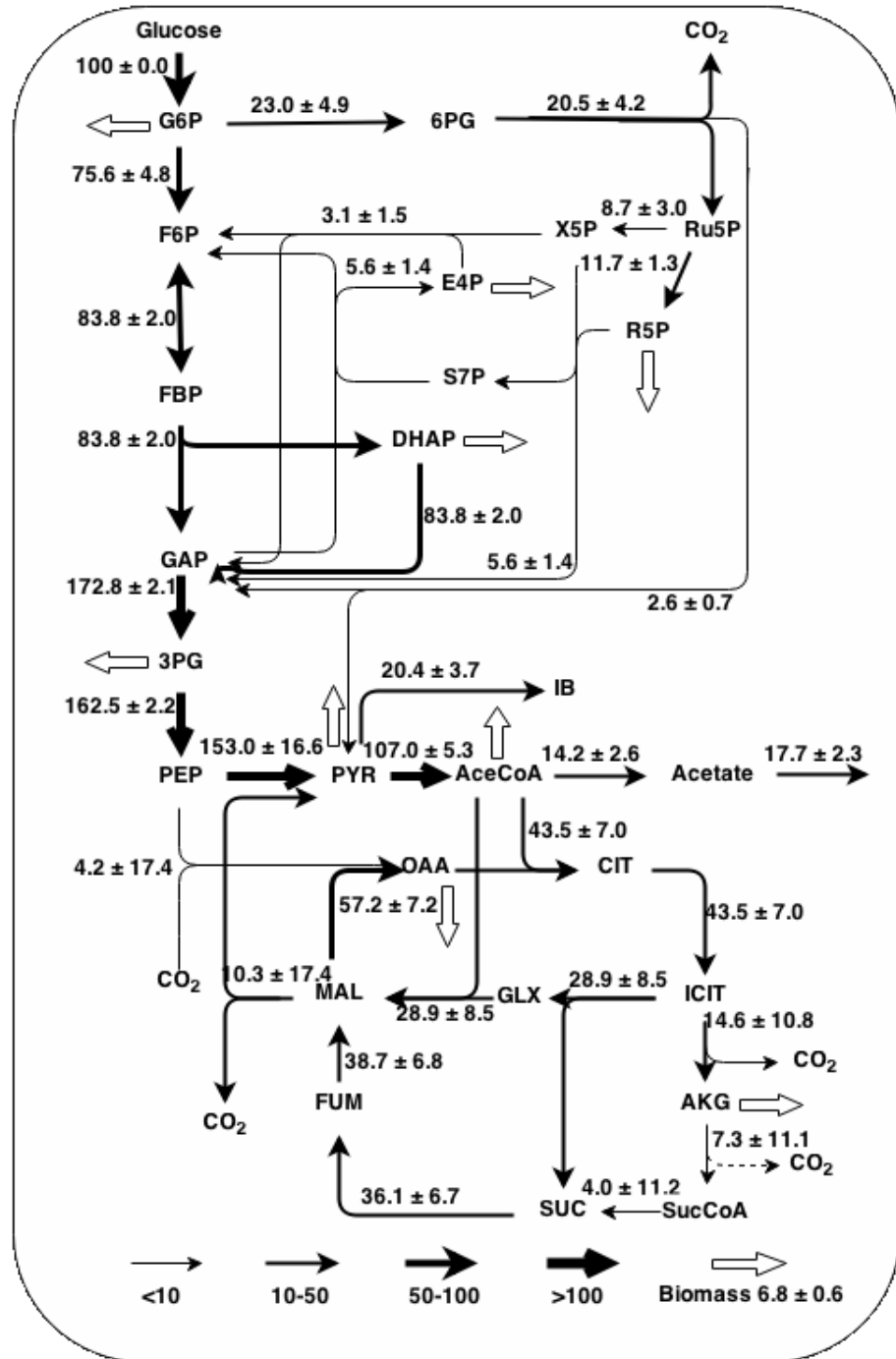
Figure 2.1 Diagram of  $^{13}\text{C}$  flux as constraints for FBA



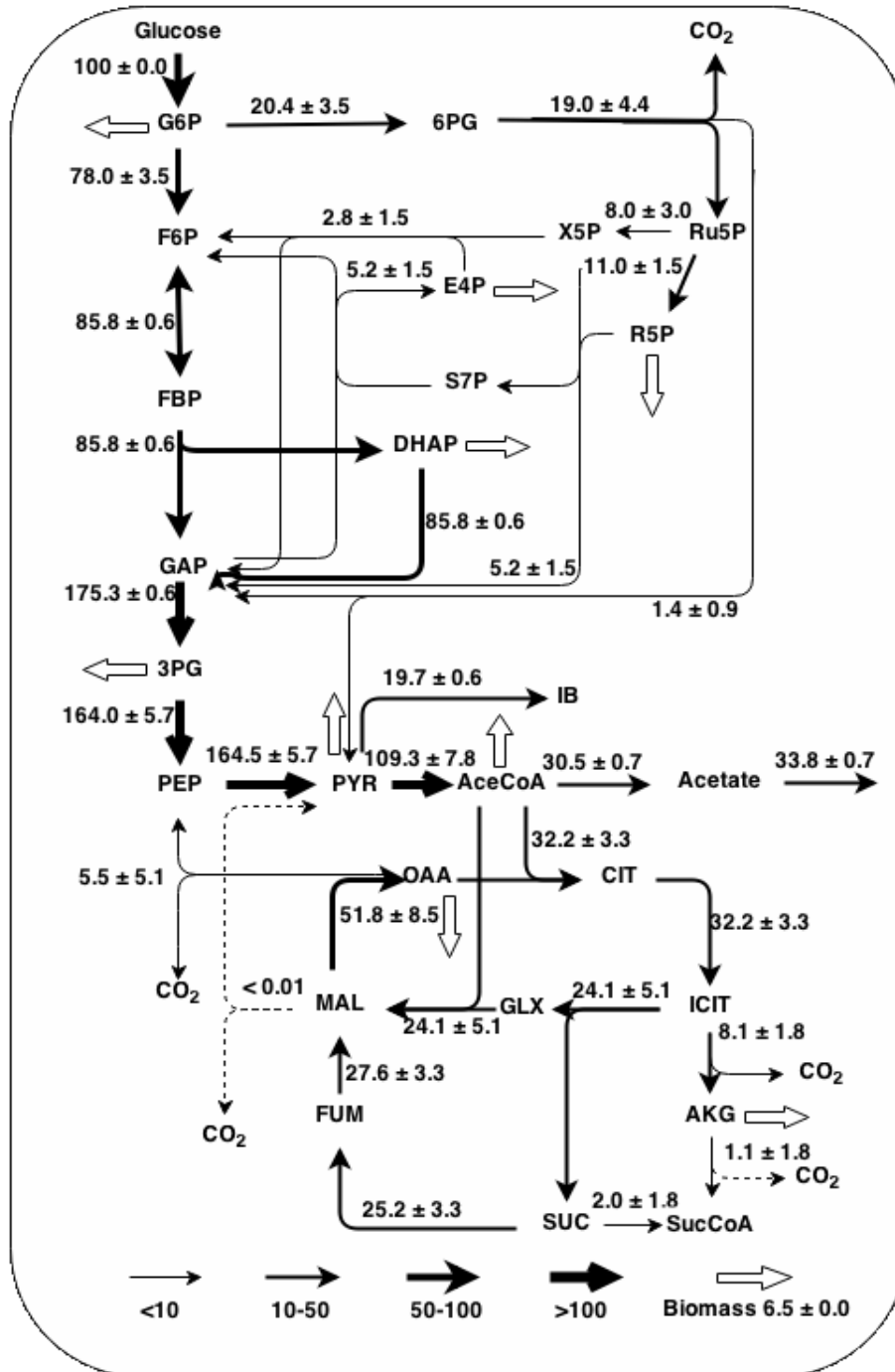
1. JCL260



2. JCL260/pSA65



3. BW25113/pSA65;



4. BW25113/( pSA65 + pSA69).

Figure 2.2. Central metabolic flux determined by <sup>13</sup>C-MFA on control strain (1. JCL260) and three engineered strains (2. JCL260/pSA65; 3. BW25113/pSA65; 4. BW25113/( pSA65 + pSA69)).

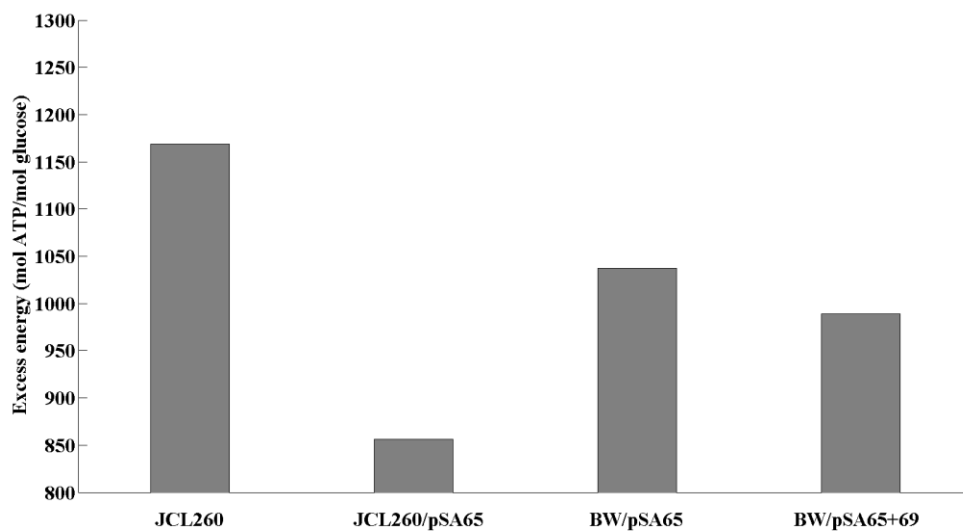


Figure 2.3a Energy analysis of four strains at ideal energy condition (P/O ratio = 3)

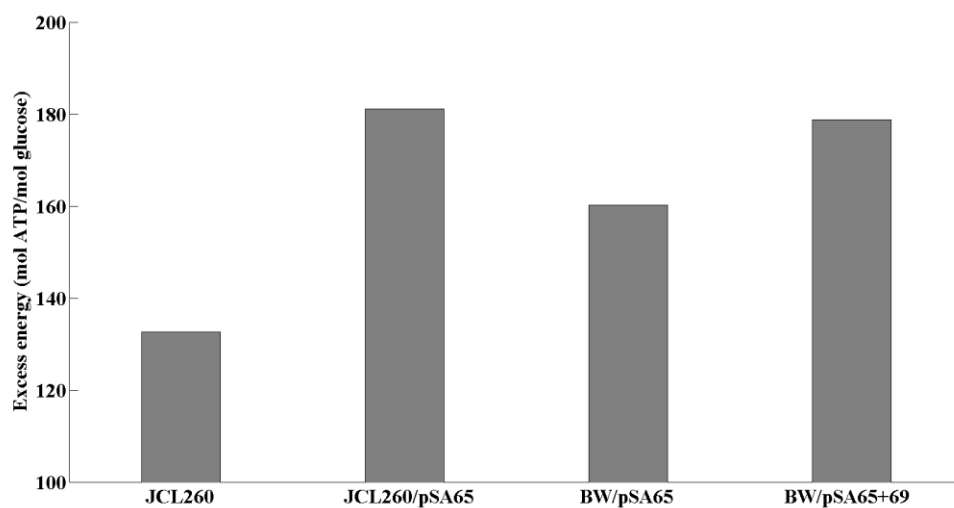


Figure 2.3b Energy analysis of four strains with low energy metabolism (P/O ratio = 1)

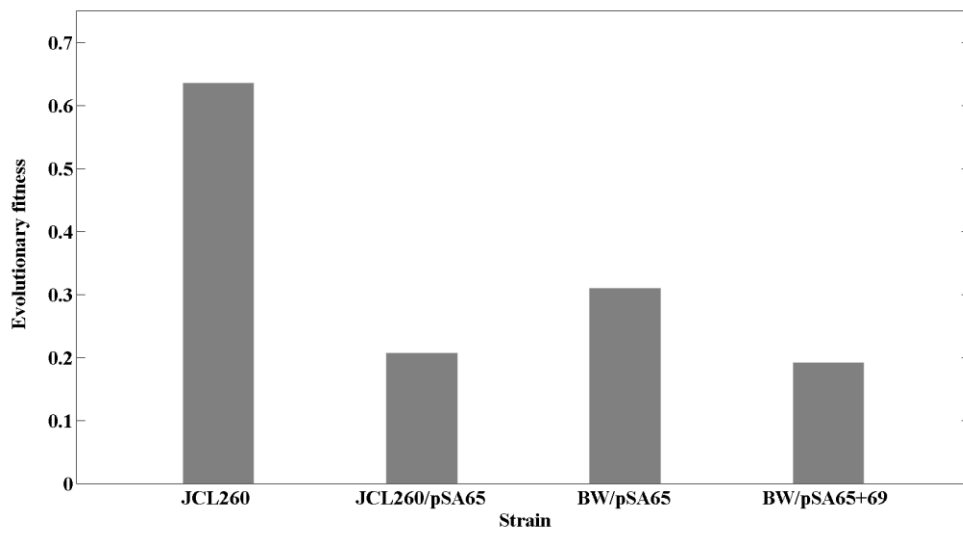
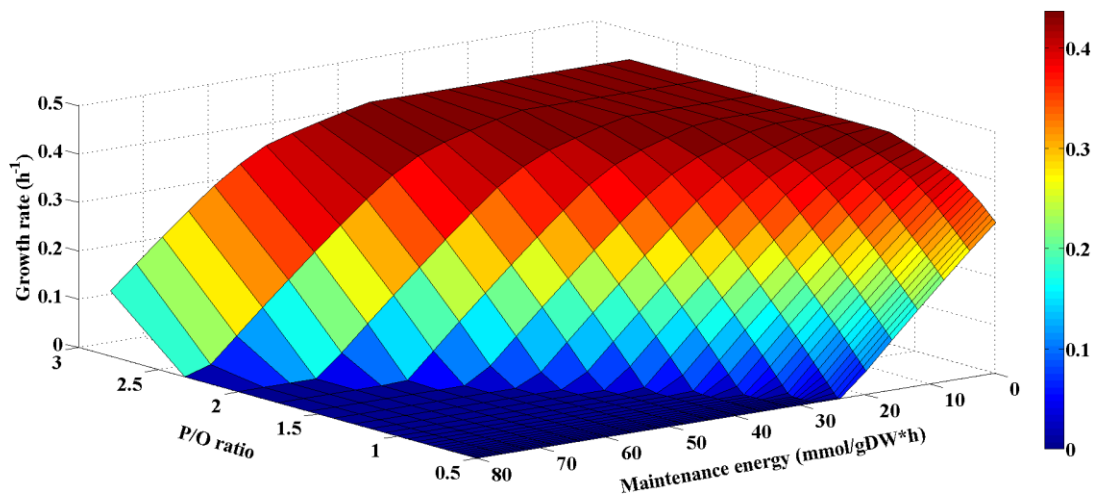
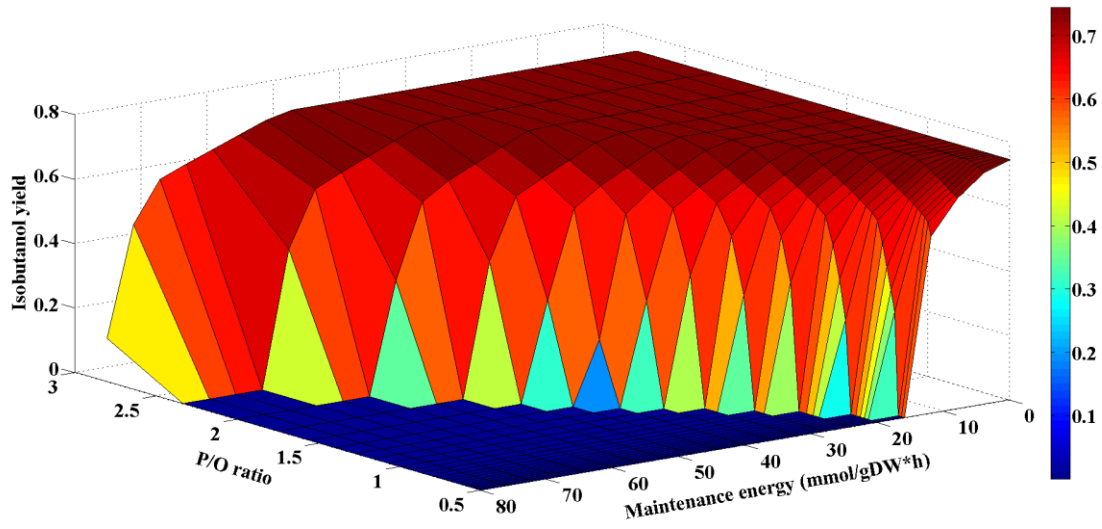


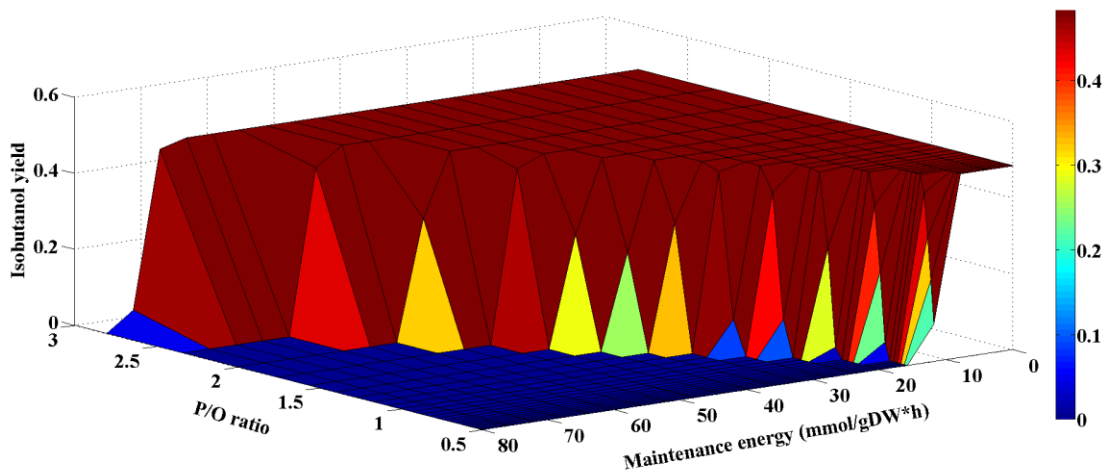
Figure 2.4 Evolutionary fitness of four strains in this study



2.5a. Influence of P/O ratio and maintenance energy on growth rate of strain 1

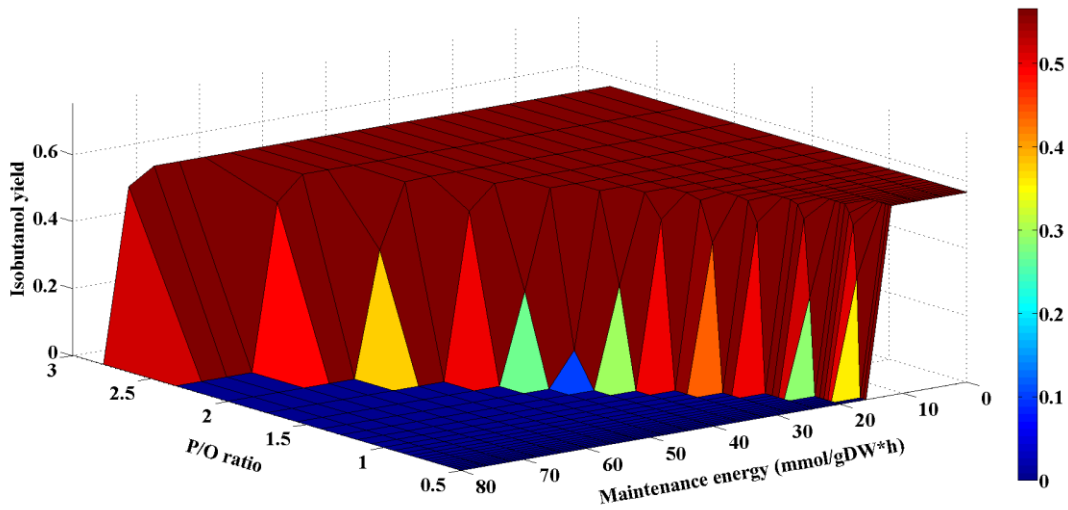


2.5b. Influence of P/O ratio and maintenance energy on isobutanol production of strain 2

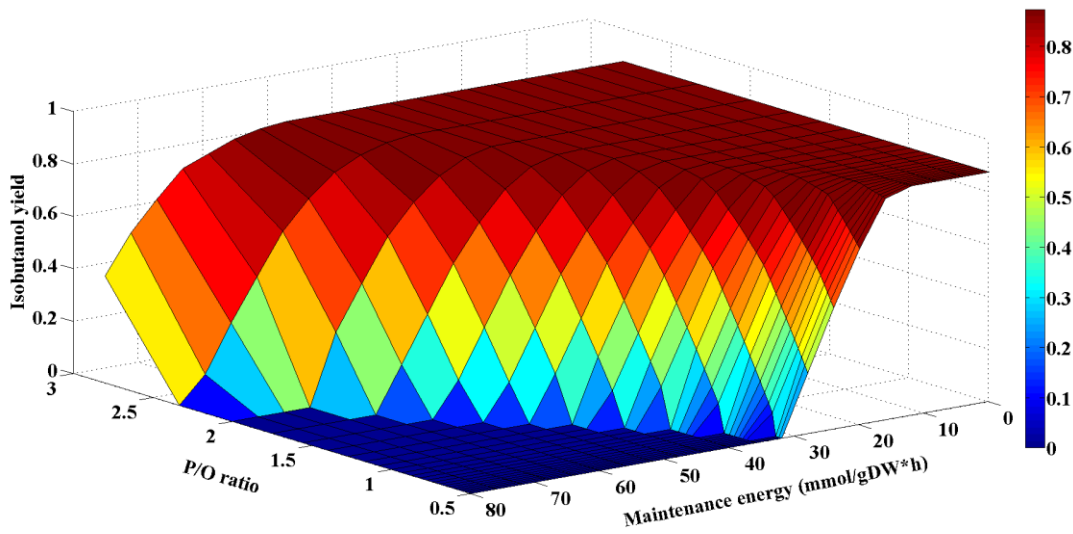


2.5c. Influence of P/O ratio and maintenance energy on isobutanol production of strain 3



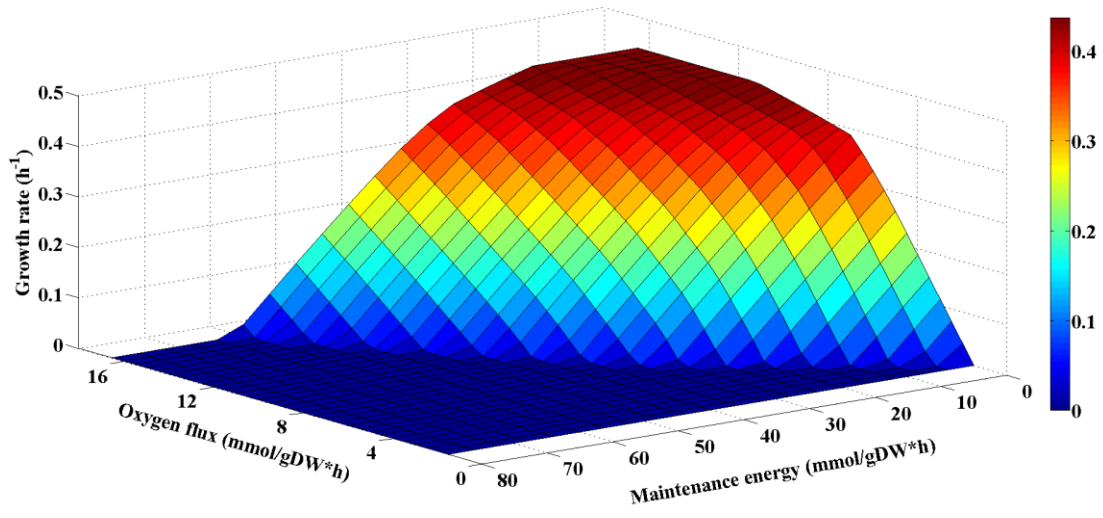


2.5d. Influence of P/O ratio and maintenance energy on isobutanol production of strain 4

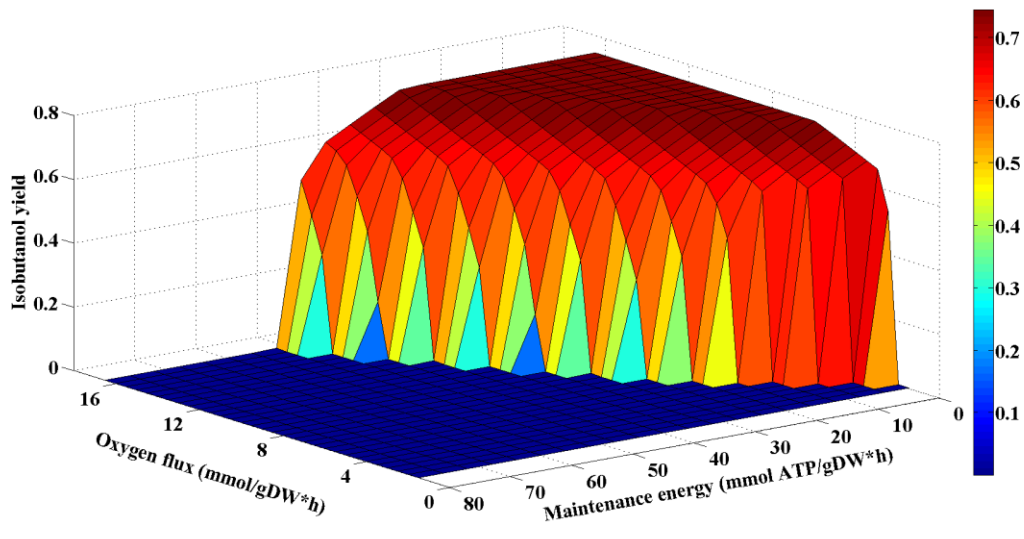


2.5e. Influence of P/O ratio and maintenance energy on isobutanol production of strain 5

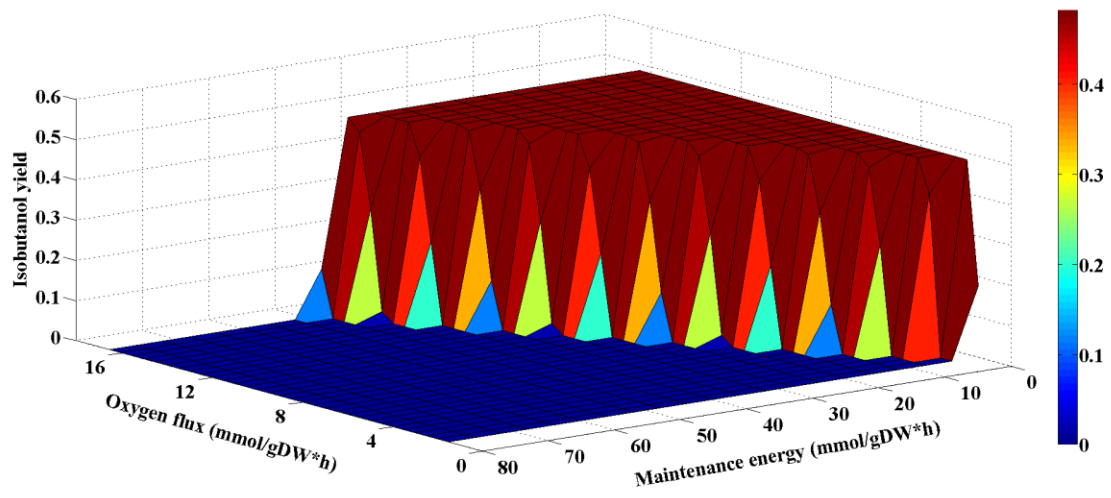
Figure 2.5a-e: Influence of P/O ratio and maintenance energy on isobutanol production potential (growth rate for strain 1, JCL260), simulated by FBA



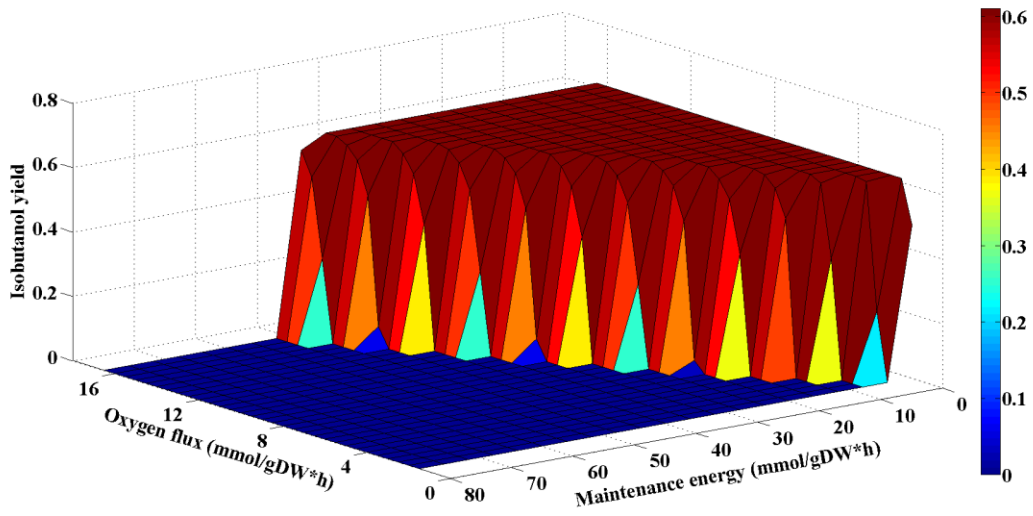
2.6a. Influence of oxygen uptake flux and maintenance energy on growth rate of strain 1



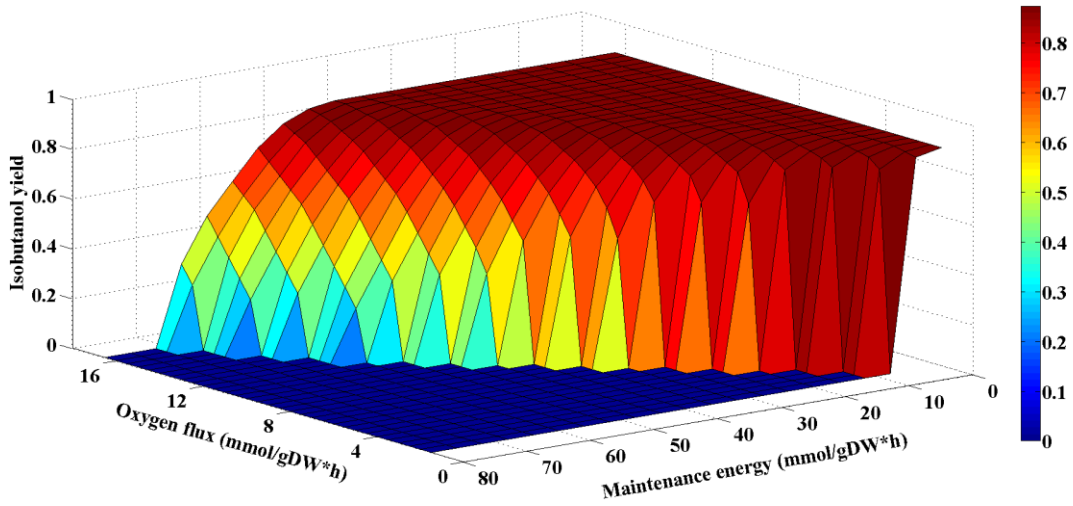
2.6b. Influence of oxygen uptake flux and maintenance energy on isobutanol production of strain 2



2.6c. Influence of oxygen uptake flux and maintenance energy on isobutanol production of strain 3



2.6d. Influence of oxygen uptake flux and maintenance energy on isobutanol production of strain 4



2.6e. Influence of oxygen uptake flux and maintenance energy on isobutanol production of strain 5

Figure 2.6.a-e: Influence of oxygen uptake flux and maintenance energy on isobutanol production potential (growth rate for strain 1, JCL260), simulated by FBA. Default P/O ratio of 1.75 is employed here.

# CHAPTER THREE

## INVESTIGATE ENERGY METABOLISM OF MICROBIAL CELL FACTORIES BY YIN-YANG THEORY

### 3.1. Abstract

In ancient Chinese philosophy, Yin-Yang describes two contrary forces that are interconnected and interdependent. This concept also holds true in microbial cell factories, where Yin represents energy metabolism in the form of ATP, and Yang represents carbon metabolism. Current biotechnology can effectively edit the microbial genome or introduce novel enzymes to redirect carbon fluxes. However, the limitation of the internal powerhouse prevents cells from achieving high carbon yields and rates. It is because that microbial metabolism could lose over 60% of free energy as heat when converting sugar into ATP; while high cell maintenance in microbial hosts further aggravates cellular ATP shortage. Via a flux balance analysis model, we further demonstrate the penalty of ATP expenditure on biofuel synthesis. To ensure cell powerhouse being sufficient for microbial cell factories, we propose five principles: 1. Take advantage of native pathways for product synthesis. 2. Pursue biosynthesis relying only on pathways or genetic parts without significant ATP burden. 3. Combine microbial production with chemical conversions (semi-biosynthesis) to reduce biosynthesis steps. 4. Create ‘minimal cells’ or use non-model microbial hosts with higher energy fitness. 5. Develop a photosynthesis chassis that can utilize light energy and cheap carbon feedstocks. Meanwhile, metabolic flux analysis can be used to quantify both carbon and energy metabolisms and determine ‘the straw that broke

the camel's back'. The fluxomics results are essential to evaluate the industrial potential of laboratory strains, avoiding false starts and dead ends during metabolic engineering.

**Key words:** ATP, energy metabolism, flux analysis, free energy, maintenance loss, semi-biosynthesis

### 3.2. Introduction

In the past decade, molecular biology tools have developed rapidly and now offer new opportunities for metabolic engineering of microbial hosts (Sun and Zhao 2013; Jiang *et al.* 2013; Pratt and MacRae 2009; Qi *et al.* 2013; Wang *et al.* 2009; Isaacs *et al.* 2011). These tools include the selection of plasmids with different copy numbers, promoter engineering, codon optimization, synthetic scaffolds, directed evolution or rational design of enzymes, ribosome binding sites editing, and competitive pathways deletion. Advanced genome engineering (e.g., CRISPRs and TALENs) and automation of conventional genetic techniques (e.g., MAGE) provide efficient capabilities for editing genomes and evolving new functions. At the same time, systems biology (e.g., genomics, transcriptomics, and proteomics) can characterize complex cell networks, mine useful genes, discover new enzymes, reveal metabolic regulations, and screen mutant phenotypes. The advent of these powerful tools seems to lead researchers into a new epoch of bioprocess industries using GMOs (genetically modified organisms) in the near future. However, that is not the whole story.

The golden age of industrial biotechnology dawned in the early 1940s, driven by the mass production of penicillin and enjoyed a fast growth in the 1950s~1980s. Microbial bioprocess has produced diverse commodity chemicals (such as ethanol, amino acids, citric acid,

and lactate) as well as recombinant proteins and antibiotics in the last century. Those commercial products mainly rely on natural strains or strains with minor genetic modifications (usually only one or few new genes). Since the recent decade, in the hope of producing chemicals at low costs and reducing greenhouse gas emissions, an enormous amount of investment has been devoted to metabolic engineering in many nations. Although modern biotechnologies can engineer cells to synthesize diverse products in laboratories, there are only a few GMO products that have become commercially promising in the past decade (e.g., artemisinic acid and 1, 4-butanediol). GMOs are particularly used for biofuel manufacturers, such as short-chain alcohols, fatty acid derived chemicals, and isoprenoid-based biofuels (Atsumi *et al.* 2009; Lindberg *et al.* 2010; Oliver *et al.* 2013). For example, Gevo and Butamax introduce the keto-acid/Ehrlich pathway into yeasts to produce isobutanol (Nielsen *et al.* 2014). Amyris extend the mevalonate pathway in *Saccharomyces cerevisiae* for branched and cyclic terpenes (e.g., farnesene) synthesis. However, these biofuel processes have not achieved strong net profit margin yet. To date, the industrial-scale biofuel is still ethanol, which is cheaply manufactured from sugar cane in Brazil. In this perspective, we will address one of the hidden constraints in microbial cell factories (i.e., Energy metabolism).

### 3.3. The energy losses in microbial cell factories

Heterotrophic organisms obtain free energy in the form of ATP by breaking organic substrates into CO<sub>2</sub> (Figure 3.1). Theoretically, oxidation of one mole of glucose to CO<sub>2</sub> ( $\Delta_c H^\ominus_{298} \approx -2.8$  MJ/mol) can generate 38 moles of ATP. Hydrolysis of these ATP to ADP ( $\Delta G^\ominus = -30.5$  kJ/mol) provide ~1.2 MJ of biochemical energy. Thereby, ~60% of energy from glucose is lost as heat during ATP synthesis (similar to a Carnot heat engine). Cell consumes ATP for diverse

activities, such as nutrient/metabolite transport, chemotaxis, chemical gradient preservation, biomass component repair, and macromolecule re-synthesis (Hoehler and Jorgensen 2013b). These maintenance costs, essential for cell survival and stress adaptation, compete for ATP resources for biomass growth and product synthesis. On the other hand, microbial hosts have not evolved towards optimal energy metabolism. Over billions of years of evolution, microbes with a higher growth rate gained a selective advantage when competing for shared energy resources, but these fast growing species have a lower yield of ATP from substrates (e.g., less than 30 ATP/glucose) (Pfeiffer *et al.* 2001). The oxidative phosphorylation (P/O) ratio represents ATP generation efficiency through substrate oxidation. Theoretically, three ATP can be obtained from the reduction of one oxygen atom (i.e.,  $P/O = 3$ ) during oxidative phosphorylation. Although slow-growing mammalian cells can achieve P/O values close to 3, bacteria and yeasts often have P/O ratios below 2.5 (note: these microbes may dissipate the proton gradient before it can be fully used for charging the ATP synthase). Secondly, microbial hosts may lose ATP yield due to byproducts synthesis, membrane leakage, removal of reactive oxygen species, or suboptimal cultivations (insufficient mixing, shear stress, or biofilm formation). Lastly, the electron transport chain for ATP generation and nutrient transporters may compete for membrane and intracellular spaces so that the capacity of the microbial powerhouse cannot be easily upgraded (Ibarra *et al.* 2002; MacLean and Gudelj 2006).

We introduce the terminology “metabolic entropy” to define the free energy in the substrates that is lost through cellular energy metabolism and becomes unavailable for biosynthesis. Metabolic entropy has gained attention from metabolic flux analysis researchers because the objective function of biomass production in FBA (flux balance analysis) always overestimates microbial growth rates. Moreover, FBA predictions highly depend on the



assumption of a fixed ATP maintenance coefficient. To address this problem, researchers developed  $^{13}\text{C}$ -metabolic flux analysis (MFA) to quantify the microbial “metabolic entropy” directly via tracer experiments. By examining *Bacillus subtilis* mutants,  $^{13}\text{C}$ -MFA has discovered that the suboptimal cell metabolism is associated with the increased energy usage in the face of environmental and random genetic perturbations (Fischer and Sauer 2005). This study suggests that mutating regulatory genes can drive carbon fluxes towards the desired pathways; however, such mutations reduce ATP availability for adaptive responses under adverse environmental conditions (i.e., metabolic engineering achieves microbial productivity by sacrificing their energy fitness).

### **3.4. The tradeoff between product yield and energy fitness**

Traditional metabolic engineering uses plasmids and heterologous enzymes to redirect carbon fluxes. Early studies have shown high copy number plasmids cause significant alterations in cell properties and strongly influence metabolic engineering endeavors (Birnbaum and Bailey 1991).  $^{13}\text{C}$ -MFA of *E. coli* strains revealed higher acetate production and  $\text{O}_2$  uptake rates in plasmid-containing strains than in the plasmid-free strains (Wang *et al.* 2006a). The presence of plasmids can increase cell maintenance, decrease growth rate and change intracellular fluxes, especially suppressing the oxidative pentose phosphate pathway (Ow *et al.* 2009). Similarly, synthetic biology parts (such as novel pathways, protein scaffolds, and genetic circuits) may also increase metabolic entropy if extra nucleic acids and proteins are required to be made by the hosts (note: elongation of one amino acid costs four ATP molecules) (Stephanopoulos *et al.* 1998a). In reality, microbial systems have frugal enzymatic machinery (each native enzyme in a single *E. coli* cell may only have dozens of molecule copies and places minimal biosynthesis burden on cell metabolism) (Taniguchi *et al.* 2010). During pathway engineering, optimizing

enzyme levels is difficult because a large portion of over-expressed enzymes may be inactivated due to protein misfolding. Considerable ATP expenditure for heterologous enzyme synthesis can trigger stress responses and alternate hosts' physiology. For example, <sup>13</sup>C-MFA has been used to examine metabolic burdens in *E. coli* during biosynthesis of recombinant proteins. The results indicate a 25% increase in the total ATP expenditure rate in the highest yielding strain (up to 45 mmol ATP/g CDW/h) (Heyland *et al.* 2011). To overcome such an energy limitation, *E. coli* has to reduce biomass synthesis and enhance oxidative phosphorylation for ATP generation. Besides, microbial hosts often suffer from increased non-growth associated maintenance as well as reduced respiration efficiency (poor P/O ratio) due to membrane stresses (Sauer and Bailey 1999; Varma and Palsson 1994). Therefore, introduction of an extended heterologous pathway into a microbial host often causes deleterious effects on cell metabolism. Ultimately, the host will lose the capability to grow in a minimal carbohydrate medium; while rich nutrients, such as yeast extract (producing 1 g of yeast extract consumes 3 g of glucose), have to be supplied to relieve the cell's energy burden (Xiao *et al.* 2012).

Our theory of energy burden can guide strain development to tolerate product stresses. For instance, an isobutanol-tolerant mutant has been isolated after serial transfers. However, the final isobutanol productivity of this evolved strain did not show improvement (Atsumi *et al.* 2010). The export systems (e.g., ABC transporters) have been engineered for recovering cell growth under biofuel stresses (Dunlop *et al.* 2011). The ATP-driven efflux pumps show limited enhancement of short-chain alcohol productivity (~10%) (Foo *et al.* 2014). On the other hand, efflux pumps work well when they are introduced into low-performance strains, in which their product titers are well below 1 g/L (Dunlop *et al.* 2011). These observations explain the fact that cell adaptation to a stressful environment may require ATP expenditure and thus induce

significant metabolic burdens on biosynthesis (Zhang and Lynd 2005). For the same reason, tolerance engineering often works well on yeast strains for ethanol production because of simple synthesis pathway and net ATP generation from glycolysis. For example, engineering transcriptional machinery or up-regulation of the potassium/proton pumps in *Saccharomyces cerevisiae* can significantly improve ethanol tolerance and the production titer (well above 100 g/L) (Alper *et al.* 2006; Lam *et al.* 2014). In conclusion, when the microbial hosts already have high biosynthesis burdens, we should focus on specific regulatory genes rather than efflux pumps. For example, a methionine biosynthesis regulator can significantly improve both biofuel tolerance and productivity in *Escherichia coli* (Foo *et al.* 2014). In yet another case, the inactivation of a histidine kinase may enhance the butanol productivity in *Clostridium acetobutylicum* by delaying cell sporulation (Xu *et al.* 2015).

### **3.5. Sensitivity analysis of the energy penalty on biofuel synthesis**

We employ a genome-scale flux balance model (iJO1366) to simulate the adverse impacts of *E. coli* energy metabolism on biofuel product yields (Figure 3.2) (Orth *et al.* 2011). Apart from the intracellular stress caused by enzyme overexpression, the release of large amounts of biofuel molecules (alcohol or fatty acid) will interfere enzymatic reactions *in vivo* and disrupt the cellular membrane's integrity, which results in reduced efficiencies of oxidative respiration (Lennen *et al.* 2011; Atsumi *et al.* 2010). Thereby, metabolic engineering approaches are effective in redirecting carbon fluxes to biosynthesis only in these low-productivity strains whose energy metabolism are not overloaded. We have FBA test the penalty of metabolic burdens (such as maintenance cost) and the decrease of P/O ratio on biofuel yields. The simulations show that microbial energy metabolism is usually abundant so that they can support

certain amount of metabolic burdens without having apparent biosynthesis deficiency (e.g., without having a slower growth). However, cell burden may increase during the routine genetic modifications. When cell powerhouse is unable to fully afford the increasing ATP expenditure, the biosynthesis yield will have a sudden drop (i.e., the straw that broke the camel's back), forming a cliff in Figure 3.2.

FBA simulations yield two insights into microbial biofuels. First, alcohol (ethanol and isobutanol) producing *E. coli* strains not only have higher carbon yields (0.67 C-product/C-glucose), but also are insensitive to P/O ratios (Figure 3.2a, b). Comparing to isobutanol, ethanol production is less sensitive to the metabolic burden (larger energy sufficient zone). Anaerobic ethanol fermentation, an ancient bioprocess from the beverage industry, does not need additional energy from O<sub>2</sub>, lowering its process costs. From a stoichiometric perspective, glycolysis generates two net ATP per glucose, which fulfills the cell energy expenditure. In addition, ethanol synthesis only needs one native enzyme, and the hosts (e.g., *Saccharomyces cerevisiae*) are naturally tolerant to alcohols. The entire ethanol synthesis pathway is always inside of the cytosol, and thus they do not have mitochondrial transport limitations. These advantages explain why ethanol fermentation (over 100 g ethanol/L) is superior to any other biofuel processes.

Second, energy metabolism may become a critical issue for synthesizing fatty acid-based compounds, which are susceptible to changes in P/O ratio, ATP maintenance loss, and oxygen uptake fluxes. Comparing to alcohol production, fatty acid based fuels (such as biodiesel) requires longer biosynthetic pathways (more enzymes to overexpress) and considerable ATP usage for product synthesis (Steen *et al.* 2010). Besides, many enzymes in fatty acid pathway are tightly regulated during cell growth, leading to growth associated bio-production. The simultaneous biomass growth and fatty acid synthesis further exaggerates ATP shortage (He *et al.*

2014). Therefore, aerobic fermentation has to be performed to enhance energy metabolism, which reduces product yield and increases the fermentation costs for aeration. Furthermore, the accumulation of fatty acid damages cell membrane and reduces oxidative phosphorylation efficiency. To demonstrate these synergistic effects on fatty acid yields, Figure 3.2c and d simulate *E. coli* fatty acid yields under different P/O ratios and metabolic burdens. As shown in Figure 3.2c, fatty acid production can achieve a similar yield as ethanol if the host's biomass growth rate (as  $0.05 \text{ h}^{-1}$ ) and energy maintenance is not high. In reality, fatty acid yield can drop to 50% or less of the theoretical maximum, which is consistent with the model prediction if we considered a practical biomass growth, extra ATP maintenance, and low P/O ratio ( $< 1.5$ ) in FBA (blue star in Figure 3.2d) (Orth and Palsson 2012). Figure 3.2d also indicates the high sensitivity of fatty acid yield in response to the P/O ratio (red star in Figure 3.2d). For instance, one unit change in P/O ratio leads to an abrupt drop in fatty acid yield -- from a theoretical maximum to zero.

### **3.6. Yin-Yang theory in metabolic engineering**

To better understand the limitations of microbial cell factories, we refer to an ancient Chinese wisdom: Yin-Yang. Yin-Yang describes both the bright side and dark side of an object in the world. Yin and Yang oppose each other but are also interdependent. In the case of metabolic engineering, the microbial metabolism is operated by thousands of enzymatic reactions and mass transport processes that involve both carbon (Yang) and energy (Yin) transformations (Figure 3.1). Through billions of years of evolution and environmental adaptations, biological systems have evolved closely interdependent carbon fluxes for biomass growth and energy fitness, which are similar to the intertwined Yin-Yang forces. Although it is easy to engineer microbial hosts to produce small amounts of diverse products, manufacturing a particular compound with titers and

rates beyond the economic break-even point is difficult. In microbial conversions of a substrate to a product, metabolic entropy always increases if more carbon flux is redirected to the final products (Figure 3.3a & b). Figure 3.3c calculates the energy loss during conversion of glucose to biofuels at their theoretical yields as well as the practical yields. Based on the stoichiometry, theoretical energy losses during the conversions of glucose to biofuel molecules (alcohols and fatty acids) are small. However, much bigger losses are observed in real cases because of the suboptimal energy metabolism in biological systems.

To leverage the “Yin-Yang” balance, early metabolic engineers tried a few effective approaches to promote energy metabolism and boost productivity. For instance, *Vitreoscilla* hemoglobin (VHb), a soluble bacterial protein, has been used to enhance energy metabolism by promoting oxygen delivery, which can significantly improve cell growth and enhance chemical production under oxygen-limited conditions (Wei and Chen 2008). Furthermore, an energy-conserving pathway in *E. coli* was developed through metabolic evolution for high production of succinate from glucose fermentation (Zhang *et al.* 2009). This study indicates that the overexpression of a phosphoenolpyruvate carboxykinase increases the net production of ATP, compared to the primary mixed acid fermentation pathway via PEP carboxylase. The extra energy supply allows *E. coli* to produce succinate close to the theoretical maximum. In another case, an ATP-consuming reaction was introduced into *S. elongatus* PCC 7942 to drive carbon flux from acetyl-CoA to 1-butanol (Lan and Liao 2012). This study of 1-butanol production further validates that the ATP coupling reaction can make engineered pathways thermodynamically more favorable. To this end, we summarize the following suggestions to overcome the energy roadblocks.

First, a clear understanding of the entire carbon and energy metabolisms in microbial species would help us to conquer the energy limitations. Using *E. coli* as an example, ATP significantly impacts the product distributions at the pyruvate node (Wang *et al.* 2010a). Understanding ATP fluxes can offer rational design of *E. coli* strains for improving product biosynthesis (Zhang *et al.* 2009; Causey *et al.* 2003). Flux balance analysis (FBA) and <sup>13</sup>C-metabolic flux analysis (MFA) are the only available tools that can quantify energy expenditures. FBA can characterize cell energy metabolism by dividing ATP cost into non-growth associated loss and growth-associated maintenance (Varma and Palsson 1994). Due to the metabolic nature of suboptimal carbon fluxes, FBA, relying on the objective functions, may overestimate the cell potential for biosynthesis capability. <sup>13</sup>C-MFA uses tracer experiments to constrain the FBA model so that it can precisely measure enzyme reaction rates. <sup>13</sup>C-MFA can profile carbon fluxes through all energy generation/consumption pathways and deduce energy flows in the cell metabolism (ATP and cofactor balancing) (He *et al.* 2014). Flux analysis not only allows us to determine the hidden Yin-Yang balance and to design rational engineering strategies, but also to characterize metabolic entropy and identify a strain's energy potential for further improvement. Although <sup>13</sup>C-MFA has not been widely accepted as a routine laboratory measurement tool to assess the engineered microbial hosts, this technology has excellent potential to reveal pathway engineering burdens (i.e., predict “the last straw” in genetic modifications). This tool can informatively tell metabolic engineers and project sponsors what can be done and what cannot be done.

Second, metabolic engineers need to exploit native pathways and avoid extensive pathway reconstruction. In history, many industrial successful cases of improved strain tolerance or productivity just relied on random mutation or evolution, leveraging Natural Selection of mutants for the best balance of ‘Yin-Yang’. Additionally, efforts should aim product synthesis at

pathways that do not require significant ATP expenditures (such as ethanol or organic acids). For example, the acetate overproduction pathway in *E. coli* generates abundant ATP, and the engineered strain performs very well even when its oxidative phosphorylation, TCA cycle and competing fermentation pathways are disrupted (Causey *et al.* 2003). When microbial hosts have low-burden biosynthesis pathways, they show robustness in industrial processes. Moreover, artificial synthetic circuits, efflux pumps, or novel pathways should be carefully considered in terms of the energy penalty. By revealing the tradeoffs behind synthetic biology parts via flux analysis approach, engineering strategies can be rationally designed.

Thirdly, although it is difficult to break the Yin-Yang balance in a natural microorganism, synthetic biologists may re-program the carbon metabolism and energy “fitness” by engineering novel microbial systems. Metabolic engineers often apply gene deletions, evolutionary engineering, or pathway overexpression to improve the strain productivity. These practices typically encounter adverse metabolic shifts due to energy imbalances. However, the creation of a “minimal or smart” cell can remove unnecessary genes in microbial hosts in effort to reduce cell burden and unlock the biosynthesis regulations (Forster and Church 2006; Trinh *et al.* 2008). Additionally, synthetic biologists try to design and assemble minimal cells using synthetic chromosomes (Gibson *et al.* 2010). These artificial biological systems do not necessarily follow the natural Yin-Yang balance evolved over billions of years, so they may have an unusually efficient energy metabolism, and thus achieve product yields close to the theoretical maximum.

Fourth, biological conversion can be integrated with non-living processes to reduce the biosynthesis burden. We can use robust microbial hosts to make simple molecules with high yields and titers, and then convert these molecules into a desired product with a complicated structure via biological and chemical processes. For example, the Keasling Lab achieved the



total synthesis of artemisinin with a two-stage semi-synthetic approach. They used the mevalonate pathway in *Saccharomyces cerevisiae* to synthesize artemisinic acid, followed by a four-step chemical conversion of artemisinic acid to artemisinin (Paddon and Keasling 2014). The Zhang Lab has made biopolymers by using engineered *E. coli* as a first step, to produce a simple molecule mevalonic acid, and then chemically converting it into biopolymers (Xiong *et al.* 2014). A significant advantage of these integrated processes is an extremely efficient bioconversion first step using a short pathway (Colletti *et al.* 2011). For instance, the titer of the semi-product mevalonic acid can reach as high as 88 g/L because its synthesis only requires three steps from the central metabolic node (acetyl-CoA) (Xiong *et al.* 2014). In another and more radical approach, an artificial cell-free system containing enzyme cocktails can mimic one or many functions of a biological system. Such systems can be used to synthesize products with near maximum theoretical yields (Hodgman and Jewett 2012; Ye *et al.* 2009). Cell-free systems can be designed to achieve optimal biosynthesis without cell maintenance cost.

Lastly, development of non-model microbial workhorses with desired traits in energy metabolism may achieve higher biosynthesis potentials, enabling the design of industrial biorefineries for the production of a broad range of products. In fact, even in the modern era of genomics, it is estimated that > 99% of all bacterial species remain unknown (Lasken and McLean 2014). Some non-model species might have a unique energetics that can facilitate product synthesis. For example, Algenol develops the engineered cyanobacteria for phototrophic ethanol production from CO<sub>2</sub> (<http://www.algenol.com/>). Moreover, cyanobacterial species have shown faster growth and higher production rate/titer by co-utilization of organic substrates (You *et al.* 2014; Atsumi *et al.* 2009). Cyanobacterial photo-fermentations, may facilitate cost-effective and large-scale biorefineries by using cheap feedstocks, CO<sub>2</sub>, and light energy. In fact,

Nature is the best synthetic biologist and may have already prepared us an excellent chassis that we have not discovered yet. When we try to out-do Nature's performance, we must first assimilate her lessons of 'Yin-Yang'.

### **3.7. Conclusions**

We have discussed the Yin-Yang concept as the underlying regulatory mechanism in cell metabolism. Biosynthesis of diverse useful products requires sophisticated genetic pathway engineering to steer a high flux to the final product while energy fitness requires the cell metabolism to be minimally changed. Since the powerhouse in microbial cell factory is not limitless, energy shortage eventually leads to metabolic shifts and reduced cell productivity in engineered microbes. The Yin-Yang balance may caution against the assumption that the host metabolism can be modified extensively to produce any desired products. By using fluxomics, we can formulate guidelines to avoid many false starts and dead ends during metabolic engineering. In addition, industrial bioprocess always faces numerous constraints and trade-offs (mass transfer limitations in fermentation, sterilization, strain stability, contaminations, and aeration costs). Feedstock selections, downstream product separation, and waste treatment are critical issues that impact product profitability. Thus, the design-build-test-learn cycle should cover both strain development and economic analysis. Nevertheless, the Yin-Yang philosophy provides general insights into biotechnology tradeoffs.

### **3.8. References**

1. Sun N, Zhao H: Transcription activator-like effector nucleases (TALENs): A highly efficient and versatile tool for genome editing. *Biotechnol Bioeng* 2013, 110:1811-1821.

2. Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA: RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat Biotech* 2013, 31:233-239.
3. Pratt AJ, MacRae IJ: The RNA-induced silencing complex: a versatile gene-silencing machine. *J Biol Chem* 2009, 284:17897-17901.
4. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA: Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell* 2013, 152:1173-1183.
5. Wang HH, Isaacs FJ, Carr PA, Sun ZZ, Xu G, Forest CR, Church GM: Programming cells by Multiplex Genome Engineering and Accelerated Evolution. *Nature* 2009, 460:894-898.
6. Isaacs FJ, Carr PA, Wang HH, Lajoie MJ, Sterling B, Kraal L, Tolonen AC, Gianoulis TA, Goodman DB, Reppas NB, *et al*: Precise Manipulation of Chromosomes *in Vivo* Enables Genome-Wide Codon Replacement. *Science* 2011, 333:348-353.
7. Atsumi S, Higashide W, Liao JC: Direct photosynthetic recycling of carbon dioxide to isobutyraldehyde. *Nat Biotech* 2009, 27:1177-1180.
8. Lindberg P, Park S, Melis A: Engineering a platform for Photosynthetic Isoprene Production in Cyanobacteria, using *Synechocystis* as the Model Organism. *Metab Eng* 2010, 12:70-79.
9. Oliver JW, Machado IM, Yoneda H, Atsumi S: Cyanobacterial Conversion of Carbon Dioxide to 2,3-butanediol. *Proc Natl Acad Sci U S A* 2013, 110:1249-1254.
10. Nielsen J, Fussenegger M, Keasling J, Lee SY, Liao JC, Prather K, Palsson B: Engineering Synergy in Biotechnology. *Nat Chem Biol* 2014, 10:319-322.
11. Hoehler TM, Jorgensen BB: Microbial life under Extreme Energy Limitation. *Nat Rev Micro* 2013, 11:83-94.

12. Pfeiffer T, Schuster S, Bonhoeffer S: Cooperation and Competition in the Evolution of ATP-Producing Pathways. *Science* 2001, 292:504-507.
13. MacLean RC, Gudelj I: Resource Competition and Social Conflict in Experimental Populations of yeast. *Nature* 2006, 441:498-501.
14. Ibarra RU, Edwards JS, Palsson BO: *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* 2002, 420:186-189.
15. Fischer E, Sauer U: Large-scale *in vivo* flux analysis shows rigidity and suboptimal performance of *Bacillus subtilis* metabolism. *Nat Genet* 2005, 37:636-640.
16. Birnbaum S, Bailey JE: Plasmid presence changes the relative levels of many host cell proteins and ribosome components in recombinant *Escherichia coli*. *Biotechnol Bioeng* 1991, 37:736-745.
17. Wang Z, Xiang L, Shao J, Wegrzyn A, Wegrzyn G: Effects of the presence of ColE1 plasmid DNA in *Escherichia coli* on the host cell metabolism. *Microb Cell Fact* 2006, 5:34.
18. Ow DS-W, Lee D-Y, Yap MG-S, Oh SK-W: Identification of cellular objective for elucidating the physiological state of plasmid-bearing *Escherichia coli* using genome-scale *in silico* analysis. *Biotechnol Prog* 2009, 25:61-67.
19. Stephanopoulos G, Aristidou A, Nielsen J: *Metabolic Engineering: Principles and Methodologies*. Academic Press; 1998.
20. Taniguchi Y, Choi PJ, Li G-W, Chen H, Babu M, Hearn J, Emili A, Xie XS: Quantifying *E. coli* Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science* 2010, 329:533-538.
21. Heyland J, Blank LM, Schmid A: Quantification of metabolic limitations during recombinant protein production in *Escherichia coli*. *J Biotechnol* 2011, 155:178-184.

22. Varma A, Palsson BO: Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl Environ Microbiol* 1994, 60:3724-3731.
23. Sauer U, Bailey JE: Estimation of P-to-O ratio in *Bacillus subtilis* and its influence on maximum riboflavin yield. *Biotechnol Bioeng* 1999, 64:750-754.
24. Xiao Y, Feng X, Varman AM, He L, Yu H, Tang YJ: Kinetic Modeling and Isotopic Investigation of Isobutanol Fermentation by Two Engineered *Escherichia coli* Strains. *Ind Eng Chem Res* 2012, 51:15855-15863.
25. Atsumi S, Wu TY, Machado IMP, Huang WC, Chen PY, Pellegrini M, Liao JC: Evolution, genomic analysis, and reconstruction of isobutanol tolerance in *Escherichia coli*. *Mol Syst Biol* 2010, 6: 449.
26. Dunlop MJ, Dossani ZY, Szmids H, Chu HC, Lee TS, Keasling JD, Hadi MZ, Mukhopadhyay A: Engineering microbial biofuel tolerance and export using efflux pumps. *Mol Syst Biol* 2011, 7: 487.
27. Foo JL, Jensen HM, Dahl RH, George K, Keasling JD, Lee TS, Leong S, Mukhopadhyay A: Improving Microbial Biogasoline Production in *Escherichia coli* Using Tolerance Engineering. *mBio* 2014, 5(6).
28. Zhang Y-HP, Lynd LR: Cellulose utilization by *Clostridium thermocellum*: Bioenergetics and hydrolysis product assimilation. *Proc Natl Acad Sci USA* 2005, 102:7321-7325.
29. Alper H, Moxley J, Nevoigt E, Fink GR, Stephanopoulos G: Engineering Yeast Transcription Machinery for Improved Ethanol Tolerance and Production. *Science* 2006, 314:1565-1568.

30. Lam FH, Ghaderi A, Fink GR, Stephanopoulos G: Engineering alcohol tolerance in yeast. *Science* 2014, 346:71-75.
31. Xu M, Zhao J, Yu L, Tang IC, Xue C, Yang S-T: Engineering *Clostridium acetobutylicum* with a histidine kinase knockout for enhanced n-butanol tolerance and production. *Appl Microbiol Biot* 2015, 99:1011-1022.
32. Orth J, Conrad T, Na J, Lerman J, Nam H, Feist A, Palsson B: A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism-2011. *Mol Syst Biol* 2011, 7:535.
33. Lennen RM, Kruziki MA, Kumar K, Zinkel RA, Burnum KE, Lipton MS, Hoover SW, Ranatunga DR, Wittkopp TM, Marner WD, Pfleger BF: Membrane Stresses Induced by Overproduction of Free Fatty Acids in *Escherichia coli*. *Appl Environ Microb* 2011, 77:8114-8128.
34. Steen EJ, Kang Y, Bokinsky G, Hu Z, Schirmer A, McClure A, del Cardayre SB, Keasling JD: Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature* 2010, 463:559-562.
35. He L, Xiao Y, Gebreselassie N, Zhang F, Antoniewicz MR, Tang YJ, Peng L: Central metabolic responses to the overproduction of fatty acids in *Escherichia coli* based on <sup>13</sup>C-metabolic flux analysis. *Biotechnol Bioeng* 2014, 111:575-585.
36. Orth J, Palsson B: Gap-filling analysis of the iJO1366 *Escherichia coli* metabolic network reconstruction for discovery of metabolic functions. *BMC Syst Biol* 2012, 6:30.
37. Wei X-X, Chen G-Q: Chapter Fifteen - Applications of the Vhb Gene vgb for Improved Microbial Fermentation Processes. In *Method Enzymol. Volume* Volume 436. Edited by Robert KP: Academic Press; 2008: 273-287

38. Zhang X, Jantama K, Moore JC, Jarboe LR, Shanmugam KT, Ingram LO: Metabolic evolution of energy-conserving pathways for succinate production in *Escherichia coli*. *Proc Natl Acad Sci* 2009, 106:20180-20185.
39. Lan EI, Liao JC: ATP drives direct photosynthetic production of 1-butanol in cyanobacteria. *Proc Natl Acad Sci* 2012, 109:6018-6023.
40. Wang Q, Ou MS, Kim Y, Ingram LO, Shanmugam KT: Metabolic Flux Control at the Pyruvate Node in an Anaerobic *Escherichia coli* Strain with an Active Pyruvate Dehydrogenase. *Appl Environ Microb* 2010, 76:2107-2114.
41. Causey TB, Zhou S, Shanmugam KT, Ingram LO: Engineering the metabolism of *Escherichia coli* W3110 for the conversion of sugar to redox-neutral and oxidized products: Homoacetate production. *Proc Natl Acad Sci* 2003, 100:825-832.
42. Forster AC, Church GM: Towards synthesis of a minimal cell. *Mol Syst Biol* 2006, 2: 45.
43. Trinh CT, Unrean P, Srienc F: Minimal *Escherichia coli* Cell for the Most Efficient Production of Ethanol from Hexoses and Pentoses. *Appl Environ Microb* 2008, 74:3634-3643.
44. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang R-Y, Algire MA, Benders GA, Montague MG, Ma L, Moodie MM, *et al*: Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science* 2010, 329:52-56.
45. Paddon CJ, Keasling JD: Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. *Nat Rev Micro* 2014, 12:355-367.
46. Xiong M, Schneiderman DK, Bates FS, Hillmyer MA, Zhang K: Scalable production of mechanically tunable block polymers from sugar. *Proc Natl Acad Sci* 2014, 111:8357-8362.

47. Colletti PF, Goyal Y, Varman AM, Feng X, Wu B, Tang YJ: Evaluating Factors That Influence Microbial Synthesis Yields by Linear Regression with Numerical and Ordinal Variables. *Biotechnol Bioeng* 2011, 108:893-901.
48. Hodgman CE, Jewett MC: Cell-free synthetic biology: Thinking outside the cell. *Metab Eng* 2012, 14:261-269.
49. Ye X, Wang Y, Hopkins RC, Adams MWW, Evans BR, Mielenz JR, Zhang YHP: Spontaneous High-Yield Production of Hydrogen from Cellulosic Materials and Water Catalyzed by Enzyme Cocktails. *ChemSusChem* 2009, 2:149-152.
50. Lasken RS, McLean JS: Recent advances in genomic DNA sequencing of microbial species from single cells. *Nat Rev Genet* 2014, 15:577-584.
51. You L, Berla B, He L, Pakrasi HB, Tang YJ: <sup>13</sup>C-MFA delineates the photomixotrophic metabolism of *Synechocystis sp.* PCC 6803 under light- and carbon-sufficient conditions. *Biotechnol J* 2014, 9:684-692.
52. Atsumi S, Hanai T, Liao JC: Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature* 2008, 451:86-89.
53. Rachman MA, Furutani Y, Nakashimada Y, Kakizono T, Nishio N: Enhanced hydrogen production in altered mixed acid fermentation of glucose by *Enterobacter aerogenes*. *J Ferment Bioeng* 1997, 83:358-363.



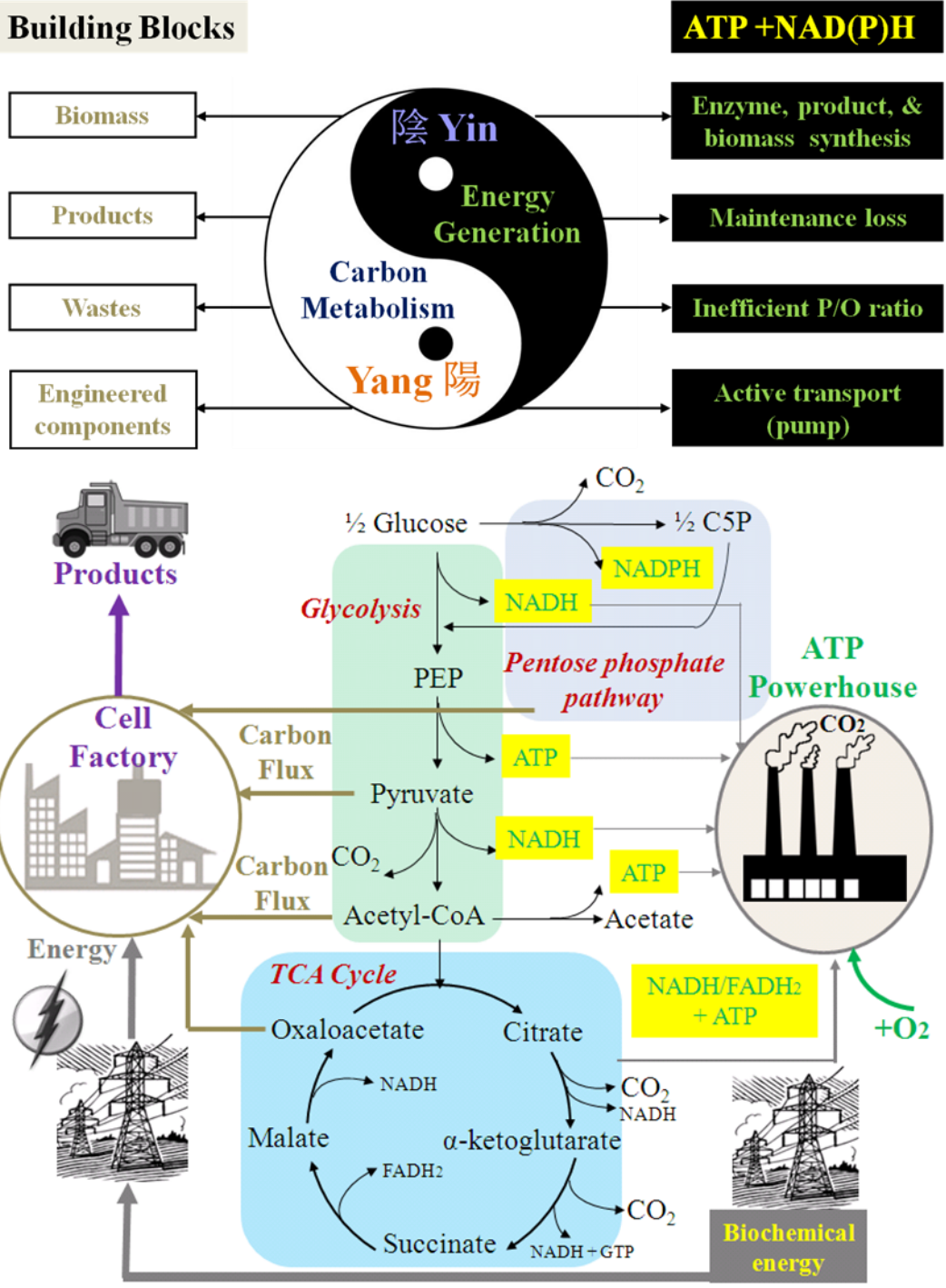


Figure 3.1. Cell carbon and energy metabolism illustrated by Yin-Yang Theory (note: engineered parts include plasmids, over-expressed enzymes, synthetic circuits, etc.)

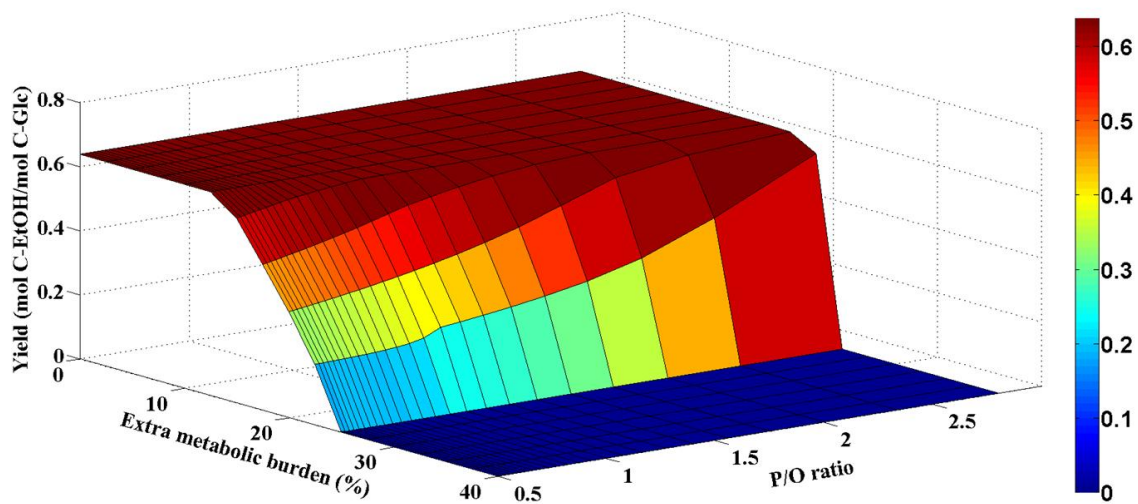


Figure 3.2a *E. coli* strains producing ethanol (growth rate =  $0.05 \text{ h}^{-1}$ )

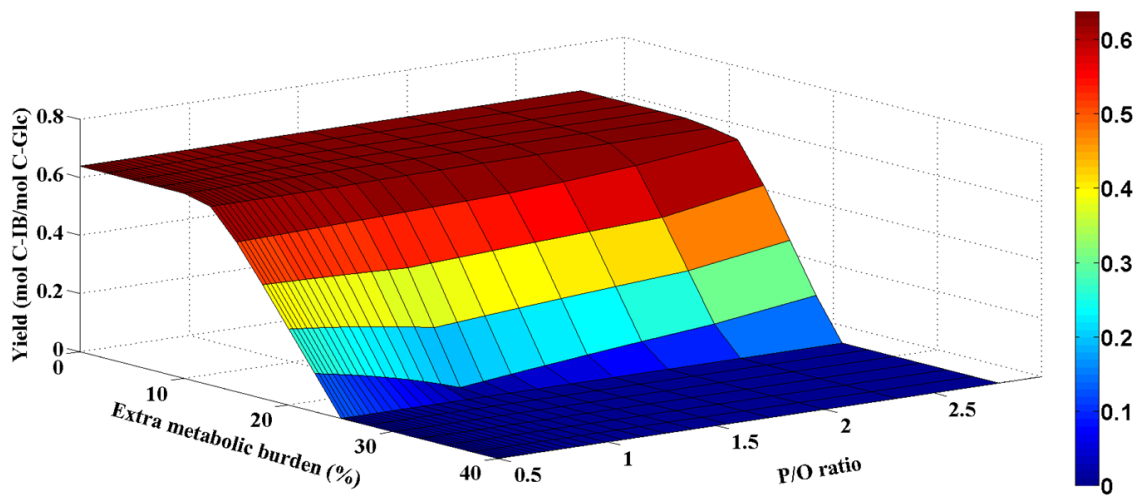


Figure 3.2b *E. coli* strains producing isobutanol (growth rate =  $0.05 \text{ h}^{-1}$ )

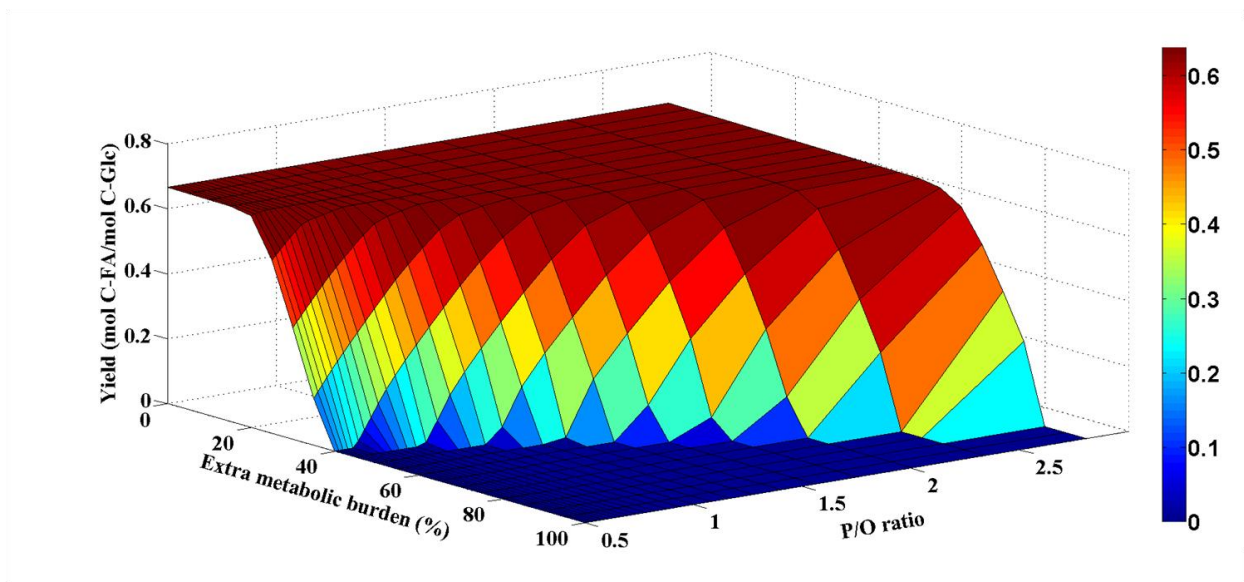


Figure 3.2c *E. coli* strains producing fatty acid (growth rate = 0.05 h<sup>-1</sup>)

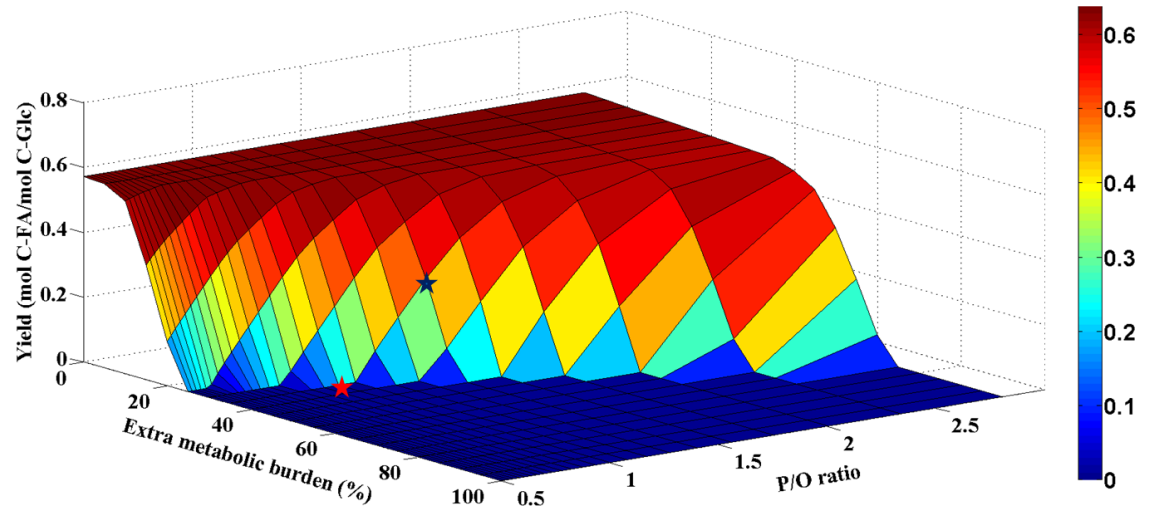
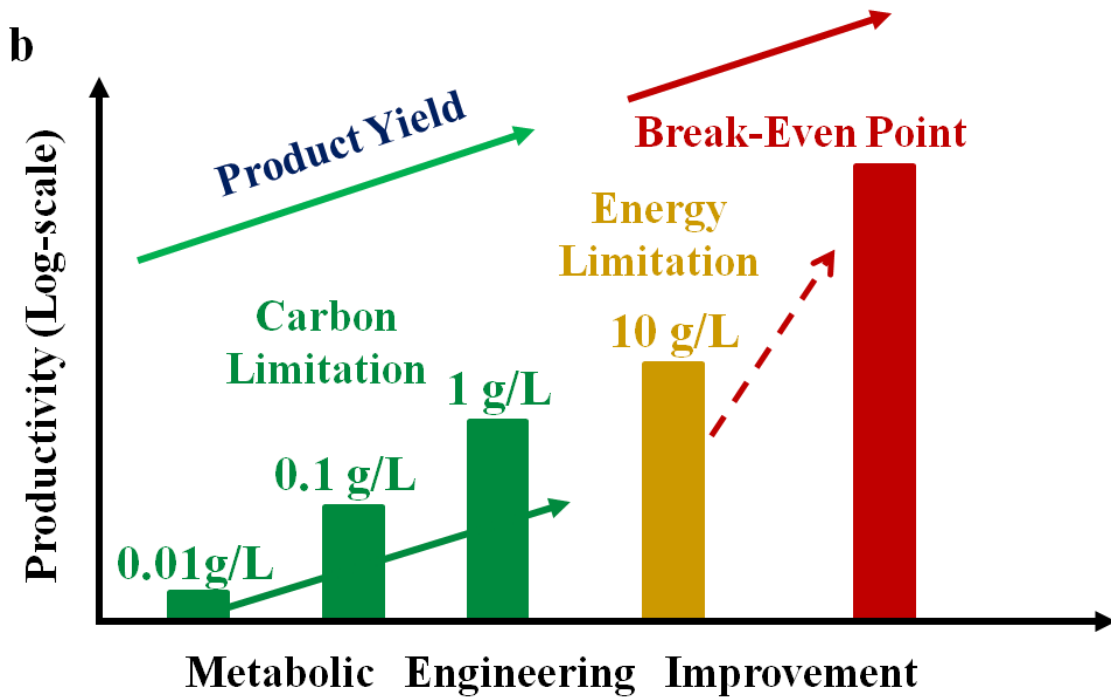
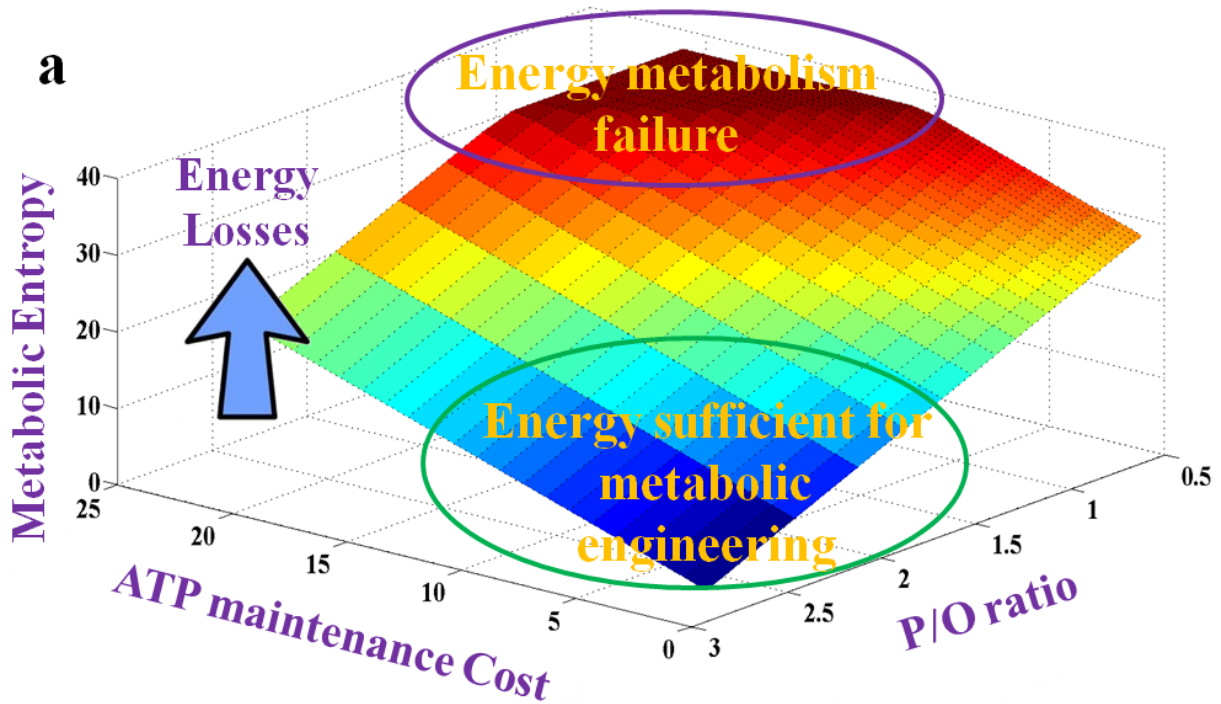


Figure 3.2d *E. coli* strains producing fatty acid (growth rate = 0.20 h<sup>-1</sup>)

Figure 3.2a-d. Genome-scale FBA models for microbial biofuel mole-carbon yields from glucose. We use an *E. coli* FBA model (iJO1366) to predict production of different biofuels from glucose. Alcohol production is simulated under the microaerobic condition ( $O_2$  influx  $\leq 1.85$  mmol/(gDW·h)), while fatty acid is under aerobic condition ( $O_2$  influx  $\leq 12$  mmol/ (gDW·h)). The medium conditions and glucose uptake rate (8 mmol/ (gDW·h)) are same for all FBAs. Extra metabolic burden includes both protein overexpression and maintenance energy increase. Here, 10% extra metabolic burden is equivalent to 10% overexpression of biomass protein plus a proportional increase of non-growth associated ATP loss by 10 mmol ATP/(gDW·h). For each case, the objective function is set as to maximize the biofuel production. Abbreviations: DW (Dry Weight); FA (Fatty acid); Glc (Glucose); IB (Isobutanol).



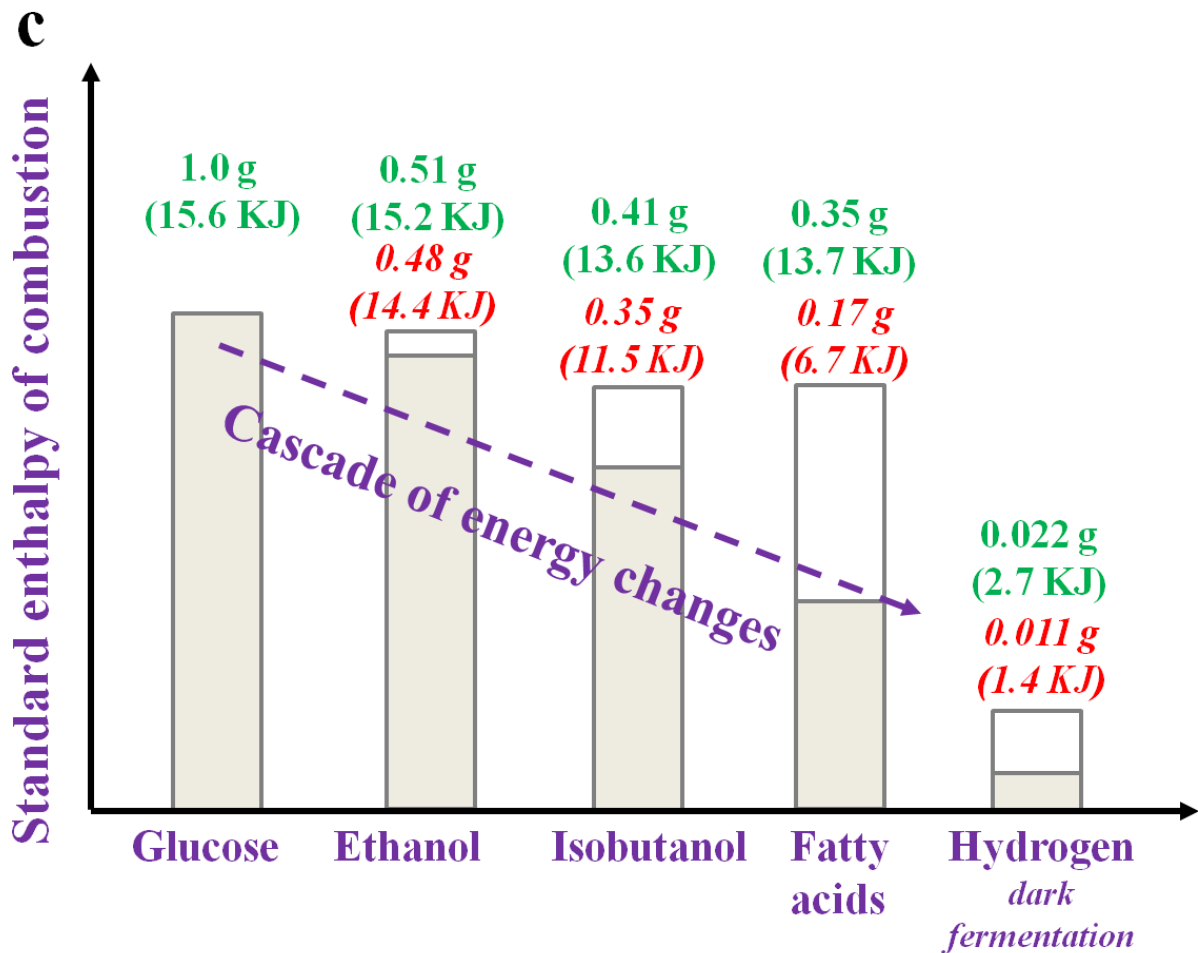


Figure 3.3 Energy fitness and productivities in microbial cell factories.

Figure 3.3a: The trend of metabolic entropy changes (unit: ATP generation per glucose). In optimal metabolism, one mole of glucose generates 38 ATP for biosynthesis. Under constraints of P/O ratios and maintenance loss, less ATP can be generated (i.e., increase of metabolic entropy).

Figure 3.3b: The transition from carbon limitation to energy limitation with the increase of product yield. In many cases, the energy limitation prevents strains from achieving the yield and titer at break-even point.

Figure 3.3c: Cascade of energy changes (Heat of combustion) during biofuel synthesis from glucose. Reported yields: ethanol -- 96% of theoretical yield (Alper *et al.* 2006), isobutanol -- 85% of theoretical yield (Atsumi *et al.* 2008a), fatty acid --50% of theoretical yield (He *et al.* 2014), and H<sub>2</sub> (dark fermentation) -- 50% of theoretical yield (Rachman *et al.* 1997).

# CHAPTER FOUR

## ENHANCE ENERGY STATE OF FATTY ACID PRODUCING STRAINS WITH *VITREOSCILLA* HEMOGLOBIN

### 4.1. Abstract

Engineered microbial species provide a sustainable platform to produce a wide range of chemicals from renewable resources. However, production of those compounds imposes significant metabolic burdens on host cells, leading to shifted metabolism, disrupted membrane, and unstable phenotype; those effects become even significant when oxygen becomes limited in the cell culture. Heterogeneous expression of *Vitreoscilla* hemoglobin (VHb) is known to enhance growth and energy efficiency of various hosts under microaerobic condition. In this study, we engineered fatty acids producing *E. coli* strain by introducing VHb and its mutant (VHb50), to solve the intracellular energy crisis. Growth and fatty production experiments indicated that the strain with VHb50 (strain GW50) achieved higher cell density and increased titer of fatty acids (50% improvement). In contrast, the benefit from wild-type VHb expression (GW1) counteracted its metabolic burden, and there is no significant difference in biomass and fatty acid titer. Further, expression of VHb50 significantly increased the ratio of unsaturated fatty acid (C16:1 and C18:1), especially oleic acid (C18:1), compared with the control strain without VHb. Lastly, we integrated the effect of VHb into flux models to simulate the responses of different host strains. The results demonstrated a different level of trade-off between the burden and the benefit from introduced genetic components, indicating the importance of specific properties of each genetic part.

### 4.2. Introduction



Metabolic engineering aims to obtain desired products through genetic modifications (Bailey 1991). The common approaches of dragging carbon fluxes to target molecules include gene knockout and heterogeneous enzyme expression. Even before the birth of ‘Metabolic engineering’, the fact that metabolic burden contributed from heterogeneous protein expression and plasmid maintenance led to decreased growth and shifted metabolism have been realized by researchers (Schaaff *et al.* 1989; Birnbaum and Bailey 1991). With the advancement of DNA technologies (e.g., PCR (Saiki *et al.* 1985), compatible multiple plasmids system (Lutz and Bujard 1997), convenient genome DNA knockout (Datsenko and Wanner 2000), and advanced sequencing) as well as the increased demand of bulk chemicals production from green approaches, metabolic engineering have greatly extended its range of product to amino acids, drugs, polymers, and most recently biofuels (Stephanopoulos *et al.* 1998). After more than twenty years of development, the competition between introduced genetic parts and desired products for carbon and energy source became even intense, leading to unstable genotypes and phenotypes (production performance), and this situation became especially severe in engineered *E. coli* strains producing biofuels (Hollinshead *et al.* 2014; Wu *et al.* 2015; He *et al.* 2014; Lennen *et al.* 2011).

Fatty acids are important precursors for productions of biodiesel, surfactants, and lubricants in the industry. Biosynthesis of fatty acids or related compounds through metabolic engineering or synthetic biology has been a hot field during recent years (Jones *et al.* 2015). A series of approaches have been employed to improve fatty acid or biodiesel production including introduction of heterogeneous enzymes (Lu *et al.* 2008), knockout degradation pathway (Lu *et al.* 2008; Steen *et al.* 2010), reversal of degradation pathway (Dellomonaco *et al.* 2011), boosting regulatory factors that activate the fatty acid pathway (Zhang *et al.* 2012a), and employment of

dynamic sensor control system to relieve toxicity from intermediates (Zhang *et al.* 2012b; Xu *et al.* 2014). However, those high-yield fatty acid producing strains have been found to be unstable even at laboratory conditions (He *et al.* 2014). This situation became worse during process scale-up when local culture conditions (e.g., pH, oxygen, toxic compounds) are unfavorable. <sup>13</sup>C-MFA has been applied to study the central metabolism of fatty acid producing strains, and the results indicated that remarkably high cellular maintenance energy was required for engineered strains producing fatty acid (He *et al.* 2014). Fatty acid production flux was found to be sensitive to P/O ratio as well as oxygen flux, based on FBA simulation (Wu *et al.* 2015). *Vitreoscilla* hemoglobin (VHb) has been well known to promote oxygen uptake and ATP production under oxygen limited condition in many hosts including *E. coli* (Khosla and Bailey 1988a). The introduction of VHb into fatty acid producing *E. coli* strains may potentially solve the problem of intracellular oxygen and energy limitation.

*Vitreoscilla* hemoglobin was first discovered by Webster and Hackett as early as 1966 (Webster and Hackett 1966). It was not realized as the first bacteria hemoglobin until its amino acid sequence was determined and showed high homology with eukaryotic hemoglobin (Wakabayashi *et al.* 1986). Heterogeneous expression of active VHb in *E. coli* was achieved two years later by two groups (Khosla and Bailey 1988a; Khosla and Bailey 1988b; Dikshit and Webster 1988). For the first time, Khosla and Bailey demonstrated that expression of VHb promoted oxygen uptake and improved the growth of host under microaerobic conditions. Later work by the same group further realized that VHb expression was able to enhance protein expression in *E. coli* under oxygen limited condition (Khosla *et al.* 1990). Considering of limited oxygen availability within cell culture at late exponential phase, the capability of VHb in enhancing growth and metabolites production have been widely applied to enhance production of

a wide range of compounds/proteins in various species (Zhang *et al.* 2007; Stark *et al.* 2015; Wei and Chen 2008). Those successful cases include: improved yields of cephalosporin C in *Acremonium chrysogenum* (DeModena *et al.* 1993), production improvement of human tissue plasminogen activator (tPA) in recombinant Chinese hamster ovary (CHO) cells (Pendse and Bailey 1994), promoted secretion/production of  $\alpha$ -amylase and neutral protease in *B. subtilis* (Kallio and Bailey 1996), increased biomass weight, chlorophyll, nicotine, and reduced germination time (half time) in transgenic *Nicotiana tabacum* (tobacco) (Holmberg *et al.* 1997), enhanced production rate and titer of erythromycin in engineered *Saccharopolyspora erythraea* (Minas *et al.* 1998). Several review papers have been published to summarize those successful applications of VHb in protein production and metabolic engineering (Zhang *et al.* 2007; Stark *et al.* 2015; Bülow *et al.* 1999; Frey and Kallio 2003; Stark *et al.* 2011).

Considering its competency and numerous successful applications of VHb, we proposed the hypothesis that VHb may also relieve the metabolic burden and intracellular stress caused by fatty acid production. To verify our hypothesis, we inserted VHb and its mutant into engineered *E. coli* strains that produced fatty acid. We also employed flux balance model to simulate the effects of VHb on fatty acid production.

### **4.3. Experimental and Methods**

#### **4.3.1. Chemicals and Strains**

All chemicals were reagent grade and purchased from either Sigma-Aldrich (St. Louis, MO, USA) or Fisher Scientific (Pittsburgh, PA, USA) unless otherwise noted. Restriction enzymes, Phusion DNA polymerase, and T4 ligase were from New England Biolabs (Ipswich, MA, USA).

The DNA clean and concentrating kit, Gel Recovery kit, and Miniprep kit were from Promega (Madison, WI, USA).

DNA sequence of *vhb* and its mutant *vhb50* was based on previous report (Andersson *et al.* 2000). Both genes were synthesized by GenScript Inc. (NJ, USA) and cloned into pUC57 vector. *E. coli* DH10B was used for plasmid manipulation. The *fadE* knockout *E. coli* DH1 strain (endA1 recA1 gyrA96 thi-1 glnV44 relA1 hsdR17 ( $r_K^- m_K^+$ )  $\lambda^-$ ) and the plasmid pA58c-TR were from Dr. Fuzhong Zhang's lab (He *et al.* 2014).

#### **4.3.2. Plasmid construction**

Primers used in this study were synthesized from Integrated DNA Technologies (Coralville, IA) and the detailed sequences are listed in Table 4.1. Plasmids GW1 and GW50 were constructed based on the plasmid pA58c-TR, to insert *vhb* or *vhb50* gene to the downstream of *tesA* gene. To ensure the expression of *vhb/vhb50* gene under the control of pLacUV5 promoter, an RBS sequence was added upstream of *vhb/vhb50* gene. (Figure 4.1) The *vhb/vhb50* genes were amplified using the primers Vh\_f and Vh\_r, while the vector pA58c-TR was amplified by the primers of pA58C\_f and pA58C\_r (as shown in Figure 4.2). The PCR products were cleaned by DNA clean and concentrating kit (Promega) and then digested by XhoI and HindIII at 37 °C for two hours. The digested DNA fragments were purified through gel purification and ligated through quick ligation kit at room temperature for 10 min. Ligation products were subsequently transformed into DH10B chemical competent cells. After incubation overnight at 37 °C, colony PCR was employed to identify positive clones. A few positive clones were incubated in Luria-Bertani (LB) medium (37 °C 220 rpm) supplied with appropriate antibiotics (30 µg/mL chloramphenicol) overnight for Miniprep. DNA sequences of both plasmids were validated by sequencing services in Genome center at Washington University School of Medicine.

### 4.3.3. Medium and culture conditions

A M9 MOPS minimal medium was employed in this study for fatty acid production experiments (Neidhardt *et al.* 1974). The detailed composition of this medium is as following (per liter): 20 g of glucose, 6.8 g of Na<sub>2</sub>HPO<sub>4</sub>, 3.0 g of KH<sub>2</sub>PO<sub>4</sub>, 3.96 g of (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>, 0.58 g of NaCl, supplemented with 50 ml of 1.5 M MOPS with pH adjusted to 7.4 with KOH, 1 ml of 1 M MgSO<sub>4</sub>, 0.1 ml of 10 mg/ml vitamin B1, 1 ml of 0.1 M CaCl<sub>2</sub>, 1 ml of 1000X micronutrients solution including 1.6 g of MnCl<sub>2</sub> 4H<sub>2</sub>O, 2.9 g of ZnSO<sub>4</sub> 7H<sub>2</sub>O, 2.5 g of H<sub>3</sub>BO<sub>3</sub>, 2.8 g of FeSO<sub>4</sub> 7H<sub>2</sub>O, 0.71 g of CoCl<sub>2</sub> 6H<sub>2</sub>O, 0.48 g of CuSO<sub>4</sub> 5H<sub>2</sub>O, and 0.37 g of (NH<sub>4</sub>)<sub>6</sub>Mo<sub>7</sub>O<sub>24</sub> 4H<sub>2</sub>O.

For fatty acid production in minimal medium, a single colony of cells from a fresh plate was used to inoculate a 5 ml of LB media for the pre-culture. The pre-culture was grown overnight at 37 °C on a rotary shaker at 225 rpm. The pre-culture was used to inoculate (2%, v/v) a 5 ml of minimal medium for overnight growth. After reaching stationary phase, the first minimal culture was subsequently inoculated into the second minimal medium (0.5%, v/v) grown on a rotary shaker in 25 ml tubes (25 x 150 mm) at 37 °C and 225 rpm. IPTG was added at appropriate concentrations when cell growth reaches an early exponential phase (OD<sub>600</sub> ~ 0.8), to induce gene expression under the control of P<sub>LacUV5</sub> promoter. The liquid cultures from each tube were centrifuged and the supernatant was separated from the biomass. Both the biomass and the supernatant samples were stored in -20 °C prior to analysis.

### 4.3.4 Fatty acid measurement

200 µl of cell culture was mixed with 300 µl of dH<sub>2</sub>O, and get acidified using 50 µl of concentrated HCl. 500 µl of EtAc spiked with C19:0 ME as internal standard was added to extract fatty acid from water phase. 400 µl of organic phase containing fatty acids separated by

centrifugation was transferred to a new tube and another 500  $\mu\text{l}$  of EtAc was added again to repeat the extraction. 100  $\mu\text{l}$  of MeOH:HCl (9:1) was added to the EtAc extract and mixed well. 100  $\mu\text{l}$  of TMS-diazomethane (2 M in hexanes) was also added under hood and the reaction was kept in the hood for 10 ~ 15 min at room temperature. The methyl esters of fatty acids were analyzed using a gas chromatograph mass spectrometer (GC-MS) (Hewlett Packard 7890A and 5975C, Agilent Technologies) equipped with a DB5-MS column (J&W Scientific). The GC-MS program was as follows: the column temperature initially held at 80  $^{\circ}\text{C}$  for 1 min, raised to 280  $^{\circ}\text{C}$  at 30  $^{\circ}\text{C min}^{-1}$  and held at 280  $^{\circ}\text{C}$  for 3 min. Helium was used as the carrier gas. The mass spectra were analyzed using the Enhanced Data Analysis software (Agilent Technologies). The fatty acids were quantified based on the standard curve of standard mixtures of methyl esters of fatty acid.

#### **4.3.5. Simulate cellular physiologies with flux balance model**

Genome scale model *iEcDH1\_1363* (includes 1363 genes, 2752 reactions, and 1950 metabolites) was employed to simulate fatty acids production in DH1 strain (Monk *et al.* 2013). A simplified flux of fatty acid (C16:0) was added as representative of fatty acid production, and the objective function was set to maximize this flux (Wu *et al.* 2015). The growth rate of engineered strain was set as 0.2  $\text{h}^{-1}$ , considering that fatty acid production is growth associated (He *et al.* 2014; Lu *et al.* 2008). Default values were employed for the boundary of all fluxes except the followings: The upper and lower boundary of flux 1032 was set to be zero for  $\Delta\text{fadE}$ ; the sensitivity of maintenance energy was tested through FBA; the lower boundary of glucose uptake flux was set as experimental value. We assumed Vhb improved the affinity of terminal oxidase (i.e.,

cytochrome *d* or cytochrome *o*), leading to increase uptake flux of glucose. We also assumed that VHb had no influence on glucose uptake based on previous reports (Tsai *et al.* 1996a).

The COBRA toolbox and LibSBML library were employed for genome scale model manipulation (Schellenberger *et al.* 2011; Bornstein *et al.* 2008); while Gurobi 5.5 linear solver (Gurobi Optimization Inc.) was utilized for FBA calculation on MATLAB 2012b.

#### **4.4. Results and discussion**

##### **4.4.1 Growth kinetics and fatty acid production**

Growth kinetics for all strains incubated in M9 minimal medium was described in Figure 4.6. Shaking tubes were employed here for obtaining microaerobic conditions. There is an obvious exponential phase rate for each strain after IPTG induction. The growth rate of exponential phase was very close to the previous report: pA58c 0.29 h<sup>-1</sup>, GW1 0.19 h<sup>-1</sup>, GW50 0.29 h<sup>-1</sup> (He *et al.* 2014). Introduction of VHb only showed apparent improvement during late exponential phase of GW50.

The trend of fatty acid production is quite similar to the biomass growth (shown in Figure 4.6): after 24 hr of induction, GW50 produced most fatty acid, while GW1 produced least fatty acid. Fatty acid titer of the control strain pA58c in this study is much lower than previous report by our lab (He *et al.* 2014), mostly due to poor oxygen supply in shaking tubes compared with baffled shaking flasks. The importance of oxygen on cell growth and free fatty acid production can also be revealed from this difference.

To interpret the profiles in biomass growth and fatty acid production, two factors may contribute most: First, wild-type VHb did improve oxygen uptake and energy metabolism in GW1 strain; however, this benefit did not counteract the metabolic burden caused by VHb. This

observation reflected the importance of specific properties for each individual cellular component, where protein engineering (e.g., direct evolution or rational design) is able to improve/change catalytic kinetics of single enzyme, leading improved yield of desired products (Leonard *et al.* 2010; Bommareddy *et al.* 2014). On the other side, the impact of VHb50 was only significant when cell density was relatively high and oxygen condition was poor, agreeing well with all previous reports regarding VHb (Zhang *et al.* 2007; Tsai *et al.* 1996b). This nature actually limits further application of VHb is, because industrial production always requires high production rate where fully aerobic condition is employed. To resolve this bottleneck, a deep understanding of VHb mechanism is necessary.

The role of VHb has been considered to be improving oxygen transfer, boosting ATP generation, and promoting intracellular energy state when extracellular oxygen concentration is low (Tsai *et al.* 1996a; Tsai *et al.* 1996b; Kallio *et al.* 1994). Compared with other hemoglobin, purified VHb has a medium affinity for oxygen binding as well as relatively slow rate for oxygen release *in vitro* (Giangiacomo *et al.* 2001), implying its possible function as assisting oxygen transport rather than transporter (structure of active VHb shown in Figure 4.2). However, later work proved that lipid-bound VHb has a significantly higher affinity with oxygen (20 fold enhancement), suggesting its potential role in oxygen transport as a membrane protein (Rinaldi *et al.* 2006). Other functions of VHb have also been reported as relieving oxidative stress, detoxifying nitric oxide (NO) *in vivo* (Frey *et al.* 2002), and enhancing intracellular level of tRNA and ribosome (Roos *et al.* 2002). Further work will provide more insights over detailed mechanism of VHb.

#### **4.4.2 Expression of VHb affects the degree of unsaturation of free fatty acid**



The profile of fatty acid also experienced a significant shift after the introduction of VHb (shown in Figure 4.6). At both early exponential (5 hr after IPTG induction) and late exponential phase (24 hr after IPTG induction), the percentage of unsaturated fatty acid (C16:1 and C18:1) increased significantly (from 0.37 to 0.46), especially for the oleic acid (C18:1). Cao *et al.* has reported that simultaneous overexpression of *fabA* and *fabB* will increase the ratio of unsaturated fatty acid (Cao *et al.* 2010). In another study, expression of *FadR* was also found to enhance the percentage of unsaturated fatty acid by activating expressions of *fabA* and *fabB* (Zhang *et al.* 2012a). However, heterogeneous expression of VHb in *Aurantiochytrium* sp. led to a decreased percentage of unsaturated fatty acids (Suen *et al.* 2014). Previous studies also proved that VHb was able to boost the intracellular expression of heterogeneous proteins in *E. coli* (Khosla *et al.* 1990). Based on those facts, we infer that VHb may promote the leaky expression of *fadR* in plasmid, leading to elevated level of *fabA* and *fabB* pathway. To further confirm this effect, evidence from transcriptional or proteomics level will be helpful.

Increased degree of unsaturation of fatty acid is desired for several reasons. One important fact is that biodiesel derived from high content of unsaturated fatty acids have lower melting temperature, which is suitable as fuels for cold environments (e.g. winter). Another feature is that uptake monounsaturated fatty acid may raise the level of high-density lipoprotein (HDL) cholesterol, which is considered as ‘good’ cholesterol. Introduction of VHb into algae may facilitate its production of omega-3 oil.

#### **4.4.3. Effect of oxygen and maintenance energy on fatty acid production**

To quantify the effect of oxygen and maintenance energy on fatty acid production, FBA simulation was employed as described in the section of methods. To simulate cellular

physiologies at different growth phases, we adopted two set of conditions describing early exponential phase ( $v_{glucose} = 7.6$  mmol/gCDW h, growth rate  $\mu = 0.29$  h<sup>-1</sup>, default P/O ratio = 1.75) and late exponential phase ( $v_{glucose} = 4$  mmol/gCDW h, growth rate  $\mu = 0.1$  h<sup>-1</sup>, decreased P/O ratio = 1). These parameters were either from previous report or experimental measurement in this study (He *et al.* 2014). Considering of membrane disruption and proton leakage caused by free fatty acid accumulation (Lennen *et al.* 2011), we assumed a decreased P/O ratio as 1 for the late exponential phase (Heyland *et al.* 2011).

The simulation results were shown in Figure 4.7a (for early exponential phase) and Figure 4.7b (for late exponential phase). At early exponential phase, theoretic maximal yield of fatty acid was ~ 0.24 g FA/g glucose (for growth rate  $\mu = 0.29$  h<sup>-1</sup>, 0.256 g FA/g glucose for growth rate  $\mu = 0.26$  h<sup>-1</sup>). In our previous report, the highest yield was 0.17 g FA/g glucose, which is already 66% of the theoretic value (He *et al.* 2014). Taken cellular maintenance energy of the fatty acid strain into consideration, 0.17 g FA/g glucose would be very close to the maximal yield at this condition. On the other side, a rough estimation of cellular maintenance energy can be made based on FBA prediction: The maximal maintenance energy for control strain pA58c was 30 mmol/gCDW h in the minimal medium when oxygen supply is sufficient. Compared with other reports on maintenance energy for engineered strains (Heyland *et al.* 2011), that value is quite reasonable. Another obvious trend was that with the decrease of oxygen flux and the increase of maintenance energy to certain level, the yield of fatty acid would meet a sudden drop – ‘energy cliff’ as we discussed in our previous report (Wu *et al.* 2015). At this period, a slight improvement in oxygen flux (e.g., from 14 to 16 mmol/gCDW h, assumed enhancement by VHB) would not make any significant difference in fatty acid yield. In late exponential phase, when cell culture reaches a relatively high density, the oxygen condition in medium becomes

microaerobic (as shown in Figure 4.7b). A minor increase in oxygen flux (e.g., from 4 to 6 mmol/gCDW h, assumed enhancement by VHb) will harvest a significant improvement in fatty acid yield. That also explained why the effect of VHb was significant at microaerobic conditions. Notably, the expression of VHb from a medium copy of plasmid would bring a certain amount of metabolic burden. This metabolic burden is not too much, however, for fatty acid producing strain with high cellular maintenance energy, this additional burden may become ‘the straw that broke the camel’s back’ if oxygen is also limited. There is no report regarding how the structure difference (i.e., amino acid mutations) leads to the function improvement in VHb50 (Andersson *et al.* 2000), compared with WT VHb. From FBA simulation, we infer the expression of VHb50 may make a larger improvement in oxygen uptake flux, compared with WT VHb.

#### **4.5. Conclusion**

In this study, our hypothesis was confirmed that the introduction of VHb was able to relieve the metabolic stress of fatty acid producing strain and improve the final titer of biomass and fatty acid. We further revealed that VHb expression would improve the ratio of unsaturated fatty acid, which may shed light to further application of VHb to produce fuel with lower melting point (Zhang *et al.* 2012a) or other valuable products with high degree of unsaturation. Lastly, we demonstrated the importance of individual VHb properties on the performance of engineered strains. To sum up, engineered components (e.g., enzyme, transporter, or circuit) would only bring expected enhancement when its benefit beats its burden.

#### **4.6. Reference**

Bailey, J. (1991). Toward a science of metabolic engineering. *Science* 252:1668-1675.

Schaaff, I., Heinisch, J. and Zimmermann, F. K. (1989). Overproduction of glycolytic enzymes in yeast. *Yeast* 5:285-290.

Birnbaum, S. and Bailey, J. E. (1991). Plasmid presence changes the relative levels of many host cell proteins and ribosome components in recombinant *Escherichia coli*. *Biotechnol Bioeng* 37:736-745.

Saiki, R., Scharf, S., Faloona, F., Mullis, K., Horn, G., Erlich, H. and Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230:1350-1354.

Lutz, R. and Bujard, H. (1997). Independent and Tight Regulation of Transcriptional Units in *Escherichia coli* Via the LacR/O, the TetR/O and AraC/I1-I2 Regulatory Elements. *Nucleic Acids Res* 25:1203-1210.

Datsenko, K. A. and Wanner, B. L. (2000). One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci USA* 97:6640-6645.

Stephanopoulos, G., Aristidou, A. and Nielsen, J. (1998). Metabolic engineering: principles and methodologies.

Hollinshead, W., He, L. and Tang, Y. (2014). Biofuel Production: an odyssey from metabolic engineering to fermentation scale-up. *Frontiers in microbiology* 5.

Wu, S., He, L., Wang, Q. and Tang, Y. (2015). An ancient Chinese wisdom for metabolic engineering: Yin-Yang. *Microb. Cell Fact.* 14:39.

He, L., Xiao, Y., Gebreselassie, N., Zhang, F., Antoniewicz, M. R., Tang, Y. J. and Peng, L. (2014). Central metabolic responses to the overproduction of fatty acids in *Escherichia coli* based on <sup>13</sup>C-metabolic flux analysis. *Biotechnol Bioeng* 111:575-585.

Lennen, R. M., Kruziki, M. A., Kumar, K., Zinkel, R. A., Burnum, K. E., Lipton, M. S., Hoover, S. W., Ranatunga, D. R., Wittkopp, T. M., Marnier, W. D. and Pfleger, B. F. (2011). Membrane Stresses Induced by Overproduction of Free Fatty Acids in *Escherichia coli*. *Appl Environ Microb* 77:8114-8128.

Jones, J. A., Toparlak, Ö. D. and Koffas, M. A. G. (2015). Metabolic pathway balancing and its role in the production of biofuels and chemicals. *Curr Opin Biotech* 33:52-59.

Lu, X., Vora, H. and Khosla, C. (2008). Overproduction of free fatty acids in *E. coli*: Implications for biodiesel production. *Metab. Eng.* 10:333-339.

Steen, E. J., Kang, Y., Bokinsky, G., Hu, Z., Schirmer, A., McClure, A., del Cardayre, S. B. and Keasling, J. D. (2010). Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature* 463:559-562.

Dellomonaco, C., Clomburg, J. M., Miller, E. N. and Gonzalez, R. (2011). Engineered reversal of the beta-oxidation cycle for the synthesis of fuels and chemicals. *Nature* 476:355-359.

Zhang, F., Ouellet, M., Batth, T. S., Adams, P. D., Petzold, C. J., Mukhopadhyay, A. and Keasling, J. D. (2012a). Enhancing fatty acid production by the expression of the regulatory transcription factor *FadR*. *Metab. Eng.* 14:653-660.

Zhang, F., Carothers, J. M. and Keasling, J. D. (2012b). Design of a dynamic sensor-regulator system for production of chemicals and fuels derived from fatty acids. *Nat Biotech* 30:354-359.

Xu, P., Li, L., Zhang, F., Stephanopoulos, G. and Koffas, M. (2014). Improving fatty acids production by engineering dynamic pathway regulation and metabolic control. *Proc. Natl. Acad. Sci.* 111:11299-11304.

Khosla, C. and Bailey, J. E. (1988a). Heterologous expression of a bacterial haemoglobin improves the growth properties of recombinant *Escherichia coli*. *Nature* 331:633-635.

Webster, D. A. and Hackett, D. P. (1966). The Purification and Properties of Cytochrome o from *Vitreoscilla*. *J Biol Chem* 241:3308-3315.

Wakabayashi, S., Matsubara, H. and Webster, D. A. (1986). Primary sequence of a dimeric bacterial haemoglobin from *Vitreoscilla*. *Nature* 322:481-483.

Khosla, C. and Bailey, J. (1988b). The *Vitreoscilla* hemoglobin gene: Molecular cloning, nucleotide sequence and genetic expression in *Escherichia coli*. *Molec. Gen. Genet.* 214:158-161.

Dikshit, K. L. and Webster, D. A. (1988). Cloning, characterization and expression of the bacterial globin gene from *Vitreoscilla* in *Escherichia coli*. *Gene* 70:377-386.

Khosla, C., Curtis, J. E., DeModena, J., Rinas, U. and Bailey, J. E. (1990). Expression of Intracellular Hemoglobin Improves Protein Synthesis in Oxygen-Limited *Escherichia coli*. *Nat Biotech* 8:849-853.

Zhang, L., Li, Y., Wang, Z., Xia, Y., Chen, W. and Tang, K. (2007). Recent developments and future prospects of *Vitreoscilla* hemoglobin application in metabolic engineering. *Biotechnol Adv* 25:123-136.

Stark, B., Pagilla, K. and Dikshit, K. (2015). Recent applications of *Vitreoscilla* hemoglobin technology in bioproduct synthesis and bioremediation. *Appl Microbiol Biot* 99:1627-1636.

Wei, X.-X. and Chen, G.-Q. (2008). Chapter Fifteen - Applications of the VHb Gene *vgb* for Improved Microbial Fermentation Processes, in *Method Enzymol*, K. P. Robert, ed., Academic Press, 273-287.

DeModena, J. A., Gutierrez, S., Velasco, J., Fernandez, F. J., Fachini, R. A., Galazzo, J. L., Hughes, D. E. and Martin, J. F. (1993). The Production of Cephalosporin C by *Acremonium chrysogenum* is Improved by the Intracellular Expression of a Bacterial Hemoglobin. *Nat Biotech* 11:926-929.

Pendse, G. J. and Bailey, J. E. (1994). Effect of *Vitreoscilla* hemoglobin expression on growth and specific tissue plasminogen activator productivity in recombinant chinese hamster ovary cells. *Biotechnol Bioeng* 44:1367-1370.

Kallio, P. T. and Bailey, J. E. (1996). Intracellular Expression of *Vitreoscilla* Hemoglobin (VHb) Enhances Total Protein Secretion and Improves the Production of  $\alpha$ -Amylase and Neutral Protease in *Bacillus subtilis*. *Biotechnol Progr* 12:31-39.

Holmberg, N., Lilius, G., Bailey, J. E. and Bulow, L. (1997). Transgenic tobacco expressing *Vitreoscilla* hemoglobin exhibits enhanced growth and altered metabolite production. *Nat Biotech* 15:244-247.

Minas, W., Brünker, P., Kallio, P. T. and Bailey, J. E. (1998). Improved Erythromycin Production in a Genetically Engineered Industrial Strain of *Saccharopolyspora erythraea*. *Biotechnol Progr* 14:561-566.

B ulow, L., Holmberg, N., Lilius, G. and Bailey, J. E. (1999). The metabolic effects of native and transgenic hemoglobins on plants. *Trends Biotechnol* 17:21-24.

Frey, A. D. and Kallio, P. T. (2003). Bacterial hemoglobins and flavohemoglobins: versatile proteins and their impact on microbiology and biotechnology. *FEMS Microbiol Rev* 27:525-545.

Stark, B., Dikshit, K. and Pagilla, K. (2011). Recent advances in understanding the structure, function, and biotechnological usefulness of the hemoglobin from the bacterium *Vitreoscilla*. *Biotechnol Lett* 33:1705-1714.

Andersson, C. I. J., Holmberg, N., Farr ́s, J., Bailey, J. E., B ulow, L. and Kallio, P. T. (2000). Error-prone PCR of *Vitreoscilla* hemoglobin (VHb) to support the growth of microaerobic *Escherichia coli*. *Biotechnol Bioeng* 70:446-455.

Neidhardt, F. C., Bloch, P. L. and Smith, D. F. (1974). Culture Medium for Enterobacteria. *J Bacteriol* 119:736-747.

Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M. and Palsson, B. Ø. (2013). Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proc. Natl. Acad. Sci.* 110:20338-20343.

Tsai, P. S., Hatzimanikatis, V. and Bailey, J. E. (1996a). Effect of Vitreoscilla hemoglobin dosage on microaerobic *Escherichia coli* carbon and energy metabolism. *Biotechnol Bioeng* 49:139-150.

Schellenberger, J., Que, R., Fleming, R., Thiele, I., Orth, J., Feist, A., Zielinski, D., Bordbar, A., Lewis, N. and Rahmanian, S. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.* 6:1290 - 1307.

Bornstein, B. J., Keating, S. M., Jouraku, A. and Hucka, M. (2008). LibSBML: an API Library for SBML. *Bioinformatics* 24:880-881.

Leonard, E., Ajikumar, P. K., Thayer, K., Xiao, W.-H., Mo, J. D., Tidor, B., Stephanopoulos, G. and Prather, K. L. J. (2010). Combining metabolic and protein engineering of a terpenoid biosynthetic pathway for overproduction and selectivity control. *Proc. Natl. Acad. Sci.* 107:13654-13659.

Bommarreddy, R. R., Chen, Z., Rappert, S. and Zeng, A.-P. (2014). A de novo NADPH generation pathway for improving lysine production of *Corynebacterium glutamicum* by rational design of the coenzyme specificity of glyceraldehyde 3-phosphate dehydrogenase. *Metab. Eng.* 25:30-37.



Tsai, P. S., Nägeli, M. and Bailey, J. E. (1996b). Intracellular expression of *Vitreoscilla* hemoglobin modifies microaerobic *Escherichia coli* metabolism through elevated concentration and specific activity of cytochrome o. *Biotechnol Bioeng* 49:151-160.

Kallio, P. T., Jin Kim, D., Tsai, P. S. and Bailey, J. E. (1994). Intracellular expression of *Vitreoscilla* hemoglobin alters *Escherichia coli* energy metabolism under oxygen-limited conditions. *Eur J Biochem* 219:201-208.

Giangiaco, L., Mattu, M., Arcovito, A., Bellenchi, G., Bolognesi, M., Ascenzi, P. and Boffi, A. (2001). Monomer–Dimer Equilibrium and Oxygen Binding Properties of Ferrous *Vitreoscilla* Hemoglobin. *Biochemistry Biochemistry* 40:9311-9316.

Rinaldi, A. C., Bonamore, A., Macone, A., Boffi, A., Bozzi, A. and Di Giulio, A. (2006). Interaction of *Vitreoscilla* Hemoglobin with Membrane Lipids. *Biochemistry Biochemistry* 45:4069-4076.

Frey, A. D., Farrés, J., Bollinger, C. J. T. and Kallio, P. T. (2002). Bacterial Hemoglobins and Flavohemoglobins for Alleviation of Nitrosative Stress in *Escherichia coli*. *Appl Environ Microb* 68:4835-4840.

Roos, V., Andersson, C. I. J., Arfvidsson, C., Wahlund, K.-G. and Bülow, L. (2002). Expression of Double *Vitreoscilla* Hemoglobin Enhances Growth and Alters Ribosome and tRNA Levels in *Escherichia coli*. *Biotechnol Progr* 18:652-656.

Cao, Y., Yang, J., Xian, M., Xu, X. and Liu, W. (2010). Increasing unsaturated fatty acid contents in *Escherichia coli* by coexpression of three different genes. *Appl Microbiol Biot* 87:271-280.

Suen, Y. L., Tang, H., Huang, J. and Chen, F. (2014). Enhanced Production of Fatty Acids and Astaxanthin in *Aurantiochytrium* sp. by the Expression of *Vitreoscilla Hemoglobin*. *J. Agr. Food Chem.* 62:12392-12398.

Heyland, J., Blank, L. M. and Schmid, A. (2011). Quantification of metabolic limitations during recombinant protein production in *Escherichia coli*. *J Biotechnol* 155:178-184.

Name of primer	Sequence (5'—3')
pA58C_f	GCAC <b><u>AAGCTT</u></b> CCAGGCATCAAATAAAACGA A
pA58C_r	CCTTA <b><u>CTCGAG</u></b> TTATG AGTCATGATTTACT
Vh_f	CG CAT <b><u>CTCGAG</u></b> TTTAAGAAGGAGATATACAT ATGTTAGACCAGCAAACCATTA
Vh_r	GCAC <b><u>AAGCTT</u></b> TTATTCAACCGCTTGAGCGTA

Table 4.1 DNA sequence of all primers used in this work

'\_\_' indicates the restriction cutting sites.

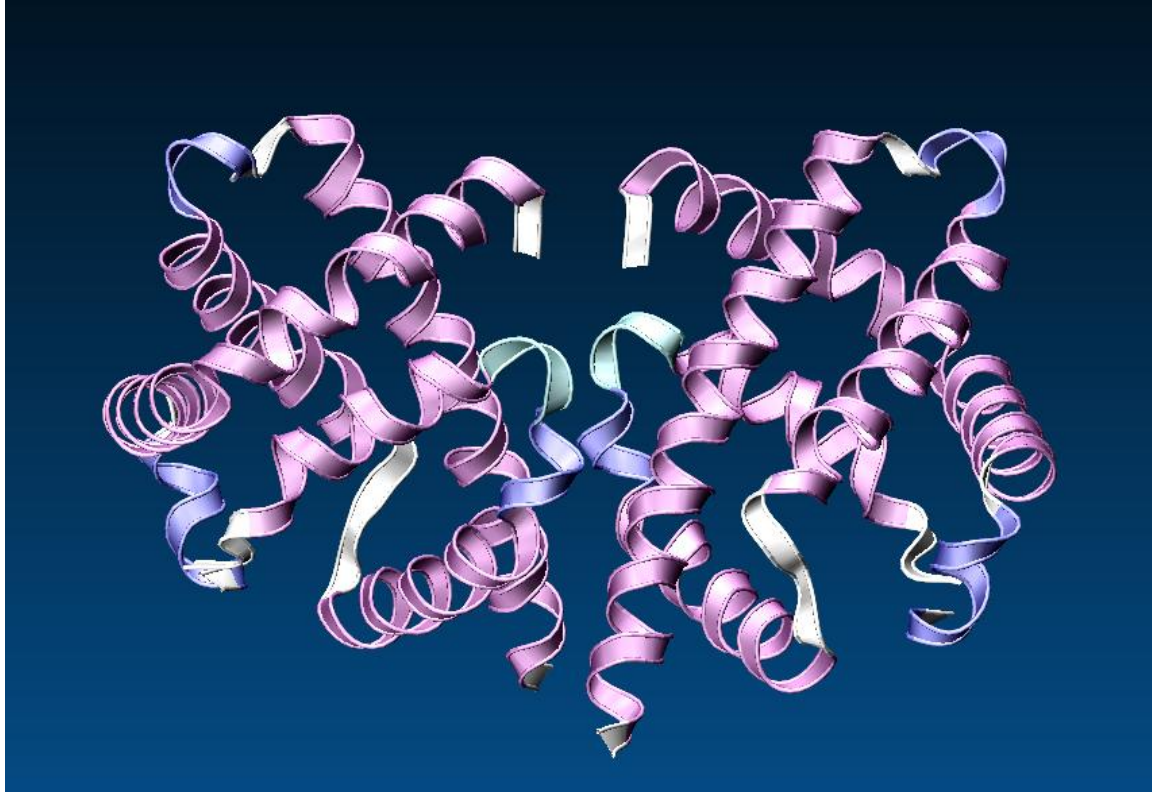
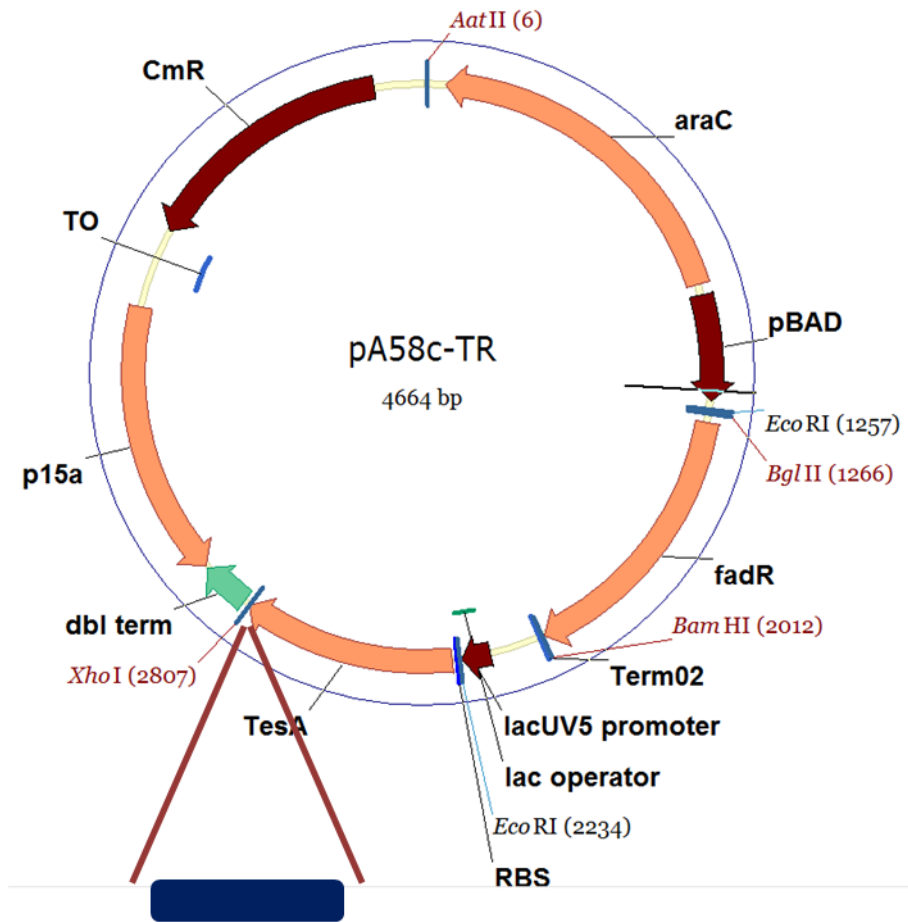


Figure 4.1 Structure of *Vitreoscilla* hemoglobin in active dimer form, simulated by VMD



VHb gene and mutants (VHb50)\*

Figure 4.2 Genetic manipulations to insert VHb into pA58c-TR

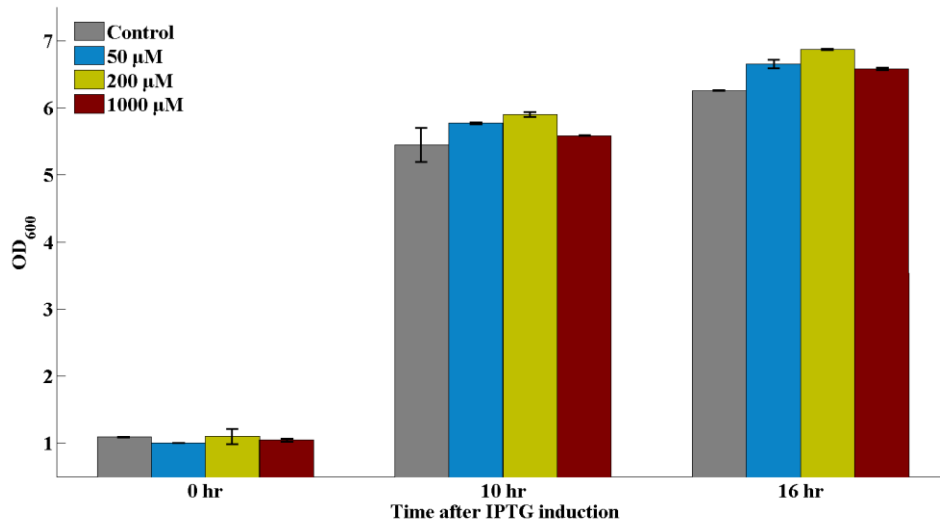


Figure 4.3 Optimization of IPTG concentrations for VHb50 expression

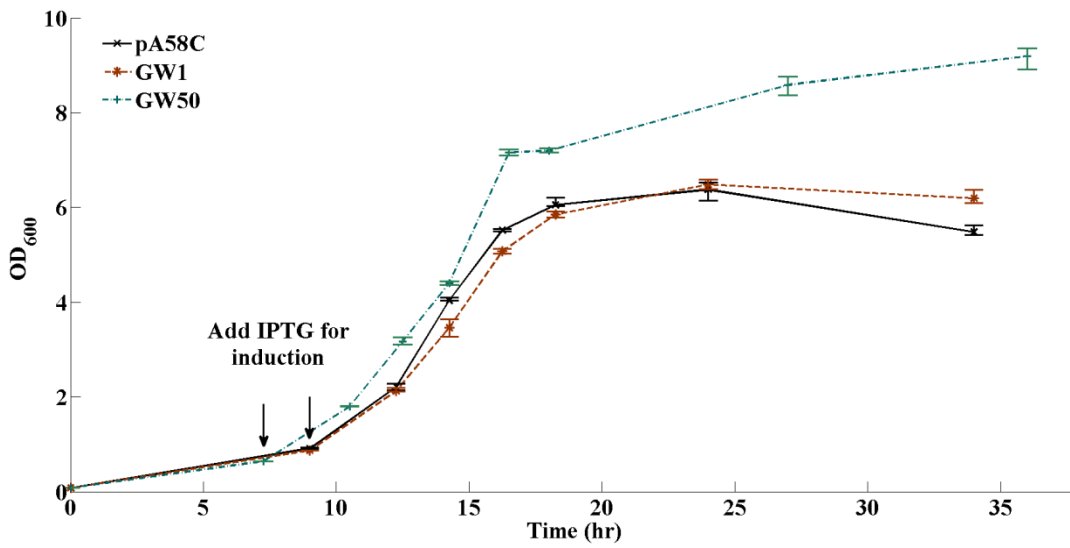


Figure 4.4 Growth curve for three fatty acid producing strains

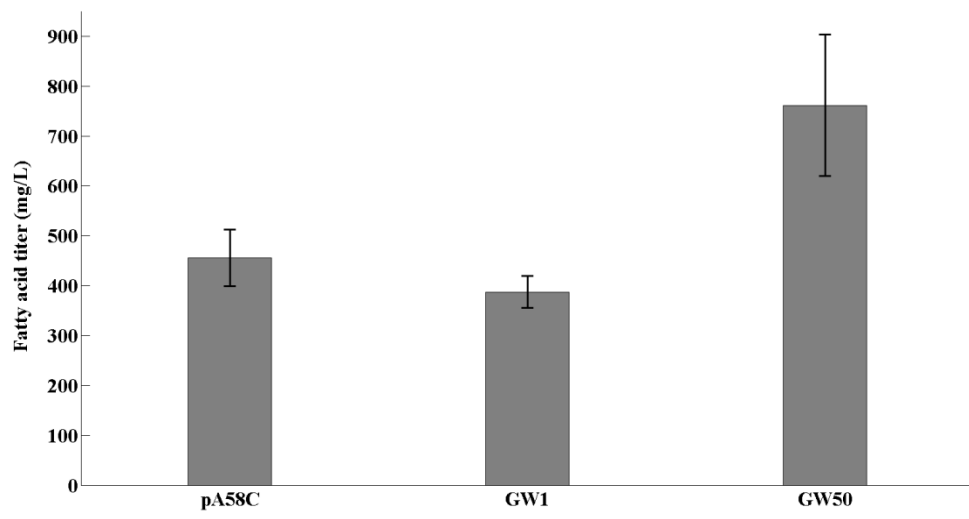


Figure 4.5 Fatty acid productions after 24 hr of IPTG Induction

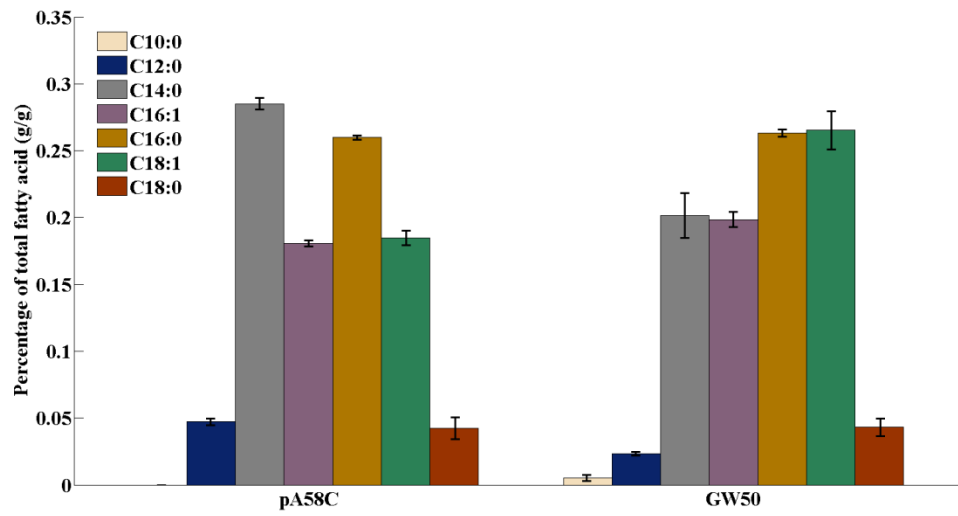


Figure 4.6a Free fatty acid production profile for control strain and VHB strain after 8 hr IPTG induction

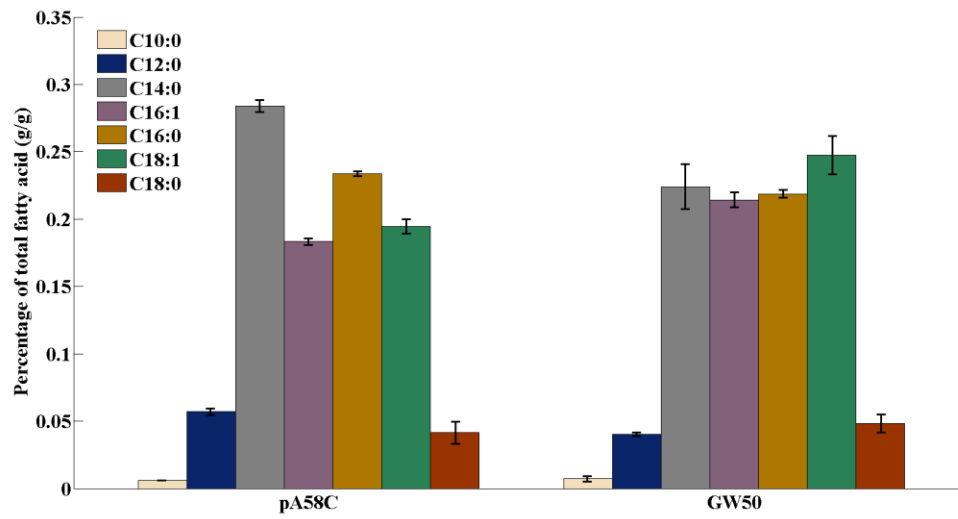


Figure 4.6b Free fatty acid production profile for control strain and VHB strain after 24 hr IPTG induction

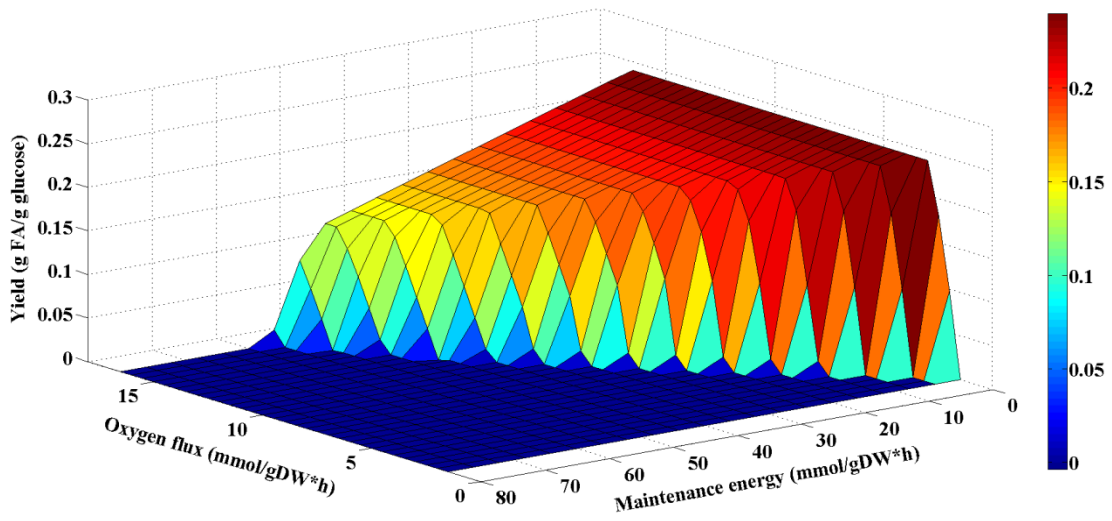


Figure 4.7a Effect of oxygen flux and maintenance energy on fatty acid yield at exponential phase

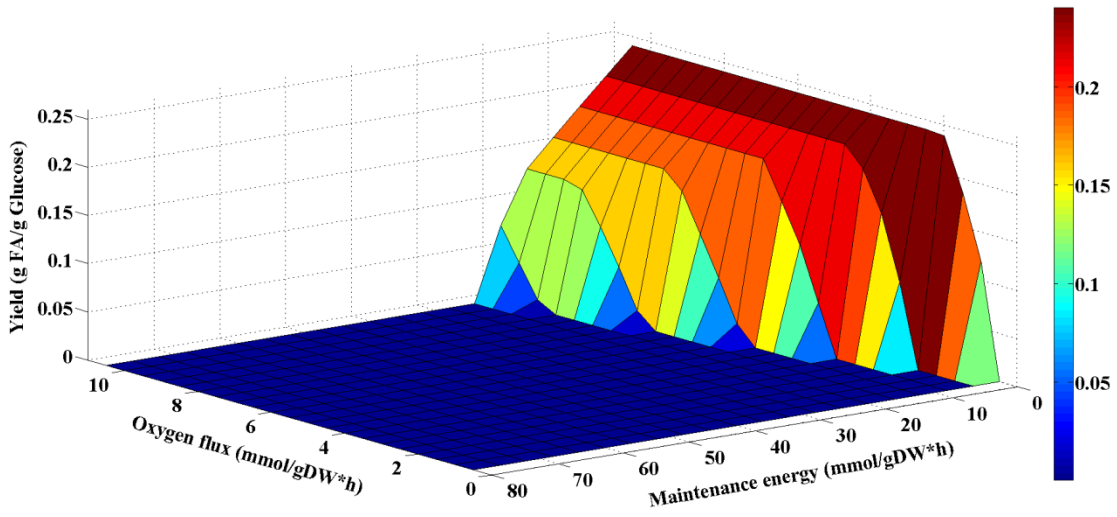


Figure 4.7b Effect of oxygen flux and maintenance energy on fatty acid yield at late exponential phase



## CHAPTER FIVE

### BUILD WEB-BASED PLATFORM FOR FLUXOMICS STUDIES:

### MICROBESFLUX REBUILD AND WEBSITE DEVELOPMENT

#### 5.1. Abstract

Metabolic flux analyses offer direct insights into cell metabolism. Metabolic flux analyses include genome-scale Flux Balance Analysis (FBA) and  $^{13}\text{C}$ -Metabolic Flux Analysis ( $^{13}\text{C}$ -MFA). To speed up fluxomics studies, we need a user-friendly platform to construct metabolic networks from genome information and perform flux calculations using  $^{13}\text{C}$  data. Four years ago (2011), Tang lab developed a web-based platform (MicrobesFlux) for reconstructing metabolic models from the KEGG database (Kyoto Encyclopedia of Genes and Genomes). The platform ran on a shared server system at Washington University in St. Louis. Unfortunately, this server was unstable and suffered from downtime occasionally. Hence, we set out to rebuild MicrobesFlux on a commercial server to make the systems more usable. The enhanced MicrobesFlux updates metabolic network information with the latest version from the KEGG database. In addition, we added MATLAB programs into the platform so that it can also provide  $^{13}\text{C}$ -MFA. The new program (called WUFlux, WU: Washington University) can carry out  $^{13}\text{C}$ -MFA of different metabolic types for prokaryote species. Furthermore, WUFlux also contains a carbon fate map of central pathways, labeling correction programs, and user manuals. The systems we developed are open-source and free to use. The new platform for fluxomics study is now available at <http://www.fluxomics.net>, and we will continue to improve the functionalities of our software for both FBA and  $^{13}\text{C}$ -MFA.

## 5.2. Introduction

Metabolic flux analysis is widely used to predict or measure *in vivo* enzyme reaction rates in microbes. FBA studies microbial metabolism based on the stoichiometry of the metabolic reactions as well as the measurement of inflow (substrate uptake) and outflow fluxes (biomass and product synthesis). FBA requires an objective function (e.g., optimization of biomass yield) to estimate the flux values. A more rigorous flux analysis combines the flux stoichiometry with  $^{13}\text{C}$  isotopic tracing, i.e.,  $^{13}\text{C}$ -MFA.  $^{13}\text{C}$ -experiments consist of feeding the cell culture with a defined  $^{13}\text{C}$ -substrate to fingerprint downstream metabolites with the labeled carbon ( $^{13}\text{C}$ ). The patterns of isotopic enrichment in metabolites, once  $^{13}\text{C}$  has reached a steady state distribution throughout the metabolic network, can be used to decipher flux distributions in the cell metabolism. The isotopomer information can discover novel pathways, resolve reversible and branched fluxes, and quantify circular metabolic routes (e.g., the TCA cycle). On the other hand,  $^{13}\text{C}$ -MFA, requiring both experimental and modeling efforts, is time-consuming and costly (Figure 5.1). In general, one  $^{13}\text{C}$ -MFA project can take several experienced researchers more than one year to accomplish (based on personal communications to  $^{13}\text{C}$ -MFA groups). Although  $^{13}\text{C}$ -MFA began as early as the 1980s, it has not been as widely used as other analytical/systems biology tools (Crown and Antoniewicz 2013). To date, published  $^{13}\text{C}$ -MFA papers are fewer than 1000 (Crown and Antoniewicz 2013). Based on Pubmed database, in 2012 and 2013, only 41 papers (related to  $^{13}\text{C}$  metabolic flux analysis) have been published. Among them, 18 were review or method papers, while 10 were research papers on bacteria and the remaining 13 papers were research on yeast or mammalian cells. The progress of flux analysis has been slow because the construction of a metabolic network based on an annotated genome,  $^{13}\text{C}$ -labeling tracing, and flux calculation can all be very time-consuming. On the other hand, the world's bacterial species

number between  $10^7$  to  $10^9$ , and  $\sim 10^5$  species have been sequenced for their 16S rRNA genes (Schloss and Handelsman 2004). There are fewer than one hundred  $^{13}\text{C}$ -MFA studies for nonmodel species. The gap between  $^{13}\text{C}$ -MFA studies and sequenced bacterial species calls for broad-scope fluxomics tools for characterization of a large amount of unknown microbial species. To reduce modeling challenges, FBA platforms have been developed to facilitate reconstructing genome-scale metabolic networks. These platforms include SuBliMinal (Swainston *et al.* 2011), SEED (Henry *et al.* 2010), RAVEN (Agren *et al.* 2013), Pathway Tools (Karp *et al.* 2002), COBRA (Becker *et al.* 2007), and FAME (Boele *et al.* 2012). These platforms have been discussed in a recent review paper (Hamilton and Reed 2014).  $^{13}\text{C}$ -MFA software platforms are also being developed, including FiatFlux (Zamboni *et al.* 2005b), iMS2Flux (Poskar *et al.* 2012), INCA (Young 2014), Metran (Yoo *et al.* 2008), OpenFLUX (Quek *et al.* 2009a), OpenMebius (Kajihata *et al.* 2014), 13CFLUX (Wiechert *et al.* 2001) and 13CFLUX2 (Weitzel *et al.* 2012). With rapid advances in genome sequencing, network reconstructions and  $^{13}\text{C}$ -based functional analysis are concurrently required for metabolic characterizations.

To augment these tools, we built an integrated platform at [www.fluxomics.net](http://www.fluxomics.net), combining the functions of both FBA -- MicrobesFlux, at ([www.microbesflux.org](http://www.microbesflux.org), as shown in Figure 5.2) and  $^{13}\text{C}$ -MFA -- WUFlux, at ([13cmfa.org](http://13cmfa.org), as shown in Figure 5.3). MicrobesFlux is a web platform to draft and reconstruct metabolic models from KEGG ([www.genome.jp/kegg/](http://www.genome.jp/kegg/)). However, the test version of MicrobesFlux ran into problems on its shared server, which suffers from instability and occasional downtimes. To resolve this issue, we ported the system to the Amazon server EC2 (as shown in Figure 5.4) and rebuilt the genome modeling system. Our tool is a template-based package for tracking carbon transition and performing isotopomer corrections and  $^{13}\text{C}$ -MFA. WUFlux consists of three parts: a) a carbon transition map (CTM),

which is the basis for building  $^{13}\text{C}$ -MFA models; b) an isotopomer analysis package for analyzing both amino acids and key free metabolites in the central metabolism (based on a MATLAB program by the Wiechert group (Wahl *et al.* 2004)); c) and open source MATLAB program files which calculates  $^{13}\text{C}$ -MFA from MID information. Our websites can be accessed via the Internet and all fluxomics tools are free for use/download and easy to expand (for WUFlux). In summary, the new version of this web-based platform offers a programming-free and user-friendly broad scope tool that supports flux analysis studies.

### **5.3. Implementation**

#### **5.3.1. MicrobesFlux update**

MicrobesFlux is a free open-source platform. Our previous paper (Feng *et al.* 2012) has described the details of this platform and provided a user manual. In a nutshell, MicrobesFlux consists of four parts: a web front-end (the user interface), a backend, a task processing system, and an optimization server. In the past, MicrobesFlux had issues with the host machine on which the task processing system and the optimization server were running. When we first built MicrobesFlux, there was no specific funding for the project. Therefore, we had to rely on a server provided by Washington University, which is shared by other users and frequently restarted, resulting in an unstable task processing system that can delay user-submitted tasks. Besides, we did not have access to the most up-to-date KEGG models due to KEGG's paid subscription model. In the summer of 2014, we received a grant to continue work on this topic, fixing problems and updating our database versions. During the process of reloading MicrobesFlux, we moved the backend to a stable, commercial server and completely rewrote the task processing system. We also added a monitoring function to the platform to better manage the task processing system.

### **5.3.2. New features of reloaded MicrobesFlux**

- a.** We have updated our backend KEGG database. Users can run FBA on any KEGG organisms as of September 2014. The new database in MicrobesFlux includes 3192 species, compared to 1304 species in the test version.
- b.** We now support the SBML output format, which broke down earlier.
- c.** Users can now store an unlimited number of models. Previously, we periodically purged models due to storage constraints. Now the backend, running on Amazon EC2, supports an unlimited number of models for any user.
- d.** We now have a robust task management system. Users are guaranteed to get their results back within 24 hours of submitting the optimization job.

### **5.3.3. Development of websites for fluxomics studies**

We have built a comprehensive web-based platform including various tools (most tools were developed by our lab) for fluxomics research (as shown in Figure 5.2). In our website (fluxomics.net, as shown in Figure 5.2), users can read the latest publications by clicking the button ‘Enter 13C Flux News’; they can also build up and calculate their genome-scale model by entering MicrobesFlux; they can further get  $^{13}\text{C}$ -MFA tools by visiting WUFlux. The use of MicrobesFlux is kept the same as it was released four years ago. For WUFlux, users can download  $^{13}\text{C}$ -MFA tools: the carbon transition map (CTM), the  $^{13}\text{C}$ -MFA software package, and the MS correction Tool (as shown in Figure 5.3) by simply clicking corresponding labels and saving packages into their PCs.

## **5.4. Results**

A comprehensive platform incorporating both functions of FBA and  $^{13}\text{C}$ -MFA has been released. The new version of MicrobesFlux has been tested thoroughly. Additionally, the newly

integrated WUFlux software has a user-friendly interface; all functions are easy to operate, and calculation can be saved at any stage. Users can choose different templates for various labeled substrates and metabolic networks. By implementing WUFlux, researchers without professional knowledge of  $^{13}\text{C}$ -MFA can easily get flux data of high quality from raw MS data. Because the MATLAB codes of all program files in WUFlux are open to researchers, users can extend or enhance its capability by editing the MATLAB program. Finally, WUFlux includes a carbon fate map and a labeling correction tool for amino acids and free metabolite analysis, which can facilitate future application of  $^{13}\text{C}$ -MFA. The completely open-source platform makes good feasibility for further development. We will continue to collect users' feedback and improve its performance in the future. We hope that our platform can not only provide broad-scope fluxomics functions for characterization of novel microbial species, but also facilitate rational metabolic engineering.

### **5.5. Availability and requirements**

- Project name: Fluxomics
- Project homepage: <http://fluxomics.net> (for each individual project, MicrobesFlux is at <http://microbesflux.org> and WUFlux is at <http://13cmfa.org>)
- Operating systems: Platform independent
- Programming language: Java, Python and MATLAB (for 2012b and later version)
- License: Both MicrobesFlux and WUFlux are freely available.
- Any restrictions to use by non-academics: none

## 5.6. References

1. Crown, S. B.; Antoniewicz, M. R., Publishing  $^{13}\text{C}$  metabolic flux analysis studies: A review and future perspectives. *Metab. Eng.* 2013, 20, (0), 42-48.
2. Schloss, P. D.; Handelsman, J., Status of the Microbial Census. *Microbiol Mol Biol R* 2004, 68, (4), 686-691.
3. Swainston, N.; Smallbone, K.; Mendes, P.; Kell, D.; Paton, N., The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks. In *J Integr Bioinform*, 2011; Vol. 8, p 186.
4. Henry, C.; DeJongh, M.; Best, A.; Frybarger, P.; Linsay, B.; Stevens, R., High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol.* 2010, 28, (9), 977-82.
5. Agren, R.; Liu, L.; Shoaie, S.; Vongsangnak, W.; Nookaew, I.; Nielsen, J., The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for *Penicillium chrysogenum*. *PLoS Comput. Biol.* 2013, 9, (3), e1002980.
6. Karp, P.; Paley, S.; Romero, P., The Pathway Tools software. *Bioinformatics* 2002, 18, S225 - 32.
7. Becker, S.; Feist, A.; Mo, M.; Hannum, G.; Palsson, B.; Herrgard, M., Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nat Protocols* 2007, 2, 727 - 738.
8. Boele, J.; Olivier, B. G.; Teusink, B., FAME, the Flux Analysis and Modeling Environment. *BMC Syst. Biol.* 2012, 6, (8).
9. Hamilton, J. J.; Reed, J. L., Software platforms to facilitate reconstructing genome-scale metabolic networks. *Environ. Microbiol.* 2014, 16, (1), 49-59.

10. Zamboni, N.; Fischer, E.; Sauer, U., FiatFlux-a software for metabolic flux analysis from  $^{13}\text{C}$ -glucose experiments. *BMC Bioinformatics* 2005, 6, 209.
11. Poskar, C. H.; Huege, J.; Krach, C.; Franke, M.; Shachar-Hill, Y.; Junker, B., iMS2Flux - a high-throughput processing tool for stable isotope labeled mass spectrometric data used for metabolic flux analysis. *BMC Bioinformatics* 2012, 13, (1), 295.
12. Young, J. D., INCA: a computational platform for isotopically non-stationary metabolic flux analysis. *Bioinformatics* 2014, 30, (9), 1333-1335.
13. Yoo, H.; Antoniewicz, M. R.; Stephanopoulos, G.; Kelleher, J. K., Quantifying Reductive Carboxylation Flux of Glutamine to Lipid in a Brown Adipocyte Cell Line. *J. Biol. Chem.* 2008, 283, (30), 20621-20627.
14. Quek, L.-E.; Wittmann, C.; Nielsen, L. K.; Krömer, J. O., OpenFLUX: efficient modelling software for  $^{13}\text{C}$ -based metabolic flux analysis. *Microb Cell Fact.* 2009, 8, (25).
15. Kajihata, S.; Furusawa, C.; Matsuda, F.; Shimizu, H., OpenMebius: An Open Source Software for Isotopically Nonstationary  $^{13}\text{C}$ -Based Metabolic Flux Analysis. *BioMed Research International* 2014, 2014, 10.
16. Wiechert, W.; Mollney, M.; Petersen, S.; de Graaf, A., A universal framework for  $^{13}\text{C}$  metabolic flux analysis. *Metab. Eng.* 2001, 3, 265 - 283.
17. Weitzel, M.; Noh, K.; Dalman, T.; Niedenfuhr, S.; Stute, B.; Wiechert, W., 13CFLUX2 - high-performance software suite for  $^{13}\text{C}$ -metabolic flux analysis. *Bioinformatics* 2012.
18. Wahl, S. A.; Dauner, M.; Wiechert, W., New tools for mass isotopomer data evaluation in  $^{13}\text{C}$  flux analysis: mass isotope correction, data consistency checking, and precursor relationships. *Biotechnol. Bioeng.* 2004, 85, (3), 259-268.



19. Feng, X.; Xu, Y.; Chen, Y.; Tang, Y., MicrobesFlux: a web platform for drafting metabolic models from the KEGG database. *BMC Syst. Biol.* 2012, 6, (1), 94.

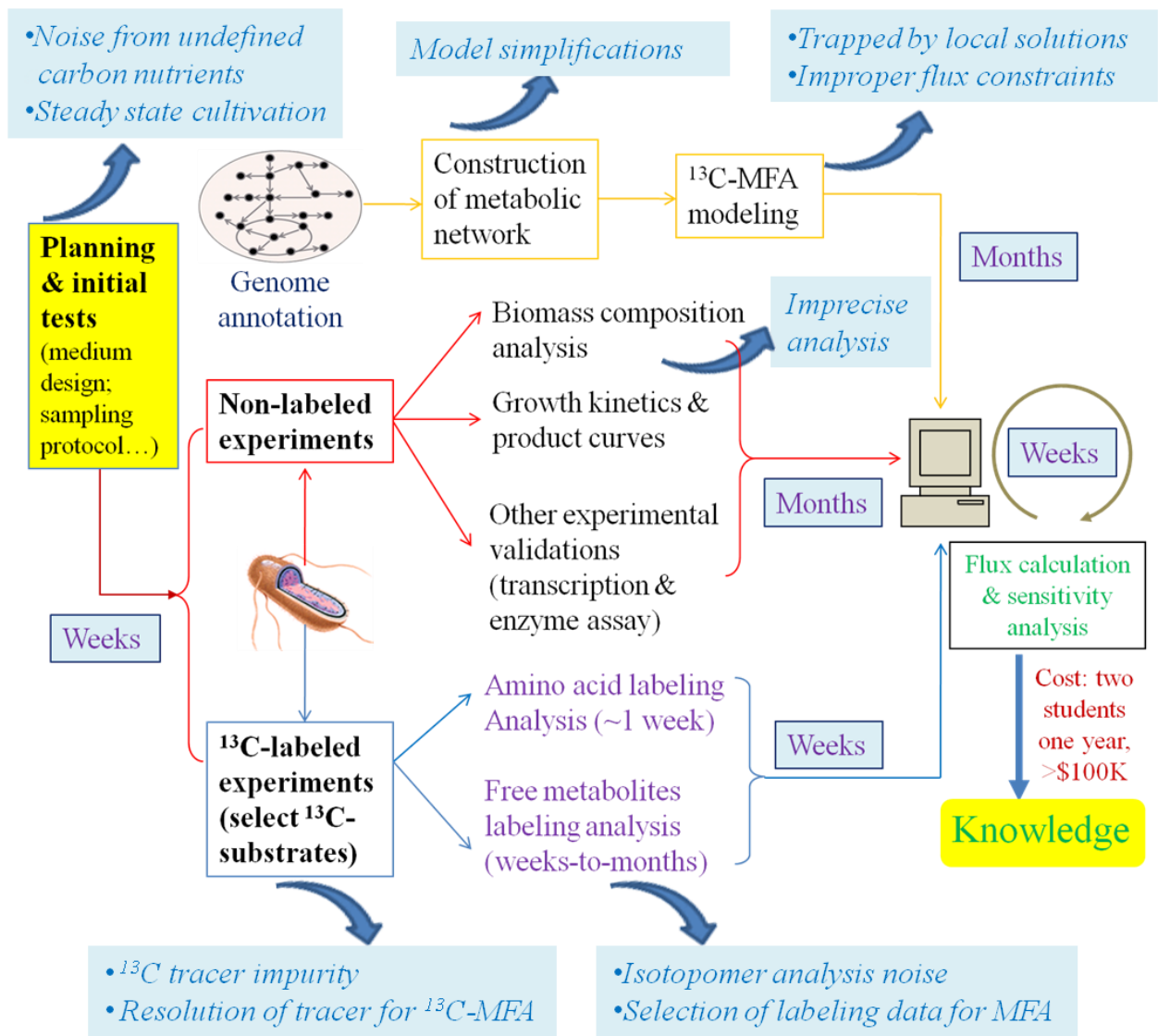


Figure 5.1. <sup>13</sup>C-MFA protocol and sources of flux analysis variance; in general, a <sup>13</sup>C-MFA requires months of work to accomplish. MFA errors (in blue boxes) can come from both experimental measurements and modeling calculations.

Fluxomics

<http://www.fluxomics.net>

Fluxomics

MicrobesFlux

WUFlux  
A Metabolic solution for <sup>13</sup>C Metabolic Flux Analysis

Fluxomics studies quantitatively describe metabolic reaction rates, therefore can be considered as a direct reflection of cellular metabolism in vivo.

This website is contributed by Tang lab members at Washington University in St. Louis. Tang lab is dedicated to developing and deploying advanced fluxomics tools into various applications.

Website Developer: Eric Xu, Gang Wu, Kevin Zhao, Shangqing Li.

Enter 13C Flux News

Fluxomics

Any questions regarding our webpage, please contact Dr. Yinjie Tang [yinjie.tang@seas.wustl.edu](mailto:yinjie.tang@seas.wustl.edu) or Mr. Gang Wu [wug@seas.wustl.edu](mailto:wug@seas.wustl.edu)

Figure 5.2. The webpage of our platform for comprehensive fluxomics studies (<http://fluxomics.net>).

# 13C MFA

RELATED PERSONNEL  
Lian He, Gang Wu, Le You

<http://13cmfa.org>

Carbon Transition Map Download

**WUFlux**  
A Model-based solution for <sup>13</sup>C Metabolic Flux Analysis

<sup>13</sup>C MFA Software Package Download

MS Correction Tool download

We provide our software 'WUFlux' here for free download.  
WUFlux consists of two parts:

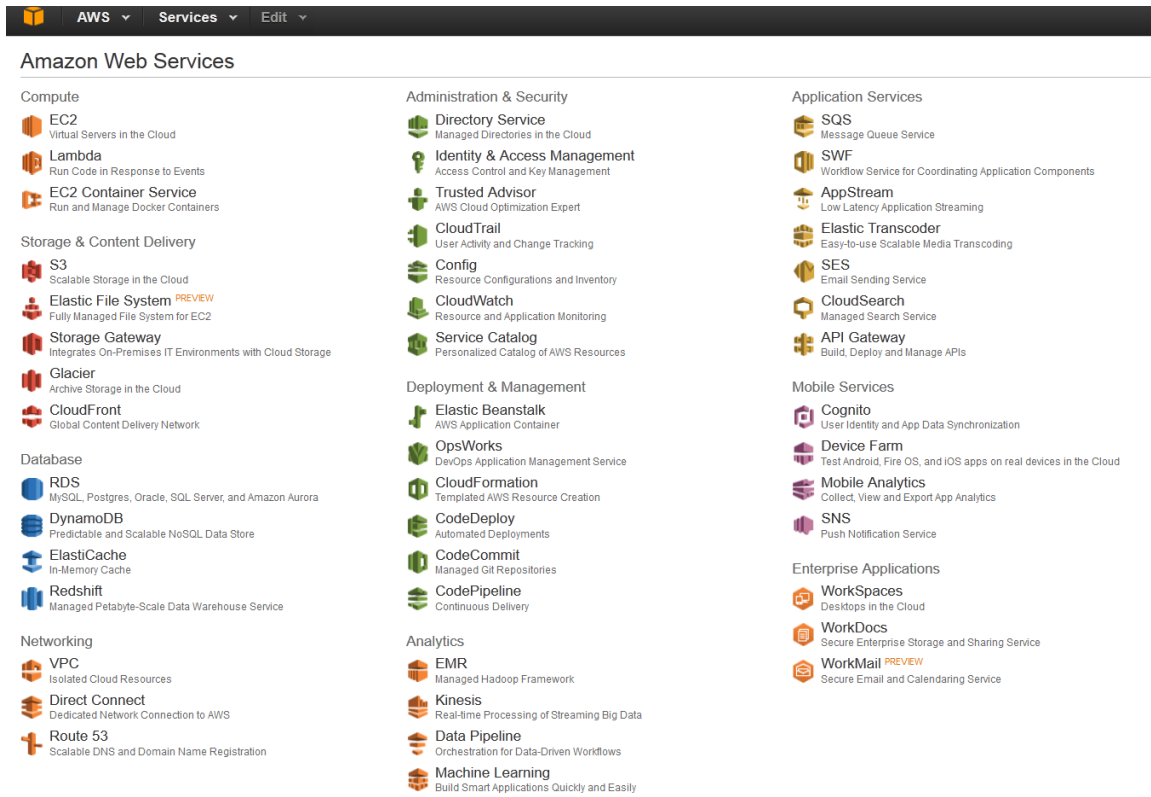
- 1) Software package for <sup>13</sup>C MFA calculation (By Lian He & Gang Wu)
- 2) Carbon transition map and MS correction tool (By Le You)

Any question, please contact Mr. Gang Wu ([gwu827@gmail.com](mailto:gwu827@gmail.com))

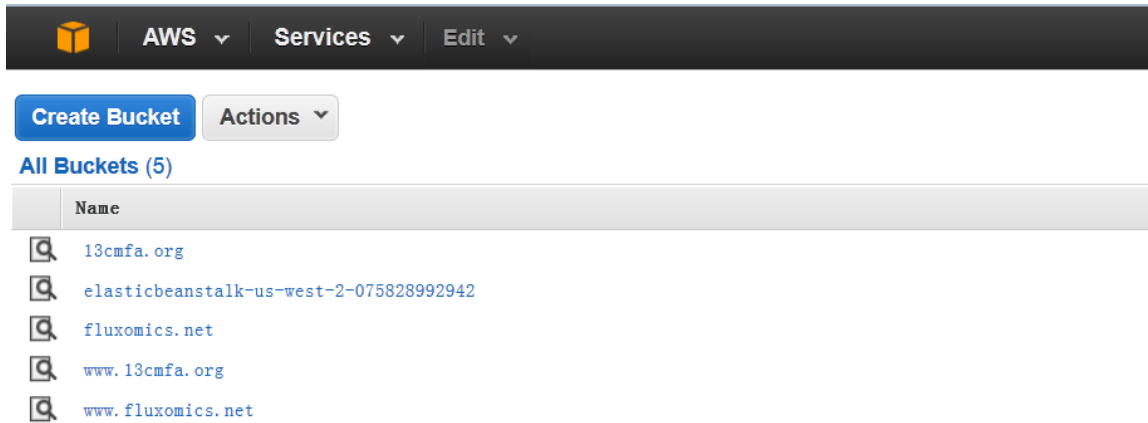
Copyright @ 2014 13cmfa.org. All Rights Reserved.

Figure 5.3. The webpage of WUFlux (<http://13cmfa.org>), which can be accessed and freely download

(a)



(b)



Figure

5.4. The webpage of Amazon server EC2, (a) all Amazon web services, (b) buckets of our websites

## CHAPTER SIX

# RAPID PREDICTION OF BACTERIAL FLUXOMICS USING MACHINE LEARNING AND CONSTRAINT PROGRAMMING

### 6.1 Abstract

Metabolic flux reflects a functional aspect of cell physiology.  $^{13}\text{C}$ -MFA ( $^{13}\text{C}$  metabolic flux analysis) has been widely used to measure *in vivo* enzyme reaction rates (i.e., metabolic flux) in microorganisms. Mining the relationship between environmental and genetic factors and metabolic fluxes hidden in existing fluxomic data will lead to predictive models that can significantly accelerate flux quantification. In this paper, we present a web-based platform (MFlux: <http://130.101.92.205/influx/>) that predicts the bacterial central metabolism via machine learning, constraint programming, and quadratic programming, leveraging data from over 100  $^{13}\text{C}$ -MFA papers on heterotrophic microbial metabolisms. Three machine learning methods, namely Support Vector Machine (SVM), k-Nearest Neighbors, and Decision Tree, were employed to study the sophisticated relationship between environmental and genetic factors and metabolic fluxes. We performed a grid search of the best parameter set for each tested algorithm and verified their performance through 10-fold cross validation. SVM yielded the highest accuracy of all three algorithms, with average error rate under 5%. Further, we employed quadratic programming to adjust flux profiles to satisfy stoichiometric constraints. Experimental results showed that MFlux can reasonably predict fluxomes as a function of bacterial species, substrate types, genetic modifications, growth rates, oxygen conditions, and cultivation methods.

## 6.2. Authors' Summary

Metabolic information is important for disease treatment, bioprocess optimization, environmental remediation, biogeochemical cycle regulation, and our understanding of life's origin and evolution. Fluxomics can quantify microbial physiology at the level of metabolic reaction rates. To speed up  $^{13}\text{C}$ -MFA, we hypothesize that genetic and environmental factors generate specific fluxome patterns that can be recognized by machine learning. Aided by constraint programming and quadratic optimization, our machine learning platform can predict meaningful metabolic information about bacterial species in their environments. Further, it can offer constraints to improve the accuracy of flux balance analysis. This study infers that the bacterial metabolic network has a certain degree of rigidity in allocating carbon fluxes, and different microbial species may share common regulatory strategies for balancing carbon and energy metabolisms. As a proof-of-principle, the use of data driven models (e.g., artificial intelligence) may assist mechanistic based models to elucidate the topology of microbial fluxomes.

## 6.3. Introduction

With the advent of systems biology tools such as genomics, transcriptomics, proteomics, and metabolomics during the last decade, the understanding of intracellular metabolisms from genotype to phenotype has been dramatically boosted. Notably,  $^{13}\text{C}$ -MFA enables the quantification of metabolic reaction rates *in vivo* [1]. It determines carbon metabolic fluxes using the mass isotopomer distribution (MID) of proteinogenic amino acids or free metabolites from  $^{13}\text{C}$  labeling experiments.  $^{13}\text{C}$ -MFA is considered as a reliable measurement of central metabolic

reaction rates [2], which has demonstrated its power in discovering novel pathways [3,4], validating gene functions [3], verifying engineered strains [5,6], and revealing energy metabolisms of host strains [7]. In the past decade, advanced parallel bioreactor systems, mass spectrometry, and computational tools resolving metabolic fluxes have been developed [8,9,10,11], which improved the precision of flux profiles [12] and extended  $^{13}\text{C}$ -MFA's application to the non-stationary metabolic phase [13,14]. On the other hand, broad applications of  $^{13}\text{C}$ -MFA are still hindered because  $^{13}\text{C}$ -experiments, biomass analysis, and flux calculations are expensive and time-consuming [15]. Moreover, some microbial systems may not be amenable to  $^{13}\text{C}$ -MFA if they require complex nutrients or their genome annotation is incomplete [16]. Before performing  $^{13}\text{C}$ -MFA on non-model species, laborious work is needed to examine extracellular metabolites, to characterize unknown pathways, and to analyze biomass compositions.

This study aimed to employ an artificial intelligence (AI) approach called machine learning (ML) to investigate bacterial fluxomics patterns. ML is a powerful tool in systems biology [17] and has demonstrated successes in omics studies [18,19]. For example, the precision of genome annotation on the model species *C. elegans* has been enhanced by employing a simplified SVM (support vector machine) method. Researchers have reached an accuracy of 75% on controversial genes [20]. At the transcriptomics level, ML approaches are used for disease identification. For instance, SVM has successfully recognized the gene expression patterns of hepatocellular carcinoma [21], diffuse large B-cell lymphoma [22], and ovarian cancer [23]. At the proteomics level, Supek *et al.*, have employed a combined approach by integrating the Principal Component Analysis (PCA) method with SVM, to enhance analytic power in identifying 'fingerprint' proteins (i.e., unique proteins in each tissue) from different horseradish



tissues (leaf, teratoma, and tumor) grown *in vitro* [24]. In metabolomics, an SVM method can resolve the NMR data of metabolites in urine samples from different groups of people (healthy vs. pneumonia) [25]. In metabolic modeling, Karp's group adopted ML algorithms to predict the existence of various pathways for metabolic network reconstruction in different organisms [19].

The general idea of ML is to statistically build a predictive *model* or an *estimator*  $\mathbb{R}^n \mapsto \mathbb{R}$  that maps an n-dimensional real number vector called the *feature vector* to a real number called the *target* or *label*. If the target takes discrete values, we call the ML model a *classifier*; otherwise, a *regressor*. A pair of a feature vector and a target forms a *sample*. Given samples, a machine learning algorithm will find such a mapping, usually through solving a numerical optimization problem, to minimize the predictive error. Samples used to train a model form the *training set* while those for testing the performance form the *testing set*. The models learned through ML are usually not analytical models that can be represented using an equation. Rather, they are numerical operators. For example, an artificial neural network (ANN) model can be represented by many matrices, and when being used to predict, the input variables will be multiplied with those matrices sequentially. A bad model can only predict well on the training set as if it 'remembers' the training samples, while a good model can learn the patterns among data and still be accurate on samples it has never 'seen'. Hence, researchers usually make the training and test sets mutually exclusive. A mechanism called *cross validation* is used to ensure the mutual exclusiveness of training and test sets while make full use of all data.

A distinct advantage of ML applications is that they can reduce the need for costly experimental supplies and time-consuming bench work. Despite the progress in utilizing ML methods in systems biology, there is no similar application in the fluxomics field to predict the flux profile. Therefore, we conceived the idea of integrating ML strategies with fluxomics

research. To efficiently employ machine learning methods, a database with a sufficient number of samples is a prerequisite. Recently, a  $^{13}\text{C}$ -MFA database ('CeCaFDB') has been constructed, which includes over 100 papers (mostly on prokaryotic metabolisms) [26]. Based on this database, we initiated five categorical and sixteen continuous features to describe the environmental and genetic factors involved in  $^{13}\text{C}$ -MFA of bacterial species. Unlike most omics projects employing ML approaches, this work built regression models rather than classifiers: Twenty-nine lumped central metabolic fluxes were adopted as the outputs to describe bacterial carbon metabolisms. A 10-fold cross validation evaluated the performance of different algorithms. Furthermore, we included a knowledge-based system to check whether user inputs were biologically meaningful. Lastly, quadratic programming was employed to adjust the fluxes predicted by ML to satisfy the stoichiometric constraints. Our web-based platform ('MFlux') provides reasonable predictions for central metabolic flux profiles on 30 bacterial species, and it can be accessed online (<http://130.101.92.205/influx/>). Although our platform is still in the early phase, our attempt to employ an AI approach in fluxome studies will have broad impacts on both systems biology and metabolic engineering fields.

## **6.4. Methods**

### **6.4.1. Data collection and preprocessing**

All the training data for MFlux comes from the literature. The total uptake rate of carbon sources is defined as 100; all other fluxes are normalized to a scale of 100. We obtained  $^{13}\text{C}$ -MFA information for bacterial species from the CeCaFDB database and added a few recent papers (total ~120 papers, as of January 2015) [26].  $^{13}\text{C}$ -MFA data related to photosynthetic bacteria was excluded in this study because of their unique fluxome topologies (such as the Calvin Cycle and the reversed TCA cycle) and insufficient sampling sizes for ML.

In heterotrophic microorganisms, interconversions between glycolysis metabolites (phosphoenolpyruvate and pyruvate) and TCA cycle metabolites (oxaloacetate and malate) involve a set of anaplerotic reactions (e.g., phosphoenolpyruvate carboxylase, phosphoenolpyruvate carboxykinase, pyruvate carboxylase, and malic enzyme), serving as a key switch points for carbon flux distribution in bacteria [27]. These reactions, balancing both carbon and cofactors, may be employed by different microbial species. For example, *E. coli* anaplerotic pathways involve phosphoenolpyruvate carboxylase and malic enzyme, while *Bacillus* species furnish pyruvate carboxylase (the pyruvate shunt). In the case of *Corynebacterium*, both phosphoenolpyruvate carboxylase and pyruvate carboxylase are functional [28,29]. These anaplerotic pathways can re-route fluxes when a central pathway such as pyruvate kinase is knocked out. To ease the machine learning efforts, the anaplerotic pathways are lumped into two routes that exchanges fluxes between the TCA cycle and the glycolysis nodes:  $MAL \leftrightarrow PYR + CO_2$  and  $OAA \leftrightarrow PEP + CO_2$ . This simplification also considered the fact that  $^{13}C$ -MFA has poor resolution on anaplerotic fluxes because various combinations of these reactions can generate similar labeling patterns in amino acids [30].

#### **6.4.2. Feature selection and scaling**

ML can make predictions with iteratively-tuned parameters and well-trained models to account for influential factors in cell metabolism. Based on published  $^{13}C$ -MFA methodologies and microbial physiologies, we proposed five categorical features: species, nutrient types, oxygen conditions, genetic background, and cultivation methods. We had two considerations during feature selection: First, genetic modifications can significantly re-organize fluxomes. To improve the predictability on mutant strains, our platform allows “turn-off” or “turn-on” certain central pathways (by manually setting the flux boundaries) in engineered strains. Second, the

factor of cultivation method aims to reveal fluxome differences between shake flask cultures (a pseudo-steady state approach) and bioreactor cultures (a well-controlled fermentation or chemostat cultivation). Meanwhile, we have sixteen continuous features: growth rate, substrate uptake rate, and the ratios of substrate co-utilizations (glucose, fructose, galactose, gluconate, glutamate, citrate, xylose, succinate, malate, lactate, pyruvate, glycerol, acetate and  $\text{NaHCO}_3$ , as shown in Figure 1). Since the training features include both categorical and continuous ones, "OneHotEncoder", a function of the Python scikit-learn module, was used to convert categorical feature values into real numbers. Each feature was then standardized into zero mean and unit variance as assumed by many ML approaches. For each predicted flux, we normalized the training dataset via the min-max method into the interval [0, 1]. In addition to the min-max method, we also tested unit-variance-zero-mean standardization for scaling flux values, and the result was quite similar.

#### **6.4.3. Machine learning algorithm selection**

The problem of predicting fluxes is formulated as a regression problem in ML, where a computer program learns from existing data to estimate continuous variables. Twenty-nine regressors were trained to predict the 29 fluxes. We tested three widely-applied ML algorithms, including k-nearest neighbors (k-NN), decision tree, and SVM. To ensure a fair comparison, we performed a grid search for the best parameter set of each algorithm. The detailed parameter sets for 29 regression models can be found in the prediction results on our web page ([http://130.101.92.205/influx/svr\\_both\\_rbf\\_shuffle.log](http://130.101.92.205/influx/svr_both_rbf_shuffle.log)). The programming language used for this project was Python 2.7; the numpy and scikit-learn modules were utilized for machine learning [31]. Program files for training the models and testing them are wrapped in Supporting Information 1.

#### **6.4.4. Error evaluation and cross validation**

To evaluate the quality of the predictive model, we used mean squared error (MSE) and root mean squared error (RMSE). Depending on different evaluation tasks, we may represent RMSE relatively (RRMSE) with respect to the dynamic range of fluxes. Considering the limited number of samples in the current database, we adopt a 10-fold cross validation. An N-fold cross validation works as follows: All samples in our database are spliced into N equal parts. In each iteration, N-1 parts are used as the training set, while the remaining as the test set. In the next iteration, the test set will be rotated to another part of the data, and the training set will consist of all other samples. This procedure will stop when all parts of the data have been incorporated into the test set exactly once, and training set exactly N-1 times. Finally, the accuracy of the model can be calculated by checking the prediction result in each sample.

#### **6.4.5. Stoichiometric constraints and boundary**

One unique feature of our method is incorporating the overall mass balance through central metabolic pathways. The stoichiometric equations in Figure 1 under steady state are summarized as follows:

$$\begin{aligned}
\text{G6P} : v_1 &= v_2 + v_{10} + vbm_{g6p} & (1) \\
\text{F6P/FBP} : v_2 + v_{15} + v_{16} + 100 \cdot ratio_{fructose} &= vbm_{f6p} + v_3 & (2) \\
\text{DHAP} : v_3 + 100 \cdot ratio_{glycerol} &= v_4 & (3) \\
\text{GAP} : v_3 + v_4 + v_{14} + v_{15} + v_{25} &= v_5 + v_{16} + vbm_{gap} & (4) \\
\text{3PG} : v_5 &= v_6 + vbm_{3pg} & (5) \\
\text{PEP} : v_6 &= v_7 + v_{28} + vbm_{pep} & (6) \\
\text{PYR} : v_7 + v_{25} + v_{29} + 100 \cdot ratio_{pyruvate} &= v_8 + v_{27} + vbm_{pyr} & (7) \\
\text{AceCoA} : v_9 + v_{17} + v_{24} + v_{26} + vbm_{accoa} &= v_8 & (8) \\
\text{Ru5P} : v_{11} &= v_{12} + v_{13} & (9) \\
\text{R5P} : v_{13} &= v_{14} + vbm_{r5p} & (10) \\
\text{E4P} : v_{15} + vbm_{e4p} &= v_{16} & (11) \\
\text{S7P} : v_{14} &= v_{16} & (12) \\
\text{X5P} : v_{12} + 100 \cdot ratio_{xylose} &= v_{14} + v_{15} & (13) \\
\text{6PG} : v_{10} + 100 \cdot ratio_{gluconate} &= v_{11} + v_{25} & (14) \\
\text{CIT} : v_{17} + 100 \cdot ratio_{citrate} &= v_{18} & (15) \\
\text{ICIT} : v_{18} &= v_{19} + v_{24} & (16) \\
\text{AKG} : v_{19} + 100 \cdot ratio_{glutamate} &= v_{20} + vbm_{akg} & (17) \\
\text{SUC} : v_{20} + v_{24} + 100 \cdot ratio_{succinate} &= v_{21} + v_{aa1} & (18) \\
\text{FUM} : v_{21} + v_{aa2} &= v_{22} & (19) \\
\text{MAL} : v_{22} + v_{24} + 100 \cdot ratio_{malate} &= v_{23} + v_{29} & (20) \\
\text{OAA} : v_{23} + v_{28} &= v_{17} + vbm_{oaa} & (21)
\end{aligned}$$

Specifically,  $v_1$  represents the flux from carbon substrate (either glucose or galactose) since both glucose and galactose can be catabolized to G6P,  $v_{aa1}$  and  $v_{aa2}$  represents fluxes involved in biomass building block synthesis [32], while  $vbm$  represents carbon fluxes going to biomass from different precursors.

A series of linear constraints can be derived from the stoichiometric equations above and used to restrain fluxes predicted by the ML methods:

$$v_1 - 100 \cdot (ratio_{glucose} + ratio_{galactose}) = 0 \quad (22)$$

$$v_3 - v_4 + 100 \cdot ratio_{glycerol} = 0 \quad (23)$$

$$v_{11} - v_{12} - v_{13} = 0 \quad (24)$$

$$v_{14} - v_{16} = 0 \quad (25)$$

$$v_{10} - v_{11} - v_{25} + 100 \cdot ratio_{gluconate} = 0 \quad (26)$$

$$-v_{17} + v_{18} - 100 \cdot ratio_{citrate} = 0 \quad (27)$$

$$-v_{12} + v_{14} + v_{15} - 100 \cdot ratio_{xylose} = 0 \quad (28)$$

$$-v_{18} + v_{19} + v_{24} = 0 \quad (29)$$

$$-v_{22} + v_{23} - v_{24} + v_{29} - 100 \cdot ratio_{malate} = 0 \quad (30)$$

Among equations listed above, Eq. 22 indicates the case for co-metabolism of both C6 sugars. Meanwhile, a list of inequality constraints can be drawn, given that all biomass fluxes are non-negative:

$$v_1 - v_2 - v_{10} \geq 0 \quad (31)$$

$$v_2 - v_3 + v_{15} + v_{16} + 100 \cdot ratio_{fructose} \geq 0 \quad (32)$$

$$v_3 + v_4 - v_5 + v_{14} + v_{15} - v_{16} + v_{25} \geq 0 \quad (33)$$

$$v_5 - v_6 \geq 0 \quad (34)$$

$$v_6 - v_7 - v_{28} \geq 0 \quad (35)$$

$$v_7 - v_8 + v_{25} - v_{27} + v_{29} + 100 \cdot ratio_{pyruvate} \geq 0 \quad (36)$$

$$v_8 - v_9 - v_{17} - v_{24} - v_{26} \geq 0 \quad (37)$$

$$v_{13} - v_{14} \geq 0 \quad (38)$$

$$-v_{15} + v_{16} \geq 0 \quad (39)$$

$$v_{19} - v_{20} + 100 \cdot ratio_{glutamate} \geq 0 \quad (40)$$

$$-v_{17} + v_{23} + v_{28} \geq 0 \quad (41)$$

$$-v_{21} + v_{22} \geq 0 \quad (42)$$

Among all inequality constraints, constraint (Eq. 39) works well except for the case of *zwf* knockout, where the directions of Eq. 39 could be reversed [33].

#### 6.4.6. Flux adjustment using stoichiometric constraints

We also adopted a quadratic programming method similar to minimization of metabolic adjustment (MOMA) [34], to adjust fluxes to satisfy the stoichiometric constraints. The CVXOPT package for Python was employed here for quadratic programming [35]. The optimization problem is modeled as

$$\begin{aligned}
& \text{Minimize } f(\mathbf{v}) = \sum_{i=1}^{29} (\text{Scaled}(v_i) - \text{Scaled}(\hat{v}_i))^2 \\
& \text{Subject to } \mathbf{S} \cdot \mathbf{v} = 0, \\
& \quad \mathbf{A} \cdot \mathbf{v} \geq 0,
\end{aligned} \tag{43}$$

where the vector  $\hat{\mathbf{v}} = [\hat{v}_1, \dots, \hat{v}_{29}]$  is the flux values predicted by ML, the vector  $\mathbf{v} = [v_1, \dots, v_{29}]$  is the flux values to be solved in this optimization problem, the function “Scaled (·)” uses Min-Max scaling to scale all fluxes into the range [0, 1], the matrix “S” is obtained from all equality constraints from Eq. 22 to Eq. 30, and the matrix A is obtained from all inequality constraints from Eq. 31 to Eq. 42. Scaling fluxes into the same range is done to avoid bias because fluxes have different dynamic ranges. The root mean squared error (RMSE) is used to evaluate the quality of flux prediction by examining the deviation of the predicted flux profile from the <sup>13</sup>C-MFA flux. The objective function  $f(\mathbf{v}')$  can be rewritten into standard quadratic programming problem using the following steps:

$$\begin{aligned}
f(\mathbf{v}) &= \sum_{i=1}^{29} (\text{Scaled}(v_i) - \text{Scaled}(\hat{v}_i))^2 = \sum_{i=1}^{29} \left( \frac{v_i - \text{Min}_i}{\text{Max}_i - \text{Min}_i} - \frac{\hat{v}_i - \text{Min}_i}{\text{Max}_i - \text{Min}_i} \right)^2 \\
&= 2 \cdot \sum_{i=1}^{29} \left( \frac{1}{2} \frac{v_i^2}{(\text{Max}_i - \text{Min}_i)^2} + \frac{-1 \cdot v_i \cdot \hat{v}_i}{(\text{Max}_i - \text{Min}_i)^2} + \frac{1}{2} \frac{\hat{v}_i^2}{(\text{Max}_i - \text{Min}_i)^2} \right)
\end{aligned} \tag{44}$$



where  $Max_i$  and  $Min_i$  are the range of the  $i^{\text{th}}$  flux. Since the last term  $\frac{1}{2} \left( \frac{\hat{v}_i}{Max_i - Min_i} \right)^2$  and the coefficient 2 are constants, they can be omitted from the objective function. Hence, Eq. 43 can be rewritten in standard quadratic programming form as

$$\begin{aligned} \text{Minimize } f(\mathbf{v}) &= \frac{1}{2} \sum_{i=1}^{29} \frac{(v_i)^2}{(Max_i - Min_i)^2} + \sum_{i=1}^{29} \frac{-1 \cdot v_i \cdot \hat{v}_i}{(Max_i - Min_i)^2} \\ \text{Subject to } \mathbf{S} \cdot \mathbf{v} &= 0, \\ \mathbf{A} \cdot \mathbf{v} &\geq 0. \end{aligned} \quad (45)$$

For the upper and lower boundaries of each flux, i.e.,  $Max_i$  and  $Min_i$ , we use the maximal and minimal values observed in multiple datasets as the default values (shown in Figure 6.2). Users can manually set desired values for the upper/lower bounds of any specific flux in the MFlux webpage, or they can opt to not use any boundaries. For instance, users can simply set the bound value of a certain flux as zero if this specific gene is knocked out.

#### 6.4.7. Constraint programming and input checking

To ensure user inputs are reasonable, MFlux first checks the satisfiability of input values. This system scans the inputs (e.g., growth rates, oxygen usage, and substrate uptake rates) and determines whether they are biologically meaningful (e.g., unrealistic high cell growth rate). If a set of inputs are suspected to be unreasonable, MFlux reports an error to warn the users.

#### 6.4.8. Overall system design

Different parts of MFlux mentioned above are put together as illustrated in Figure 3. The prediction on 29 fluxes is done via an RBF-kernel SVM, whose outcome will be tuned by constraint programming to generate final prediction. Users can set boundary constraints to

represent information about genes that are knocked out on the species, and such information will be used by constraint programming. If parameter inputs by users are not biologically meaningful, a warning message will be attached in the final result. In the future, users will also have the option to enter fluxes and settings of their own experiment to enrich our database and improve the prediction accuracy of MFlux.

## **6.5. Results and Discussion**

### **6.5.1. Pathway map and statistical analysis results**

The core metabolism of bacteria is summarized into a pathway map in Figure 1. Considering the availability of information, 29 major fluxes with 14 potential substrates were used to represent a universal heterotrophic carbon metabolism for non-photosynthetic prokaryotic species, which includes glycolysis, the tricarboxylic acid (TCA) cycle, the pentose phosphate (PP) pathway, the Entner–Doudoroff (ED) pathway, the glyoxylate shunt, and the anaplerotic pathways. The anaplerotic pathway fluxes cannot be determined when [1-<sup>13</sup>C] glucose is fed as a sole labeled substrate [36]. Information on the anaplerotic pathway is either incomplete or not precise in many publications in our database. Consequently, we simplified the anaplerotic pathway into two reversible fluxes. Similarly, we ignored several overflow fluxes which occasionally appear in <sup>13</sup>C-MFA of anaerobic metabolisms (e.g., the secretion of formate, butyrate, or pyruvate), because of lacking sufficient samples for efficient machine learning. Omission of those fluxes can also partially explain the high prediction error in specific fluxes (e.g., v8: Pyruvate → Acetyl-CoA).

By statistical analysis, we determined the variation between each flux profile and the average flux profile from our <sup>13</sup>C-MFA database. The average value, the range, and the 95%

confidence interval for each individual flux are shown in Figure 2. The most conservative fluxes include the non-oxidative pentose phosphate pathway and the glyoxylate shunt. The former pathway supplies precursors for bio-synthesizing amino acids (i.e., histidine, phenylalanine, tryptophan, and tyrosine) and nucleotides. The latter acts as an alternative carbon-conserving path to the TCA cycle and is inhibited by the presence of glucose (most  $^{13}\text{C}$ -MFA is based on the glucose metabolism). All 29 fluxes were found to have a narrow 95% confidence interval (compared to possible flux ranges), suggesting that fluxes of bacterial species in our database varies in a relatively small range. This is because most  $^{13}\text{C}$ -MFA studies are focusing on models species (e.g., *E. coli* and *B. subtilis*) and glucose-based metabolism, while there are much less MFA efforts to study non-model species or metabolism of carbon substrates other than sugars (i.e., bias of fluxome research across).

### **6.5.2. Optimization of algorithm and parameters**

To decide the most suitable ML algorithm, we first performed a grid search on the parameter space, based on the collection of a wild type (WT) database for initial screening. After one week of running the search program on the server, the best results of the three different algorithms (for SVM, only the linear kernel was considered here) were presented in Figure 4. Evidently, SVM made better predictions than either decision tree or k-NN on most fluxes. After this step, we carried out a second round of grid searching to optimize parameters and improve the performance of SVM on the whole phenotype (WP) database (both WT and engineered). Both the linear kernel and radial bias function (RBF) kernel were included in this round of searching.

Better cross-validation results were expected from the SVM model trained on the WT database, rather than on the WP database, considering that sophisticated genetic variations are

not included in the WT database. However, cross-validation results refuted our initial thought: the models from the WP database demonstrated better performance than those trained on the WT database (Figure 5). This result can be interpreted as meaning that the size of the training dataset is a major factor in determining the model quality, especially when the training database size is relatively small (the size of WT dataset is 154, and the size of WP dataset is 450). We also compared the SVM results with the linear kernel and the RBF kernel, and the RBF kernel showed slightly better performance (Figure 6). We finalized the parameter setting of MFlux by taking the parameter set which output the best cross-validation result. For all the algorithms tested, v11 (the second step of the oxidative PP pathway) and v24 (the glyoxylate shunt) are insensitive in terms of RRMSE. Two factors may contribute to this problem: v11 and v24 have relatively narrow numerical ranges, and consequently even small numerical variations will generate larger relative errors for both fluxes. Meanwhile, genetic modifications may influence both v11 (e.g., *zwf* knockout [37]) and v24 (e.g., *ppc* knockout [38,39]). For instance, knocking out *zwf* in *E. coli* will cause a zero flux in v10 (the oxidative pentose phosphate pathway, OPP pathway). However, lack of sufficient information on flux re-organization mechanisms in engineered microbes reduced ML predictability. This is because most engineered microbial fluxomics studies are focused on a few model species such as *E. coli*. To resolve this problem, the MFlux platform allows users to manually set the boundaries of central fluxes to improve prediction quality (e.g., give a zero flux through the OPP pathway for *E. coli zwf* mutant).

### **6.5.3. Flux correction by quadratic programming**

After parameter optimization, the SVM models equipped with the best parameter sets can predict with relatively small variation. However, the flux profile predicted by the ML method

does not necessarily satisfy the inherent stoichiometric constraints of metabolic networks because the ML method does not consider them. Sometimes the situation could get even worse: Specific fluxes predicted by the ML algorithm may go beyond a reasonable range (e.g., the predicted glyoxylate shunt may have a negative value). To address those issues, we employed quadratic programming for flux correction, as described in the Methods section. More rational results with improved accuracy are expected after flux correction. An essential assumption of this step is that ML predictions are relatively close to the real values reported in the literature. This hypothesis is backed up by our cross-validation results and further validated in the following case studies. Notably, biomass equations may have differences among MFA papers (e.g., equation 18 and 19). Considering the variations in biomass fluxes, the revised quadratic programming didn't include constraints from succinate mass balance equation (i.e., succinyl-CoA flux towards biomass synthesis).

#### 6.5.4. MFlux case studies

To demonstrate the functionality of MFlux, we carried out tests on twenty cases, and the results are illustrated in Figure 7. General information for each case is listed in Table 1, and comprehensive results are included in Supporting Information 2. In general, MFlux can achieve decent flux predictions. Here we will demonstrate two cases which are case 8 and 16 in Supporting Information 2. In case 8, *B. subtilis* **strain uptakes the mixed substrates succinate and glutamate.**

To illustrate mixed substrates co-metabolisms, we tested MFlux with <sup>13</sup>C-MFA data of *B. subtilis* strain reported by Chubukov *et al.* [40]. Microbial fermentation fed with multiple substrates of low price is promising for the biotechnology industry. However, there are few quantitative analyses of this topic. In this test, we adopted the same set of parameters found in

the literature (Supporting Information 2, Case 8) as the inputs of MFlux. For flux correction, we directly took the default boundary setting for quadratic programming. A comparison of flux profiles reported by  $^{13}\text{C}$ -MFA, predicted by ML, and predicted by MFlux is illustrated in Figure 8. ML and MFlux produce good predictions on most fluxes, closely matching the  $^{13}\text{C}$ -MFA flux profile (the RMSE is less than 5). For ML, the predictions have large variation on specific fluxes (e.g., v11 - oxidative PP pathway and v19 – TCA cycle). Quadratic programming can further adjust flux profiles and reduce deviations of flux predictions. The corrected flux profiles also meet the basic stoichiometric relationship of the metabolic network. The final prediction from MFlux shows improvement, with RMSE reduced to 3.2.

In case 16, *G. thermoglucosidasius* grows under microaerobic conditions. *G. thermoglucosidasius* is a thermophilic and ethanol-tolerant bacterium which can convert both hexose and pentose into ethanol [30]. To predict its central fluxomes, the parameter set we used is listed in Supporting Information 2 (with the default boundary setting for flux correction). A heat map compares  $^{13}\text{C}$ -MFA fluxes with ML-only fluxes and MFlux results (Figure 9). The results are encouraging: ML alone gives an RMSE of 4.0, while MFlux uses both ML and quadratic programming to improve the prediction to an RMSE of only 3.0. According to 20 case studies, the average flux set has very large variations (average RMSE of 33.5) from actual  $^{13}\text{C}$ -MFA fluxes (Supporting Information 2). In this case, MFlux reduces the deviations of predicted fluxes from  $^{13}\text{C}$ -MFA values.

For species with genetic modifications in major pathways (cases 2, 3, 4, 12, and 13, *E. coli* and *C. glutamicum*), MFlux predictions have an average RMSE between 5 and 10, higher than the RMSE for prediction of wild type strains. Since MFlux is currently unable to capture complex regulatory mechanisms of flux reorganization, Human-Computer Interaction can be

employed by manually tuning boundary values of certain fluxes to improve flux prediction quality. For example, knocking out *ppc* on *E. coli* may activate the glyoxylate shunt [38,39], so users can assign a non-zero lower boundary of the glyoxylate shunt when running MFlux.

### **6.5.5. Compare flux balance analysis with MFlux for *E. coli* metabolism**

Stoichiometry-based flux balance analysis (FBA) is an important tool to predict unknown cell metabolism. Accurate FBA prediction relies highly on appropriate setting the objective function and the flux constraints appropriately (based on thermodynamics or experimental analysis) [41]. Here, we compare FBA with MFlux for predicting *E. coli* metabolisms. The latest version of *E. coli* iJO1366 genome-scale model (2583 fluxes) was used [42,43]. Two comparative case studies were performed on *E. coli* fluxomes: One case for glucose based <sup>13</sup>C-MFA via parallel labeling experiments [12], the other case for glucose and glycerol co-utilization (unpublished data from the Shimizu Group). Neither of the test cases was included in the training database of MFlux. Given <sup>13</sup>C-MFA results as the control, MFlux results have smaller RMSEs than FBA predictions. In the first case, the FBA has an RMSE of 11.3, while MFlux has an RMSE of 6.5 (Figure 10a). In the second case, the FBA has an RMSE of 22.5, while MFlux has an RMSE of 5.1 (Figure 10b). To circumvent variations caused by alternative solutions in FBA, we also employed pFBA and geometricFBA in cases study [44,45] (results were included in Supporting Information 3). In general, pFBA didn't show better results compared with FBA for either case, while geometricFBA did not converge during our calculation.

FBA alone has been shown to give good predictions of growth rate as well as input and output fluxes, but not of intercellular fluxes [2,46]. It is difficult to obtain actual P/O ratios, the non-growth associated maintenance energy, the oxygen flux, and the transhydrogenase activities [47]. These energy/cofactor variables strongly affect the fluxes in the oxidative PP pathway

(NADPH generation) and the TCA cycle (NADH, NADPH, and FADH<sub>2</sub> generation). Without proper flux constraints and objective functions, it is challenging for FBA to narrowly determine intracellular fluxomes in suboptimal metabolisms, especially for co-metabolism dual substrates (i.e., there are large solution spaces for the cell metabolism to optimize biomass growth using two substrates). As a complementary tool, MFlux may offer a quick metabolic overview and provide reasonable flux boundaries to reduce FBA solution spaces when proper constraints for FBA are not available.

#### **6.5.6. Perspective of metabolic robustness and machine learning of fluxome patterns**

‘Robustness’ was originally defined as closed-loop process stability under perturbations in the control field. This definition is applicable to biochemical networks. To maintain the physiological output (i.e., the fluxome) within a desired range, microorganisms employ sophisticated control disciplines at different architecture levels, from the genome to the phenotype [48]. In contrast to chaotic transcriptional profiles, the microbial fluxome shows robustness so that cells can survive in constantly-altering environments or in response to genetic mutations [49,50,51]. Metabolic rigidity at the flux level was first reported by Stephanopoulos in the early 1990s [52,53]: NADPH is important for anabolism in the exponential growth phase, and the flux ratio around glucose-6-phosphate is rigid to form NADPH at the oxidative PP pathway [53]. Moreover, 12 precursors from the central metabolism are required for biomass formation, which all have relatively small variations (mainly dependent on biomass compositions). Due to both thermodynamic and mass balance constraints, cell metabolism aims to minimize variations in flux ratios under environmental perturbations. This rule also works for engineered microbes with moderately over-expressed pathways or strains from random



mutations. Those metabolic patterns can be identified by computational intelligence methods to facilitate fluxome prediction.

Flux pattern recognition enables MFlux to predict metabolism of new species by learning from a small set of fluxome information from the same genus. For example, the metabolisms of *Pseudomonas aeruginosa*, *Pseudomonas fluorescens*, and *Pseudomonas putida* have been studied by <sup>13</sup>C-MFA in the past decade [54,55,56,57,58]. The results show that different *Pseudomonas* species employ remarkably identical fluxomics types: They employ a highly active ED pathway for glycolytic metabolism and keep a low flux on the PP pathway for biomass synthesis, due to the lack of the *pfk* gene [59]. The ED pathway has less cost for protein formation than the Embden–Meyerhof–Parnas (EMP) pathway, yet only one ATP is formed per glucose [60,61]. *Pseudomonas* species have slow cell growth rates, and their aerobic metabolisms do not yield any by-products. They also demonstrate a very active pyruvate shunt (malate → pyruvate) and NADPH overproduction flux (a benefit for counteracting oxidative stress). On the other hand, the TCA cycle in *Pseudomonas* species shows plasticity under genetic and environmental variations [62], and can respond to increased ATP and NADH demands under stress conditions [63].

For different bacterial families (e.g., *E. coli* and *Bacillus*), their fluxomes (e.g., glucose metabolisms) can also be similar, because central fluxes in catabolism are regulated by energy and building block requirements that show much smaller variations than genome or transcriptional differences. On the other hand, change of carbon substrates may alternate flux distributions. For example, co-utilization of glucose and glycerol (case study 3) in *E. coli* cause significant re-organization of fluxomes. In a same microbial strain, different fluxome patterns can be employed for metabolizing different substrates (e.g., glucose based fluxome vs acetate

based fluxomes). Recognizing these metabolic patterns allows the use of a relatively small training database to perform a decent metabolic prediction of diverse metabolic types. Consequently, these common principles of certain classes of microorganisms can be captured by machine learning for fluxome predictions.

### **6.5.7. Limitations of machine learning**

There are still several major challenges regarding MFlux. First, the  $^{13}\text{C}$ -MFA flux in literature database may have errors and bias, which would be included in the learning/training process of MFlux and lead to further variations. For example, current  $^{13}\text{C}$ -MFA studies are not evenly distributed among a broad scope of microbial genus. Most reported MFAs are concentrated in a few model microbial species using glucose as substrates, while there are much fewer papers on non-model species or metabolism of diverse substrates other than sugar. Such problem (bias of fluxome in the database) can be resolved after more papers on  $^{13}\text{C}$ -MFA can be published for non-model species.

Second, the predictability of ML is limited to species and pathways that are already included in learning. More information and effort are required to deal with cases of strains with engineered pathways that hijack flux for synthesis of diverse commodity chemicals [13]. In future versions of MFlux, new metabolic knowledge and rules should be applied for flux corrections.

Third, it is still difficult to incorporate regulation mechanisms into the current model due to insufficient  $^{13}\text{C}$ -MFA studies. For instance, various catabolite repression mechanisms regulate the cell fluxome in the presence of multiple substrates (e.g., glucose shows catabolite repression for fast growing *E. coli* when both glucose and glycerol are available, Figure 10) [64]. These

hierarchy regulations among substrate utilization can be dependent on growth rates or can differ among microbial species (*E. coli*, *Bacillus*, and *Corynebacterium*).

Fourth, when oxygen is not available, fast sugar utilization will activate mixed acid fermentation (e.g., by utilizing lactate dehydrogenase and pyruvate formate lyase) to produce complicated overflow metabolites. This mechanism is also furnished in yeast and mammalian cells. However,  $^{13}\text{C}$ -MFA studies on anaerobic metabolisms are much less frequent than on aerobic metabolisms. MFlux cannot predict the complicated patterns of overflow fluxes at this stage.

Lastly, ML cannot directly estimate fluxes for carbon sources which are not part of the learning dataset. To predict fluxomes for new substrates, users need to make assumption that similar entry-points of carbon sources into the central metabolic network may cause similar flux distributions (e.g., sucrose has to be treated as a combination of glucose and fructose).

## 6.6. Conclusion

This proof-of-principle study demonstrates that AI methods can facilitate fluxomics research with reasonable precision.  $^{13}\text{C}$ -MFA is a very small field: There are just hundreds of MFA research papers on microbial species published in the past two decades. In the long term, ML methods may solve this problem: With a large and reliable fluxomics dataset and more information from  $^{13}\text{C}$ -MFA and AI scientists, the future model can make broad-scope metabolism predictions. To sum up, MFlux presents the first platform that incorporates machine learning, constraint programming, and quadratic programming in the field of fluxomics. It will inspire the development of similar computational tools to advance omics and metabolic engineering fields [47,65].

## 6.7. Supporting information

**Appendix II S1** MFlux Computer Program (Source codes).

**Appendix II S2** Results of 20 case studies: Detailed information for 20 cases studies using MFlux, including literature references, input conditions,  $^{13}\text{C}$ -MFA fluxes, the flux profiles predicted by only Machine Learning, and the flux profiles predicted by MFlux with additional constraints.

## 6.8. References

1. Winter G, Krömer JO (2013) Fluxomics – connecting ‘omics’ analysis and phenotypes. *Environmental Microbiology* 15: 1901-1916.
2. Chen X, Alonso AP, Allen DK, Reed JL, Shachar-Hill Y (2011) Synergy between  $^{13}\text{C}$ -metabolic flux analysis and flux balance analysis for understanding metabolic adaption to anaerobiosis in *E. coli*. *Metabolic Engineering* 13: 38-48.
3. Tang YJ, Chakraborty R, Martín HG, Chu J, Hazen TC, *et al.* (2007) Flux Analysis of Central Metabolic Pathways in *Geobacter metallireducens* during Reduction of Soluble Fe(III)-Nitrilotriacetic Acid. *Applied And Environmental Microbiology* 73: 3859-3864.
4. Tang JK-H, You L, Blankenship RE, Tang YJ (2012) Recent advances in mapping environmental microbial metabolisms through  $^{13}\text{C}$  isotopic fingerprints. *Journal of the Royal Society Interface* 9: 2767-2780.
5. Yim H, Haselbeck R, Niu W, Pujol-Baxley C, Burgard A, *et al.* (2011) Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. *Nature Chemical Biology* 7: 445-452.

6. Becker J, Zelder O, Häfner S, Schröder H, Wittmann C (2011) From zero to hero—Design-based systems metabolic engineering of *Corynebacterium glutamicum* for l-lysine production. *Metabolic Engineering* 13: 159-168.
7. He L, Xiao Y, Gebreselassie N, Zhang F, Antoniewicz MR, *et al.* (2014) Central metabolic responses to the overproduction of fatty acids in *Escherichia coli* based on  $^{13}\text{C}$ -metabolic flux analysis. *Biotechnology And Bioengineering* 111: 575-585.
8. Antoniewicz MR, Kelleher JK, Stephanopoulos G (2007) Elementary metabolite units (EMU): A novel framework for modeling isotopic distributions. *Metabolic Engineering* 9: 68-86.
9. Weitzel M, Noh K, Dalman T, Niedenfuhr S, Stute B, *et al.* (2012) 13CFLUX2 - high-performance software suite for  $^{13}\text{C}$ -metabolic flux analysis. *Bioinformatics*.
10. Quek L-E, Wittmann C, Nielsen L, Kromer J (2009) OpenFLUX: efficient modelling software for  $^{13}\text{C}$ -based metabolic flux analysis. *Microbial Cell Factories* 8:25.
11. Zamboni N, Fischer E, Sauer U (2005) FiatFlux - a software for metabolic flux analysis from  $^{13}\text{C}$ -glucose experiments. *BMC Bioinformatics* 6: 209.
12. Crown SB, Long CP, Antoniewicz MR (2015) Integrated  $^{13}\text{C}$ -metabolic flux analysis of 14 parallel labeling experiments in *Escherichia coli*. *Metabolic Engineering* 28: 151-158.
13. Antoniewicz MR, Kraynie DF, Laffend LA, González-Lergier J, Kelleher JK, *et al.* (2007) Metabolic flux analysis in a nonstationary system: Fed-batch fermentation of a high yielding strain of *E. coli* producing 1,3-propanediol. *Metabolic Engineering* 9: 277-292.
14. Nöh K, Grönke K, Luo B, Takors R, Oldiges M, *et al.* (2007) Metabolic flux analysis at ultra short time scale: Isotopically non-stationary  $^{13}\text{C}$  labeling experiments. *Journal of Biotechnology* 129: 249-267.

15. Tang YJ, Martin HG, Myers S, Rodriguez S, Baidoo EEK, *et al.* (2009) Advances in analysis of microbial metabolic fluxes via  $^{13}\text{C}$  isotopic labeling. *Mass Spectrometry Reviews* 28: 362-375.
16. Zhuang W-Q, Yi S, Bill M, Brisson VL, Feng X, *et al.* (2014) Incomplete Wood–Ljungdahl pathway facilitates one-carbon metabolism in organohalide-respiring *Dehalococcoides mccartyi*. *Proceedings of the National Academy of Sciences* 111: 6419-6424.
17. Tarca AL, Carey VJ, Chen X-W, Romero R, Drăghici S (2007) Machine Learning and Its Applications to Biology. *PLoS Computational Biology* 3: e116.
18. Kell DB (2006) Metabolomics, modelling and machine learning in systems biology – towards an understanding of the languages of cells. *FEBS Journal* 273: 873-894.
19. Dale J, Popescu L, Karp P (2010) Machine learning methods for metabolic pathway prediction. *BMC Bioinformatics* 11: 15.
20. Räscher G, Sonnenburg S, Srinivasan J, Witte H, Müller K-R, *et al.* (2007) Improving the *Caenorhabditis elegans* Genome Annotation Using Machine Learning. *PLoS Computational Biology* 3: e20.
21. Ye Q-H, Qin L-X, Forgues M, He P, Kim JW, *et al.* (2003) Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. *Nature Medicine* 9: 416-423.
22. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, *et al.* (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine* 8: 68-74.

23. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, *et al.* (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16: 906-914.
24. Supek F, Peharec P, Krsnik-Rasol M, Šmuc T (2008) Enhanced analytical power of SDS-PAGE using machine learning algorithms. *Proteomics* 8: 28-31.
25. Mahadevan S, Shah SL, Marrie TJ, Slupsky CM (2008) Analysis of Metabolomic Data Using Support Vector Machines. *Analytical Chemistry* 80: 7562-7570.
26. Zhang Z, Shen T, Rui B, Zhou W, Zhou X, *et al.* (2015) CeCaFDB: a curated database for the documentation, visualization and comparative analysis of central carbon metabolic flux distributions explored by <sup>13</sup>C-fluxomics. *Nucleic Acids Research*. D549-57
27. Sauer U, Eikmanns BJ (2005) The PEP—pyruvate—oxaloacetate node as the switch point for carbon flux distribution in bacteria. *FEMS Microbiology Reviews* 29: 765-794.
28. Tang YJ, Sapra R, Joyner D, Hazen TC, Myers S, *et al.* (2009) Analysis of metabolic pathways and fluxes in a newly discovered thermophilic and ethanol-tolerant *Geobacillus* strain. *Biotechnology and Bioengineering* 102: 1377-1386.
29. Peters-Wendisch PG, Kreutzer C, Kalinowski J, Pátek M, Sahm H, *et al.* (1998) Pyruvate carboxylase from *Corynebacterium glutamicum*: characterization, expression and inactivation of the *pyc* gene. *Microbiology* 144: 915-927.
30. Toya Y, Ishii N, Nakahigashi K, Hirasawa T, Soga T, *et al.* (2010) <sup>13</sup>C-metabolic flux analysis for batch culture of *Escherichia coli* and its *pyk* and *pgi* gene knockout mutants based on mass isotopomer distribution of intracellular metabolites. *Biotechnology Progress* 26: 975-992.

31. Pedregosa F, Ga, Varoquaux I, Gramfort A, *et al.* (2011) Scikit-learn: Machine Learning in Python. *Journal Of Machine Learning Research* 12: 2825-2830.
32. Zhao J, Baba T, Mori H, Shimizu K (2004) Effect of *zwf* gene knockout on the metabolism of *Escherichia coli* grown on glucose or acetate. *Metabolic Engineering* 6: 164-174.
33. Towell GG, Shavlik JW (1994) Knowledge-based artificial neural networks. *Artificial Intelligence* 70: 119-165.
34. Segrè D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences* 99: 15112-15117.
35. Fischer E, Zamboni N, Sauer U (2004) High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry derived  $^{13}\text{C}$  constraints. *Analytical Biochemistry* 325: 308-316.
36. Zhao J, Baba T, Mori H, Shimizu K (2004) Global metabolic response of *Escherichia coli* to *gnd* or *zwf* gene-knockout, based on  $^{13}\text{C}$ -labeling experiments and the measurement of enzyme activities. *Applied Microbiology And Biotechnology* 64: 91-98.
37. Fong SS, Nanchen A, Palsson BO, Sauer U (2006) Latent Pathway Activation and Increased Pathway Capacity Enable *Escherichia coli* Adaptation to Loss of Key Metabolic Enzymes. *Journal Of Biological Chemistry* 281: 8024-8033.
38. Peng L, Arauzo-Bravo MJ, Shimizu K (2004) Metabolic flux analysis for a *ppc* mutant *Escherichia coli* based on  $^{13}\text{C}$ -labelling experiments together with enzyme activity assays and intracellular metabolite measurements. *FEMS Microbiology Letters* 235: 17-23.



39. Chubukov V, Uhr M, Le Chat L, Kleijn RJ, Jules M, *et al.* (2013) Transcriptional regulation is insufficient to explain substrate-induced flux changes in *Bacillus subtilis*. *Molecular Systems Biology* 9: 709.
40. Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? *Nature Biotechnology* 28: 245-248.
41. Orth JD, Conrad T, Na J, Lerman J, Nam H, *et al.* (2011) A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism-2011. *Molecular Systems Biology* 7: 535.
42. Wu S, He L, Wang Q, Tang Y (2015) An ancient Chinese wisdom for metabolic engineering: Yin-Yang. *Microbial Cell Factories* 14: 39.
43. Stelling J, Sauer U, Szallasi Z, Doyle Iii FJ, Doyle J (2004) Robustness of Cellular Functions. *Cell* 118: 675-685
44. Fischer E, Sauer U (2005) Large-scale *in vivo* flux analysis shows rigidity and suboptimal performance of *Bacillus subtilis* metabolism. *Nature Genetics* 37: 636-640.
45. Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Molecular Systems Biology* 3:119.
46. Tang YJ, Martin HG, Deutschbauer A, Feng X, Huang R, *et al.* (2009) Invariability of central metabolic flux distribution in *Shewanella oneidensis* MR-1 under environmental or genetic perturbations. *Biotechnology Progress* 25: 1254-1259.
47. Stephanopoulos G (1999) Metabolic Fluxes and Metabolic Engineering. *Metabolic Engineering* 1: 1-11.
48. Stephanopoulos G, Vallino J (1991) Network rigidity and metabolic engineering in metabolite overproduction. *Science* 252: 1675-1681.

49. Lien S, Niedenfuhr S, Sletta H, Noh K, Bruheim P (2015) Fluxome study of *Pseudomonas fluorescens* reveals major reorganisation of carbon flux through central metabolic pathways in response to inactivation of the anti-sigma factor MucA. *BMC Systems Biology* 9: 6.
50. Fuhrer T, Fischer E, Sauer U (2005) Experimental Identification and Quantification of Glucose Metabolism in Seven Bacterial Species. *Journal of Bacteriology* 187: 1581-1590.
51. Wierckx N, Ruijsenaars HJ, de Winde JH, Schmid A, Blank LM (2009) Metabolic flux analysis of a phenol producing mutant of *Pseudomonas putida* S12: Verification and complementation of hypotheses derived from transcriptomics. *Journal of Biotechnology* 143: 124-129.
52. del Castillo T, Ramos JL, Rodríguez-Herva JJ, Fuhrer T, Sauer U, *et al.* (2007) Convergent Peripheral Pathways Catalyze Initial Glucose Catabolism in *Pseudomonas putida*: Genomic and Flux Analysis. *Journal of Bacteriology* 189: 5142-5152.
53. Blank LM, Ionidis G, Ebert BE, Bühler B, Schmid A (2008) Metabolic response of *Pseudomonas putida* during redox biocatalysis in the presence of a second octanol phase. *FEBS Journal* 275: 5173-5190.
54. Conway T (1992) The Entner-Doudoroff pathway: history, physiology and molecular biology. *FEMS Microbiology Letters* 103: 1-27.
55. Bar-Even A, Flamholz A, Noor E, Milo R (2012) Rethinking glycolysis: on the biochemical logic of metabolic pathways. *Nature Chemical Biology* 8: 509-517.
56. Flamholz A, Noor E, Bar-Even A, Liebermeister W, Milo R (2013) Glycolytic strategy as a tradeoff between energy yield and protein cost. *Proceedings of the National Academy of Sciences* 110: 10039-10044.

57. Berger A, Dohnt K, Tielen P, Jahn D, Becker J, *et al.* (2014) Robustness and Plasticity of Metabolic Pathway Flux among Uropathogenic Isolates of *Pseudomonas aeruginosa*. PLoS ONE 9: e88368.
58. Ebert BE, Kurth F, Grund M, Blank LM, Schmid A (2011) Response of *Pseudomonas putida* KT2440 to Increased NADH and ATP Demand. Applied And Environmental Microbiology 77: 6597-6605.
59. Yao R, Hirose Y, Sarkar D, Nakahigashi K, Ye Q, *et al.* (2011) Catabolic regulation analysis of *Escherichia coli* and its *crp*, *mlc*, *mgsA*, *pgi* and *ptsG* mutants. Microbial Cell Factories 10: 67.
60. Ishii N, Nakahigashi K, Baba T, Robert M, Soga T, *et al.* (2007) Multiple High-Throughput Analyses Monitor the Response of *E. coli* to Perturbations. Science 316: 593-597.
61. Tannler S, Decasper S, Sauer U (2008) Maintenance metabolism and carbon fluxes in *Bacillus* species. Microbial Cell Factories 7: 19.
62. van Ooyen J, Noack S, Bott M, Reth A, Eggeling L (2012) Improved L-lysine production with *Corynebacterium glutamicum* and systemic insight into citrate synthase flux and activity. Biotechnology And Bioengineering 109: 2070-2081.
63. Bommareddy RR, Chen Z, Rappert S, Zeng A-P (2014) A *de novo* NADPH generation pathway for improving lysine production of *Corynebacterium glutamicum* by rational design of the coenzyme specificity of glyceraldehyde 3-phosphate dehydrogenase. Metabolic Engineering 25: 30-37.
64. Wang Z-J, Wang P, Liu Y-W, Zhang Y-M, Chu J, *et al.* (2012) Metabolic flux analysis of the central carbon metabolism of the industrial vitamin B12 producing strain *Pseudomonas*

*denitrificans* using  $^{13}\text{C}$ -labeled glucose. Journal of the Taiwan Institute of Chemical Engineers 43: 181-187.

65. Hemme C, Fields M, He Q, Deng Y, Lin L, *et al.* (2011) Correlation of genomic and physiological traits of thermoanaerobacter species with biofuel yields. Applied and Environmental Microbiology 77: 7998 - 8008.

66. Tang Y, Pingitore F, Mukhopadhyay A, Phan R, Hazen TC, *et al.* (2007) Pathway Confirmation and Flux Analysis of Central Metabolic Pathways in *Desulfovibrio vulgaris Hildenborough* using Gas Chromatography-Mass Spectrometry and Fourier Transform-Ion Cyclotron Resonance Mass Spectrometry. Journal of Bacteriology 189: 940-949.

**Table 6.1.** Summary of 20 cases of study

Species	Carbon source	Oxygen condition	Reactor	Genetics	Case	Reference
<i>E. coli</i>	Glc	aerobic	shake tube	WT	1	(Crown <i>et al.</i> 2015)
<i>E. coli</i>	Glc	aerobic	shake flask	<i>ppc</i> KO	2 - 4	(Fong <i>et al.</i> 2006)
<i>B. subtilis</i>	Glc	aerobic	shake flask, CSTR	WT, <i>spo0A</i> KO	5 - 7	(Tannler <i>et al.</i> 2008)
<i>B. subtilis</i>	Multiple substrates	aerobic	shake flask	mutant	8 -11	(Chubukov <i>et al.</i> 2013)
<i>C. glutamicum</i>	Glc	aerobic	shake flask	WT	12	(van Ooyen <i>et al.</i> 2012)
<i>C. glutamicum</i>	Glc	aerobic	shake flask	mutant	13	(Bommareddy <i>et al.</i> 2014)
<i>P. denitrificans</i>	Glc	aerobic, microaerobic	fermentor	WT	14, 15	(Wang <i>et al.</i> 2012)
<i>G. thermoglucosidasius</i>	Glc	microaerobic	shake flask	WT	16	(Tang <i>et al.</i> 2009c)
<i>Thermoanaerobacter sp.</i>	Xyl	anaerobic	Sealed bottle	WT	17, 18	(Hemme <i>et al.</i> 2011)
<i>D. vulgaris</i>	Lac	anaerobic	Sealed bottle	WT	19	(Tang <i>et al.</i> 2007b)
<i>G. metallireducens</i>	Ace	anaerobic	Sealed bottle	WT	20	(Tang <i>et al.</i> 2007a)

Table notes: Glc: glucose, Xyl: xylose, Lac: lactate, Ace: acetate, KO: knockout

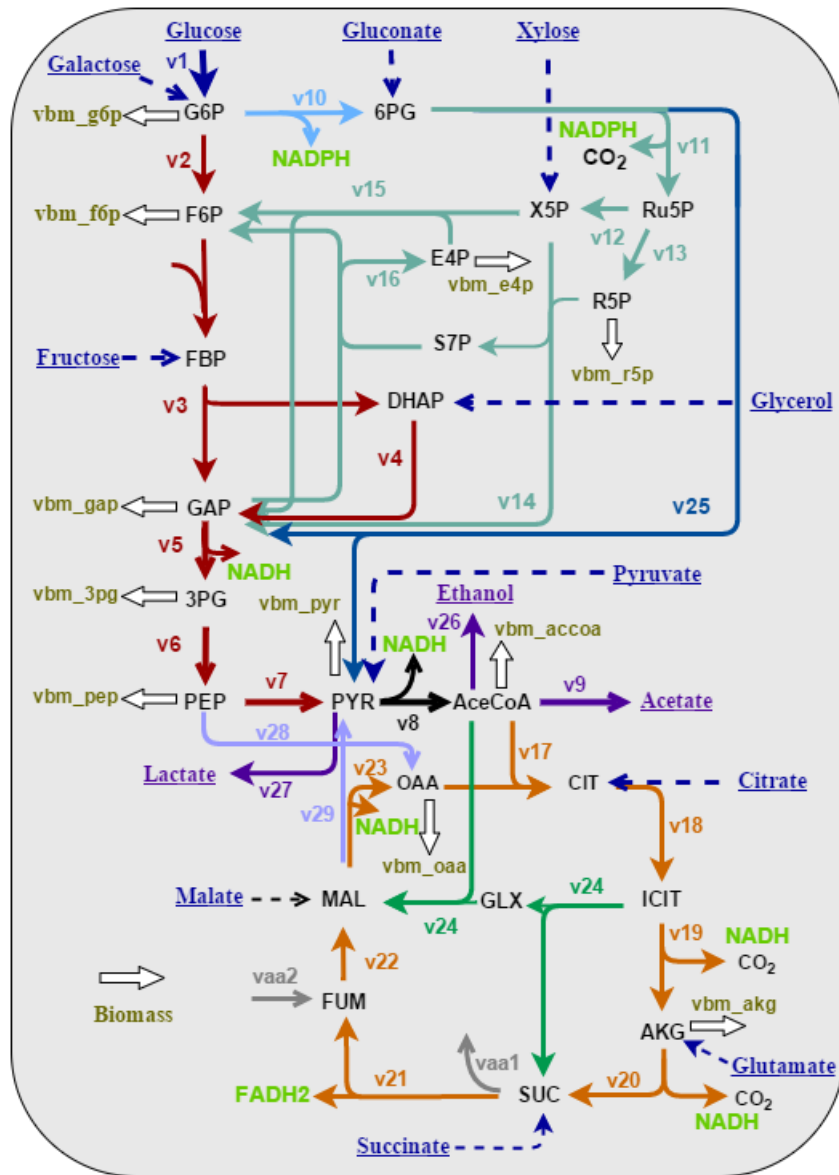


Figure 6.1: A universal central metabolic pathway for bacteria: The central carbon metabolic pathway is simplified into 29 fluxes in MFlux.

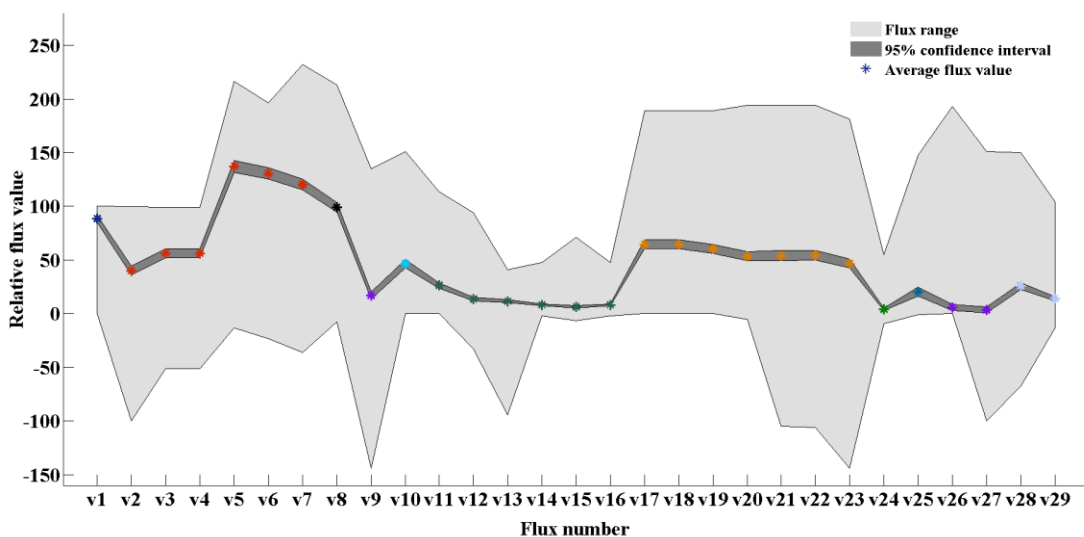


Figure 6.2. Statistical analysis of central metabolic fluxes collected in our database. “Flux range” represents variations of each fluxes among  $^{13}\text{C}$ -MFA database; “95% confidence interval” represents 95% of flux data were within a small range; “Average flux value” are the mean of flux values from  $^{13}\text{C}$ -MFA database.

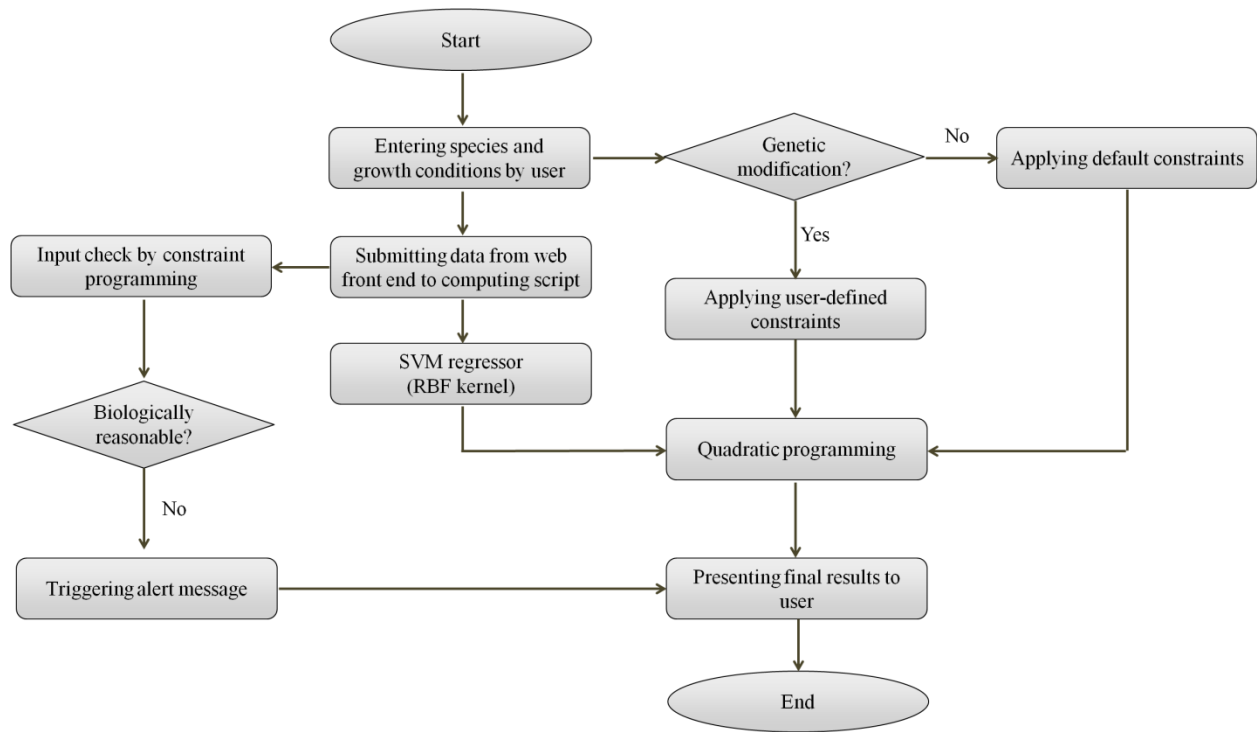


Figure 6.3. Flow chart of MFlux algorithm. This diagram is to illustrate the detailed procedures for our algorithm.



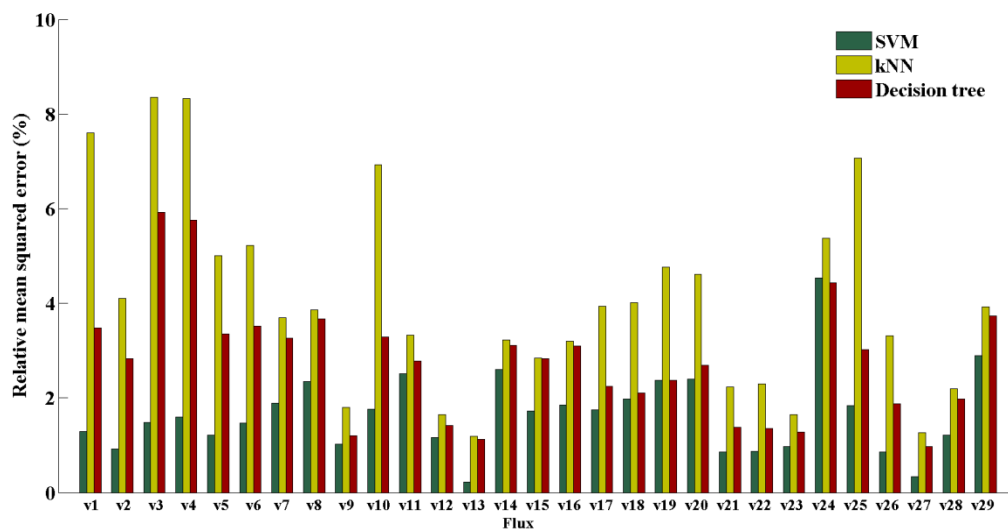


Figure 6.4. A comparison of three different algorithms: SVM, kNN, and decision tree: The best cross-validation results on 29 fluxes are compared. All tests in this step were performed only on the WT database.

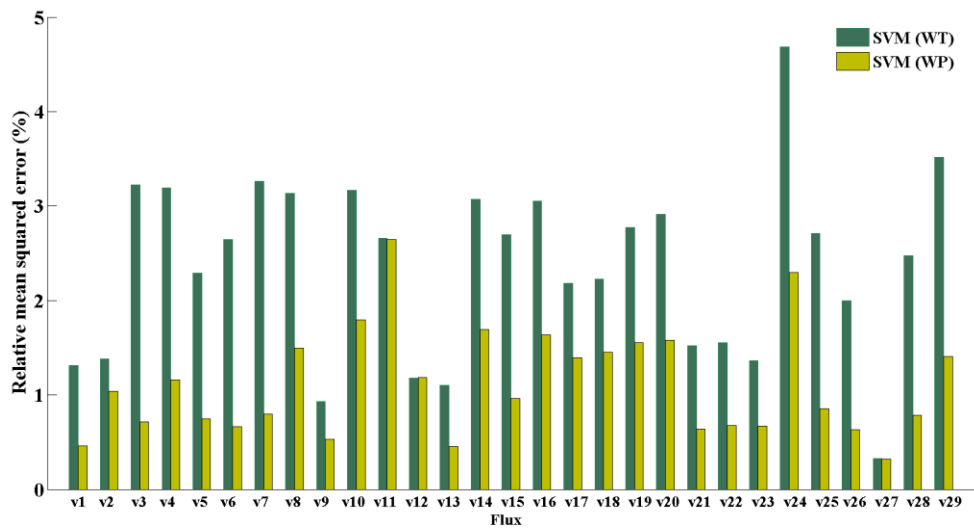


Figure 6.5. Best results by SVM for WT and WP databases. Both the linear and the RBF kernels are considered in a grid search, and the results from WP database is much better than from the WT database

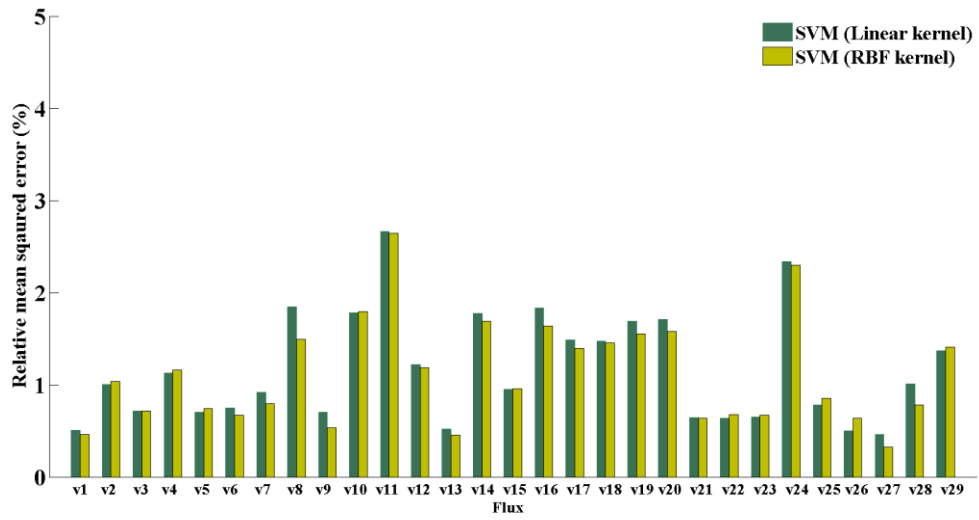


Figure 6.6. A comparison between the linear kernel and the RBF kernel for SVM. The results are quite similar.

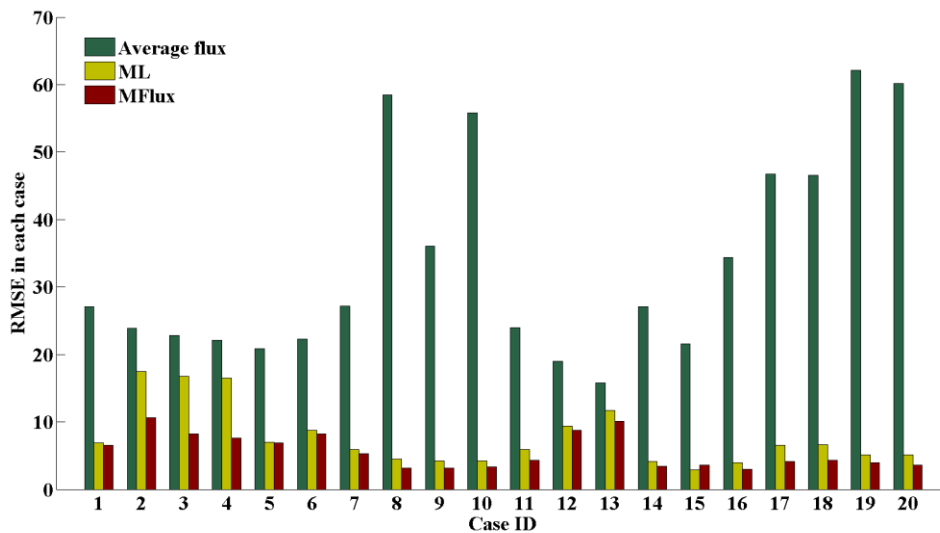


Figure 6.7. Summary of root mean squared error (RMSE) from 20 case studies: averaged flux from  $^{13}\text{C}$ -MFA database; machine learning, and MFlux. The average RMSE is 7.7 from machine

learning alone and 5.6 from MFlux. The RMSE is calculated by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{29} (v_i - \hat{v}_i)^2}{29}}$$

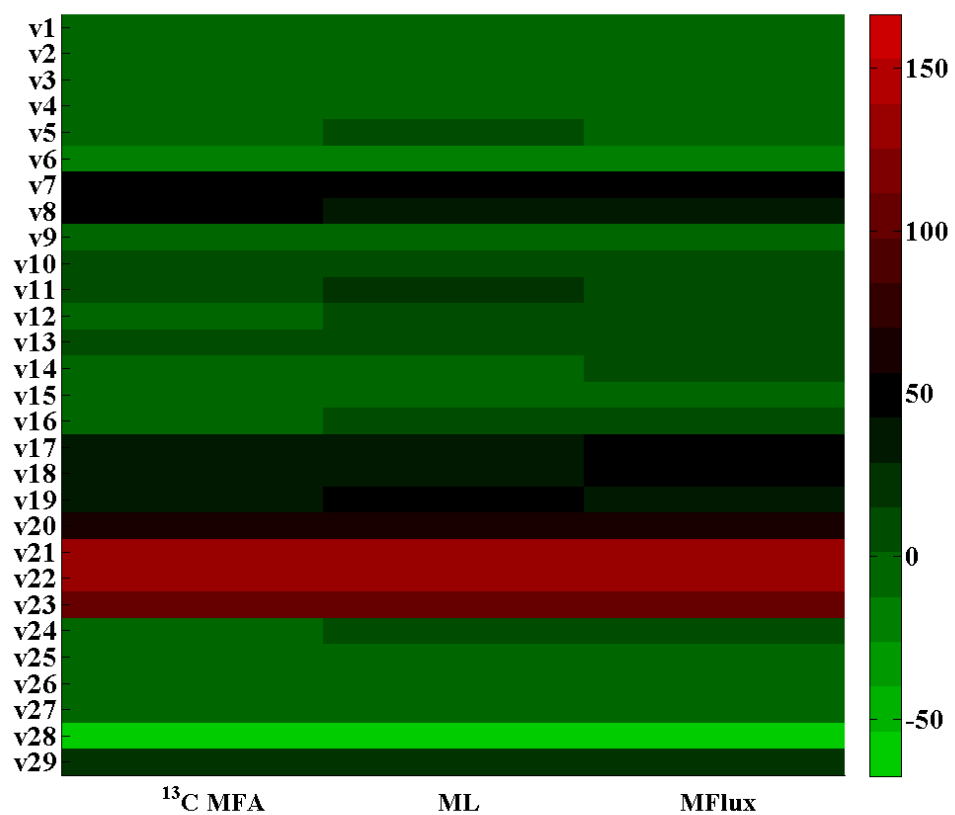


Figure 6.8. A comparison of  $^{13}\text{C}$ -MFA, the flux predicted by ML, and the flux predicted by MFlux in case 8. *B. subtilis* was incubated in a shake flask (37 °C, 300 rpm, aerobic condition), and supplied with labeled succinate and glutamate as carbon sources in M9 minimal medium. The detailed information is in supporting file 2.

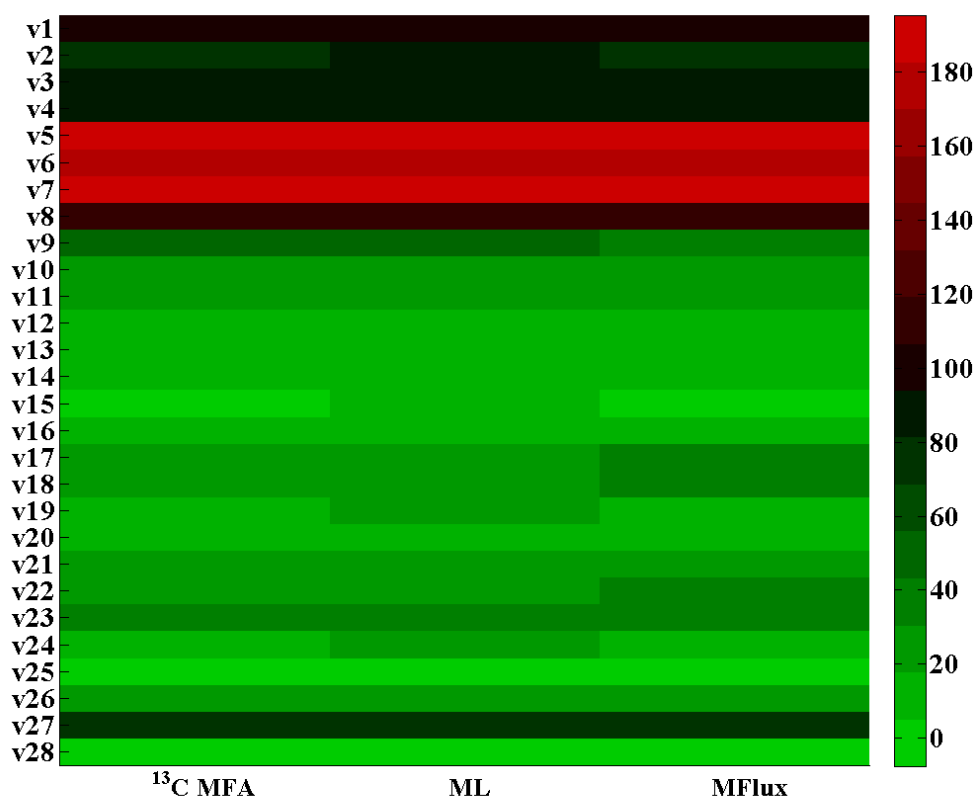


Figure 6.9. A comparison of  $^{13}\text{C}$ -MFA, the flux predicted by ML flux, and the flux predicted by MFlux. *G. thermoglucosidasius* M10EXG was incubated in sealed bottles (micro-aerobic condition), supplied with glucose as a carbon source.  $\text{RMSE}_{\text{ML}} = 4.0$ ,  $\text{RMSE}_{\text{MFlux}} = 3.0$ . The detailed information is in supporting file 2.

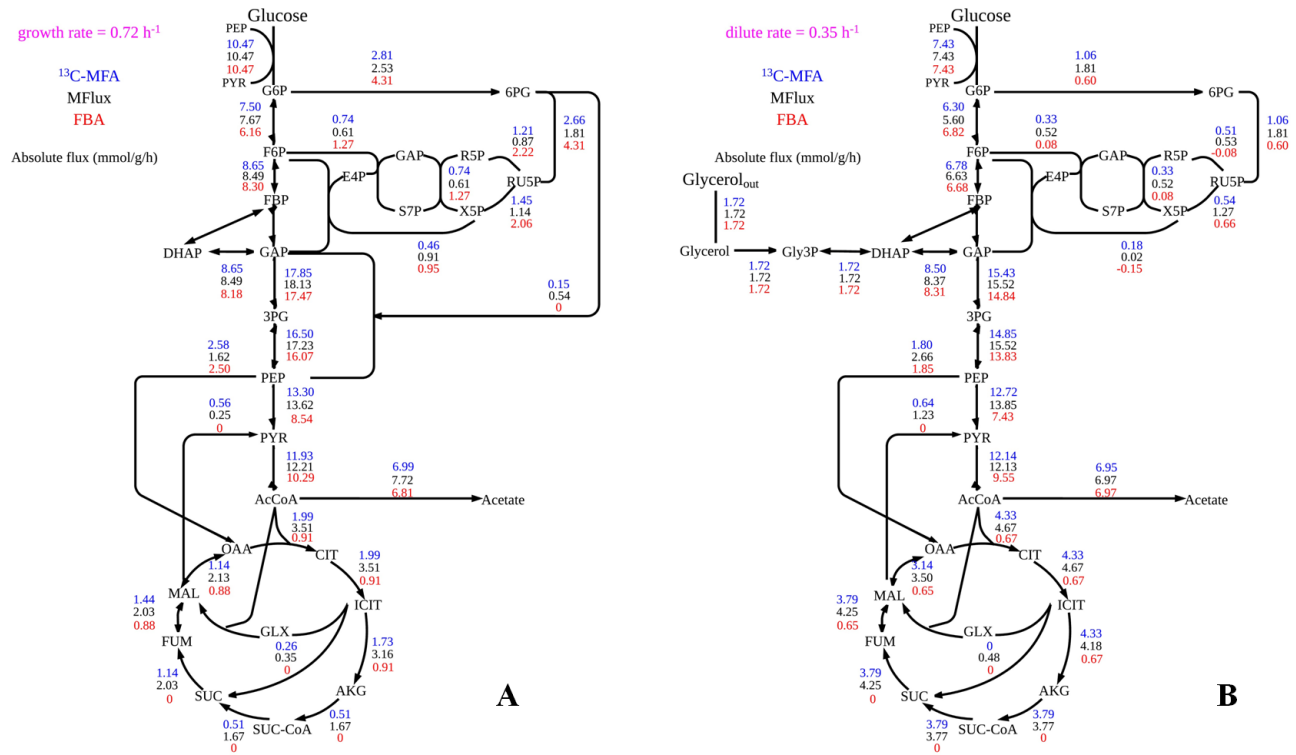


Figure 6.10. A comparison of <sup>13</sup>C-MFA, MFlux and the flux predicted by FBA. FBA Analysis is simulated by *E. coli* iJO1366 model with defaulted boundary settings from the reference (Orth *et al.* 2011). (A) *E. coli* fluxome of glucose metabolism was precisely measured via parallel labeling experiments (a recent paper not in our database) (Crown *et al.* 2015).  $RMSE_{FBA} = 11.3$ ,  $RMSE_{MFlux} = 6.5$ . (B) *E. coli* fluxome of glycerol and glucose co-metabolism were measured by Dr. Yao and Dr. Shimizu (unpublished data). *E. coli* strain was cultured in chemostat fermentor with a working volume of 1 L (37°C). The dilution rates in the continuous culture were 0.35 h<sup>-1</sup>. [1-<sup>13</sup>C] glucose and [1,3-<sup>13</sup>C] glycerol were used for tracer experiments. The flux calculation is based on previous method (Fong *et al.* 2006; Peng *et al.* 2004).  $RMSE_{FBA} = 22.5$ ,  $RMSE_{MFlux} = 5.1$

## CHAPTER SEVEN

# ENABLE FAST LITERATURE ANALYSIS BY TEXT MINING AND BIG DATA TECHNOLOGY

### 7.1. Abstract

Information acquisition by human being is severely limited by the speed and time of reading, as well as the background of readers. To gain a deep understanding on a specific research subject requires even longer time of training and learning. With information gradually digitalized, text mining method provides an automatic way for literature analysis. In this study, we built up a workflow integrating text mining and the emerging Big Data technology for fast literature analysis. We also performed several case studies to demonstrate its functionality. A comparison between ‘metabolic engineering’ and ‘synthetic biology’ finds that energy metabolism terms (i.e., ‘NADH’, ‘NADPH’, ‘ATP’) and non-glucose carbon substrate (i.e., ‘xylose’, ‘acetate’, ‘glycerol’) have significant higher frequency for ‘metabolic engineering’, while researchers on ‘synthetic biology’ talk more on regulatory modules such as ‘circuit’, ‘loop’, ‘switch’, ‘IPTG’, ‘LacI’. In another case the transition of ‘metabolic engineering’ between ‘2000 - 2009’ and ‘2010 - 2015’ has also been identified: more focuses were put on model species such as ‘*B. subtilis*’ and ‘*P. pastoris*’ as microbial cell factories and ‘hydrogen’ during 2000 ~ 2009, while the focuses have been shifted to photosynthetic species ‘cyanobacteria’ as well as ‘butanol’, ‘isobutanol’, ‘lipid’. Each comparison takes less than three minutes and is easily extended with various specific searching settings. To sum up, this proof-of-principle study demonstrates that Big Data technique can quickly capture similarity/difference and provide quantitative information,



which enable us have more reflections over the past and make more reasonable choices in the future.

## **7.2. Introduction**

As a cornerstone of human society, technology development is accompanied with emerging fields and changing focuses in different field. Understanding those similarities and differences between various fields, as well as the same field of different time periods will not only bring novel insights to researchers, but also provide invaluable historical experiences for a broader audience (e.g., industry, government). To gain a deep understanding over a specific subject, information from Wikipedia is far away from sufficient; extensive reading over thousands of papers is a perquisite, which is very time-consuming yet difficult to avoid bias.

The trend of information digitalization was emerged with wide adoptions of personal computer and Internet. (Arms 2000) It has revolutionized the manner of information record and storage in human civilization: tons of information can be stored in small hard disks and can be easily duplicated and spread. Based on the increasing availability of digital information, text mining provides an automatic approach for information analysis (Dorre *et al.* 1999). Text mining have been widely adopted in many fields such as biomedical literature analysis (Cohen and Hersh 2005), systems biology (Ananiadou *et al.* 2010), and human phenome (van Driel *et al.* 2006).

As the size of data and information is increasing with time, the concept of ‘Big Data’ also comes out several years ago. Correspondingly, analyzing large amount of data comes across the limitation of hardware. For instance, searching a specific term from 60 GB of data in txt formate

takes about 10 hours in a powerful working station, which is unacceptable if similar search actions are carried out all the time. Two major approaches have been developed by Google, to deal with those challenges: the first one is MapReduce, which is suitable for batch processing of large datasets; the other one is BigQuery, which works well with interactive analysis over large datasets (Tigani and Naidu 2014).

In this study, we propose to build up a workflow which provides rapid literature analysis based on an integrated platform of text mining and BigQuery. Development of such a workflow will bring a novel approach for fast acquisition of professional knowledge.

### **7.3. Methods**

#### **7.3.1. Database availability and record structure**

All full-text papers are downloaded from NCBI PMC database (<http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/>). PMC database contains ~1.1 million full-text papers (without images) from ~5600 different journals with a focus on the biomedical research. This database is updated weekly; hence, the total number of papers and journals are still increasing (shown in Figure 7.1.).

There are two basic formats of papers in PMC database: .txt type and .nxml type. Txt files have to lose lots of essential contents of original papers, due to their format limitation. In contrast, nxml files contain most important information of papers except images. Therefore, we choose nxml files for further literature analysis.

Nxml files contain structured information extracted from papers. In general, there are three parts for each nxml file as shown in Figure 7.2.:

- (1) Front: which contains information such as paper title, pmcid, author name, author email, and author affiliation.
- (2) Body: the major part of the manuscript, including the section of methods, results, discussions, supporting information.
- (3) Back: list of references

Most information illustrating research topics of papers normally is included in the part of 'body'.

### **7.3.2. Text mining methods**

To extract information from each paper (nxml file), a powerful language processing toolbox is indispensable. The most common Natural Language Processing libraries includes NLTK (in Python) (Bird 2006; Bird *et al.* 2009), Stanford NLP Toolbox (in Java), OpenNLP (in Java), Gate (in Java). Different libraries have their advantages and disadvantages; for nxml format files, BeautifulSoup library (in Python) can work well with them. Therefore, we employed BeautifulSoup and NLTK library together in this study. All programming language is Python 2.7.

To demonstrate the functionality of our workflow, we only extract PMC id, year, whole body, unique words in body, and frequency of each word in this work. A simple procedure of text mining is listed below:

1. A paper in nxml format is read in by BeautifulSoup library
2. The PMC id, article name, as well as the body, are extracted by using BeautifulSoup library with following codes:

```
pmc_id = soup.find_all('article-id', attrs={"pub-id-type": "pmc"})
```

```

body_raw = soup.find_all('body');
front_raw = soup.find_all('front');
back_raw = soup.find_all('back');

body_text = body_raw[0].get_text();
front_text = front_raw[0].get_text();
back_text = back_raw[0].get_text();
real_body =
(body_text.replace(back_text, '').replace(front_text, '').encode(
'ascii', 'ignore')).lower();

```

To our surprise, the part of body extracted by Beautiful Soup function still contains the front and the back part. Hence, we artificially remove strings of both parts by using ‘replace’ function.

3. We remove all symbols in the body and tokenize body text by the following code:

```

real_text = ((((((real_body.replace("\n", " ").replace(',',' '
')).replace('.', ' ')).replace('(',' ' ')).replace(')',' '
')).replace(':', ' ')).replace(';',' '));

paper_word = nltk.word_tokenize(real_text);

```

4. We employ the library of common English words (stopwords) such as ‘I’, ‘and’, and add a few more. The final library of common words is about 200. Since those words do not have any exact meanings related with research topics, we can remove them from the text. The total size of text strings can also be reduced to about 55% of their original size through step 2 - 4.

5. We convert all words to be lowercase, and then extract all unique words and their respective relative frequency. Considering of significant difference in text size of different papers (the average word number for papers is about 5,800; a technical note can be as short as 1,500 words; while a reviewer paper can be as along as 14,000 words), to equalize the impact of each paper, we use relative frequency rather than absolute frequency in our analysis.

6. We store the following information in a CSV file with five columns:

PMC id, year, body, unique word, relative word frequency

### 7.3.3. Fast search via BigQuery

The final CSV file generated from text mining is around 21.7 GB, about 30% volume of their original files. This file contains more than 1 million records, searching a single term takes at least several hours using traditional methods. To accelerate this process, we employed Google BigQuery for the search process. BigQuery is the commercialized product of Dremel, the search engine for Google insiders. (Tigani and Naidu 2014) Equipped with columnar structure data, BigQuery performs a searching task in parallel (on several thousands of servers), and can finish a searching task over several billion of records within ten seconds.

First, we need create a bucket in Google Cloud and upload the large CSV file into the bucket. A stable internet connection is necessary because the uploading process may take several hours.

Second, we create a project under BigQuery and create a Big Table under this project. During table creation process, we need to get the link address of the big CSV file in Google Cloud and use for data uploading. Also, we need to define the names of columns, as well as their type and mode in the step of 'Edit Schema'.

Third, to enable script based Google BigQuery search, we need to download a secret key 'client\_secret.json' from Google Big Query. This file stores a personalized keyword linked with your Google client account. Thus, each BigQuery task can directly charge your account via this information. After that, we can download the demo code program (in Python) from BigQuery and perform the test in local PC. Once client verification is finished in local PC, we can modify the searching program to perform any search tasks.

#### **7.3.4. Data analysis**

For a specific term, we need to input its lowercase string in Python program for BigQuery search. Information of any records contained the specific term will be recorded locally upon searching

request. This local record can be processed through Python program and converted into cumulative relative frequencies of unique words. Through simple sorting over cumulative relative frequencies, word list with a descending order is the output. For more specified search, for instance, search ‘metabolic engineering’ during 2010 ~ 2015, we can modify the program and include more searching parameters in BigQuery.

To visualize top 300 related words of a specific searching term, we employ the library of word cloud in R to display them. For simple quantification, we define the word of highest cumulative relative frequency as 100. Frequencies of other words are normalized in a scale of 100.

To identify the similarity and the difference between two different search term, we perform a simple match between top 500 words for different terms, to determine the percentage of similarity and difference. We also extract those words of difference as an output.

## **7.4. Results and Discussions**

To demonstrate the functionality of our workflow, we carried out several case studies and put the results as below.

### **7.4.1. Most related words of ‘metabolic engineering’, ‘environmental engineering’, ‘synthetic biology’, ‘systems biology’, and ‘metabolic flux’**

The first function of this Big Data workflow is to define related words of a specific term. To run this function, we first search this term within the column ‘body’ via BigQuery. After word list of top cumulative frequency is given, we use R to display their word cloud.

We performed case studies on several terms including ‘metabolic engineering’, ‘environmental engineering’, ‘synthetic biology’, ‘systems biology’, and ‘metabolic flux’, and the results of their word cloud are listed as Figure 7.3. – 7.7. The results provide us much meaningful information:

For instance, the word with highest frequency related with ‘metabolic engineering’ is ‘production’, which reflects the aim of ‘metabolic engineering’ field is to develop production processes through metabolism of microbial cell factories (Stephanopoulos 1999).

In another instance, the most frequent words related with ‘environmental engineering’ is ‘health’, which reveals the motivation of researches on environmental engineering is the health of human beings.

#### **7.4.2. Compare the difference and similarity between two different terms**

The second function we want to demonstrate is the comparison of two different terms. This time, we take ‘metabolic engineering’ and ‘synthetic biology’ as a case study, because researchers have different opinions over those two concepts (Nielsen and Keasling 2011; Carothers *et al.* 2009; Lee *et al.* 2008; Church *et al.* 2014). Through this Big Data workflow, we can provide a comparative analysis based on published papers in PMC database.

The result is shown in Table 7.1.: they have a similarity of 69.6% -- indicating that both terms have a large range of scope overlapped, such as ‘gene’, ‘PCR’, ‘*E. coli*’. The difference between ‘metabolic engineering’ and ‘synthetic biology’ is also apparent: The word of highest frequency for ‘metabolic engineering’ is ‘production’; and there are several words with significantly higher cumulative frequency related with ‘metabolic engineering’, including FBA, NADH, NADPH, ATP, xylose, acetate, glycerol, HPLC, transport, and tolerance. For ‘synthetic

biology’, the most frequent word is ‘gene’, words with significantly higher cumulative frequency are: circuit, loop, IPTG, lacI, RBS, Switch egfp, Phage, and virus. To explain this difference, we can refer to their background and origins: ‘metabolic engineering’ comes from biochemical engineering at the early 1990s (Bailey 1991). With the development of genetic modification tools (e.g., PCR, restriction enzymes), as well as successful commercialization of heterogeneous protein expression (e.g., insulin by Genentech), researchers tried to expand the product scope of biochemical engineering through extensive genetic modifications over microbial metabolism. Thereby, metabolic engineering focus more on energy metabolism (NADH, NADPH, and ATP), substrate utilization (xylose, acetate, and glycerol), engineering and model (transport and FBA), product measurement (HPLC), and microbial physiology (tolerance) (Stephanopoulos *et al.* 1998b). In contrast, ‘synthetic biology’ first appeared in early the 2000s, coming from the background of biophysics and electric engineering. Researchers tried to redefine biological modules via the standards and rules applied in electrical circuit and chips. Successful demonstration of simple logic parts such as ‘toggle switch’ and ‘oscillator’ in biological systems motivated further work in a similar manner (Gardner *et al.* 2000; Elowitz and Leibler 2000). With a broad spread of iGEM (International Genetically Engineered Machine) (Smolke 2009; Brown 2007), ‘synthetic biology’ has been widely recognized. And it focuses more on electrical engineering concepts (circuit, switch, and loop), and genetic demonstration tools (IPTG, lacI, RBS, egfp, Phage, and virus)(Canton *et al.* 2008).

#### **7.4.3. Identify the developing trend of a specific term**

Another function we want to show is the comparison with a time specification. This function is quite similar to the second function, except that we need to specify a period during BigQuery



search. We perform two case studies here; one is ‘metabolic engineering’, and the other is ‘biofuel’, and the results are presented in Table 7.2 and 7.3.

Case study on ‘metabolic engineering’:

From Table 7.2., we can see two obvious trends in metabolic engineering from 2000~ to 2010~. The first trend is that the focus of microbial working horse has been changed to from ‘*B. subtilis*’ and ‘*P. Pastoris*’ to ‘cyanobacteria’ and ‘*C. glutamicum*’: *B. subtilis* is a gram-positive model species widely used to produce vitamin and enzymes, however, the disadvantages for *B. subtilis* as a host include high maintenance energy (Tannler *et al.* 2008), existence of many proteases (Zhang *et al.* 2005). *P. Pastoris* is a methylotrophic yeast species widely used for recombinant protein expression in industry because of its powerful secretion system and ease to reach high cell density and to avoid contamination in large-scale fermentation (with methanol) (Cregg *et al.* 2009). However, the disadvantage for *P. Pastoris* also lies in its methylotrophic property: methanol is volatile and highly toxic; it is highly risky for students with little experience to work on this strain. Researchers turned to species ease to manipulate such as cyanobacteria and *C. glutamicum*’ during recent years. The second trend is that the list of hot products has been slightly changed: publications on hydrogen, PHB, and lysine go down, and lipid, butanol, and isobutanol gain more focus. Biosynthesis of PHB and lysine has been successfully commercialized, and most research work supported by the industrial funding will not send for publication. The bottleneck of hydrogen is storage, rather than synthesis. Novel biofuels (butanol, isobutanol, and lipid) with higher energy density than ethanol and ease-to-storage and utilization have gained attentions from both scientific and industrial fields during recent years. With funding pouring into those topics, the outcomes – the number of related papers increased.

Case study on 'biofuel':

The development trend of 'biofuel' is fascinating, the trend have changed significantly from 'Lignocellulosic', 'hemicellulose', 'Emissions', 'policy', 'Economic', 'market', 'Feedstock' from 2000 to 2009, to 'Glycerol', 'acetate', 'xylan', 'Algae', 'microalgae', 'Cyanobacteria', 'Lipid', 'Chromosome', 'cDNA' within recent years. Lots of researches were focusing on degradation and pretreatment of cellulosic material (e.g., lignocelluloses, hemicellulose) to sugar during 2000 to 2009, also there are lots of comments and perspectives on economy/market analysis, and carbon neutral economy. With the support of DOE (Department of Energy) on three energy centers (2007-currently), as well as production of biofuels through metabolic engineering (Atsumi *et al.* 2008a; Atsumi *et al.* 2009; Steen *et al.* 2010), the trend has been shifted to microbial substrates utilization (Glycerol, acetate, and xylan), and photosynthetic hosts (microalgae and cyanobacteria).

#### **7.4.4. Advantages and Limitations of Big Data workflow**

Current workflow provides fast, reproducible, and quantitative analysis over literature database. The whole process takes less than 3 min from search to the comparison, and can be easily modified and extended with the addition of more sophisticated functions through programming. For instance, we can easily extract the author information or the institute information in data mining process and store in Big Table. Moreover, such information can be used to track publication records related with specific authors or specific institutes.

The major limitation of the current workflow is the information available in the database. Although PMC is the largest full-text database online, the papers it includes are only 1.1 million currently. Considering the total number of papers available, which counts over 65 million

currently and this number is increasing now. The large gap between the resource we can access and the total number of papers available can be explained by two reasons: first, lots of old papers are still not digitized. Lots of efforts are needed to put all those information into the digital library; second, accessing lots of digital libraries are charged (Hull *et al.* 2008). There will be copyright issues, as well as conflict of interests to make digital libraries freely access now. Further, data storage and Big Query search will lead to some cost, but is relatively low (\$5/TB for either storage or query), which will not be a big issue for further development.

## **7.5. Conclusion**

We have successfully built up a literature analysis workflow based on text mining and Big Data technology. The capability of this workflow has been demonstrated through case studies and can be further enhanced by integrating with other information sources. Considering of its fast speed, reproducible results, scalability with other databases, and ease-to-modify, we believe the further development of this platform will provide deeper insights into literature as well as bringing more benefits for researchers.

## **7.6. References**

Arms, W. Y. (2000). *Digital libraries*. Boston, MIT Press

Dorre, J., Gerstl, P. and Seiffert, R. (1999). Text mining: finding nuggets in mountains of textual data, in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, San Diego, California, USA, 398-401.

Cohen, A. M. and Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Brief. Bioinform.* 6:57-71.

Ananiadou, S., Pyysalo, S., Tsujii, J. i. and Kell, D. B. (2010). Event extraction for systems biology by text mining the literature. *Trends Biotechnol* 28:381-390.

van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G. and Leunissen, J. A. M. (2006). A text-mining analysis of the human phenome. *Eur J Hum Genet* 14:535-542.

Tigani, J. and Naidu, S. (2014). *Google BigQuery Analytics*. Wiley Publishing.

Bird, S. (2006). NLTK: the natural language toolkit, in *Proceedings of the COLING/ACL on Interactive presentation sessions*, Association for Computational Linguistics, Sydney, Australia, 69-72.

Bird, S., Klein, E. and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc.

Tigani, J. and Naidu, S. (2014). *Google BigQuery Analytics*. Wiley Publishing.

Stephanopoulos, G. (1999). Metabolic Fluxes and Metabolic Engineering. *Metab. Eng.* 1:1-11.

Nielsen, J. and Keasling, J. D. (2011). Synergies between synthetic biology and metabolic engineering. *Nat. Biotech.* 29:693-695.

Carothers, J. M., Goler, J. A. and Keasling, J. D. (2009). Chemical synthesis using synthetic biology. *Curr. Opin. Biotech.* 20:498-503.

Lee, S. K., Chou, H., Ham, T. S., Lee, T. S. and Keasling, J. D. (2008). Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. *Curr. Opin. Biotech.* 19:556-563.

Church, G. M., Elowitz, M. B., Smolke, C. D., Voigt, C. A. and Weiss, R. (2014). Realizing the potential of synthetic biology. *Nat. Rev. Mol. Cell. Bio.* 15:289-294.

- Bailey, J. (1991). Toward a science of metabolic engineering. *Science* 252:1668-1675.
- Stephanopoulos, G., Aristidou, A. and Nielsen, J. (1998). Metabolic engineering: principles and methodologies.
- Gardner, T. S., Cantor, C. R. and Collins, J. J. (2000). Construction of a genetic toggle switch in *Escherichia coli*. *Nature* 403:339-342.
- Elowitz, M. B. and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature* 403:335-338.
- Smolke, C. D. (2009). Building outside of the box: iGEM and the BioBricks Foundation. *Nat. Biotech.* 27:1099-1102.
- Brown, J. (2007). The iGEM competition: building with biology, in *IET Synthetic Biology*, 3-6.
- Canton, B., Labno, A. and Endy, D. (2008). Refinement and standardization of synthetic biological parts and devices. *Nat. Biotech.* 26:787-793.
- Tannler, S., Decasper, S. and Sauer, U. (2008). Maintenance metabolism and carbon fluxes in *Bacillus* species. *Microb. Cell Fact.* 7:19.
- Zhang, X.-Z., Cui, Z.-L., Hong, Q. and Li, S.-P. (2005). High-Level Expression and Secretion of Methyl Parathion Hydrolase in *Bacillus subtilis* WB800. *Appl Environ Microb* 71:4101-4103.
- Cregg, J. M., Tolstorukov, I., Kusari, A., Sunga, J., Madden, K. and Chappell, T. (2009). Chapter 13 Expression in the Yeast *Pichia pastoris*, in *Method Enzymol*, R. B. Richard and P. D. Murray, eds., Academic Press, 169-189.
- Atsumi, S., Hanai, T. and Liao, J. C. (2008). Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature* 451:86-89.

Steen, E. J., Kang, Y., Bokinsky, G., Hu, Z., Schirmer, A., McClure, A., del Cardayre, S. B. and Keasling, J. D. (2010). Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature* 463:559-562.

Atsumi, S., Higashide, W. and Liao, J. C. (2009). Direct photosynthetic recycling of carbon dioxide to isobutyraldehyde. *Nat Biotech* 27:1177-1180.

Hull, D., Pettifer, S. R. and Kell, D. B. (2008). Defrosting the Digital Library: Bibliographic Tools for the Next Generation Web. *PLoS Comput. Biol.* 4:e1000204.

<b>Terms</b>	<b>Metabolic engineering</b>	<b>Synthetic biology</b>
<b>Most frequent words</b>	Production	Gene
<b>Major differences in 500 most frequent words</b>	FBA (flux model) NADH/NADPH (cofactor) ATP (energy) Xylose/acetate/glycerol HPLC transport/tolerance	circuit/loop IPTG/lacI (inducer) RBS Switch (regulation) egfp Phage/virus

Table 7.1. Comparison of ‘metabolic engineering’ and ‘synthetic biology’, similarity 69.6%

Terms	Metabolic engineering	
	2000 ~	2010 ~
<b>Most frequent words</b>	Metabolic	Production
<b>Major differences in 500 most frequent words</b>	PHB/Lysine PTS <i>B. subtilis</i> hydrogen <i>P. Pastoris</i>	butanol/isobutanol HPLC cyanobacteria Lipid <i>C. glutamicum</i>

Table 7.2. Development trend of ‘metabolic engineering’ during 2000 ~ 2009 and 2010 ~ 2015, similarity 81%



<b>Terms</b>	<b>Biofuel</b>	
	2000 ~	2010 ~
<b>Most frequent words</b>	Production	Genes
<b>Major differences in 500 most frequent words</b>	Lignocellulosic hemicellulose Emissions policy Economic/market Feedstock	Glycerol/acetate xylan Algae/microalgae Cyanobacteria Lipid Chromosome/cDNA

Table 7.3. Development trend of 'biofuel' during 2000 ~ 2009 and 2010 ~ 2015 similarity 72.4%

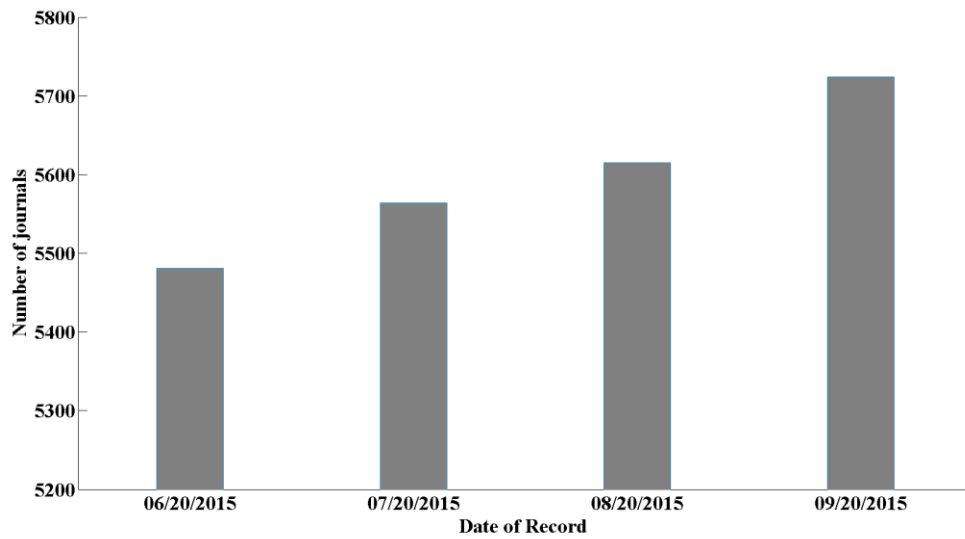


Figure 7.1a. Total number of journals in the database at different time

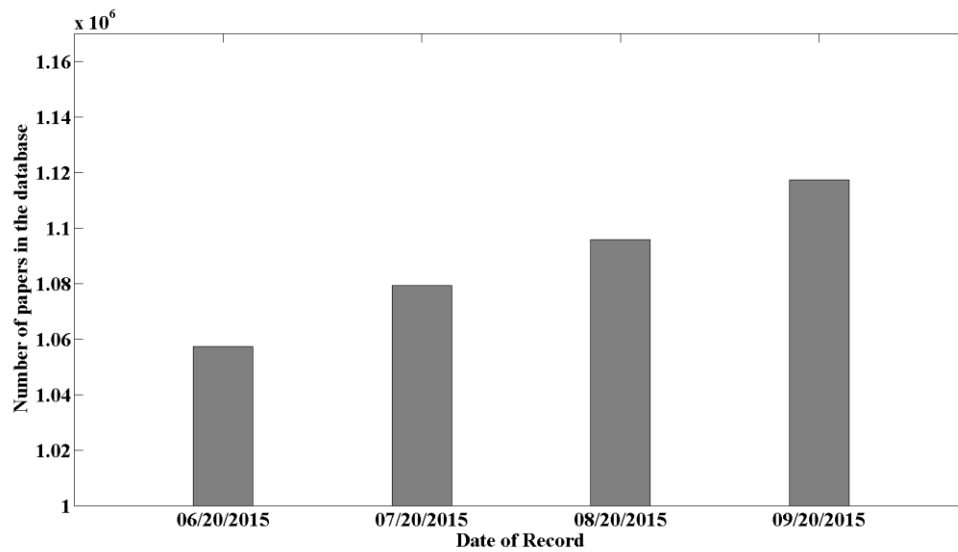


Figure 7.1b. Total number of papers in the database increase at different time

```
- <front>
  + <journal-meta>
  + <article-meta>
</front>
- <body>
  + <sec>
  + <sec sec-type="results">
  + <sec sec-type="discussion">
  + <sec sec-type="conclusions">
  + <sec sec-type="methods">
  + <sec>
  + <sec>
  + <sec sec-type="supplementary-material">
</body>
+ <back>
```

Figure 7.2. Structure of nxml file



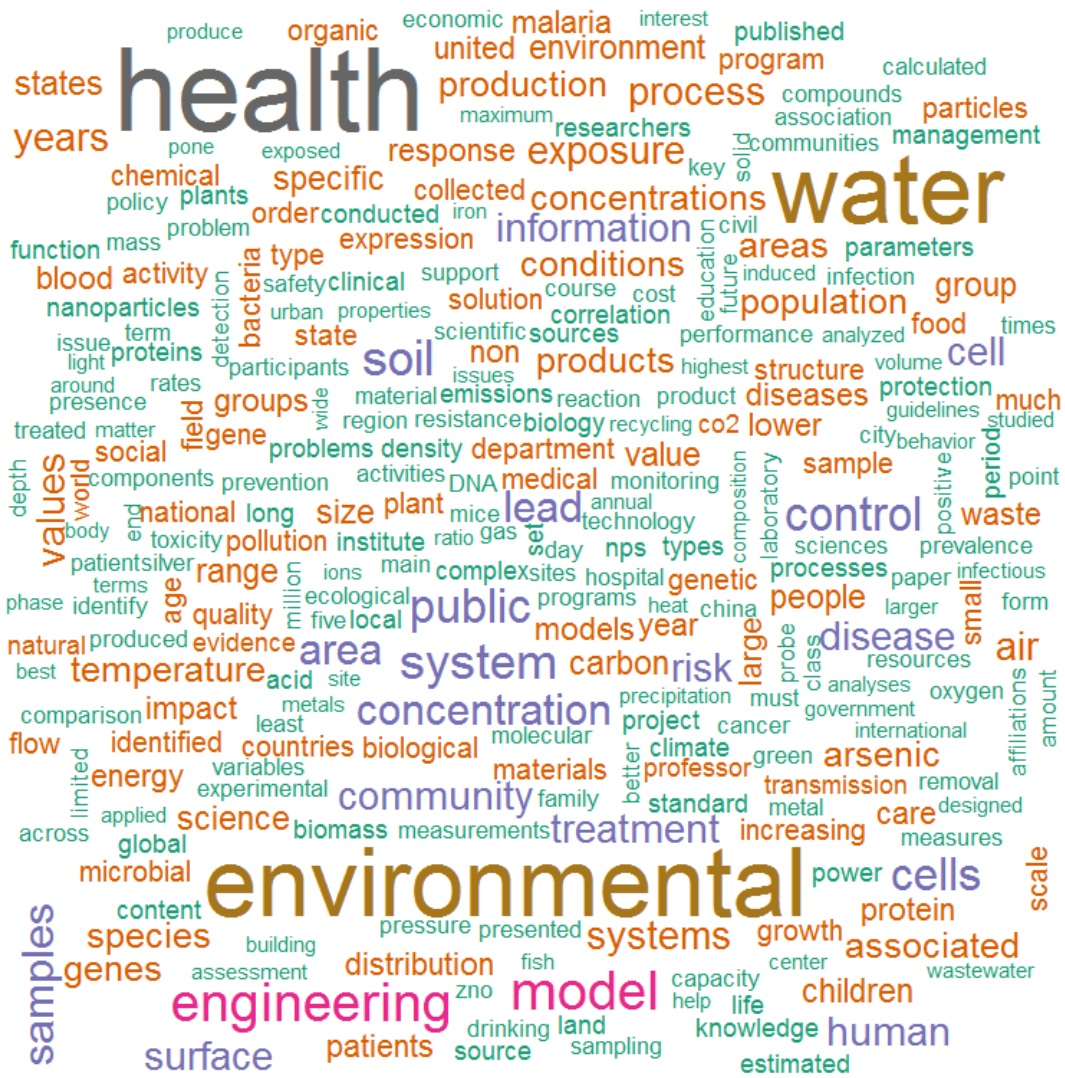


Figure 7.4. Word cloud of ‘environmental engineering’











## CHAPTER EIGHT

### CONCLUSIONS AND FUTURE PERSPECTIVES

#### 8.1 Conclusions

In this dissertation, I first employed fluxomics tools (FBA and  $^{13}\text{C}$ -MFA) to investigate the metabolisms of isobutanol-producing *E. coli* strains in chapter two.  $^{13}\text{C}$ -MFA results indicated that isobutanol production reshaped the central metabolism with increased activities in Pentose Phosphate pathway (supply more NADPH for isobutanol synthesis), and glyoxylate shunt (reserve more carbon). To gain an overview on how the factor of oxygen concentration, P/O ratio, and maintenance energy affects isobutanol production, we set up an integrated flux model (use  $^{13}\text{C}$ -MFA flux values to constrain genome scale FBA) to test the sensitivity of each factor. The simulation result revealed that the maintenance energy played a more important role in affecting isobutanol yield than P/O ratio and oxygen flux. Further, we found a ‘cliff’ in isobutanol yield landscape which can be triggered by increased maintenance energy, decreased P/O ratio or decreased oxygen supply. Discovery of this cliff explained many failures observed in isobutanol experiments and scale-up processes. Besides, we also quantify the impacts of yeast extract and the results show that supply of rich nutrients such as yeast extract can efficiently relieve the intracellular crisis in carbon and energy resource, thereby can significantly boost isobutanol production without directly contributing to the product. The limitation of this work is industrial isobutanol production mainly focus on the non-growth phase, which central metabolites tracing, fast quenching, as well as dynamics  $^{13}\text{C}$ -MFA are required to resolve cellular metabolism at this phase.

We expanded the analysis to other biofuel producing cases in chapter three. FBA simulations indicated that alcohol (i.e., ethanol, isobutanol) production is more robust to P/O ratio change (efficiency of energy metabolism) than fatty acid production in *E. coli* strains. In addition, we took an overview of those successful and failure cases in biotechnology industry and realized that avoiding extensive genetic modification is important for high yield and strain stability. Heterogeneous plasmid or enzyme expression introduces extra metabolic burden, leading to shifted central metabolism, and even imbalance between carbon and energy metabolism (Yin-Yang Balance). We proposed several approaches to solve this problem, such as employment of fluxomics tools (i.e.,  $^{13}\text{C}$ -MFA) to decipher energy metabolism, relying on native pathways, performing minimal engineering, *etc.* Based on the Yin-Yang theory, we tried to insert *Vitreoscilla* hemoglobin (*vhb*) gene into fatty acid producing strains to improve its oxygen uptake ability and thus fatty acid production, which is discussed in chapter four.

#### Chapter four

Besides energy metabolism related projects, I also developed a series of computational tools for fluxomics studies and literature analysis. In chapter five, I rebuilt the MicrobesFlux platform, moved it from a share server at Washington University to a commercial server (Amazon EC2 server). In addition, I enhanced the functionality of MicrobesFlux by including much more species (previously: 1304 species; currently: 3192 species), supporting the SBML file format, having unlimited storage space. Further, I participated in the development of MATLAB based open-source  $^{13}\text{C}$ -MFA software (WUFlux). In addition, I designed and built several website platforms (fluxomics.net, 13cmfa.org) for fluxomics studies, also we shared all fluxomics tools used in our group on the website for free.

In chapter six, we built a computational platform that can calculate microbial metabolism very quickly based on user input. This computational platform named MFlux (MFlux) is developed based on an integration of machine learning, constraint programming, and quadratic programming. Through grid searching, we chose Support Vector Machine as machine learning algorithm and optimize its parameters. Quadratic programming can adjust predicted flux profile to satisfy the stoichiometry constraints and constraint programming is used to avoid non-sense inputs. To sum up, we build up the correlations between genetic/environmental factors and central metabolic fluxes.

In chapter seven, I developed a Big Data based framework which can perform fast literature analysis based on user inputs. We first employed the text mining approach to extract information from full text papers. Subsequently, we upload all information into Google Cloud server and set up a fast search tool through Google BigQuery. Finally, through BigQuery search plus searching results processing, we get meaningful information from literature database. We performed several case studies, and the results turned to be fast (whole process less than 2 min), reliable (results repeatable), and informative. The limitation of this work is that current literature database only contains 1.1 million papers focusing on biomedical research; and we need more information to avoid bias in analysis.

## **8.2. Future directions to solve intracellular energy crisis**

To circumvent the energy bottlenecks within microbial cell factories, we propose the following approaches in addition to those already mentioned in Chapter three:

### **1. Improve energy efficiency by engineering respiration metabolism**

In aerobic metabolism, energy generation mainly comes from ATP synthesis through oxidative phosphorylation (respiration). However, in real cases, respiration rates in many strains are far below theoretical maximum (Varma and Palsson 1994; Wu et al. 2015; Sauer and Bailey 1999). Thereby, improving oxidative respiration efficiency is an efficient manner to enhance energy supply. The first successful case of knocking out inefficient respiration metabolism was reported by Zamboni and Sauer, to enhance riboflavin production in *Bacillus subtilis* (Zamboni et al. 2003). In this work, knockout of cytochrome *bd* oxidase also leads to a 40% reduction of cellular maintenance energy. Afterwards, this strategy has been adopted and combined with other strategies, to enhance the yield of other products (e.g., N-acetylglucosamine) in *B. subtilis* (Liu et al. 2014). Observations of reduced maintenance energy were also reported in model species *E. coli* (*ndh* knockout) (Calhoun et al. 1993) and *C. glutamicum* (*cydAB* knockout) after eliminating the energy metabolism component of low efficiency (Kabus et al. 2007). Further, the cytochrome *bd* knockout mutant was reported to enhance lysine production by ~12% in *Corynebacterium glutamicum* (Kabus et al. 2007). This strategy may apply to species with several sets of respiration chains with different efficiency. Notably, successful applications of this strategy are also closely related with other factors, such as oxygen concentration, medium composition and *etc.* (Kabashima et al. 2009).

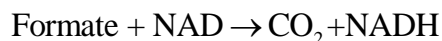
## **2. Utilization of other energy sources**

As a traditional energy source, hydrogen can be utilized by a broad range of microbes (B Friedrich and Schwartz 1993; Petersen et al. 2011), Utilization of hydrogen as the energy supplier in industrial fermentation is not preferred, albeit the well-known syngas fermentation. This is due to those undesired properties of hydrogen gas such as low solubility, easy-to-leak,

and explosive. Compared with hydrogen, formate is a better source of energy supply in terms of the uptake efficiency.

The utilization of formate by microbes as extra energy source was discovered more than thirty years ago. In 1983, Bruinenberg *et al.* discovered that *Candida utilis* can uptake formate as an additional energy source in the presence of glucose (Bruinenberg *et al.* 1983). Utilization of formate leads to an increased biomass yield. Meanwhile, similar phenomena were also observed in *Hansenula polymorpha* by Babel *et al.* and in *Pichia pastoris* by Hazeu *et al.* (Babel *et al.* 1983; Hazeu and Donker 1983). *Saccharomyces cerevisiae* CBS 8066 also joined this list of formate utilization species. However, formate utilization did not make any improvement in biomass yield of *S. cerevisiae*. (Bruinenberg *et al.* 1985) It was then realized that utilization of formate normally requires the functionality NAD-dependent formate dehydrogenase (FDH, EC 1.2.1.1). The strategy of employing formate as extra energy source has been extended to other species with FDH, such as oleaginous yeasts (*Cryptococcus curvatus*, *Rhodotorula glutinis*, and *Lipomyces starkeyi*) for improving lipid production (Lian *et al.* 2012), *Penicillium chrysogenum* for enhancing penicillin G productivity (Harris *et al.* 2007), and *Bacillus thuringiensis* for promoting thuringiensin yield (Zhi *et al.* 2007). Formate can also be generated through an electrochemical process, to feed engineered *Ralstonia eutropha* H16 strain to produce isobutanol and 3-Methyl-1-butanol (Li *et al.* 2012). On the other side, heterologous expression of an *fdh* gene enables formate usage in those species without this gene. For instance, after insertion of the *fdh* gene into the chromosome, *Corynebacterium glutamicum* was able to utilize formate and produce 20% more succinate anaerobically in the presence of glucose. Formate was used as the NADH and CO<sub>2</sub> donor in this case (Litsanov *et al.* 2012). In yet another case, *fdh* was introduced into a succinate-producing *E. coli* strain, leading to significantly reduced byproduct

formate and improve succinate yield and productivity (Balzer et al. 2013). Metabolic burden of *fdh* overexpression may offset its benefit; thus, careful consideration is necessary at the stage of strain design.



Photosynthetic microorganisms (cyanobacteria and microalgae) convert CO<sub>2</sub> to useful products with light as the energy source. The list of their products has been greatly boosted with the advent of advanced genetic tools (Wijffels et al. 2013). However, scale-up photosynthetic microbial process to industrial production has been severely hindered by a limited range of light penetration, which leads to a series of problems such as high-cost photobioreactor, low cell density, and cost-inefficient harvesting. To circumvent these problems, a photomixotrophic strategy has been proposed (You et al. 2015). Under light and glucose sufficient condition, *Synechocystis* sp. PCC 6803 is able to consume both CO<sub>2</sub> and glucose for biomass production which potentially leads to higher biomass density.

### **3. Minimize maintenance energy in host strain**

During industrial fermentation, cellular maintenance energy is released out as heat, leading to increased temperature of fermentation broth, which is undesired for the whole process. Further, low maintenance energy requirement indicates more energy allocated on biomass & product synthesis (Sauer et al. 1996). Thereby, hosts of low maintenance energy are preferred. In a comparative study, cellular metabolisms of several bacilli strains (both wide type and mutant) close to *Bacillus subtilis* were investigated. And the result show that the *B. licheniformis* T380B strain has the lowest maintenance energy (0.20 mmol/g\*h) of all strains analyzed in that work. Therefore, the authors consider it as a potential host to replace *B. subtilis* (maintenance energy 0.39 mmol/g\*h) for vitamin and other chemical production (Tannler et al. 2008).

On the other side, cellular maintenance energy is positively correlated with incubation temperature (Price and Sowers 2004; Lever et al. 2015). From energy point of view, lower temperature is preferred for fermentation process, in the sake of lower maintenance energy and enhanced strain stability. However, temperature is an essential factor affecting many processes within cellular metabolism (e.g., enzymes kinetics, regulation). In practice, a trade-off between production rate and strain stability requires careful consideration during temperature control (Dunn-Coleman et al. 1992).

### **8.3 Personal views on the future of microbial cell factories**

#### **8.3.1. Limitation of microbial systems**

Each closed system has its limitation. For instance, enzymatic systems are facing the trade-off between its kinetic efficiency ( $K_{cat}/K_m$ ) and thermostability: it is very challenging to obtain enzymes with both good thermostability and high catalytic efficiency (Ye et al. 2012; Romero and Arnold 2009). In a similar manner, microbial systems have a series of physical limitations: the trade-off between cell membrane surface area/volume and material exchange efficiency, the limited cell membrane surface area decides the upper limit of nutrient and oxygen uptake rate (Zhuang et al. 2011), pose a further restriction on the total carbon/energy available within a cell. Taken all limiting factors together before the design, we are able to realize what is impossible even before many failures after extensive engineering.

#### **8.3.2. The success of distributed system in computational system shed lights on microbial systems**

Computational systems have their limitation: the CPU frequency is strictly limited by the speed of signal travel (light of speed), therefore, after we had 3GHz CPU more than ten years (2002), it

is difficult to get more improvements in CPU clock. Expensive multi-core high performance working stations were the first choice to handle those extensive computation, however, are being replaced by distributed computational systems (GDFS, Google Distributed File System or HDFS, Hadoop Distributed File System, both are based on MapReduce algorithm) due to high costs and low robustness to error. Reliability (robustness) is an important factor deciding the commercial values of any systems, and that is the reason distributed computational systems wins the battle with working stations. Similarly, extensive genetic modification on microbial cells bring extra burdens on cell metabolism which leads to increased instability of cell itself, as well as its poor reliability, which has been verified by many failed bioprocess scale-up. In a similar manner, microbial distributed systems (e.g., coculture (Zhou et al. 2015), integrated bio-chem process (Xiong et al. 2014)) are still in its infant phase. The key point for successful microbial distributed systems is to have a universal framework (similar as MapReduce) to improve the reliability of the system substantially, which current synthetic biology or metabolic engineering does not solves.

#### **8.4. Reference**

Varma, A. and Palsson, B. (1994). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl Environ Microb* 60:3724 - 3731.

Wu, S., He, L., Wang, Q. and Tang, Y. (2015). An ancient Chinese wisdom for metabolic engineering: Yin-Yang. *Microb. Cell Fact.* 14:39.



- Sauer, U. and Bailey, J. E. (1999). Estimation of P-to-O ratio in *Bacillus subtilis* and its influence on maximum riboflavin yield. *Biotechnol Bioeng* 64:750-754.
- Zamboni, N., Mouncey, N., Hohmann, H.-P. and Sauer, U. (2003). Reducing maintenance metabolism by metabolic engineering of respiration improves riboflavin production by *Bacillus subtilis*. *Metab. Eng.* 5:49-55.
- Liu, Y., Zhu, Y., Ma, W., Shin, H.-d., Li, J., Liu, L., Du, G. and Chen, J. (2014). Spatial modulation of key pathway enzymes by DNA-guided scaffold system and respiration chain engineering for improved N-acetylglucosamine production by *Bacillus subtilis*. *Metab. Eng.* 24:61-69.
- Calhoun, M. W., Oden, K. L., Gennis, R. B., de Mattos, M. J. and Neijssel, O. M. (1993). Energetic efficiency of *Escherichia coli*: effects of mutations in components of the aerobic respiratory chain. *J Bacteriol* 175:3020-3025.
- Kabus, A., Niebisch, A. and Bott, M. (2007). Role of Cytochrome bd Oxidase from *Corynebacterium glutamicum* in Growth and Lysine Production. *Appl Environ Microb* 73:861-868.
- Kabashima, Y., Kishikawa, J.-i., Kurokawa, T. and Sakamoto, J. (2009). Correlation Between Proton Translocation and Growth: Genetic Analysis of the Respiratory Chain of *Corynebacterium glutamicum*. *J Biochem* 146:845-855.
- B Friedrich, a. and Schwartz, E. (1993). Molecular Biology of Hydrogen Utilization in Aerobic Chemolithotrophs. *Annu Rev Microbiol* 47:351-383.
- Petersen, J. M., Zielinski, F. U., Pape, T., Seifert, R., Moraru, C., Amann, R., Hourdez, S., Girguis, P. R., Wankel, S. D., Barbe, V., Pelletier, E., Fink, D., Borowski, C., Bach, W. and

Dubilier, N. (2011). Hydrogen is an energy source for hydrothermal vent symbioses. *Nature* 476:176-180.

Bruinenberg, P. M., Van Dijken, J. P. and Scheffers, W. A. (1983). An Enzymic Analysis of NADPH Production and Consumption in *Candida utilis*. *Microbiology* 129:965-971.

Babel, W., Müller, R. and Markuske, K. (1983). Improvement of growth yield of yeast on glucose to the maximum by using an additional energy source. *Arch Microbiol* 136:203-208.

Hazeu, W. and Donker, R. A. (1983). A continuous culture study of methanol and formate utilization by the yeast *Pichia pastoris*. *Biotechnol Lett* 5:399-404.

Bruinenberg, P., Jonker, R., van Dijken, J. and Scheffers, W. A. (1985). Utilization of formate as an additional energy source by glucose-limited chemostat cultures of *Candida utilis* CBS 621 and *Saccharomyces cerevisiae* CBS 8066. *Arch Microbiol* 142:302-306.

Lian, J., Garcia-Perez, M., Coates, R., Wu, H. and Chen, S. (2012). Yeast fermentation of carboxylic acids obtained from pyrolytic aqueous phases for lipid production. *Bioresource Technol* 118:177-186.

Harris, D. M., van der Krogt, Z. A., van Gulik, W. M., van Dijken, J. P. and Pronk, J. T. (2007). Formate as an Auxiliary Substrate for Glucose-Limited Cultivation of *Penicillium chrysogenum*: Impact on Penicillin G Production and Biomass Yield. *Appl Environ Microb* 73:5020-5025.

Zhi, W., Shouwen, C., Lifang, R., Ming, S. and Ziniu, Y. (2007). A fundamental regulatory role of formate on thuringiensin production by resting cell of *Bacillus thuringiensis* YBT-032. *Bioprocess Biosyst. Eng.* 30:225-229.

Li, H., Opgenorth, P. H., Wernick, D. G., Rogers, S., Wu, T.-Y., Higashide, W., Malati, P., Huo, Y.-X., Cho, K. M. and Liao, J. C. (2012). Integrated Electromicrobial Conversion of CO<sub>2</sub> to Higher Alcohols. *Science* 335:1596.

Litsanov, B., Brocker, M. and Bott, M. (2012). Toward Homosuccinate Fermentation: Metabolic Engineering of *Corynebacterium glutamicum* for Anaerobic Production of Succinate from Glucose and Formate. *Appl Environ Microb* 78:3325-3337.

Balzer, G. J., Thakker, C., Bennett, G. N. and San, K.-Y. (2013). Metabolic engineering of *Escherichia coli* to minimize byproduct formate and improving succinate productivity through increasing NADH availability by heterologous expression of NAD<sup>+</sup>-dependent formate dehydrogenase. *Metab. Eng.* 20:1-8.

Wijffels, R. H., Kruse, O. and Hellingwerf, K. J. (2013). Potential of industrial biotechnology with cyanobacteria and eukaryotic microalgae. *Curr Opin Biotech* 24:405-413.

You, L., He, L. and Tang, Y. J. (2015). Photoheterotrophic Fluxome in *Synechocystis* sp. Strain PCC 6803 and Its Implications for Cyanobacterial Bioenergetics. *J Bacteriol* 197:943-950.

Sauer, U., Hatzimanikatis, V., Hohmann, H. P., Manneberg, M., van Loon, A. P. and Bailey, J. E. (1996). Physiology and metabolic fluxes of wild-type and riboflavin-producing *Bacillus subtilis*. *Appl Environ Microb* 62:3687-3696.

Tannler, S., Decasper, S. and Sauer, U. (2008). Maintenance metabolism and carbon fluxes in *Bacillus* species. *Microb. Cell Fact.* 7:19.

Price, P. B. and Sowers, T. (2004). Temperature dependence of metabolic rates for microbial growth, maintenance, and survival. *Proc Natl Acad Sci USA* 101:4631-4636.

Lever, M. A., Rogers, K. L., Lloyd, K. G., Overmann, J., Schink, B., Thauer, R. K., Hoehler, T. M. and Jørgensen, B. B. (2015). Life under extreme energy limitation: a synthesis of laboratory- and field-based investigations. *Fems Microbiol Rev.*

Dunn-Coleman, N., Bodie, E., Carter, G. and Armstrong, G. (1992). Stability of recombinant strains under fermentation conditions. *Applied Molecular Genetics of Filamentous Fungi*:152-174.

Ye, X., Zhang, C. and Zhang, Y. H. P. (2012). Engineering a large protein by combined rational and random approaches: stabilizing the *Clostridium thermocellum* cellobiose phosphorylase. *Mol. BioSyst.* 8:1815-1823.

Romero, P. A. and Arnold, F. H. (2009). Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Bio* 10:866-876.

Zhuang, K., Vemuri, G. N. and Mahadevan, R. (2011). Economics of membrane occupancy and respiro-fermentation. *Mol. Syst. Biol.* 7:500.

Zhou, K., Qiao, K., Edgar, S. and Stephanopoulos, G. (2015). Distributing a metabolic pathway among a microbial consortium enhances production of natural products. *Nat Biotech* 33:377-383.

Xiong, M., Schneiderman, D. K., Bates, F. S., Hillmyer, M. A. and Zhang, K. (2014). Scalable production of mechanically tunable block polymers from sugar. *Proc. Natl. Acad. Sci.* 111:8357-8362.

## **Appendix I**

### **1. List of courses taken and to be taken with grades**

Courses finished:

EECE501 Transport Phenomena in Energy, Environmental, and Chemical Engineering

EECE5404 Combustion Phenomena

EECE548 Environmental Organic Chemistry

EECE503 Kinetics and Reaction Engineering Principles

EECE596 Metabolic Engineering

EECE517 Partial Differential Equations

BIOL5014 Biotech Industry Innovators

EECE590 Energy and Environmental Economic Decision-Making

Courses to be transferred from master program (Virginia Tech):

CHE5984 Engineering Mathematics

STAT5605 Biometry 1

MATH5515 Mathematical Methods for Modeling and Simulation of Biological Systems

### **2. TA experiences**

CHE473A Chemical Engineering Laboratory (12 fall)

EECE262 Introduction to Environmental Engineering (13 spring)

CHE478A Process & Product Design (14 spring)

Workshop certification:

Grading and responding to students' concerns about grades

Improving presentation skills

### **3. Academic achievements**

Conference experience:

2014 Annual Meeting of American Institute of Chemical Engineers (AIChE) Poster

2012 Annual Conference of Sustainable Nanotechnology Organization (SNO) Poster

2012 Annual Meeting of American Institute of Chemical Engineers (AIChE) Poster

2011 Annual Mid-American Environmental Engineering Conference (MAEEC) Oral

Publication:

L. He, **S. G. Wu**, and Y. J. Tang, WUFlux: an open-source MATLAB based software for  $^{13}\text{C}$ -MFA, under review by *BMC Bioinformatics*

**S. G. Wu**, A. Varman, L. He, and Y. J. Tang, Evaluating physiological state of engineered e. coli strains by isotopomer constrained flux balance analysis, *in preparation*

**S. G. Wu**, Y. Wang, W. Jiang, T. Oyetunde, R. Yao, K. Shimizu, Y. J. Tang, and F. S. Bao, Rapid prediction of bacterial fluxomics using machine learning and constraint programming. *PLOS Computational Biology*, accepted

**S. G. Wu**, K. Shimizu, J. K-H Tang, and Y. J. Tang, *Collaboration between Synthetic Biology & Metabolic Engineering with the Aid of Machine Learning for Industrial Biotechnology*, *ChemBioEng Reviews*, accepted.

L. He, **S. G. Wu**, N. Wan, R. Adrienne, and Y. J. Tang, Simulating cyanobacterial phenotypes by coupling flux balance analysis, kinetics, and a light distribution function, *Microbial Cell Factories* 14: 206

**S. G. Wu**, L. He, Q. Wang, Y. J. Tang (2015) An ancient Chinese wisdom for metabolic engineering: Yin-Yang. *Microbial Cell Factories* 14: 39.

**S. G. Wu**, L. Huang, J. Head, M. Ball, Y. J. Tang, and D-R. Chen, Electrospray Facilitates the Germination of Plant Seeds, *Aerosol Air Qual. Res.* 14 (3), 632-641, 2014

Y. Xiao, Z. Ruan, Z. Liu, **S. G. Wu**, A. M. Varman, Y. Liu, and Y. J. Tang, Engineering *Escherichia coli* to convert acetic acid to free fatty acids, *Biochem. Eng. J.* 76: 60-69, 2013

**S. G. Wu**, L. Huang, J. Head, D-R. Chen, In-Chul Kong and Y. J. Tang, and, Phytotoxicity of Metal Oxide Nanoparticles is Related to Both Dissolved Metals Ions and Adsorption of Particles on Seed Surfaces, *J. Pet. Environ. Biotechnol.* 3:126, 2012

Awards:

National Institute for Mathematical and Biological Synthesis (NIMBioS) Visiting Graduate Fellowship, 2013

## Appendix II Supporting information of Chapter 6

### S1. Source code

---

#### clp.py

```
def process_species_db(File):
    """Load the species database which is Supplement Information I

    Format
    =====
    Fields separated by tab

    Species Spcies name Oxygen condition Substrate uptake rate
    upper bound (mmol/gDW*mol ) 1 2 3 4 5 6 7 8 9 10
    11 12 13 14 Growth rate upper bound (h-1) Reference
    1 Escherichia coli 1,3,2 20 Y Y Y Y Y Y Y Y Y
    Y Y Y Y Y N 1.2 1
    2 Corynebacterium glutamicum 1,3,2 40 Y Y Y Y Y Y
    Y Y N Y N Y Y N 1 2

    Returns
    =====
    DB: A list of tuples. Each tuple is (species, Oxygen, rate,
    Carbon1, Carbon 2, ..., Carbon 14, Growth_rate_upper)
    Oxygen itself is a string, e.g., "1,2,3"

    """
    Carbon_sub = {"Y":True, "N":False}
    DB = []
    with open(File, "r") as F:
        F.readline() # skip the header
        for Line in F:
            Field = Line.split("\t")
            # print Field
            [Species, Substrate_rate] = map(int, [Field[0], Field[3]])
            Oxygen = Field[2] #map(int, Field[2].split(","))
            Carbon_src = [ Carbon_sub.get(x, False) for x in
            Field[4:4+13+1] ]
            Growth_rate_upper = Field[4+14]
            DB.append(tuple([Species, Oxygen, Substrate_rate]+
            Carbon_src + [Growth_rate_upper]))

    return DB
```

```

def species_db_to_constraints(DB, Debug=False):
    """Turn the species DB into a CSP problem (constraints only, no
    variable ranges)

    Parameters
    =====
    DB: list of tuples
        Each tuple is (species, Oxygen, rate, Carbon1, Carbon 2, ...,
        Carbon 14)
        Oxygen itself is a tuple, e.g., (1,3)

    Returns
    =====
    problem: an instance of python-constraint
        containing only constraints but no variable domains

    Notes
    =====
    the problem has a solution if any of the rules set in species
    database is VIOLATED.
    In other words, if the problem has solution, then the input does
    NOT make sense.

    """
    import constraint # python-constraint module
    problem = constraint.Problem() # initialize the CSP problem

    # create variables
    # problem.addVariable("Species", range(1,41+1))
    # problem.addVariable("Substrate_rate", range(0, 100+1))
    # problem.addVariable("Oxygen", [1,2,3])
    # for i in xrange(1, 14+1):
    #     problem.addVariable("Carbon"+str(i), [True, False]) # create
one variable for each carbon source
    # This part should be from user input

    # add constraints, where each entry in DB is a constraint.
    # create the lambda functions
    All_vars= ["Species", "Substrate_rate", "Oxygen"] +
["Carbon"+str(i) for i in xrange(1, 14+1)] + ["Growth_rate_upper"]
    for Entry in DB:
        Oxygen_values = Entry[1] # as string
        Foo = "lambda "
        Foo += ", ".join(All_vars) # done with listing all variables
        Foo += ": "
        Logic_exp = ["Substrate_rate<=" + str(Entry[2]), "Species==" +
str(Entry[0]), "Growth_rate_upper<=" + str(Entry[4+13])]

        for i in range(3, 3+14): # carbon sources
            if not Entry[i]: # only use false ones to create the
constraint

```







```

print """\
<html>
<head><title>Result of Influx analysis </title></head>

<body>
<h2>Parameters entered:</h2>
"""\

# Process the form values
for Feature_name in Feature_names:
    Feature_value = form.getfirst(Feature_name)
    Feature_value = cgi.escape(Feature_value)
    Features[Feature_name] = float(Feature_value) # convert all string
to numbers

    print """\
    %s is %s,
    """ % (Feature_name, Feature_value)

import libflux
Vector, Substrates = libflux.process_input(Features)
Boundary_dict = libflux.process_boundaries(form, Substrates)

#libflux.test("hello, world")
Influxes = libflux.predict(Vector, Substrates, Boundary_dict) # use
the feature vector to predict influx values

print """\
<p><a href="index.html">Go back to submission page</a></p>

<hr>
<p>
This project is supported by National Science Foundation. <a
href="http://www.nsf.gov/awardsearch/showAward?AWD_ID=1356669">More
info</a> <br>
Information on this website only reflects the perspectives of the
individuals.<br>
</p>
</body>
</html>
"""\

```

---

### **get\_model.py**

```

# extract the training data from spreadsheet

from collections import defaultdict

```

```

import cPickle

import numpy
from sklearn import cross_validation, preprocessing, grid_search
from sklearn.neighbors import KNeighborsRegressor
from sklearn.svm import SVR
from sklearn.tree import DecisionTreeRegressor

class RegressionModel(object):
    """A help class to store model's name and the actual model."""

    def __init__(self, name, **kwargs):
        self.name = name
        self.model = eval(name)(**kwargs)

    def __str__(self):
        return self.name

class RegressionModelFactory(object):
    """A factory to create new instances of RegressionModel."""

    def __init__(self, name, **kwargs):
        self.name = name
        self.kwargs = kwargs

    def __str__(self):
        return "{} ({}).format(self.name, self.kwargs)

    def __call__(self):
        """Create a new RegressionModel instance each time."""
        return RegressionModel(self.name, **self.kwargs)

def shuffle_data(Training_data):
    """Shuffle the order of data in training data

    Shuffle by scrambling the index

    (New)_training_data: a dict, keys are EMPs (e.g., v1, v2, etc.),
        values are 2-tuples (Feature, Label), where
        Feature is a 2-D list, each sublist is 24-D feature
vector for one sample
        and
        Label is a 1-D list, labels for all samples.

    """
    import random
    New_training_data = {}
    for i, (Feature_vector, Labels) in Training_data.iteritems():
        Num_Samples = len(Feature_vector)

```

```

        if Num_Samples != len(Labels):
            print "Error! Inconsistent numbers of Features Vectors and
Labels"
        Shuffled_index = range(Num_Samples)
        random.shuffle(Shuffled_index)
        New_Feature_vector = [Feature_vector[j] for j in
Shuffled_index]
        New_Label = [Labels[j] for j in Shuffled_index]

        New_training_data[i] = ([New_Feature_vector, New_Label])

    return New_training_data

def read_spreadsheet(filename):
    """Turn spreadsheet into matrixes for training

    Returns
    =====

    training_data: a dict, keys are EMPs (e.g., v1, v2, etc.),
        values are 2-tuples (Feature, Label), where
        Feature is a 2-D list, each sublist is 24-D feature
vector for one sample
        and
        Label is a 1-D list, labels for all samples.

    Notes
    =====
    EMPs are N.A. for some samples, training features were dropped for
them.
    That's why we need one training feature matrix for each EMP.

    We have 29 influxes values to predict and thus the index/key for
training_data goes from 1 to 29

    AA is 26 for 0-index
    BA is 32 for 0-index
    """

    training_data = {}
    for i in range(1, 29+1):# prepare the data structure
        training_data[i] = ([],[]) # the 1st list is the features and
the 2nd the labels for i-th influx

    reports = defaultdict(list)
    with open(filename, 'r') as f:
        for i, line in enumerate(f.readlines(), 1):
            line = line.strip()
            line = line.split("\t")
            vector = line[2:26+1] # training vector, from Species (C)
to Other carbon (AA).

```

```

                                # one empty column
key = ", ".join(vector)
reports[key].append(i)

if "" in vector:
    vector.remove("")
if not vector :
    print line
    exit()

labels = line[26+3: 26+3+26+5] # AD to BF, v1 to v29
#
    print Labels

    try:
        vector = map(float, vector)
    except ValueError:
        print vector

# Now create the dictionaries we need, one dictionary for
each influx
    for i in range(1, 29+1):
        label = labels[i-1]
        try:
            label = float(label)
        except ValueError:
            #
            #
            #
            print Label, "=>"
            print Line
            continue # this label for this influx is not
numeric

        training_data[i][0].append(vector) # add a row to
feature vectors
        training_data[i][1].append(float(label)) # add one
label

    print("checking duplicate lines...")
    for k, v in reports.iteritems():
        if len(v) > 1:
            print("line number: {}".format(v))
    print("Done.")
    return training_data

def one_hot_encode_features(training_data):
    """Use one-hot encoder to represent categorical features

    Feature from 1 to 7 are categorical features:
    Species, reactor, nutrient, oxygen, engineering method, MFA and
    extra energy

    """
    import numpy

```

```

    encoded_training_data, encoders = {}, {}
    for vid, (vectors, targets) in training_data.iteritems():
        encoder = preprocessing.OneHotEncoder()
        vectors = numpy.array(vectors) # 2-D array
        encoded_categorical_features =
encoder.fit_transform(vectors[:, 0:6+1])
        encoded_categorical_features =
encoded_categorical_features.toarray()
        encoded_vectors =
numpy.hstack((encoded_categorical_features, vectors[:, 6+1:]))
        encoded_training_data[vid] = (encoded_vectors, targets)
        encoders[vid] = encoder
    return encoded_training_data, encoders

def standardize_features(training_data):
    """Standarize feature vectors for each influx

    Later, a new feature vector X for i-th influx can be normalized as:
    Scalers[i].transform(X)

    """
    std_training_data, scalers = {}, {}
    for vid, (vectors, labels) in training_data.iteritems():
        vectors_scaled = preprocessing.scale(vectors)
        std_training_data[vid] = (vectors_scaled, labels)

        scalers[vid] = preprocessing.StandardScaler().fit(vectors)

    return std_training_data, scalers

def train_model(training_data, Parameters):
    """Train a regression model for each of the 29 influxes

    Returns
    =====
    Models: dict, keys are influx indexes and values are regression
models
    parameters: dict, keys are intergers 1 to 29, values are dicts,
such as
                'epsilon': 0.01, 'c': 100.0, 'gamma': 0.001, 'kernel':
'rbf'

    Notes
    =====
    Parameters are not in use. Now use same parameters for all v's.

    """
    models = {}
    for i in range(1, 29+1):
        vectors, label = training_data[i]

```





```

    """Do a grid search to find best params for the given model.

    :param training_data: A dict with keys as v, and values as
    [vectors, label].
    :param model_gen: A RegressionModel generator.
    :param params: All parameters the grid search needs to find. It's
    a subset
        of all the optional params on each model. i.e. for
    KNeighborsRegressor
        model, it's a subset of
        ...
        {
            "n_neighbors": [1, 5, 10, ...],
            "weights": ["uniform", ...],
            "algorithm": ["auto", ...],
            "leaf_size": [30, 50, ...],
            "p": [2, 5, ...],
            "metric": ["minkowski", ...],
            "metric_params": [...],
        }
        ...

    """
    print("model: {}".format(model_gen))
    print("\v\tscoring\tbest_score\tbest_params")
    for i in range(1, 29 + 1):
        vectors, label = training_data[i]
        model = model_gen()
        for scoring in SCORINGS:
            clf = grid_search.GridSearchCV(model.model, params,
scoring=scoring, n_jobs=CORE_NUM, cv=FOLDS)
            clf.fit(vectors, label)
            print("{}\t{}\t{}\t{}".format(i, scoring, clf.best_score_,
                                            clf.best_params_))

def grid_search_tasks(std_training_data):
    """One function to run grid search on different regressors

    CORE_NUM: int, number of CPU cores to be used
    FOLDS: int, number of folds for cross validate

    """
    import numpy
    knn_model_gen = RegressionModelFactory("KNeighborsRegressor",
n_neighbors=10, weights="distance")
    svr_model_gen = RegressionModelFactory("SVR", kernel="linear",
C=10, epsilon=0.2)
    dtree_model_gen = RegressionModelFactory("DecisionTreeRegressor",
random_state=0)

    KNN_PARAMS = {

```

```

    "n_neighbors": range(1, 16),
    "weights": ["distance", "uniform"],
    "algorithm": ["ball_tree", "kd_tree", "brute"],
    "metric": ["euclidean", "chebyshev", "minkowski", ]
}

SVR_PARAMS = {
    "C": 10.0 ** numpy.arange(-4,4),
    "epsilon": [0., 0.0001, 0.001, 0.01, 0.1], # experience:
epsilon>=0.1 is not good.
    "kernel": [
        "linear",
#         "rbf",
#         "poly", # polynomial kernel sucks. Never use it.
#         "sigmoid",
        # "precomputed"
    ],
#     "degree": [5,], # because polynomial kernel sucks. Never use
it.
    "gamma": 10.0 ** numpy.arange(-4, 4),
}

DTREE_PARAMS = {
#     "criterion": ["mse"],
    "splitter": ["best", "random"],
    "min_samples_split": range(2, 16),
    "min_samples_leaf": range(1, 16),
    "max_features": ["sqrt", "log2"],
#     "random_state": [0, 1, 10, 100],
}

SCORINGS = ["mean_squared_error",
#           "mean_absolute_error"
]

TRAINING_PARAMS = [
#     (knn_model_gen, KNN_PARAMS),
    (svr_model_gen, SVR_PARAMS),
#     (dtree_model_gen, DTREE_PARAMS),
]

FOLDS = 10
CORE_NUM = 32

[grid_search_cv(std_training_data, k, v, SCORINGS, CORE_NUM, FOLDS)
for k, v in TRAINING_PARAMS]

def cv_tasks(std_training_data, Folds, N_jobs, Label_scalers,
Parameters):
    """Cross-validation on all v's

```

```

:param Folds: number of CV folds
:param N_jobs: number of CPU cores
:param label_scaler: dict, keys are fluxes and values are sklearn
scaler objects
:param Parameters: dict, keys are fluxes and values are parameters
for all fluxes

"""
import sklearn
knn_model_gen = RegressionModelFactory("KNeighborsRegressor",
n_neighbors=10, weights="distance")
# dtree_model_gen = RegressionModelFactory("DecisionTreeRegressor",
random_state=0)

if Parameters != None: # need to create one instances for one flux
    svr_model_gen = {}
    for i in xrange(1, 29+1):
        svr_model_gen[i] = RegressionModelFactory("SVR",
** (Parameters[i]))
    else: # same set of parameters for all SVR models.
        svr_model_gen = RegressionModelFactory("SVR", kernel="linear",
C=0.1, epsilon=0.01)

Classifier_models = [
#     knn_model_gen,
    svr_model_gen,
#     dtree_model_gen,
]

[cross_validation_model(std_training_data, m, Folds, N_jobs) for m
in Classifier_models]

def svr_training_test(std_training_data, Parameters,
Label_scalers=None):
    """Test SVR training accuracy

    Parameters
    =====
    std_training_data: dict, keys are vID, values are tuples (vector,
label)
                                each vector is 2-D array and label is a 1-D
array

    Parameters: dict, keys are intergers 1 to 29, values are dicts,
such as
                'epsilon': 0.01, 'c': 100.0, 'gamma': 0.001, 'kernel':
'rbf'

    Label_scalers: dict, keys are int 1 to 29, value sare sklearn
scaler objects

```

```

"""
from numpy import square, mean, sqrt
Models = train_model(std_training_data, Parameters)
Influxes = {}

for vID, Model in Models.iteritems():
    (Vectors_for_this_v, Label_for_this_v) = std_training_data[vID]
    Label_predict = Model.predict(Vectors_for_this_v)
    if Label_scalers != None:
        Label_predict =
Label_scalers[vID].inverse_transform(Label_predict)
        Label_for_this_v =
Label_scalers[vID].inverse_transform(Label_for_this_v)

    MSE = Label_predict - Label_for_this_v
#     if vID==2:
#         print Label_predict
#         print Label_for_this_v
#         print MSE
#     MSE = sqrt(mean(square(MSE)))

    print "\t&\t".join(map(str, [vID, MSE
, max(Label_for_this_v), min(Label_for_this_v
)]) + "\t\\\\"

#     for i, j in enumerate(list(MSE)):
#         print i+1, j
#         print list(square(MSE))
#         print Label_predict
#         break

def _validate_training_data(training_data):
    reports = []
    for _, d in training_data.iteritems():
        report = defaultdict(list)
        vectors = d[0]
        for i, v in enumerate(vectors):
            key = ", ".join(map(str, v))
            report[key].append(i)
        # only keep duplicated rows
        report_ = {k: v for k, v in report.iteritems() if len(v) > 1}
        reports.append(report_)

    return reports

def label_std(Training_data, Method="Norm"):
    """standardize the labels in training data
    training_data: a dict, keys are EMPs (e.g., v1, v2, etc.),
        values are 2-tuples (Feature, Label), where
        Feature is a 2-D list, each sublist is 24-D feature
vector for one sample
        and
        Label is a 1-D list, labels for all samples.

```

Label\_scalers: dict, keys are vIDs and values are sklearn.preprocessing.MinMaxScaler instances for 29 influxes

sklearn's preprocessing MixMaxScaler does column-wise Minmax scaling.

Since influxes have different number of intances, we must loop thru the 29.

```
"""
import sklearn
Label_scaled_data = {}
Label_scalers = {}
if Method == "None": # No label std needed
    return Training_data, None

for vID, (Vector, Label) in Training_data.iteritems():
    if Method == "Norm":
        Label_scaler =
sklearn.preprocessing.StandardScaler().fit(Label)
#           Label_scaled = sklearn.preprocessing.scale(Label) #
option 1 of standarization
        elif Method == "MinMax":
            Label_scaler =
sklearn.preprocessing.MinMaxScaler().fit(Label)
        else:
            print "Unrecognized label standarization method "
            Label_scaled = Label_scaler.transform(Label) # Option 2,
MinMax scaler

            Label_scaled_data[vID] = (Vector, Label_scaled)
            Label_scalers[vID] = Label_scaler

return Label_scaled_data, Label_scalers

def load_parameters(File):
    """Load a parameter file from grid search print out

    The format of grid search print out:
    checking duplicate lines
    model: SVR ({'epsilon': 0.2, 'C': 10, 'kernel': 'linear'})
    v      scoring      best_score best_params
    1      mean_squared_error      -0.00462588529703      {'epsilon': 0.01,
'C': 100.0, 'gamma': 0.001, 'kernel': 'rbf'}
    2      mean_squared_error      -0.0103708930608 {'epsilon': 0.01, 'C':
1000.0, 'gamma': 0.0001, 'kernel': 'rbf'}
    3      mean_squared_error      -0.00713773093885      {'epsilon': 0.01,
'C': 1000.0, 'gamma': 0.0001, 'kernel': 'rbf'}
    4      mean_squared_error      -0.0115793576617 {'epsilon': 0.001,
'C': 1000.0, 'gamma': 0.0001, 'kernel': 'rbf'}
"""
```

```

import re
Parameters = {}
with open(File, 'r') as F:
    F.readline() # Skip first line
    F.readline() # Skip second line
    F.readline() # Skip 3rd line
    for Line in F.readlines():
        [v, _, _, Parameter] = Line.split("\t")
        v = int(v)
        exec "Parameter = " + Parameter
        Parameters[v] = Parameter

return Parameters

def test_label_std():
    """Test the accuracy on labels using different label std methods

    The 3 methods are: no std, normalization, MinMax.
    We will study RMSE under different normalization
    """
    training_data = read_spreadsheet("wild_and_mutant.csv")
    training_data = shuffle_data(training_data)
    encoded_training_data, encoders =
one_hot_encode_features(training_data)
    std_training_data, Feature_scalers =
standardize_features(encoded_training_data)

    Parameters = load_parameters("svr_both_rbf_shuffle.log")

    for Std_method in ["None", "Norm", "MinMax"]:
        final_training_data, Label_scalers =
label_std(std_training_data, Method=Std_method) # standarize the
labels/targets as well.

#     grid_search_tasks(std_training_data)
#     cv_tasks(std_training_data, 10, 32)
#     svr_training_test(final_training_data, Parameters,
Label_scalers=Label_scalers)

def prepare_data(Datasheet, Parameter_file=None,
Label_std_method="MinMax"):
    """Prepare all data including scaling

    Patermeters
    =====
    Datasheet: str, full path to database spreadsheet file
    Parameters_file: str, full path to file that defines best
parameters for different v.
    Label_std_method: str, label preprocessing method, one in ["None",
"Norm", "MinMax"]
    Feature_std_method: str, feature preprocessing method, currentllyl
not used

```

```

"""
    Training_data = read_spreadsheet("wild_and_mutant.csv")
    Training_data = shuffle_data(Training_data)
    Encoded_training_data, Encoders =
one_hot_encode_features(Training_data)
    Std_training_data, Feature_scalers =
standardize_features(Encoded_training_data)

    if Parameter_file != None:
        Parameters = load_parameters(Parameter_file)
    else:
        Parameters = None

    Final_training_data, Label_scalers = label_std(Std_training_data,
Method=Label_std_method) # standarize the labels/targets as well.

    return Final_training_data, Feature_scalers, Label_scalers,
Encoders, Parameters

if __name__ == "__main__":
#     test_label_std()
#     exit()

    Datasheet = "wild_and_mutant.csv"
    Parameter_file = "svr_both_rbf_shuffle.log"
    Training_data, Feature_scalers, Label_scalers, Encoders,
Parameters\
    = prepare_data(Datasheet, Parameter_file=Parameter_file,
Label_std_method="MinMax")

#     grid_search_tasks(std_training_data)
#     cv_tasks(Training_data, 10, 4, Label_scalers, Parameters)

#     reports = _validate_training_data(std_training_data)
#     for i, report in enumerate(reports, 1):
#         print("v = {}, duplicate data index = {}".format(i,
report.values()))

    models = train_model(Training_data, Parameters)
    cPickle.dump(models, open("models_svm.p", "wb"))
    cPickle.dump(Feature_scalers, open("feature_scalers.p", "wb"))
    cPickle.dump(Encoders, open("encoders.p", "wb"))
    cPickle.dump(Label_scalers, open("label_scalers.p", "wb"))

#     cPickle.dump(training_data, open("training_data.p", "wb"))
#     cPickle.dump(encoded_training_data,
open("encoded_training_data.p", "wb"))
#     cPickle.dump(std_training_data, open("std_training_data.p",
"wb"))

```

---

## libflux.py

```
def quadprog_adjust(Substrates, Fluxes, Boundary_dict, Debug=False,
Label_scalers=None):
    """adjust values from ML

    Parameters
    =====
    Substrates: OrderedDict, keys as integers and values as floats,
e.g., {1:0.25, 2:0, 3:0.75, ...}
    Fluxes: Dict, keys as integers and values as floats, e.g., {1:99.5,
2:1.1, ...}
    Debug: Boolean, True for showing debug info and False (default)
for no.
    Label_scaler: sklearn.preprocessing.standardScaler
or .MinMaxScaler
                Forward transform is from fluxes in true range to
scaled range
                Inverse transform is from scaled range to true range
    Boundary_dict: Upper boundaries and lower boundaries for 29 fluxes,
depending on user inputs,
                e.g., {"lb29":999, "ub8":50}, populate ub and lb
inequalities from them

    Returns
    =====
    Solution: Dict, keys as integers and values as floats, e.g.,
{1:99.5, 2:1.1, ...}

    Notes
    =====
    In Substrates, the mapping from keys to real chemicals is as
follows:
        1. Glucose
        2. Fructose
        3. Galactose
        4. Gluconate
        5. Glutamate
        6. Citrate
        7. Xylose
        8. Succinate
        9. Malate
        10. Lactate
        11. Pyruvate
        12. Glycerol
        13. Acetate
        14. NaHCO3
```



Formulation of quadratic problems in MATLAB optimization toolbox are different from that in cvxopt.  
 Here is a mapping between variables  
 \* H => P (the quadratic terms in objective function)  
 \* f => q (the linear terms in objective function)  
 \* A and eyes for boundaries => G (coefficients for linear terms in inequality constraints)  
 \* b, -lb, ub => h (coefficients for constant terms in inequality constraints)  
 \* Aeq => A  
 \* Beq => b

Unimplemented features:

1. Using scaled values for quadprog

Example

```

=====
>>> Substrates = {1:1, 2:0, 3:0, 4:0, 5:0, 6:0, 7:0, 8:0, 9:0,
10:0, 11:0, 12:0, 13:0, 14:0}
>>> Fluxes = {1: 100.0, 2: -2.7159, 3: 15.2254, 4: 17.7016, 5:
110.9973, 6: 91.8578, 7: 137.7961, 8: 91.1558, 9: -0.7373, 10: 94.1518,
11: 24.1126, 12: 21.231, 13: 2.8816, 14: 11.0324, 15: 10.1986, 16:
11.0324, 17: 79.4203, 18: 79.4203, 19: 67.9442, 20: 67.8806, 21:
79.3567, 22: 79.3567, 23: 64.0876, 24: 11.4761, 25: 70.0392, 26: -
1.2424, 27: 0.0059, 28: 23.2159, 29: 26.7451}
>>> import libflux
>>> libflux.quadprog_adjust(Substrates, Fluxes, {}, Debug=True)
>>> import cPickle
>>> Label_scalers = cPickle.load(open("label_scalers.p", "r"))
>>> libflux.quadprog_adjust(Substrates, Fluxes, {}, Debug=True,
Label_scalers = Label_scalers)
>>> libflux.quadprog_adjust(Substrates, Fluxes, {"ub1":50},
Debug=True, Label_scalers = Label_scalers)

```

"""

```

import numpy
import cvxopt, cvxopt.solvers

```

```

Substrate2Index= {"glucose":1, "galactose":3, "fructose":2,
"gluconate":4, "glutamate":5, "citrate":6, "xylose":7, "succinate":8,
"malate":9, "lactate":10, "pyruvate":11, "glycerol":12, "acetate":13}

```

```

Ubs = numpy.array([[100,99.5,99.3,99.3,216.6,
196.2,232,213.1,135,151.4,
113.7,94.1,41.2,47.5,71,
47.5,189,189,189,194,
194,194,181.5,55,148,
193.2,151,149.8,104.2043714]])

```

```

Ubs = Ubs.transpose() # turn it into column vector, 29x1

```

```

Lbs = numpy.array([[0, -99.9, -51.5, -51.5, -13.5,
                  -23.3, -36, -7.9, -144, 0,
                  0, -33, -94.4, -2, -6.6,
                  -2, 0, -0.1, -0.1, 0,
                  -105, -106, -144.3, 0, 0,
                  0, -100, -67.60986805, -13.5]])
Lbs = Lbs.transpose() # turn it into column vector, 29x1

Aineq_bound, Bineq_bound =
populate_boundary_inequalities(Boundary_dict)

Aineq = numpy.zeros((12+1, 29+1)) # the plus 1 is to tackle MATLAB
1-index
Aineq[1,1] = 1; Aineq[1,2] = -1; Aineq[1,10] = -1;
# Aineq[2,2] = 1; Aineq[2,3] = -1; Aineq[2,15] = 1; Aineq[2,16] = 1;
# Aineq[3,3] = 1; Aineq[3,4] = 1; Aineq[3,5] = -1; Aineq[3,14] = 1;
Aineq[3,15] = 1; Aineq[3,16] = 1; Aineq[3,25] = 1;
Aineq[2,2] = 1; Aineq[2,3] = -1; Aineq[2,16] = 2;
Aineq[3,3] = 1; Aineq[3,4] = 1; Aineq[3,5] = -1; Aineq[3,14] = 1;
Aineq[3,25] = 1;
Aineq[4,5] = 1; Aineq[4,6] = -1;
# Aineq[5,6] = 1; Aineq[5,7] = -1; Aineq[5,28] = -1;
Aineq[6,7] = 1; Aineq[6,8] = -1; Aineq[6,25] = 1; Aineq[6,27] = -1;
Aineq[6,29] = 1;
Aineq[7,8] = 1; Aineq[7,9] = -1; Aineq[7,17] = -1; Aineq[7,24] = -1;
Aineq[7,26] = -1;
Aineq[8,13] = 1; Aineq[8,14] = -1;
Aineq[9,16] = 1; Aineq[9,15] = -1;
Aineq[10,19] = 1; Aineq[10,20] = -1;
Aineq[11,23] = 1; Aineq[11,17] = -1; Aineq[11,28] = 1;
Aineq[12,21] = -1; Aineq[12,22] = 1;
Aineq = Aineq[1:, 1:] # convert 1-index to 0-index
Aineq = -1 * Aineq # because in standardized formulation, it's
Ax<=b but in our paper it is Ax>=b

# if Label_scalers == None: # if flux in their true range instead
of scaled range
# Aineq = numpy.vstack([Aineq, -numpy.eye(29), numpy.eye(29)])
# add eye matrixes for Lbs and Ubs

if not Aineq_bound == None :
    Aineq = numpy.vstack([Aineq, Aineq_bound])
else:
    Aineq = numpy.matrix(Aineq)

bineq = numpy.zeros((12+1, 1+1))
bineq[2,1]= 100 * Substrates[Substrate2Index["fructose"]]
bineq[6,1]= 100 * Substrates[Substrate2Index["pyruvate"]]
bineq[10,1] = 100 * Substrates[Substrate2Index["glutamate"]]
bineq = bineq[1:, 1:] # convert 1-index to 0-index

```

```

#   if Label_scalers == None: # if flux in their true range instead
of scaled range
#       bineq = numpy.vstack([bineq, -Lbs, Ubs])
    if not Bineq_bound == None:
        bineq = numpy.vstack([bineq, Bineq_bound])
    else:
        bineq = numpy.matrix(bineq)

Aeq = numpy.zeros((10+1, 29+1))
Aeq[1,1] = 1;
Aeq[2,3] = 1; Aeq[2,4] = -1;
Aeq[3,11] = 1; Aeq[3,12] = -1; Aeq[3,13] = -1;
Aeq[4,14] = 1; Aeq[4,16] = -1;
Aeq[5,10] = 1; Aeq[5,11] = -1; Aeq[5,25] = -1;
Aeq[6,18] = 1; Aeq[6,17] = -1;
Aeq[7,15] = 1; Aeq[7,12] = -1; Aeq[7,14] = 1;
Aeq[8,24] = 1; Aeq[8,18] = -1; Aeq[8,19] = 1;
Aeq[9,22] = -1; Aeq[9,23] = 1; Aeq[9,24] = -1; Aeq[9,29] = 1;
Aeq[10,20] = 1; Aeq[10,24] = 1; Aeq[10,21] = -1;
Aeq = Aeq[1:, 1:] # convert 1-index to 0-index
Aeq = numpy.matrix(Aeq)
#   Aeq = Aeq.transpose().tolist()

    beq = numpy.zeros((10+1,1+1))
    beq[1,1] = 100 * (Substrates[Substrate2Index["glucose"]] +
Substrates[Substrate2Index["galactose"]])
    beq[2,1] = -100 * Substrates[Substrate2Index["glycerol"]]
    beq[5,1] = -100 * Substrates[Substrate2Index["gluconate"]]
    beq[6,1] = 100 * Substrates[Substrate2Index["citrate"]]
    beq[7,1] = 100 * Substrates[Substrate2Index["xylose"]]
    beq[9,1] = 100 * Substrates[Substrate2Index["malate"]]
    beq[10,1] = -100 * Substrates[Substrate2Index["succinate"]]
    beq = beq[1:, 1:] # convert 1-index to 0-index
    beq = numpy.matrix(beq)

    if Label_scalers == None:
        P = numpy.eye((29))
        q = [[Fluxes[i] for i in range(1, 29+1)]]
    else: # convert non-scaled fluxes into [0,1]
        P = numpy.square(numpy.diag([Label_scalers[i].scale_ for i in
range(1, 29+1)]))
        q = [[Label_scalers[i].scale_**2 * Fluxes[i] for i in range(1,
29+1)]]

        if Debug:
#           print P
            for i in range(1,29+1):
                pass

        q = -1*numpy.array((q)).transpose() # -1 is because -v_i but f.T*x
in standard quadprog formalization

#   print map(numpy.shape, [Aineq, bineq, Aeq, beq, P, q])

```

```

# print map(type, [Aineq, bineq, Aeq, beq, P, q])
# [bineq] = map(cvxopt.matrix, [bineq])
# [beq] = map(cvxopt.matrix, [beq])
# [Aineq, bineq, Aeq, beq] = map(cvxopt.matrix, [Aineq, bineq, Aeq,
beq])

[Aineq, bineq, Aeq, beq, P, q] = map(cvxopt.matrix, [Aineq, bineq,
Aeq, beq, P, q])

cvxopt.solvers.options['show_progress'] = False

Solv = cvxopt.solvers.qp(P, q, Aineq, bineq, Aeq, beq)

Solution = Solv['x']

Solution = numpy.array(Solution)[: ,0] # conversion from cvxopt's
matrix to numpy array

if Debug:

    numpy.set_printoptions(precision=4, suppress=True)

    print "<pre>"
    print "".join([" V", " Adjusted ", " Predicted ", " Diff
", " Diff% ", " Diff%Rg  "])
    for Idx, Value in enumerate(Solution):
#         print type((Ubs-Lbs)[Idx][0])
        Diff = Value-Fluxes[Idx+1]
        print
"{0:2d}{1:10.3f}{2:10.3f}{3:10.3f}{4:8.1f}{5:8.1f}".\
            format(Idx+1, Value, Fluxes[Idx+1], Diff,
Diff/Fluxes[Idx+1]*100, Diff/((Ubs-Lbs)[Idx][0])*100) # convert from
0-index to 1-index
        print "</pre>"

    Solution = {i+1: Solution[i] for i in xrange(29)} # turn from
numpy array to dict

    return Solution

def test(S):
    print S

def print_influxes(Influxes):
    """Print influxes

    Influxes: dict, keys are influx id, values are floats
    """

```

```

import sys
sys.stderr = sys.stdout

# print Influxes
print """\
<h2>Influx values based on given parameters:</h2>
"""># % len(Vector) #"\t".join(map(str, Vector))

print """\
<table border=0 border-spacing=5px>
  <tr>
    <td>

    ""

# for x in range(5):
#     print x

for ID, Value in Influxes.iteritems():
    print """\
    v%s = %.4f, <br>
    """ % (ID, Value)

print """\
  </td>
  <td>
    
  </td>
</tr>
</table>
"""

# for ID, Value in Influxes.iteritems():
#     print """\
#     v%s = %.4f, <br>
#     """ % (ID, Value)

def populate_boundary_inequalities(Boundary_dict, Debug=False):
    """
    Boundary_dict: Upper boundaries and lower boundaries for 29 fluxes,
    depending on user inputs,
                    e.g., {"lb29":999, "ub8":50}, populate ub and lb
    inequalities from them

    Aineq: X-by-29 binary matrix, where N is the number of Ubs and Lbs
    set by user
    Bineq: X-by-1 column vector

    for any  $v_j \leq p$ , there is  $A_{ineq}[i][j] == 1$  and  $B_{ineq}[j] == P$ 
    for any  $v_j \geq p$ , there is  $A_{ineq}[i][j] == -1$  and  $B_{ineq}[j] == -p$ 
    Note the inequalities are:  $Ax \leq B$ 

```

```

"""
import numpy
if Boundary_dict == {}:
    return None, None

Row_vectors = [] # must be 29 columns and X rows where X is the
number of Ubs and Lbs set by user
Boundary_column_vectors = [] # X rows and 1 column
for Polarity_Id, Bound_value in Boundary_dict.iteritems():
    Bound_type, Flux_ID = Polarity_Id[:2], int(Polarity_Id[2:])
    Row_vector = numpy.zeros(29)
    if Bound_type == "lb":
        Bound_value = -1*Bound_value
        Row_vector[Flux_ID-1] = -1.
    elif Bound_type == "ub":
        Row_vector[Flux_ID-1] = 1.
    else:
        print "wrong boundary"
    Row_vectors.append(Row_vector)
    Boundary_column_vectors.append(Bound_value)
#    print "<br>", Bound_type, Flux_ID, Bound_value

Aineq = numpy.vstack(Row_vectors)
Bineq = numpy.vstack(Boundary_column_vectors)

if Debug:
    print "<pre>"
    print Aineq
    print Bineq
    print "</pre>"

return Aineq, Bineq

def process_boundaries(Form, Substrates):
    """Extract boundaries for fluxes from user input

    Form: cgi object
    Features: {}, empty dictionary by default

    Notes
    =====
    In Substrates, the mapping from keys to real chemicals is as
    follows:
        1. Glucose
        2. Fructose
        3. Galactose
        4. Gluconate
        5. Glutamate
        6. Citrate
        7. Xylose
        8. Succinate
        9. Malate

```

```

10. Lactate
11. Pyruvate
12. Glycerol
13. Acetate
14. NaHCO3

"""
import itertools
import cgi
Substrate2Index= {"glucose":1, "galactose":3, "fructose":2,
"gluconate":4, "glutamate":5, "citrate":6, "xylose":7, "succinate":8,
"malate":9, "lactate":10, "pyruvate":11, "glycerol":12, "acetate":13}
Feature_names = ["".join([Bound, ID]) for (Bound, ID) in
itertools.product(["lb", "ub"], map(str, range(1, 29+1))) ]

Features= {}
for Feature_name in Feature_names:
    Feature_value = Form.getfirst(Feature_name)
    if Feature_value:
#         print Feature_name, Feature_value
        Feature_value = cgi.escape(Feature_value)
        Features[Feature_name] = float(Feature_value) # convert
all string to numbers

if Substrates[Substrate2Index["acetate"]] == 0:
    Features["lb9"] = 0
if Substrates[Substrate2Index["lactate"]] == 0:
    Features["lb27"] = 0

for Feature_name in Feature_names:
    print """\
%s is %s,
"" % (Feature_name, Features[Feature_name])

return Features

def process_input(Features):
    """Process the result from CGI parsing to form feature vector
including substrate matrixi

Substrates: OrderedDict, keys as integers and values as floats
1. Glucose
2. Fructose
3. Galactose
4. Gluconate
5. Glutamate
6. Citrate
7. Xylose
8. Succinate
9. Malate
10. Lactate
11. Pyruvate

```

12. Glycerol
13. Acetate
14. NaHCO<sub>3</sub>

Feature vectors order: [Species, Reactor, Nutrient, Oxygen, Method, MFA, Energy, Growth\_rate, Substrate\_uptake\_rate] + ratio of 14 carbon sources in the order above

```

"""

Num_substrates = 14 # excluding other carbon
# Generate substrate matrix
import collections
Substrates = collections.OrderedDict([(i,0) for i in range(1,
Num_substrates+1)]) # substrate values, initialization
Substrates[int(Features["Substrate_first"])] +=
Features["Ratio_first"]
Substrates[int(Features["Substrate_sec"])] += Features["Ratio_sec"]

# Form the feature vector
Vector = [Features[Feature_name] for Feature_name in ["Species",
"Reactor", "Nutrient", "Oxygen", "Method", "MFA", "Energy",
"Growth_rate", "Substrate_uptake_rate"]]
Vector += [Substrates[i] for i in range(1, Num_substrates+1)]
Vector.append(Features["Substrate_other"]) # Other carbon source

# Print input check
import clp
DB = clp.process_species_db("SI_1_species_db.csv")
P = clp.species_db_to_constraints(DB)
if not clp.input_ok(P, Vector):
    print "<p><font color=\"red\">The input data might violate the
oxygen, substrate uptake rate or carbon sources of the selected
species. Therefore, the following prediction may not be biologically
meaningful. Please check your inputs!</font></p>"

# Print debug info

Substrate_names = ["glucose", "fructose", "galactose", "gluconate",
"glutamate", "citrate", "xylose", "succinate", "malate", "lactate",
"pyruvate", "glycerol", "acetate", "NaHCO3"]
Substrate_dict = collections.OrderedDict([(i+1,Name) for i, Name
in enumerate(Substrate_names)])
print "<p>Feature Vector (pre-one-hot-encoding and pre-scaling):",
Vector, "</br>"
print "in which the substrates ratios are:",
[(Substrate_dict[Index],Ratio) for Index, Ratio in
Substrates.iteritems()],
print "<br>Feature vector size is ", len(Vector), "</p>"

return Vector, Substrates

```



```

def rule_adjust(Influxes, Substrates):
    """Adjust influxes values using rules
    """

    Substrate2Index= {"glucose":1, "galactose":3, "fructose":2,
"gluconate":4, "glutamate":5, "citrate":6, "xylose":7, "succinate":8,
"malate":9, "lactate":10, "pyruvate":11, "glycerol":12, "acetate":13}

    #Step 1: Compute dependent influxes
    #   Influxes[1] = 100 * Substrates[Substrate2Index["glucose"]]
    #   Influxes[13] = Influxes[11] - Influxes[12]
    #   Influxes[16] = Influxes[14]
    #   Influxes[25] = Influxes[10] - Influxes[11] + 100 *
Substrates[Substrate2Index["gluconate"]]
    #   Influxes[18] = Influxes[17] + 100 *
Substrates[Substrate2Index["citrate"]]
    #   Influxes[15] = Influxes[12] - Influxes[14] + 100 *
Substrates[Substrate2Index["xylose"]]
    #   Influxes[24] = Influxes[18] - Influxes[19]
    #   Influxes[21] = Influxes[20] + Influxes[24] + 100 *
Substrates[Substrate2Index["succinate"]]
    #   Influxes[22] = Influxes[21]
    #   Influxes[29] = Influxes[22] + Influxes[24] - Influxes[23] + 100 *
Substrates[Substrate2Index["malate"]]

    # Step 2: Correct flux values
    if Substrates[Substrate2Index["acetate"]] != 0:
        Influxes[9] = -100 * Substrates[Substrate2Index["acetate"]]
    if Substrates[Substrate2Index["lactate"]] != 0:
        Influxes[27] = -100 * Substrates[Substrate2Index["lactate"]]

    return Influxes

def predict(Vector, Substrates, Boundary_dict):
    """ Predict and adjust all influx values

    Vector: 1-D list of floats, the feature vector, including
substrate matrix, size = 24
    Substrates: dict of floats, 1-indexed part of Feature_vector,
ratio of substrates
    Boundary_dict: Upper boundaries and lower boundaries for 29 fluxes,
depending on user inputs,
        e.g., {"lb29":999, "ub8":50}, populate ub and lb
inequalities from them
        If no boundary set by user, it can be an empty
dictionary
    Models: dict of models, 1-indexed, 29 models for 29 influxes.

    Calls adjust_influxes() to compute dependent influxes.
    """
    import cPickle
    import time

```

```

import collections
import sys

Models = cPickle.load(open("models_svm.p", "r"))
Feature_scalers = cPickle.load(open("feature_scalers.p", "r"))
Encoders = cPickle.load(open("encoders.p", "r"))
Label_scalers = cPickle.load(open("label_scalers.p", "r"))

print "<p>Models, feature and label Scalers and one-hot Encoder
loaded.</p>"
# Models: dict, keys are influx indexes and values are regression
models

T = time.clock()
Influxes = {}
# Influxes =
{vID:Model.predict(Scalers[vID].transform(Vector))[0] for vID,
Model in Models.iteritems()}# use dictionary because influx IDs are
not consecutive

print "Standardized (zero mean and unit variance) influx
prediction from ML:"
for vID, Model in Models.iteritems():
    Vector_local = list(Vector) # make a copy; o/w Vector will be
changed in one-hot encoding and standarization for different models
    One_hot_encoding_of_categorical_features =
Encoders[vID].transform([Vector[:6+1]]).toarray().tolist()[0] # one-
hot encoding for categorical features
# print len(One_hot_encoding_of_categorical_features), "\n"
    Vector_local = One_hot_encoding_of_categorical_features +
Vector_local[6+1:] # combine one-hot-encoded categorical features with
continuous features (including substrate matrix)
# print Vector_local, len(Vector_local)
    Vector_local = Feature_scalers[vID].transform(Vector_local) #
standarization of features
# print Vector_local
    Influx_local = Model.predict(Vector_local)[0] # prediction
    print "v{0:d}={1:.5f}, ".format(vID, Influx_local)
    Influx_local =
Label_scalers[vID].inverse_transform([Influx_local])[0]
    Influxes[vID] = Influx_local

Influxes = quadprog_adjust(Substrates, Influxes, Boundary_dict,
Label_scalers=Label_scalers, Debug=True)
Influxes = rule_adjust(Influxes, Substrates)

T = time.clock() -T

print_influxes(Influxes)

print ""</p>\

```

Using RBF kernel SVM as regressor. Parameters vary for different fluxes. For details, refer to [this document generated by grid search on SVM parameters](svr_both_rbf_shuffle.log).

Standardization and Regression done in %s seconds.

""" % T

return Influxes

---

## S2: Detailed information of case study (20 cases)

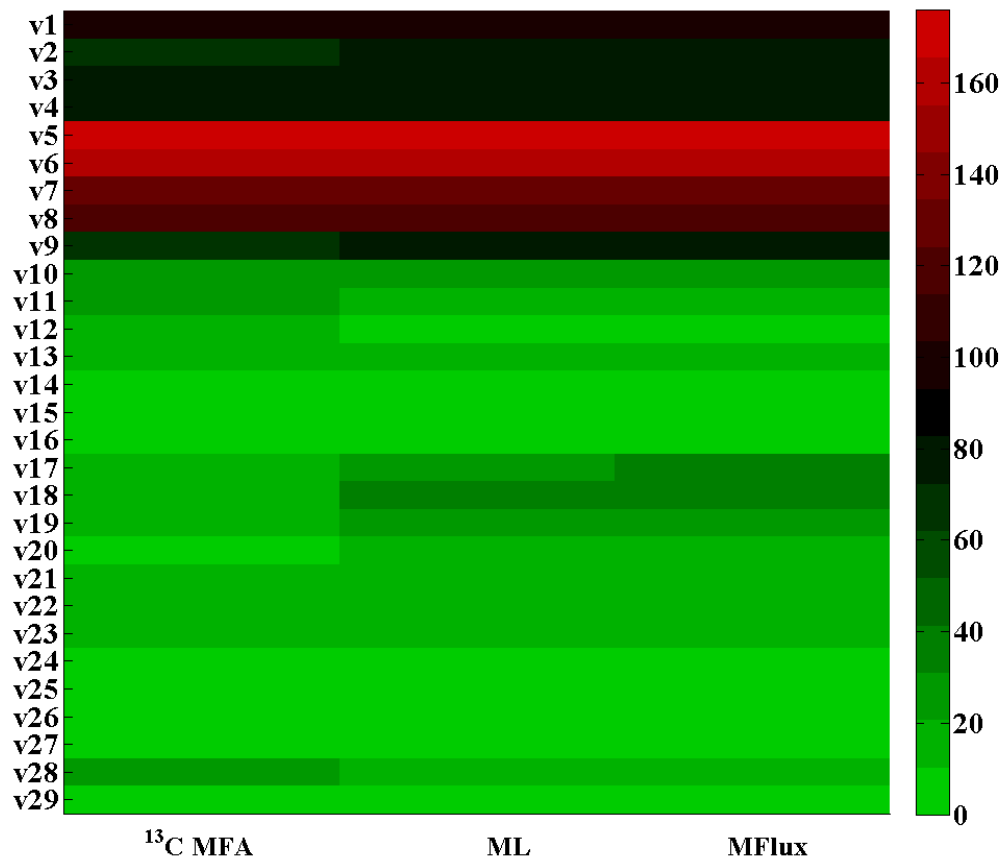
### Summary of 20 cases

Case number	Root mean squared error (RMSE)		
	Average flux	ML	MFlux
1	27.1	6.9	6.5
2	23.9	17.5	10.6
3	22.8	16.8	8.2
4	22.1	16.5	7.6
5	20.9	7.0	6.9
6	22.3	8.8	8.2
7	27.2	5.9	5.3
8	58.5	4.5	3.2

9	36.1	4.2	3.2
10	55.8	4.2	3.3
11	24.0	5.9	4.3
12	14.7	9.4	8.8
13	15.8	11.7	10.1
14	27.1	4.1	3.4
15	21.6	2.9	3.6
16	34.4	4.0	3.0
17	46.7	6.5	4.1
18	46.6	6.6	4.3
19	62.1	5.1	4.0
20	60.2	5.1	3.6
average	33.5	7.7	5.6

### Case 1

Reference: Crown SB, Long CP, Antoniewicz MR (2015) Integrated  $^{13}\text{C}$ -metabolic flux analysis of 14 parallel labeling experiments in *Escherichia coli*. *Metabolic Engineering* 28: 151-158.



Case 1 heat map

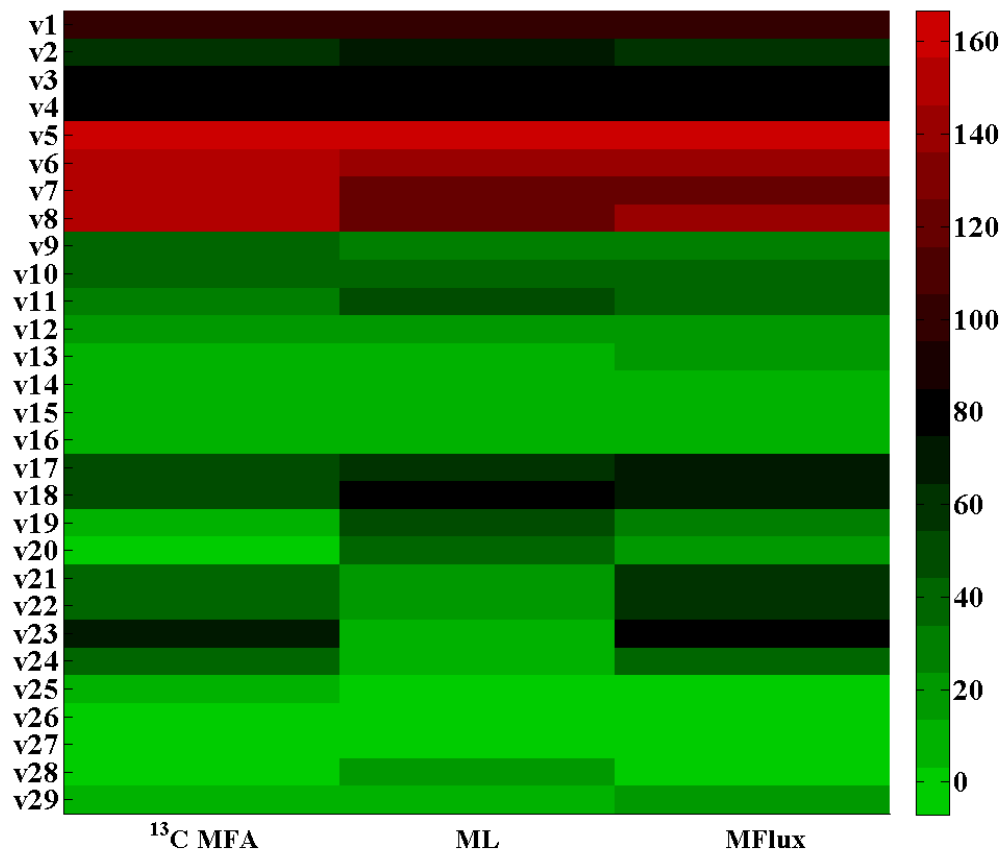
Case 1 results

	13C-flux	ML	MFlux
v1 EMP	100.0	100.0	100.0

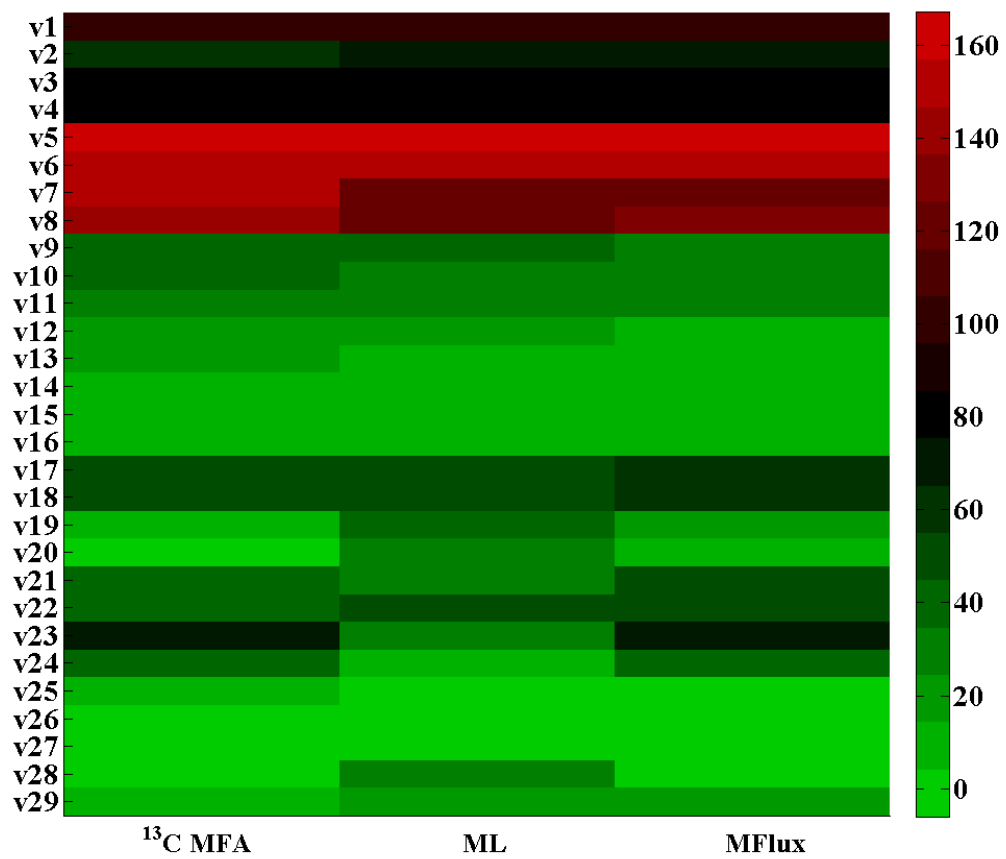
v2 EMP	71.6	73.3	73.3
v3 EMP	82.6	80.0	81.1
v4 EMP	82.6	79.9	81.1
v5 EMP	170.5	175.8	173.2
v6 EMP	157.6	164.6	164.6
v7 EMP	127.1	130.1	130.1
v8	114.0	116.6	116.6
v9	66.8	73.8	73.8
v10	26.8	26.8	24.2
v11	25.4	17.2	19.1
v12	11.5	8.7	8.3
v13	13.9	11.5	10.8
v14	7.1	5.8	5.8
v15	4.4	2.5	2.5
v16	7.1	5.7	5.8
v17 TCA cycle	19.0	25.6	33.5
v18 TCA cycle	19.0	41.1	33.5
v19 TCA cycle	16.6	30.5	30.2
v20 TCA cycle	4.9	19.3	16.0
v21 TCA cycle	10.9	19.0	19.3
v22 TCA cycle	13.7	18.5	19.3
v23 TCA cycle	10.9	12.8	20.3
v24 Glyoxylate	2.5	4.0	3.4
v25 ED	1.4	1.5	5.1
v26 ETOH	0.0	0.1	0.2
v27 LAC	0.0	0.0	0.0
v28	24.7	15.5	15.5
v29	5.3	1.4	2.4
RMSE		6.9	6.5

#### Case 2-4

Reference: Fong SS, Nanchen A, Palsson BO, Sauer U (2006) Latent Pathway Activation and Increased Pathway Capacity Enable *Escherichia coli* Adaptation to Loss of Key Metabolic Enzymes. Journal Of Biological Chemistry 281: 8024-8033.

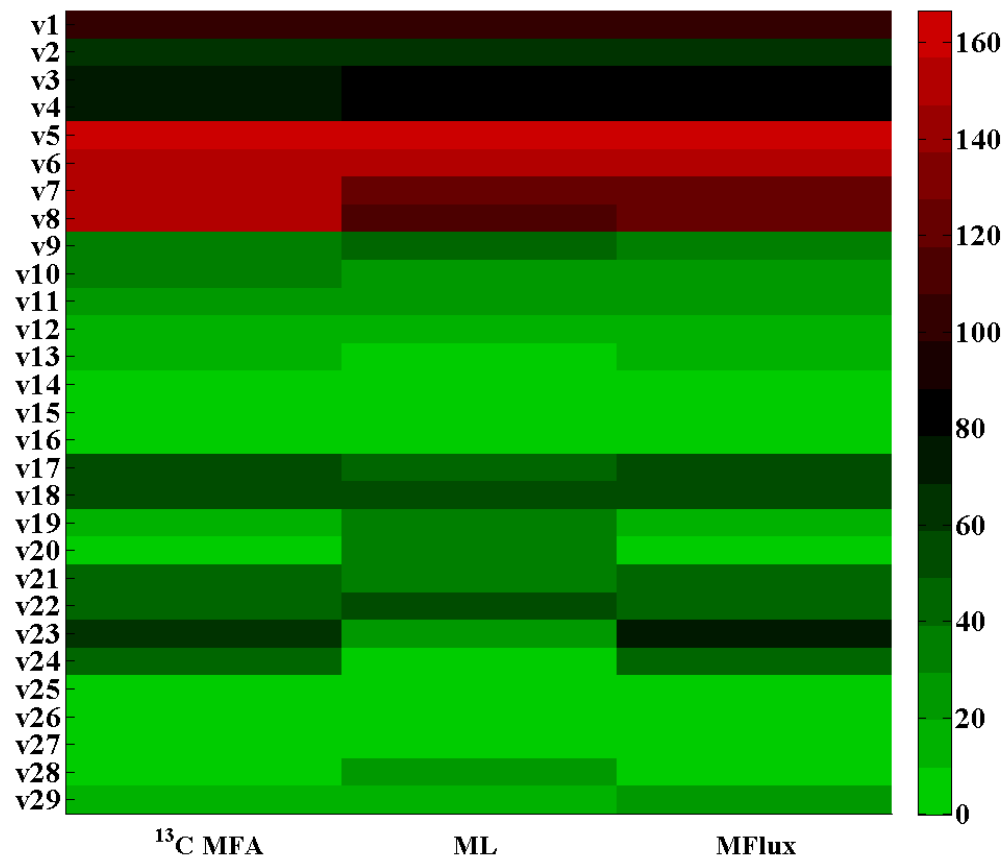


Case 2 heat map



Case 3 heat map





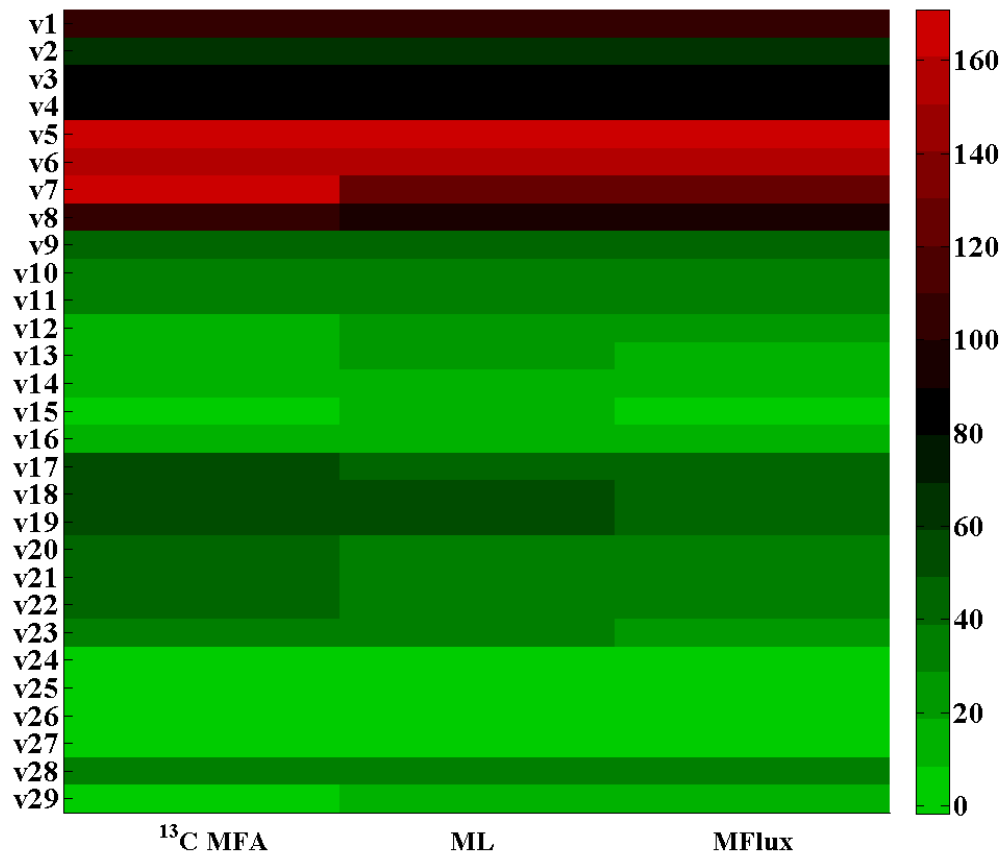
Case 4 heat map

## Case 2 - 4 results

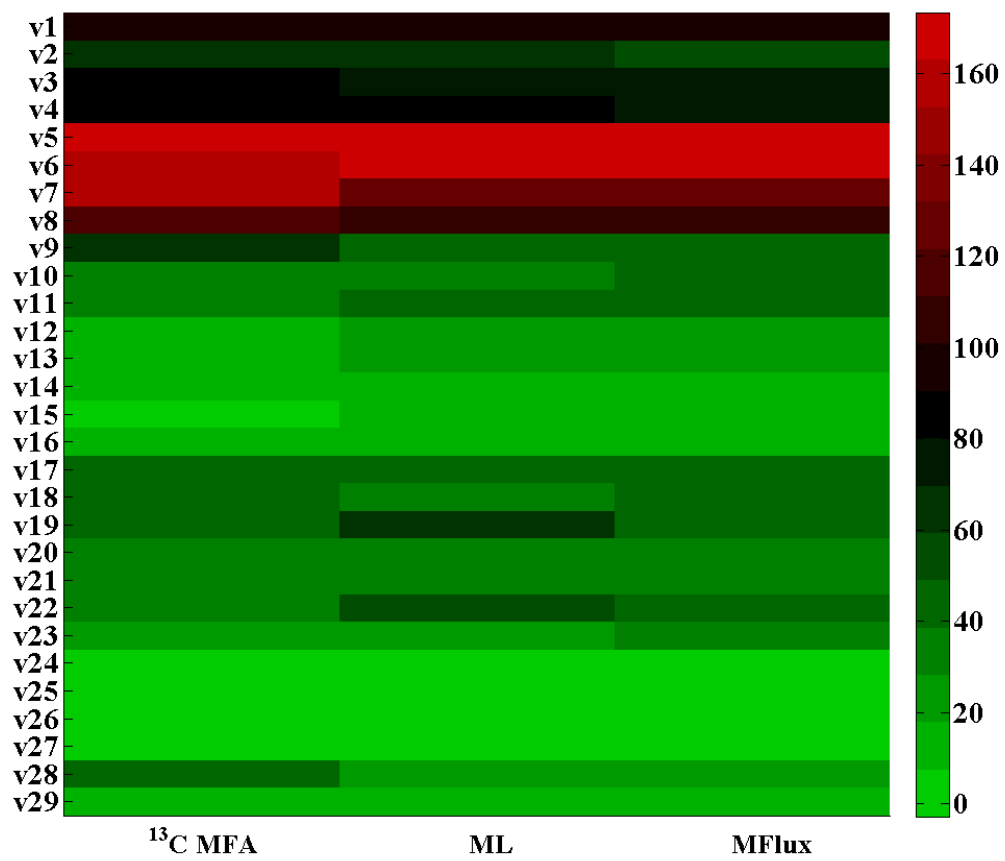
	case 2			case 3			case 4		
	13C-flux	ML	MFlux	13C-flux	ML	MFlux	13C-flux	ML	MFlux
v1 EMP	100.0	100.2	100.0	100.0	100.0	100.0	100.0	100.1	100.0
v2 EMP	62.0	65.0	61.0	61.0	67.6	67.6	61.0	67.8	67.8
v3 EMP	76.7	79.1	79.1	76.3	79.0	79.4	74.7	79.2	79.6
v4 EMP	76.7	79.2	79.1	76.3	79.8	79.4	74.7	80.0	79.6
v5 EMP	166.7	161.7	161.7	165.0	167.0	167.0	162.7	166.6	166.6
v6 EMP	156.0	142.0	142.0	151.0	156.1	156.1	149.0	155.4	155.4
v7 EMP	156.0	125.4	125.4	150.0	122.4	122.4	147.0	121.1	121.1
v8	149.0	118.7	139.5	144.0	117.8	128.1	147.0	116.3	126.0
v9	37.0	31.2	30.0	35.0	42.1	30.0	32.0	40.4	30.0
v10	38.0	35.1	39.0	39.0	29.0	27.4	38.0	28.9	27.3
v11	30.0	53.0	39.0	31.0	26.7	25.0	29.0	26.2	25.0
v12	16.7	18.9	18.2	15.0	16.1	12.8	13.3	16.1	12.7
v13	13.3	5.9	20.8	16.3	8.7	12.3	16.0	9.1	12.3
v14	10.0	7.5	9.1	8.8	5.9	6.4	8.0	5.9	6.4
v15	6.7	7.0	9.1	6.3	5.3	6.4	5.3	5.3	6.4
v16	10.0	7.4	9.1	8.8	5.3	6.4	8.0	5.3	6.4
v17 TCA cycle	53.0	61.5	69.5	51.0	48.8	58.1	54.0	47.9	56.0
v18 TCA cycle	53.0	78.1	69.5	51.0	53.0	58.1	54.0	49.4	56.0
v19 TCA cycle	11.0	44.1	29.5	11.0	40.0	18.1	11.0	37.8	16.0
v20 TCA cycle	0.0	39.3	17.0	0.0	34.0	9.9	0.0	33.0	8.9
v21 TCA cycle	42.0	49.5	57.0	40.0	32.3	49.9	43.0	31.4	48.9
v22 TCA cycle	42.0	46.3	57.0	40.0	50.0	49.9	43.0	49.1	48.9
v23 TCA cycle	73.0	40.9	79.0	70.0	25.8	69.4	67.0	25.3	69.1
v24 Glyoxylate	42.0	7.2	40.0	40.0	5.4	40.0	43.0	5.6	40.0
v25 ED	8.0	1.7	0.0	7.0	0.9	2.4	9.0	0.7	2.3
v26 ETOH	0.0	-0.1	0.0	0.0	0.2	0.0	0.0	0.3	0.0
v27 LAC	0.0	-0.1	-0.1	0.0	-0.1	-0.1	0.0	-0.1	-0.1
v28	-7.0	14.3	0.0	-6.0	26.3	0.0	0.0	25.6	0.0
v29	11.0	13.0	18.0	10.0	14.7	20.4	19.0	14.1	19.8
<b>RMSE</b>		<b>17.5</b>	<b>10.6</b>		<b>16.8</b>	<b>8.2</b>		<b>16.5</b>	<b>7.6</b>

## Case 5-7

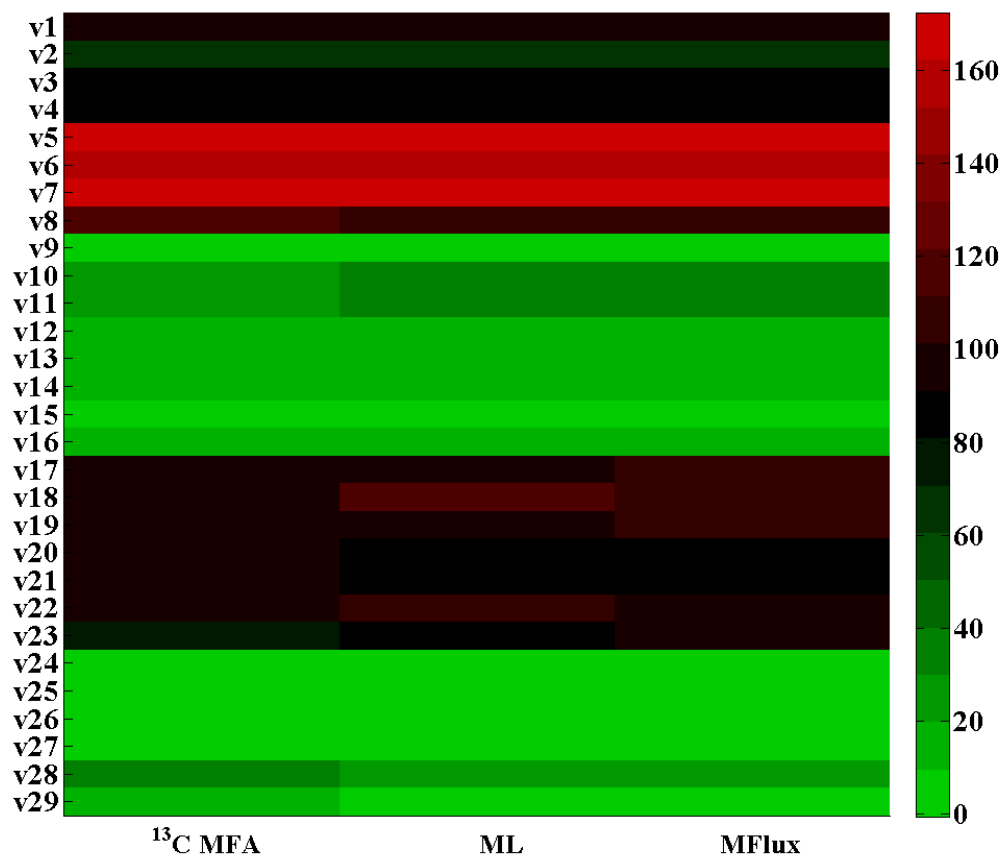
Reference: Tannler S, Decasper S, Sauer U (2008) Maintenance metabolism and carbon fluxes in *Bacillus* species. *Microbial Cell Factories* 7: 19.



Case 5 heat map



Case 6 heat map



Case 7 heat map

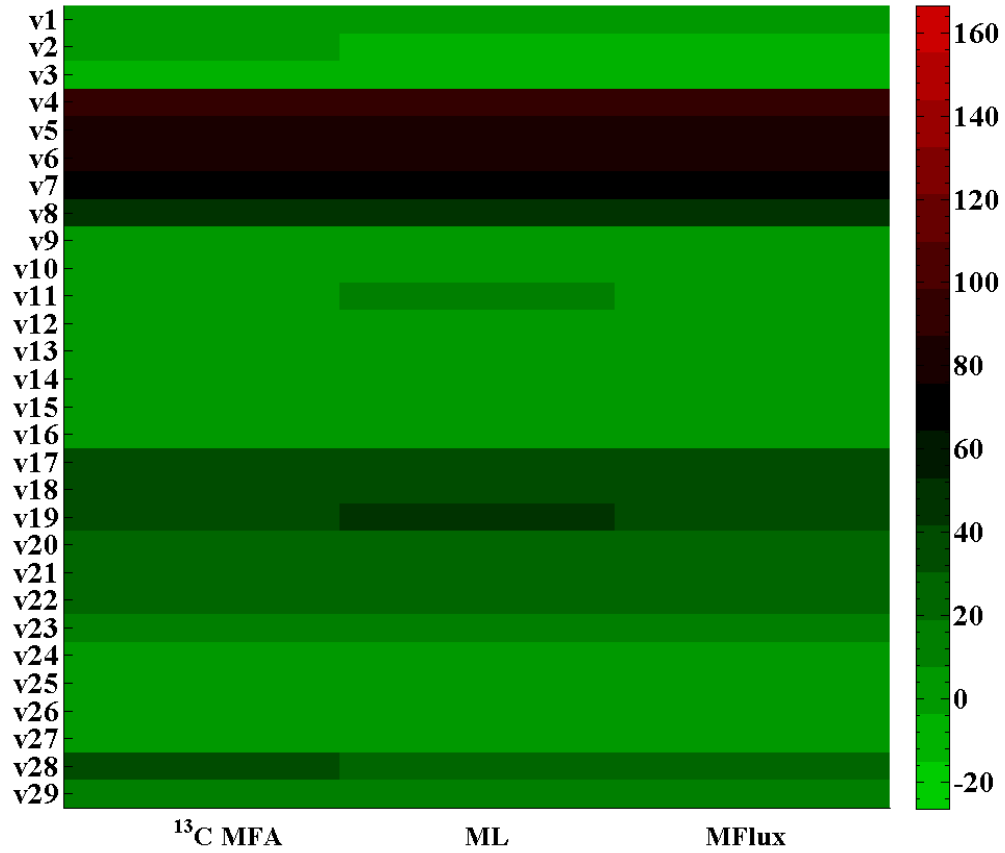
		case 5			case 6			case 7		
	13C-flux	ML	MFlux	13C-flux	ML	MFlux	13C-flux	ML	MFlux	
v1 EMP	100.0	99.4	100.0	100.0	98.9	100.0	100.0	99.0	100.0	
v2 EMP	66.3	63.1	63.1	66.8	60.6	58.6	66.4	65.3	65.3	
v3 EMP	82.6	81.1	81.1	83.6	79.2	80.0	82.8	83.6	82.9	
v4 EMP	82.6	81.0	81.1	83.6	80.2	80.0	82.8	82.1	82.9	
v5 EMP	170.9	170.2	170.2	173.4	172.4	171.7	169.9	167.7	167.7	
v6 EMP	158.8	158.5	158.5	162.6	167.1	167.1	161.2	158.5	158.5	
v7 EMP	162.0	130.2	130.2	159.2	128.8	128.8	168.8	172.5	172.5	
v8	103.3	94.6	94.6	113.1	107.2	107.2	113.8	104.9	105.4	
v9	45.8	43.2	43.2	62.1	47.6	47.6	0.0	0.9	0.1	
v10	31.9	34.2	36.2	31.8	37.3	41.4	29.2	33.9	34.1	
v11	31.9	35.7	36.2	31.8	42.6	41.4	29.2	36.7	34.1	
v12	16.3	20.4	19.7	16.7	20.2	19.8	16.4	16.2	18.8	
v13	15.6	18.8	16.5	15.0	24.0	21.6	12.9	12.0	15.4	
v14	10.0	11.0	11.4	10.0	10.5	11.7	9.6	11.1	11.0	
v15	6.4	8.9	8.4	6.8	8.7	8.1	6.8	7.6	7.8	
v16	10.0	11.9	11.4	10.0	13.2	11.7	9.6	10.9	11.0	
v17 TCA cycle	49.5	42.5	47.8	44.3	45.9	47.5	100.7	100.4	105.2	
v18 TCA cycle	49.5	49.4	47.8	44.3	36.8	47.5	100.7	116.0	105.2	
v19 TCA cycle	49.5	51.5	47.8	44.3	59.7	47.5	100.7	99.5	105.2	
v20 TCA cycle	39.8	31.1	31.2	35.7	38.6	37.0	92.2	89.8	87.8	
v21 TCA cycle	39.8	31.6	31.2	35.7	33.5	37.0	92.2	83.2	87.8	
v22 TCA cycle	39.8	33.0	35.6	35.7	50.5	45.7	92.2	103.4	97.7	
v23 TCA cycle	31.8	29.2	26.2	19.2	26.1	31.7	77.9	87.6	94.3	
v24 Glyoxylate	0.0	-1.7	0.0	0.0	-2.8	0.0	0.0	-0.6	0.0	
v25 ED	0.0	-1.0	0.0	0.0	-0.6	0.0	0.0	-0.1	0.0	
v26 ETOH	0.0	0.0	0.1	0.0	-0.2	0.0	0.0	0.5	0.1	
v27 LAC	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	
v28	33.0	30.9	30.9	38.8	27.9	27.9	35.2	23.5	23.5	
v29	8.1	9.8	9.4	16.5	13.3	14.0	14.4	2.6	3.4	
<b>RMSE</b>		<b>7.0</b>	<b>6.9</b>		<b>8.8</b>	<b>8.2</b>		<b>5.9</b>	<b>5.3</b>	

## Case 8 – 11

Reference: Chubukov V, Uhr M, Le Chat L, Kleijn RJ, Jules M, *et al.* (2013) Transcriptional regulation is insufficient to explain substrate-induced flux changes in *Bacillus subtilis*. *Molecular Systems Biology* 9

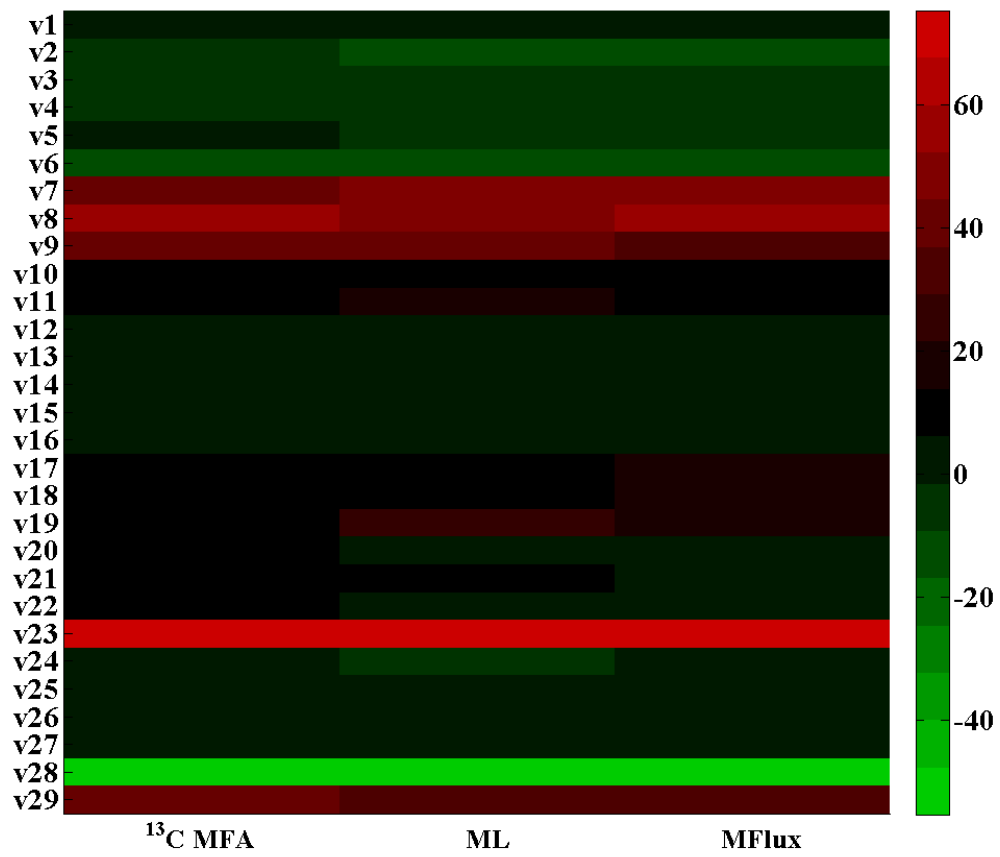
	succinate + glutamate case 8			glycerol case 9			Malate case 10			Fructose case 11		
	13C- flux	ML	MFlux	13C- flux	ML	MFlux	13C- flux	ML	MFlux	13C- flux	ML	MFlux
v1 EMP	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0
v2 EMP	-7.1	-5.0	-8.2	-2.4	-4.4	-5.7	-8.3	10.3	-10.6	-26.5	-16.4	-19.5
v3 EMP	-6.5	-5.0	-4.3	-5.1	-4.2	-4.8	-4.8	-4.2	-4.4	83.9	90.5	89.7
v4 EMP	-6.5	-6.3	-4.3	94.9	94.7	95.2	-4.8	-4.7	-4.4	83.9	88.9	89.7
v5 EMP	0.0	2.2	-6.5	76.9	79.2	79.2	0.0	-2.3	-5.7	152.2	149.9	152.3
v6 EMP	-23.3	-21.0	-21.0	76.9	79.2	79.1	-13.5	11.3	-11.3	152.2	154.4	152.3
v7 EMP	45.8	46.0	46.0	72.2	74.9	74.9	43.7	46.4	46.4	143.7	146.5	146.5
v8	44.0	41.9	41.9	41.7	43.9	43.9	52.7	50.6	52.7	88.7	90.9	90.9
v9	0.0	-2.9	-2.9	1.8	4.6	4.6	39.1	39.8	36.5	11.0	8.2	8.2
v10	4.2	4.2	8.2	0.1	0.2	5.7	7.5	7.5	10.6	23.4	23.4	19.5
v11	4.2	15.6	8.2	0.1	11.5	5.7	7.5	15.3	10.6	23.4	16.3	19.5
v12	0.6	1.9	2.6	-2.6	-1.3	-0.4	3.5	4.7	5.1	10.4	12.1	9.6
v13	3.6	3.7	5.7	2.8	2.9	6.1	4.0	3.9	5.5	13.0	12.9	9.9
v14	1.2	0.7	2.0	-0.2	0.3	0.5	2.3	2.5	3.1	7.2	7.7	7.2
v15	-0.6	0.2	0.6	-2.4	-1.6	-0.9	1.2	1.6	2.0	3.2	2.4	2.4
v16	1.2	1.7	2.0	-0.2	0.3	0.5	2.3	2.7	3.1	7.2	6.7	7.2
v17 TCA cycle	35.3	35.3	42.8	32.2	32.2	38.4	10.7	10.7	16.2	66.3	66.3	72.9
v18 TCA cycle	35.3	37.2	42.8	32.2	34.1	38.4	10.7	12.6	16.2	66.3	67.2	72.9
v19 TCA cycle	35.3	54.2	41.2	32.2	48.9	38.4	10.7	26.8	16.2	66.3	85.2	72.9
v20 TCA cycle	69.3	67.3	65.5	26.0	25.2	26.4	7.6	5.6	5.5	55.3	57.3	59.3
v21 TCA cycle	129.5	126.6	127.4	26.0	29.1	26.4	7.6	8.5	5.5	55.3	63.9	59.3
v22 TCA cycle	129.5	126.5	127.4	26.0	29.0	26.7	7.6	5.5	5.5	55.3	58.4	63.5
v23 TCA cycle	111.4	108.1	110.6	10.0	13.3	16.0	68.5	71.7	75.4	37.7	41.0	34.9
v24 Glyoxylate	0.0	3.4	1.6	0.0	-0.4	0.0	0.0	-6.5	0.0	0.0	-2.4	0.0
v25 ED	0.0	-1.5	0.0	0.0	0.4	0.0	0.0	0.9	0.0	0.0	-12.0	0.0
v26 ETOH	0.0	0.1	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.1
v27 LAC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
v28	-67.6	-65.5	-65.5	31.6	29.5	29.5	-53.1	-	-55.3	45.4	47.7	47.7

v29	18.1	18.1	18.4	16.0	10.4	10.7	39.0	55.3	29.6	30.1	17.6	29.4	28.6
RMSE		4.5	3.2		4.2	3.2			4.2	3.3		5.9	4.3

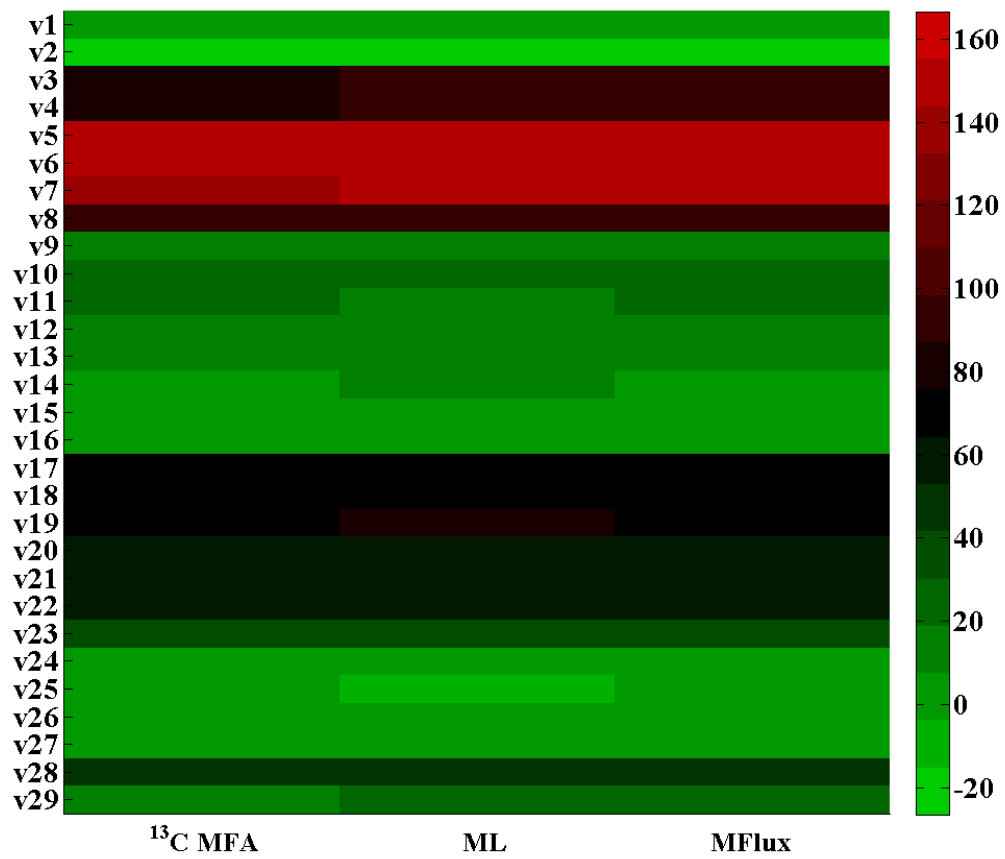


Case 9 heat map





Case 10 heat map

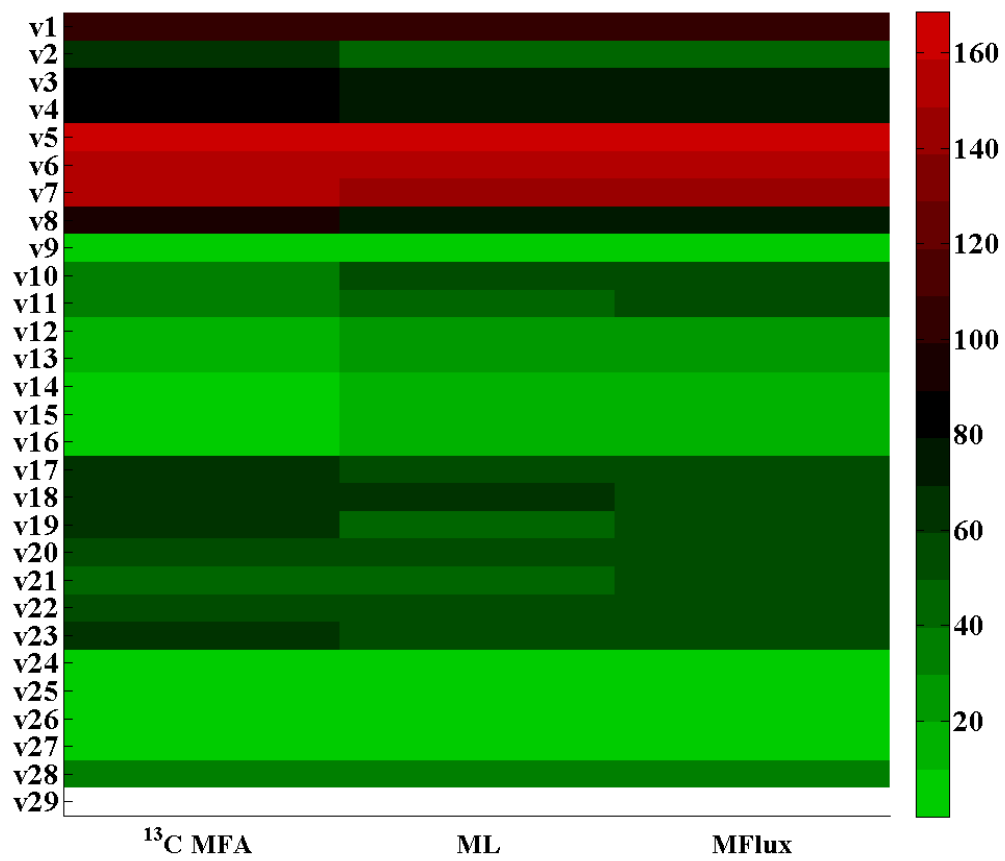


Case 11 heat map

## Case 12

Reference: Van Ooyen J, Noack S, Bott M, Reth A, Eggeling L (2012) Improved L-lysine production with *Corynebacterium glutamicum* and systemic insight into citrate synthase flux and activity. *Biotechnology And Bioengineering* 109: 2070-2081.

	case 12		
	<sup>13</sup> C-flux	ML	MFlux
v1 EMP	100.0	100.0	100.0
v2 EMP	66.9	45.3	45.3
v3 EMP	81.4	76.3	76.4
v4 EMP	81.4	76.5	76.4
v5 EMP	168.5	166.1	166.1
v6 EMP	155.6	153.4	153.4
v7 EMP	150.4	147.6	147.6
v8	89.4	78.4	78.4
v9	0.0	0.9	0.9
v10	31.3	57.9	50.5
v11	31.3	44.8	50.5
v12	15.1	27.9	27.9
v13	16.3	24.9	22.6
v14	8.7	14.6	16.0
v15	6.4	12.6	11.9
v16	8.7	17.7	16.0
v17 TCA cycle	64.0	58.6	55.3
v18 TCA cycle	64.0	62.2	55.3
v19 TCA cycle	64.0	45.0	55.3
v20 TCA cycle	53.3	51.2	50.2
v21 TCA cycle	43.1	47.8	50.2
v22 TCA cycle	55.5	52.4	55.6
v23 TCA cycle	64.1	55.7	51.9
v24 Glyoxylate	0.0	5.0	0.0
v25 ED	0.0	0.5	0.0
v26 ETOH	0.0	0.0	0.1
v27 LAC	0.0	0.0	0.0
v28	32.7	30.8	30.8
v29	N.A.	4.2	3.7
<b>RMSE</b>		<b>9.4</b>	<b>8.8</b>

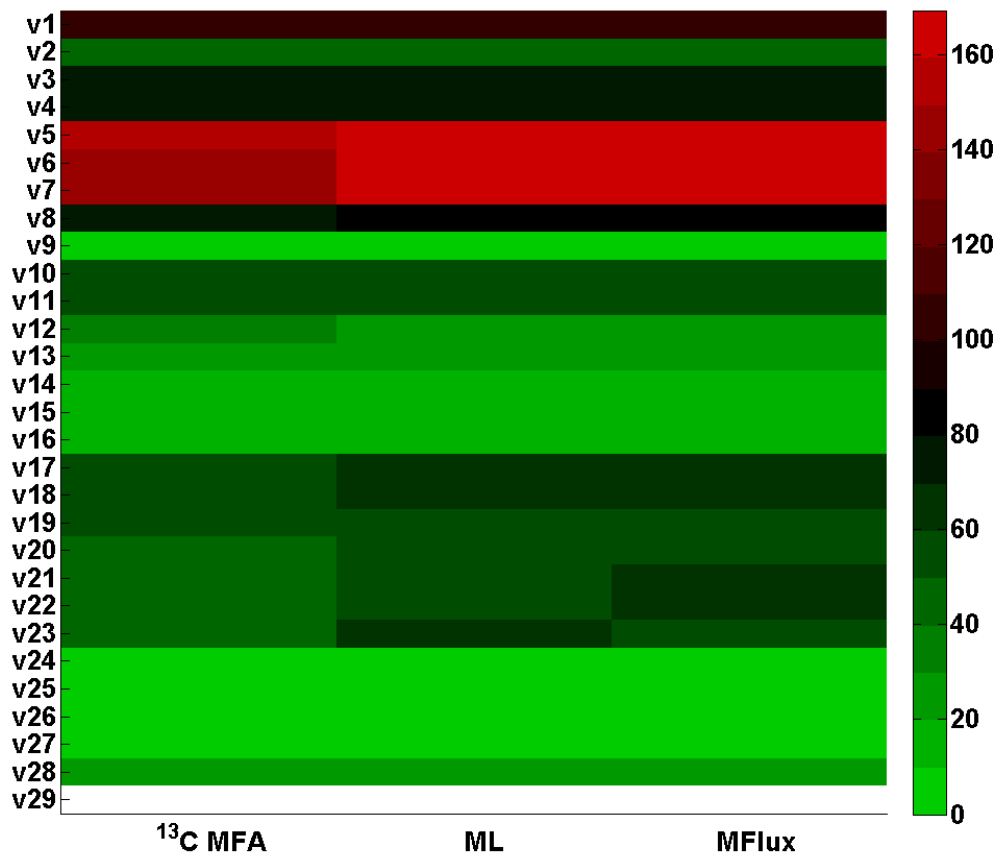


Case 12 heat map

Case 13

Reference: Bommareddy RR, Chen Z, Rappert S, Zeng A-P (2014) A de novo NADPH generation pathway for improving lysine production of *Corynebacterium glutamicum* by rational design of the coenzyme specificity of glyceraldehyde 3-phosphate dehydrogenase. *Metabolic Engineering* 25: 30-37.

	case 13		
	13C-flux	ML	MFlux
v1 EMP	100	99.7	100.0
v2 EMP	40.2	43.8	43.1
v3 EMP	70.9	75.3	75.4
v4 EMP	70.9	75.5	75.4
v5 EMP	155.9	169.3	169.3
v6 EMP	147.6	164.3	164.3
v7 EMP	147.8	165.0	165.0
v8	70.6	84.3	84.3
v9	1.1	6.6	6.6
v10	55.1	58.9	56.9
v11	55.1	53.0	53.7
v12	32.5	28.5	29.1
v13	22.6	24.3	24.6
v14	17.8	14.7	16.3
v15	14.7	13.0	12.9
v16	17.8	17.9	16.3
v17 TCA cycle	50.2	65.4	63.9
v18 TCA cycle	50.2	66.6	63.9
v19 TCA cycle	50.2	54.9	59.2
v20 TCA cycle	42.1	59.1	57.2
v21 TCA cycle	42.1	58.2	61.9
v22 TCA cycle	42.1	59.6	61.9
v23 TCA cycle	42.1	61.3	59.4
v24			
Glyoxylate	0	4.3	4.7
v25 ED	0	1.6	3.2
v26 ETOH	0	0.0	0.1
v27 LAC	2.2	0.2	0.2
v28	28.5	26.4	26.4
v29	N.A.	7.4	7.1
RMSE		11.7	10.1

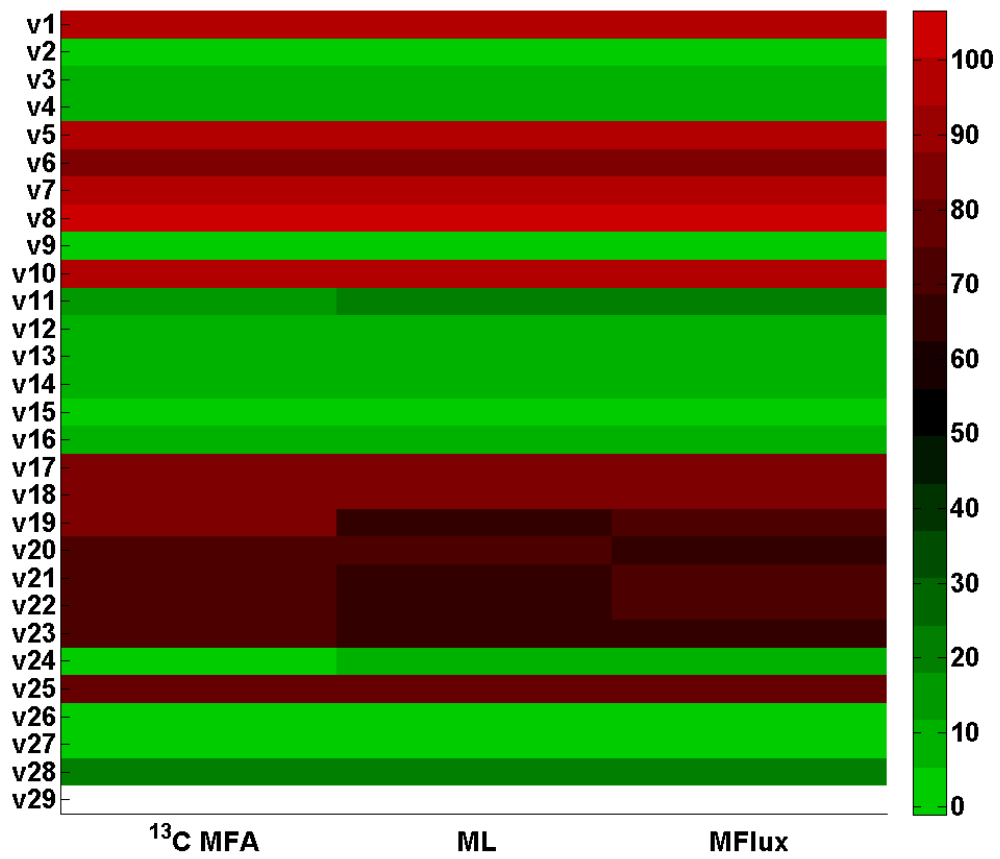


Case 13 heat map

Case 14, 15

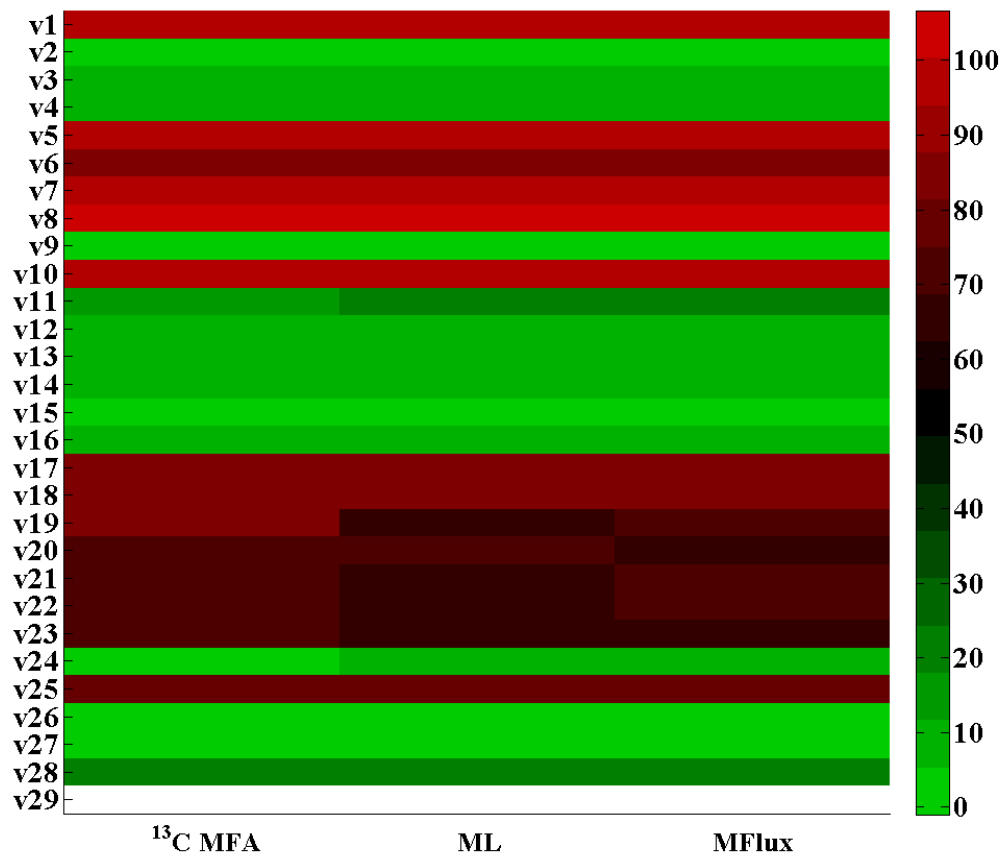
Reference: Wang Z-J, Wang P, Liu Y-W, Zhang Y-M, Chu J, *et al.* (2012) Metabolic flux analysis of the central carbon metabolism of the industrial vitamin B12 producing strain *Pseudomonas denitrificans* using 13C-labeled glucose. Journal of the Taiwan Institute of Chemical Engineers 43: 181-187.

	case 14			case 15		
	13C-flux	ML	MFlux	13C-flux	ML	MFlux
v1 EMP	100.0	99.5	100.0	100.0	99.0	100.0
v2 EMP	1.1	3.2	2.4	1.8	-0.2	-0.2
v3 EMP	7.6	9.1	8.5	18.6	17.1	17.7
v4 EMP	7.6	7.8	8.5	18.6	18.4	17.7
v5 EMP	96.7	99.0	99.0	109.5	107.3	107.3
v6 EMP	84.5	86.7	86.7	90.7	88.5	88.5
v7 EMP	97.2	100.0	100.0	123.8	121.1	121.1
v8	106.7	104.5	104.5	96.9	99.2	99.2
v9	0.0	-1.0	-1.0	0.0	2.7	2.7
v10	97.7	97.7	97.6	97.1	97.1	96.3
v11	18.0	20.3	19.7	33.1	25.1	29.9
v12	9.1	10.4	10.4	21.6	20.3	18.7
v13	8.9	8.7	9.3	11.5	11.6	11.1
v14	5.6	6.1	5.9	11.8	11.3	11.1
v15	3.5	4.3	4.5	9.8	9.0	7.6
v16	5.6	5.6	5.9	11.8	12.3	11.1
v17 TCA cycle	86.4	86.4	81.5	78.6	78.6	77.9
v18 TCA cycle	86.4	84.5	81.5	78.6	76.8	77.9
v19 TCA cycle	86.0	67.1	75.0	79.0	72.7	72.3
v20 TCA cycle	73.1	71.1	66.0	65.0	67.1	60.9
v21 TCA cycle	70.0	67.0	72.6	59.0	62.1	66.5
v22 TCA cycle	70.0	67.0	72.6	59.0	62.0	66.5
v23 TCA cycle	70.0	66.8	67.4	59.0	62.4	68.0
v24 Glyoxylate	0.0	6.2	6.6	0.0	6.5	5.6
v25 ED	79.7	78.2	77.9	64.1	65.7	66.5
v26 ETOH	0.0	0.1	0.1	0.0	0.1	0.2
v27 LAC	0.0	-0.1	-0.1	0.0	0.0	0.0
v28	23.7	21.5	21.5	27.3	25.2	25.2
v29	NA	11.7	11.7	NA	3.3	4.0
RMSE		4.1	3.4		2.9	3.6



Case 14 heat map





Case 15 heat map

Case 16

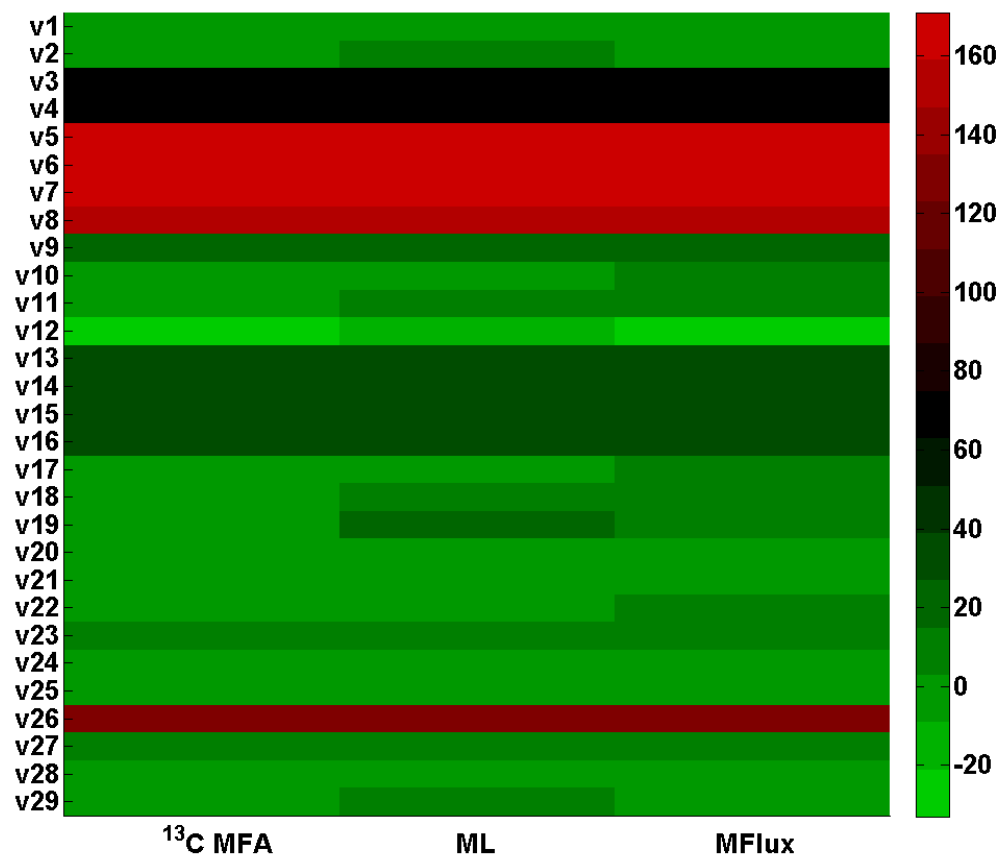
Reference: Tang YJ, Sapra R, Joyner D, Hazen TC, Myers S, *et al.* (2009) Analysis of metabolic pathways and fluxes in a newly discovered thermophilic and ethanol-tolerant *Geobacillus* strain. *Biotechnology And Bioengineering* 102: 1377-1386.

	case 16		
	13C-flux	ML	MFlux
v1 EMP	100	99.4	100.0
v2 EMP	80	82.0	80.5
v3 EMP	91	92.6	91.5
v4 EMP	91	91.2	91.5
v5 EMP	187	184.7	184.7
v6 EMP	181	178.9	178.9
v7 EMP	192.5	195.3	195.3
v8	112	109.8	114.1
v9	46	48.9	42.1
v10	19.5	19.5	19.5
v11	19.5	20.3	19.4
v12	10.5	9.3	10.4
v13	9	8.9	9.0
v14	6	5.5	5.5
v15	4.5	5.3	4.9
v16	6	5.5	5.5
v17 TCA cycle	25	25.0	31.1
v18 TCA cycle	25	27.0	31.1
v19 TCA cycle	10.5	29.5	15.0
v20 TCA cycle	7	7.6	6.0
v21 TCA cycle	21.5	18.5	22.1
v22 TCA cycle	21.5	24.6	30.8
v23 TCA cycle	36	39.3	35.1
v24 Glyoxylate	14.5	18.0	16.1
v25 ED	0	1.6	0.1
v26 ETOH	28	28.0	24.8
v27 LAC	73	72.9	72.9
v28	-7.5	-5.3	-3.9
v29	NA	12.75	11.80
RMSE		4.0	3.0

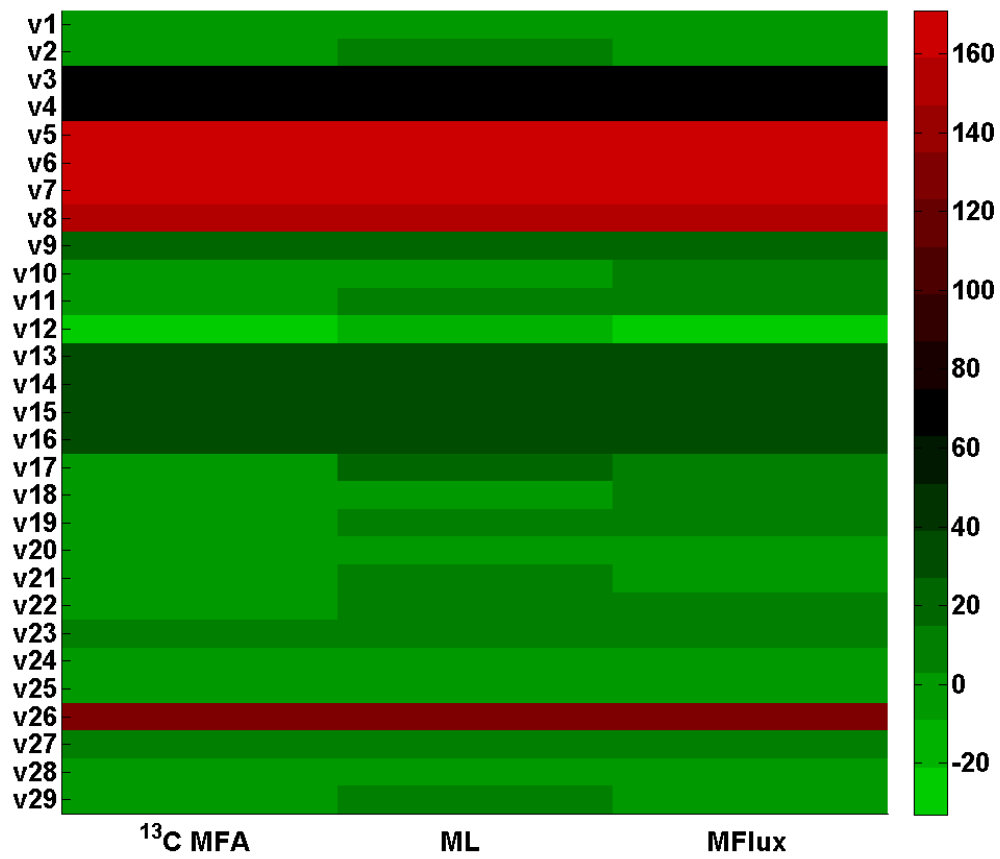
Case 17-18

Reference: Hemme CL, Fields MW, He Q, Deng Y, Lin L, *et al.* (2011) Correlation of genomic and physiological traits to biofuel yields in *Thermoanaerobacter* species. Applied And Environmental Microbiology.

	case 17			case 18		
	X514			39E		
	13C-flux	ML	MFlux	13C-flux	ML	MFlux
v1 EMP	0.0	1.0	0.0	0.0	1.1	0.0
v2 EMP	-1.0	3.0	-4.9	-1.0	3.0	-5.2
v3 EMP	65.0	66.4	65.0	65.0	66.6	64.9
v4 EMP	65.0	65.1	65.0	65.0	65.2	64.9
v5 EMP	163.0	160.8	160.9	163.0	160.9	162.0
v6 EMP	163.0	160.7	160.7	163.0	160.9	162.0
v7 EMP	171.0	169.5	169.5	171.0	168.4	165.4
v8	152.0	154.2	155.7	153.0	153.3	153.8
v9	21.0	23.8	21.4	24.0	21.3	20.5
v10	0.0	-0.4	4.9	0.0	0.0	5.2
v11	0.0	11.2	4.9	0.0	11.3	5.2
v12	-33.0	-10.3	-30.1	-33.0	-10.0	-29.8
v13	33.0	32.9	35.0	33.0	33.1	35.1
v14	33.0	32.5	35.0	33.0	32.7	35.1
v15	33.0	32.2	35.0	33.0	32.4	35.1
v16	33.0	32.5	35.0	33.0	32.8	35.1
v17 TCA cycle	1.0	1.0	6.5	1.0	1.3	6.8
v18 TCA cycle	1.0	3.0	6.5	1.0	2.8	6.8
v19 TCA cycle	1.0	19.5	6.5	1.0	19.9	6.8
v20 TCA cycle	0.0	1.8	2.0	0.0	2.0	2.3
v21 TCA cycle	0.0	2.6	2.0	0.0	3.1	2.3
v22 TCA cycle	0.0	2.6	11.1	0.0	3.0	12.4
v23 TCA cycle	8.0	10.4	8.7	8.0	11.3	10.2
v24 Glyoxylate	0.0	-4.7	0.0	0.0	-4.4	0.0
v25 ED	0.0	-1.4	0.0	0.0	-1.6	0.0
v26 ETOH	129.0	128.9	127.8	127.0	126.9	126.6
v27 LAC	10.0	10.0	10.0	8.0	8.0	8.0
v28	-8.0	-6.0	-2.2	-8.0	-5.9	-3.4
v29	-8.0	3.8	2.4	-8.0	3.7	2.2
<b>RMSE</b>		<b>6.5</b>	<b>4.1</b>		<b>6.6</b>	<b>4.3</b>



Case 17 heat map

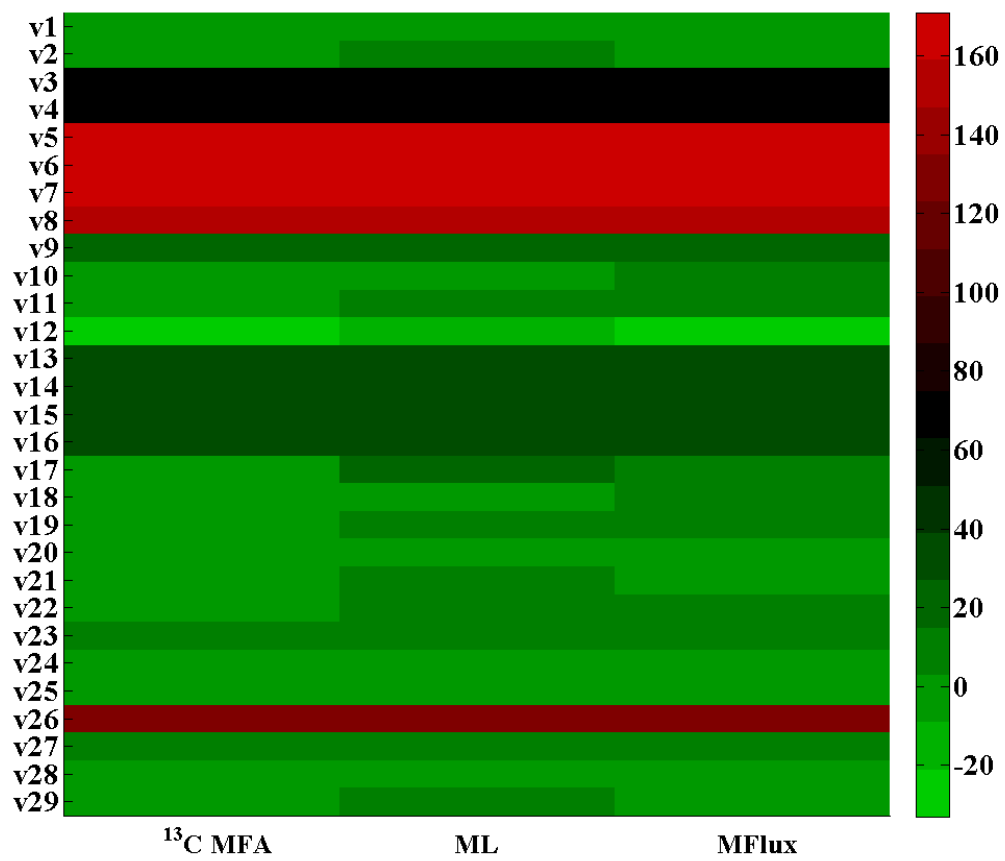


Case 18 heat map

## Case 19

Reference: Tang Y, Pingitore F, Mukhopadhyay A, Phan R, Hazen TC, *et al.* (2007) Pathway Confirmation and Flux Analysis of Central Metabolic Pathways in *Desulfovibrio vulgaris* Hildenborough using Gas Chromatography-Mass

	case 19		
	13C-flux	ML	MFlux
v1 EMP	0.0	1.0	0.0
v2 EMP	-0.7	1.4	-2.8
v3 EMP	-2.2	-0.7	-2.0
v4 EMP	-2.2	-2.1	-2.0
v5 EMP	-2.5	-0.1	-3.7
v6 EMP	-3.4	-1.1	-3.7
v7 EMP	-3.8	-1.1	-1.1
v8	91.1	88.8	90.8
v9	84.0	81.1	78.0
v10	0.0	0.0	2.8
v11	0.0	11.4	2.8
v12	-0.4	-1.6	-0.3
v13	0.4	0.5	3.1
v14	-0.1	-0.6	0.4
v15	-0.3	-1.1	-0.7
v16	-0.1	-0.6	0.4
v17 TCA cycle	0.7	0.8	8.2
v18 TCA cycle	0.7	2.6	8.2
v19 TCA cycle	0.7	19.7	3.6
v20 TCA cycle	0.0	2.1	0.7
v21 TCA cycle	-0.8	2.2	5.2
v22 TCA cycle	-0.8	2.3	7.6
v23 TCA cycle	0.0	3.3	2.0
v24 Glyoxylate	0.0	6.5	4.6
v25 ED	0.0	0.9	0.0
v26 ETOH	0.0	0.0	0.0
v27 LAC	-100.0	-100.0	-99.9
v28	1.8	4.0	6.2
v29	-0.8	11.0	10.1
RMSE		5.1	4.0



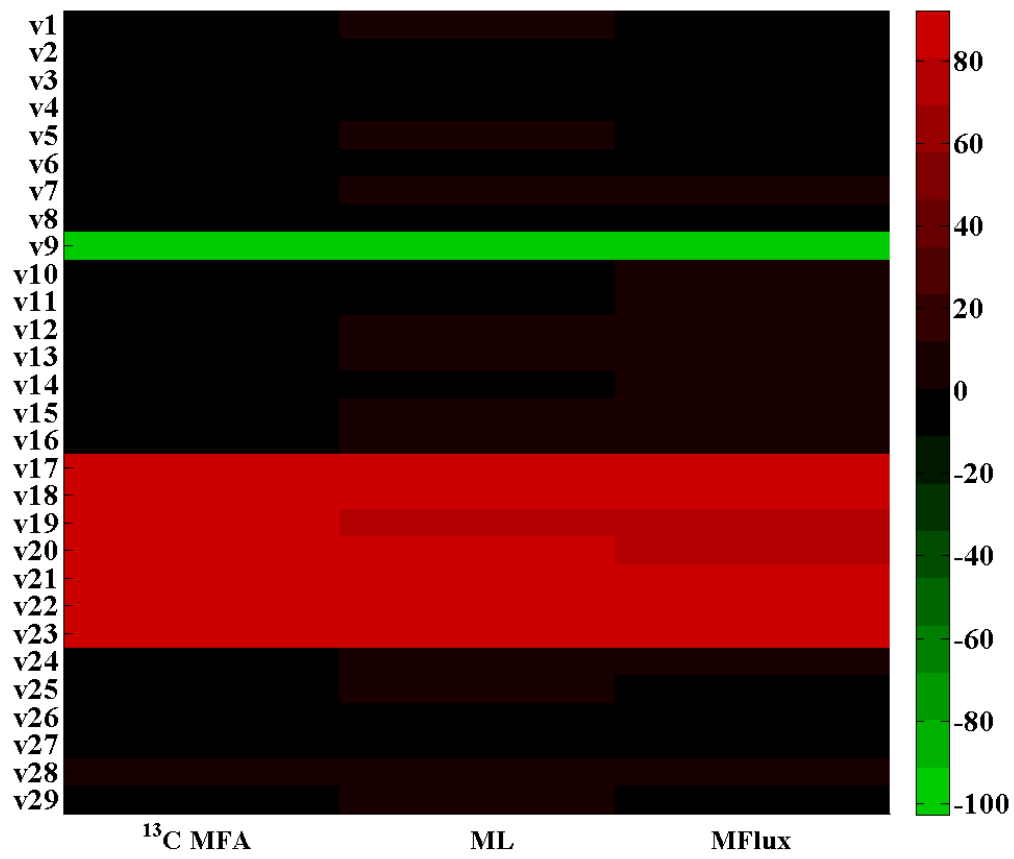
Case 19 heat map

## Case 20

Reference: Tang YJ, Chakraborty R, Martin HG, Chu J, Hazen TC, *et al.* (2007) Flux Analysis of Central Metabolic Pathways in *Geobacter metallireducens* during Reduction of Soluble Fe(III)-Nitrilotriacetic Acid. *Applied And Environmental Microbiology* 73: 3859-3864.

	case 20		
	13C-flux	ML	MFlux
v1 EMP	0.0	1.0	0.0
v2 EMP	-0.7	-2.6	-4.6
v3 EMP	-1.6	-3.1	-2.0
v4 EMP	-1.6	-1.8	-2.0
v5 EMP	-1.7	0.6	-2.7
v6 EMP	-2.3	-0.1	-2.7
v7 EMP	-1.9	1.0	1.0
v8	-7.9	-9.9	-7.8
v9	-100.0	-102.7	-100.0
v10	0.3	0.3	4.6
v11	0.3	11.7	4.6
v12	0.1	1.4	2.4
v13	0.2	0.2	2.2
v14	0.1	0.6	1.3
v15	0.0	0.8	1.1
v16	0.1	0.6	1.3
v17 TCA cycle	90.1	90.1	85.9
v18 TCA cycle	90.1	88.2	85.9
v19 TCA cycle	90.1	71.2	79.6
v20 TCA cycle	88.5	86.5	79.6
v21 TCA cycle	88.5	85.7	85.9
v22 TCA cycle	88.5	85.4	85.9
v23 TCA cycle	88.5	85.2	92.2
v24 Glyoxylate	0.0	6.5	6.3
v25 ED	0.0	1.4	0.0
v26 ETOH	0.0	0.0	0.0
v27 LAC	0.0	0.0	0.0
v28	4.2	6.4	6.4
v29	0.0	11.8	0.0
RMSE		5.1	3.6





Case 20 heat map

Attachment is two research papers on nanotechnology and electrospray published as first author separately.

**Phytotoxicity of metal oxide nanoparticles is related to both dissolved metal ions and adsorption of particles on seed surfaces**

Stephen G. Wu<sup>1,\*</sup>, Li Huang<sup>1,\*</sup>, Jennifer Head<sup>1</sup>, Daren Chen<sup>1</sup>, In Chul Kong<sup>#,2</sup>, and Yinjie J. Tang<sup>#,1</sup>

1. Department of Energy, Environmental and Chemical Engineering, Washington University, St. Louis, Missouri 63130, USA

2. Department of Environmental Engineering, Yeungnam University, Kyongsan City, Kyungbuk 712-749, Republic of Korea

**Running title: Investigation of metal oxide NP Phytotoxicity**

**\*: Wu and Huang contributed equally to this study.**

Corresponding author

ICK: Email: [ickong@ynu.ac.kr](mailto:ickong@ynu.ac.kr), Tel: 82-53-810-2546; Fax: 82-53-810-4624.

YJT: Email: [yinjie.tang@seas.wustl.edu](mailto:yinjie.tang@seas.wustl.edu), Tel: 1-314-935-3441, Fax: 1-314-935-5464.

## **Abstract**

This study assesses the biological effects of nanoparticles (NPs) based on seed germination and root elongation tests. Lettuce, radish and cucumber seeds were incubated with various metal oxide NPs (CuO, NiO, TiO<sub>2</sub>, Fe<sub>2</sub>O<sub>3</sub>, Co<sub>3</sub>O<sub>4</sub>), of which only CuO and NiO showed deleterious impacts on the activities of all three seeds. The measured EC<sub>50</sub> for seed germinations were: lettuce seed (NiO: 28 mg/L; CuO: 13 mg/L), radish seed (NiO: 401 mg/L; CuO: 398 mg/L), and cucumber seed (NiO: 175 mg/L; CuO: 228 mg/L). Phytotoxicity of TiO<sub>2</sub>, Fe<sub>2</sub>O<sub>3</sub> and Co<sub>3</sub>O<sub>4</sub> to the tested seeds was not significant, while Co<sub>3</sub>O<sub>4</sub> NP solution (5 g/L) was shown to improve root elongation of radish seedling. Metal oxide NPs tended to adsorb on seed surfaces in the aqueous medium and released metal ions near the seeds. Therefore, metal oxide NPs had higher phytotoxicity than free metal ions of the equivalent concentrations. Further, the surface area-to-volume ratio of seeds may also affect NPs phytotoxicity, whereby small seeds (i.e., lettuce) were the most sensitive to toxic CuO and NiO NPs in our experiments.

**Key words:** CuO, EC<sub>50</sub>, NiO, root elongation, metal ions, seeds germination

## Introduction

As applications for metal oxide nanoparticles (NPs) are employed by industry, the release of nanomaterials into the environment may pose severe threats for ecological systems and human health [(Roco 2005; Lin and Xing 2008; Klaine *et al.* 2008; Marambio-Jones and Hoek 2010)]. Risk assessments of nano-toxicities have already attracted public attention [(Roco 2005)]. Toxic effects of NPs on microorganisms and animals have also been reported [(Marambio-Jones and Hoek 2010; Wang *et al.* 2010b; Ji *et al.* 2011; Ma *et al.* 2010a; Wang *et al.* 2006b; Navarro *et al.* 2008a; Menard *et al.* 2011)], where metal oxide nanoparticles are the most extensively studied. Their toxicities are attributed to three mechanisms: 1. Generation of reactive oxygen species (ROS), which can damage the cell membrane; 2. Penetration of nanoparticles into the cell where they interfere with intracellular metabolism (nano-Trojan horse effect) [(Limbach *et al.* 2007)]; 3. Release of metal ions that hinder enzyme functions. Moreover, the phytotoxicity profile of NPs has also been investigated by researchers via seed germination and root elongation tests which evaluate the acute effects of NPs on plant physiologies [(Di Salvatore *et al.* 2008)]. For instance, alumina and zinc oxide NPs have been applied to different plant species [(Lin and Xing 2007; Yang and Watts 2005)]. Inhibition of seed germination and root elongation has been found to be highly dependent on both plant type and NP properties. This paper explores the impacts of additional metal oxide NPs on seed activities. In particular, we investigate three common vegetable seeds after they were incubated in aqueous NP-containing solutions: lettuce (*Lactuca sativa*) seed (length/width: 3 mm /1 mm); radish (*Raphanus sativus*) seed (length/width: 3 mm /3 mm) and cucumber (*Cucumis sativus*) seed (length/width: 8 mm /6 mm). This work aims to increase understanding of both NPs phytotoxicity on various edible plants and the potential impact of NPs on agricultural processes [(Mondal *et al.* 2011; Rico *et al.* 2011)].

## Materials and Methods

**Chemicals.** All chemicals used were reagent grade and purchased from Sigma (St. Louis, MO, US) or Fisher (Pittsburg, PA, US). TiO<sub>2</sub> NP (30-50 nm). Fe<sub>2</sub>O<sub>3</sub> NPs (20-40 nm), CuO NPs (30-50 nm), NiO NPs (30 nm) and Co<sub>3</sub>O<sub>4</sub> NPs (10-30 nm) were obtained from Nanostructured & Amorphous Materials, Inc. (Houston, TX, US). The pH of germination solutions (containing deionized water and NP suspensions) was adjusted to 7 for all toxicity studies done in aqueous phases.

**Seed Germination and Root Elongation Assay.** Lettuce, radish and cucumber seeds purchased from Ferry-Morse Seed Co. (Fulton, KY, US) were used in this study (Lettuce, Black Seeded Simpson, 2846; Radish, Icicle, Short Top, 3236; Cucumber, Marketmore 76, 2646). All three species are commonly used and recommended for phytotoxicity tests [(Rivetta *et al.* 1997; Wang *et al.* 2001; U.S.EPA 1996)]. Seeds were first sterilized by soaking them in 3% H<sub>2</sub>O<sub>2</sub> solution for 1 min and then rinsing twice with deionized water (dH<sub>2</sub>O). After, seeds were placed into dH<sub>2</sub>O (control) or certain NP solutions and shaken gently for two-hours [(Lin and Xing 2007)]. All seeds were subsequently transferred into 15 mm × 100 mm Petri dishes containing one piece of filter paper (90 mm in diameter, Whatman NO.1). 10 seeds of radish and cucumber or 15 seeds of lettuce were evenly spaced on top of the filter paper in each Petri dish. The dishes were filled with 5 ml of dH<sub>2</sub>O or NP solutions and sealed before being incubated at 25 °C in dark conditions [(Reddy and Singh 1992; El-Temsah and Joner 2010)]. After 3 days of incubation, the root length of each seed was measured. Experimental procedures are summarized in Figure 1. In this study, root length greater than 1 cm for lettuce seeds and 2 cm for radish and cucumber seeds was considered positive for germination based on our preliminary experiments. For each

condition, experiments were conducted in triplicate, from which standard deviations were calculated.

**Data Analysis.** Three parameters were adopted in this analysis to evaluate the conditions of seed germination: Relative germination rate, Germination Index and EC<sub>50</sub> value. They were calculated based on the following equations according to previous reports [(Barrena *et al.* 2009; Thompson *et al.* 2001)]:

$$\text{Relative germination rate} = \frac{\text{Seeds germinated in test sample}}{\text{Seeds germinated in control}} \times 100$$

$$\text{Relative root elongation} = \frac{\text{Mean root length in test sample}}{\text{Mean root length in control}} \times 100$$

$$\text{Germination Index} = \frac{\text{Relative germination rate} \times \text{Relative root elongation}}{100}$$

EC<sub>50</sub> is defined as the effective concentration of a certain drug/chemical that reaches half of its maximal effects or reduces growth of the control by 50%. We employed the software provided by the USEPA ([19] <http://www.epa.gov/eerd/stat2.htm#tsk>) which utilizes the Trimmed Spearman-Kärber Method to calculate EC<sub>50</sub> values for different chemicals [(Hamilton *et al.* 1977)]. Student's t-test was performed to analyze the variations in root length and germination rate between different treatments and control groups. Statistics Toolbox of Matlab (MathWorks, MA, US) was employed to conduct all statistical analyses and statistically significant was defined at the level of  $P < 0.05$ .

**Determination of metal ions released from NP suspensions.** To measure the concentration of metal ions released from NP solutions, aliquots of all five NP suspensions were drawn after the suspensions were incubated at room temperature for 2 hours. The extracts were centrifuged at 19,000 g for 20 min, and supernatants were collected and filtered with 0.22 μm nylon filters (GE Water & Process Technologies, CT, US). Inductively coupled plasma mass spectroscopy (ICP-MS, Agilent, CA, US) was used to conduct concentration assays of metal ions, and duplicated samples were measured for each condition.

**Protocols for Scanning Electron Microscope (SEM) and Dynamic Lighting Scattering (DLS).** Seeds sprayed with NPs or incubated with NP suspensions were dried overnight in a fume hood. They were then coated with gold nanoparticles by a low vacuum sputter coater (SPI supplies, PA, US) prior to image taking. Images of seed surfaces were taken with a scanning electron microscope (SEM) (Nova 2300 FEI, OR, US) and Zeta potential of NP suspensions was determined by dynamic lighting scattering (Malvern Instruments, Worcestershire, UK) after 30 minutes of incubation in room temperature.

## **Results and Discussion**



The toxicities of different metal oxide NPs at various concentrations on lettuce, radish, and cucumber seeds were tested. Seeds incubated in dH<sub>2</sub>O (pH = 7) were considered as the control upon which all statistical analysis was performed. From results shown in Table 1 and Figure 2, CuO and NiO NPs were far more toxic than the other three NPs on all three species of seeds, while lettuce seeds were the most sensitive to NPs in terms of germination. Our results showed that the toxicities of the NPs were also dependent upon the plant species, which was in accordance with a previous report [(Lin and Xing 2007)]. The relative toxicities based on the germination index (combined seed germination and root elongation) for the tested NPs were listed below:

Lettuce CuO > NiO >> Fe<sub>2</sub>O<sub>3</sub> > TiO<sub>2</sub> ≈ Co<sub>3</sub>O<sub>4</sub>

Radish NiO > CuO >> TiO<sub>2</sub> > Fe<sub>2</sub>O<sub>3</sub> > Co<sub>3</sub>O<sub>4</sub>

Cucumber NiO > CuO >> Fe<sub>2</sub>O<sub>3</sub> > TiO<sub>2</sub> > Co<sub>3</sub>O<sub>4</sub>

Interestingly, Co<sub>3</sub>O<sub>4</sub> NP solution did not inhibit the germination of cucumber seeds and even improve root elongation of radish seedling at high concentrations (5 g/L). Previous studies have provided similar reports of the positive effects of NPs on germination and growth of plants. For example, TiO<sub>2</sub> and SiO<sub>2</sub> NPs are found to enhance both the germination and growth of *Glycine max* seeds [(Lu *et al.* 2002)], carbon nanotubes (CNT) are discovered to improve germination and root elongation of tomato seeds (Khodakovskaya *et al.* 2009)], and Nano-Al are shown to augment root elongation of radish and rape seedling (Lin and Xing 2007)]. Such observations are likely due to an increased water uptake by seeds in the presence of high concentrations of NPs (Nair *et al.* 2010)].

The biological effects of NPs in aqueous solutions are closely associated to the concentration of released metal ions [(Ji *et al.* 2011; Navarro *et al.* 2008b)]. In this study, we measured the concentrations of metal ions released from all five types of NPs. We did not detect any metal ions released from TiO<sub>2</sub> NP solution, while Fe<sub>2</sub>O<sub>3</sub> and Co<sub>3</sub>O<sub>4</sub> NPs both released trace metal ions. For example, the aqueous solution with Co<sub>3</sub>O<sub>4</sub> NPs contained ~2 mg/L cobalt ion, but its inhibition of seed activity was minimal. Similarly, both Cu and Ni ions were released from the metal oxide NPs during incubation with the seeds (Table 2). To compare phytotoxicity between metal ions and NPs, we assessed seed activity in copper chloride and nickel chloride solutions and determined their EC<sub>50</sub> values. When CuCl<sub>2</sub> or NiCl<sub>2</sub> solutions were used to treat seeds (Table 2), the EC<sub>50</sub> concentrations of Cu<sup>2+</sup> and Ni<sup>2+</sup> were 5 ~ 8 mg/L and 9 ~ 19 mg/L, respectively. However, at their EC<sub>50</sub> concentrations, CuO or NiO NPs released much lower free metal ions (less than 2 mg/L). For example, a 13 mg/L CuO NP solution was able to strongly inhibit lettuce seed germination, while the released Cu<sup>2+</sup> concentration in the culture medium was only ~ 0.2 mg/L. Therefore, the phytotoxicity of metal oxide NPs is not only due to their dissolved metals ions, but also to their interactions with the seed/root surface.

It has been widely accepted that smaller NPs would have higher surface energy and thus cause more toxic to the cell [(Krug and Wick 2011)]. However, metal oxide NPs often agglomerate in the aqueous phase to minimize surface energy, and disaggregating is extremely difficult [(Lin and Xing 2007; Yang and Watts 2005)]. The actual size of our tested NPs in the aqueous solution was therefore up to 1 micrometer due to agglomeration (Table 3 and Figure 3). Previous studies reported that increasing the size of particle aggregates would reduce the toxic effect of the metal oxide particles [(Lin and Xing 2007; Yang and Watts 2005)]. On the other hand, suspended metal oxide NPs tend to agglomerate and accumulate on root/seed surfaces

[(Nair *et al.* 2010)], and phytotoxicity in our tests was not likely caused by mono-dispersed NPs. Instead, we observed that a large amount of NPs (e.g., TiO<sub>2</sub> or CuO) adsorbed on the surface of the seeds in all experiments (Figure 3). The main factors contributing to such adsorption can be concluded as increased surface area due to a rough seed surface, surface charges of NP agglomeration (e.g., 1000 mg/L of CuO NPs:  $-23.5 \pm 94.5$  mV, determined by DLS) and hydrophobic interactions between the NPs and the seed coat. Variations in the ratio of lipid to fatty acid content and the wax to fatty acid layer of the seed coat would affect the strength of such hydrophobic interactions, and thus NPs phytotoxicity [(Zhu *et al.* 2005; Zeng *et al.* 2005; Hu *et al.* 1994)]. The adsorption of NPs on the seed surface can enhance the effect of locally concentrated ions (released from NPs) on seed activities. The adsorption of metal oxide NPs on the seeds' surface also explains why small-size lettuce seeds are particularly sensitive to NP phytotoxicity. Because of the relatively high ratio of surface area to volume, more NPs per unit volume can be absorbed on the seed surface, thus increasing their toxic effect (Krug and Wick 2011; Stark 2011) . Therefore, the toxic NPs are more inhibitory on the germination of lettuce seeds (Figure 4).

## **Conclusion**

Our experiments determined the impact of five different nanoparticles on common plant seeds. It was discovered that smaller sized seeds, such as lettuce seeds, are more sensitive to toxic NPs. Additionally, this study shows that engineered metal oxide nanoparticles may hold significant potential applications in agriculture and gardening, as they may selectively inhibit unwanted plants (such as weeds), kill harmful fungi and bacteria in plant fields, and release essential metal elements for plant growth.

## Acknowledgments

We appreciate valuable suggestions given by Dr. Hui Wei from the University of Illinois at Urbana-Champaign. This research was supported by the National Research Foundation of Korea (2011-0026754) through the Ministry of Education, Science and Technology given to I. C. Kong. This research was also supported by Washington University MAGEEP program.

## References

1. Tang, Y. J.; Martin, H. G.; Myers, S.; Rodriguez, S.; Baidoo, E. E. K.; Keasling, J. D., Advances in analysis of microbial metabolic fluxes via  $^{13}\text{C}$  isotopic labeling. *Mass Spectrom Rev* **2009**, 28, (2), 362-375.
2. Orth, J.; Conrad, T.; Na, J.; Lerman, J.; Nam, H.; Feist, A.; Palsson, B., A comprehensive genome-scale reconstruction of Escherichia coli metabolism-2011. *Mol. Syst. Biol.* **2011**, 7, 535.
3. Monk, J. M.; Charusanti, P.; Aziz, R. K.; Lerman, J. A.; Premyodhin, N.; Orth, J. D.; Feist, A. M.; Palsson, B. Ø., Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proc. Natl. Acad. Sci.* **2013**, 110, (50), 20338-20343.
4. Kauffman, K. J.; Prakash, P.; Edwards, J. S., Advances in flux balance analysis. *Curr Opin Biotech* **2003**, 14, (5), 491-496.
5. Edwards, J.; Covert, M.; Palsson, B., Metabolic modeling of microbes: the flux-balance approach. *Environ Microbiol* **2002**, 4, (3), 133 - 140.
6. Orth, J.; Thiele, I.; Palsson, B., What is flux balance analysis? *Nat Biotechnol* **2010**, 28, (3), 245 - 248.
7. Becker, S.; Feist, A.; Mo, M.; Hannum, G.; Palsson, B.; Herrgard, M., Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nat Protocols* **2007**, 2, 727 - 738.
8. Bordbar, A.; Monk, J. M.; King, Z. A.; Palsson, B. O., Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet* **2014**, 15, (2), 107-120.
9. Gowen, C. M.; Fong, S. S., Genome-scale metabolic model integrated with RNAseq data to identify metabolic states of Clostridium thermocellum. *Biotechnol. J.* **2010**, 5, (7), 759-767.
10. Schellenberger, J.; Que, R.; Fleming, R.; Thiele, I.; Orth, J.; Feist, A.; Zielinski, D.; Bordbar, A.; Lewis, N.; Rahmanian, S., Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.* **2011**, 6, (9), 1290 - 1307.
11. Åkesson, M.; Förster, J.; Nielsen, J., Integration of gene expression data into genome-scale metabolic models. *Metab. Eng.* **2004**, 6, (4), 285-293.

12. Coquin, L.; Feala, J. D.; McCulloch, A. D.; Paternostro, G., Metabolomic and flux - balance analysis of age - related decline of hypoxia tolerance in *Drosophila* muscle tissue. *Mol. Syst. Biol.* **2008**, 4, (1).
13. Winter, G.; Krömer, J. O., Fluxomics – connecting ‘omics analysis and phenotypes. *Environ Microbiol* **2013**, 15, (7), 1901-1916.
14. Khannapho, C.; Zhao, H.; Bonde, B. K.; Kierzek, A. M.; Avignone-Rossa, C. A.; Bushell, M. E., Selection of objective function in genome scale flux balance analysis for process feed development in antibiotic production. *Metab. Eng.* **2008**, 10, (5), 227-233.
15. Vallino, J. J.; Stephanopoulos, G., Metabolic flux distributions in *Corynebacterium glutamicum* during growth and lysine overproduction. *Biotechnol Bioeng* **1993**, 41, (6), 633-646.
16. Yen, J. Y.; Nazem-Bokae, H.; Freedman, B. G.; Athamneh, A. I. M.; Senger, R. S., Deriving metabolic engineering strategies from genome-scale modeling with flux ratio constraints. *Biotechnol. J.* **2013**, 8, (5), 581-594.
17. McAnulty, M.; Yen, J.; Freedman, B.; Senger, R., Genome-scale modeling using flux ratio constraints to enable metabolic engineering of clostridial metabolism in silico. *BMC Syst. Biol.* **2012**, 6, (1), 42.
18. Senger, R. S.; Papoutsakis, E. T., Genome-scale model for *Clostridium acetobutylicum*: Part II. Development of specific proton flux states and numerically determined sub-systems. *Biotechnol Bioeng* **2008**, 101, (5), 1053-1071.
19. Fischer, E.; Sauer, U., Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using GC-MS. *Eur J Biochem* **2003**, 270, (5), 880 - 891f.
20. Zamboni, N.; Fendt, S.; Ruhl, M.; Sauer, U., <sup>13</sup>C-based metabolic flux analysis. *Nat. Protoc.* **2009**, 4, 878 - 892.
21. Möllney, M.; Wiechert, W.; Kownatzki, D.; de Graaf, A. A., Bidirectional reaction steps in metabolic networks: IV. Optimal design of isotopomer labeling experiments. *Biotechnol Bioeng* **1999**, 66, (2), 86-103.
22. Forbes, N. S.; Clark, D. S.; Blanch, H. W., Using isotopomer path tracing to quantify metabolic fluxes in pathway models containing reversible reactions. *Biotechnol Bioeng* **2001**, 74, (3), 196-211.
23. Antoniewicz, M. R.; Kelleher, J. K.; Stephanopoulos, G., Elementary metabolite units (EMU): A novel framework for modeling isotopic distributions. *Metab. Eng.* **2007**, 9, (1), 68-86.
24. Young, J. D.; Walther, J. L.; Antoniewicz, M. R.; Yoo, H.; Stephanopoulos, G., An elementary metabolite unit (EMU) based method of isotopically nonstationary flux analysis. *Biotechnol Bioeng* **2008**, 99, (3), 686-699.
25. Tang, Y. J.; Chakraborty, R.; Mart ın, H. G.; Chu, J.; Hazen, T. C.; Keasling, J. D., Flux Analysis of Central Metabolic Pathways in *Geobacter metallireducens* during Reduction of Soluble Fe(III)-Nitrilotriacetic Acid. *Appl Environ Microb* **2007**, 73, (12), 3859-3864.
26. Zhuang, W.-Q.; Yi, S.; Feng, X.; Zinder, S. H.; Tang, Y. J.; Alvarez-Cohen, L., Selective Utilization of Exogenous Amino Acids by *Dehalococcoides ethenogenes* strain 195 and the Effects on Growth and Dechlorination Activity. *Appl Environ Microb* **2011**, AEM.05676-11.
27. He, L.; Xiao, Y.; Gebreselassie, N.; Zhang, F.; Antoniewicz, M. R.; Tang, Y. J.; Peng, L., Central metabolic responses to the overproduction of fatty acids in *Escherichia coli* based on <sup>13</sup>C-metabolic flux analysis. *Biotechnol Bioeng* **2014**, 111, (3), 575-585.
28. Tang, J. K.-H.; You, L.; Blankenship, R. E.; Tang, Y. J., Recent advances in mapping environmental microbial metabolisms through <sup>13</sup>C isotopic fingerprints. *J. Royal Soc. Interface* **2012**, 9, (76), 2767-2780.

29. Bandini, S.; Manzoni, S.; Vizzari, G., Agent Based Modeling and Simulation: An Informatics Perspective. *Journal of Artificial Societies and Social Simulation* **2009**, 12, (4), 4.
30. Bouchaud, J.-P., Economics needs a scientific revolution. *Nature* **2008**, 455, (7217), 1181-1181.
31. Bonabeau, E., Agent-based modeling: Methods and techniques for simulating human systems. *Proc. Natl. Acad. Sci.* **2002**, 99, (suppl 3), 7280-7287.
32. Lardon, L. A.; Merkey, B. V.; Martins, S.; Dötsch, A.; Picioreanu, C.; Kreft, J.-U.; Smets, B. F., iDynoMiCS: next-generation individual-based modelling of biofilms. *Environ Microbiol* **2011**, 13, (9), 2416-2434.
33. Biggs, M. B.; Papin, J. A., Novel Multiscale Modeling Tool Applied to *Pseudomonas aeruginosa* Biofilm Formation. *PLoS ONE* **2013**, 8, (10), e78011.
34. Papagianni, M., Fungal morphology and metabolite production in submerged mycelial processes. *Biotechnol Adv* **2004**, 22, (3), 189-259.
35. Metz, B.; Kossen, N., The growth of molds in the form of pellets—a literature review. *Biotechnol Bioeng* **1979**, 19, (6), 781-799.
36. Atsumi, S.; Liao, J. C., Metabolic engineering for advanced biofuels production from *Escherichia coli*. *Curr Opin Biotech* **2008**, 19, (5), 414-419.
37. Baba, T.; Ara, T.; Hasegawa, M.; Takai, Y.; Okumura, Y.; Baba, M.; Datsenko, K.; Tomita, M.; Wanner, B.; Mori, H., Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2006**, 2, 2006.0008.
38. Wiechert, W.; Mollney, M.; Petersen, S.; de Graaf, A., A universal framework for 13C metabolic flux analysis. *Metab. Eng.* **2001**, 3, 265 - 283.
39. Zamboni, N.; Fischer, E.; Sauer, U., FiatFlux - a software for metabolic flux analysis from 13C-glucose experiments. *BMC Bioinformatics* **2005**, 6, (1), 209.
40. Longacre, A.; Reimers, J. M.; Gannon, J. E.; Wright, B. E., Flux Analysis of Glucose Metabolism in *Rhizopus oryzae* for the Purpose of Increasing Lactate Yields. *Fungal Genet Biol* **1997**, 21, (1), 30-39.
41. Bai, D.-M.; Zhao, X.-M.; Li, X.-G.; Xu, S.-M., Strain improvement of *Rhizopus oryzae* for over-production of l(+)-lactic acid and metabolic flux analysis of mutants. *Biochem Eng J* **2004**, 18, (1), 41-48.
42. Sun, Y.; Li, Y. L.; Bai, S., Modeling of continuous L(+)-lactic acid production with immobilized *R. oryzae* in an airlift bioreactor. *Biochem Eng J* **1999**, 3, (1), 87-90.
43. Mor, G. K.; Shankar, K.; Paulose, M.; Varghese, O. K.; Grimes, C. A., Use of Highly-Ordered TiO<sub>2</sub> Nanotube Arrays in Dye-Sensitized Solar Cells. *Nano Lett.* **2005**, 6, (2), 215-218.
44. Wang, H.; Yang, Y.; Liang, Y.; Cui, L.-F.; Sanchez Casalongue, H.; Li, Y.; Hong, G.; Cui, Y.; Dai, H., LiMn<sub>1-x</sub>FexPO<sub>4</sub> Nanorods Grown on Graphene Sheets for Ultrahigh-Rate-Performance Lithium Ion Batteries. *Angew. Chem., Int. Ed.* **2011**, 50, (32), 7364-7368.
45. Benn, T. M.; Westerhoff, P., Nanoparticle Silver Released into Water from Commercially Available Sock Fabrics. *Environ. Sci. Technol.* **2008**, 42, (11), 4133-4139.
46. Chen, J.; Liu, M.; Zhang, L.; Zhang, J.; Jin, L., Application of nano TiO<sub>2</sub> towards polluted water treatment combined with electro-photochemical method. *Water Res.* **2003**, 37, (16), 3815-3820.
47. Schoen, D. T.; Schoen, A. P.; Hu, L.; Kim, H. S.; Heilshorn, S. C.; Cui, Y., High Speed Water Sterilization Using One-Dimensional Nanostructures. *Nano Lett.* **2010**, 10, (9), 3628-3632.

48. Tian, B.; Cohen-Karni, T.; Qing, Q.; Duan, X.; Xie, P.; Lieber, C. M., Three-Dimensional, Flexible Nanoscale Field-Effect Transistors as Localized Bioprobes. *Science* **2010**, 329, (5993), 830-834.
49. Chen, J.; Patil, S.; Seal, S.; McGinnis, J. F., Rare earth nanoparticles prevent retinal degeneration induced by intracellular peroxides. *Nat. Nano.* **2006**, 1, (2), 142-150.
50. Karakoti, A. S.; Tsigkou, O.; Yue, S.; Lee, P. D.; Stevens, M. M.; Jones, J. R.; Seal, S., Rare earth oxides as nanoadditives in 3-D nanocomposite scaffolds for bone regeneration. *J. Mater. Chem.* **2010**, 20, (40), 8912-8919.
51. Wang, Y.; Brown, P.; Xia, Y., Nanomedicine: Swarming towards the target. *Nat. Mater.* **2011**, 10, (7), 482-483.
52. Copley, C. M.; Chen, J.; Cho, E. C.; Wang, L. V.; Xia, Y., Gold nanostructures: a class of multifunctional materials for biomedical applications. *Chem. Soc. Rev.* **2011**, 40, (1), 44-56.
53. Roco, M. C., Environmentally Responsible Development of Nanotechnology. *Environ. Sci. Technol.* **2005**, 39, (5), 106A-112A.
54. Lin, D.; Xing, B., Root Uptake and Phytotoxicity of ZnO Nanoparticles. *Environ. Sci. Technol.* **2008**, 42, (15), 5580-5585.
55. Klaine, S. J.; Alvarez, P. J. J.; Batley, G. E.; Fernandes, T. F.; Handy, R. D.; Lyon, D. Y.; Mahendra, S.; McLaughlin, M. J.; Lead, J. R., Nanomaterials in the environment: Behavior, fate, bioavailability, and effects. *Environ. Toxicol. Chem.* **2008**, 27, (9), 1825-1851.
56. Marambio-Jones, C.; Hoek, E., A review of the antibacterial effects of silver nanomaterials and potential implications for human health and the environment. *J. Nanoparticle Res.* **2010**, 12, (5), 1531-1551.
57. Wang, Z.; Lee, Y.-H.; Wu, B.; Horst, A.; Kang, Y.; Tang, Y. J.; Chen, D.-R., Anti-microbial activities of aerosolized transition metal oxide nanoparticles. *Chemosphere* **2010**, 80, (5), 525-529.
58. Ji, J.; Long, Z.; Lin, D., Toxicity of oxide nanoparticles to the green algae *Chlorella* sp. *Chem. Eng. J.* **2011**, 170, (2-3), 525-530.
59. Ma, Y.; Kuang, L.; He, X.; Bai, W.; Ding, Y.; Zhang, Z.; Zhao, Y.; Chai, Z., Effects of rare earth oxide nanoparticles on root elongation of plants. *Chemosphere* **2010**, 78, (3), 273-279.
60. Wang, B.; Feng, W.-Y.; Wang, T.-C.; Jia, G.; Wang, M.; Shi, J.-W.; Zhang, F.; Zhao, Y.-L.; Chai, Z.-F., Acute toxicity of nano- and micro-scale zinc powder in healthy adult mice. *Toxicol. Lett.* **2006**, 161, (2), 115-123.
61. Navarro, E.; Baun, A.; Behra, R.; Hartmann, N.; Filser, J.; Miao, A.-J.; Quigg, A.; Santschi, P.; Sigg, L., Environmental behavior and ecotoxicity of engineered nanoparticles to algae, plants, and fungi. *Ecotoxicology* **2008**, 17, (5), 372-386.
62. Menard, A.; Drobne, D.; Jemec, A., Ecotoxicity of nanosized TiO<sub>2</sub>. Review of in vivo data. *Environ. Pollut.* **2011**, 159, (3), 677-684.
63. Limbach, L. K.; Wick, P.; Manser, P.; Grass, R. N.; Bruinink, A.; Stark, W. J., Exposure of Engineered Nanoparticles to Human Lung Epithelial Cells: Influence of Chemical Composition and Catalytic Activity on Oxidative Stress. *Environ. Sci. Technol.* **2007**, 41, (11), 4158-4163.
64. Di Salvatore, M.; Carafa, A. M.; Carratù, G., Assessment of heavy metals phytotoxicity using seed germination and root elongation tests: A comparison of two growth substrates. *Chemosphere* **2008**, 73, (9), 1461-1464.
65. Lin, D.; Xing, B., Phytotoxicity of nanoparticles: Inhibition of seed germination and root growth. *Environ. Pollut.* **2007**, 150, (2), 243-250.

66. Yang, L.; Watts, D. J., Particle surface characteristics may play an important role in phytotoxicity of alumina nanoparticles. *Toxicol. Lett.* **2005**, 158, (2), 122-132.
67. Mondal, A.; Basu, R.; Das, S.; Nandy, P., Beneficial role of carbon nanotubes on mustard plant growth: an agricultural prospect. *J. Nanoparticle Res.* **2011**, 13, (10), 4519-4528.
68. Rico, C. M.; Majumdar, S.; Duarte-Gardea, M.; Peralta-Videa, J. R.; Gardea-Torresdey, J. L., Interaction of Nanoparticles with Edible Plants and Their Possible Implications in the Food Chain. *J. Agr. Food Chem.* **2011**, 59, (8), 3485-3498.
69. Rivetta, A.; Negrini, N.; Cocucci, M., Involvement of Ca<sup>2+</sup>-calmodulin in Cd<sup>2+</sup> toxicity during the early phases of radish (*Raphanus sativus* L.) seed germination. *Plant, Cell Environ.* **1997**, 20, (5), 600-608.
70. Wang, X.; Sun, C.; Gao, S.; Wang, L.; Shuokui, H., Validation of germination rate and root elongation as indicator to assess phytotoxicity with *Cucumis sativus*. *Chemosphere* **2001**, 44, (8), 1711-1721.
71. U.S.EPA, Ecological Effects Test Guidelines (OPPTS 850.4200): Seed Germination/Root Elongation Toxicity Test. In Agency, E. P., Ed. 1996.
72. Reddy, K. N.; Singh, M., Germination and Emergence of Hairy Beggarticks (*Bidens pilosa*). *Weed Sci.* **1992**, 40, (2), 195-199.
73. El-Temseh, Y. S.; Joner, E. J., Impact of Fe and Ag nanoparticles on seed germination and differences in bioavailability during exposure in aqueous suspension and soil. *Environ Toxicol* **2010**, n/a-n/a.
74. Barrena, R.; Casals, E.; Colón, J.; Font, X.; Sánchez, A.; Puentes, V., Evaluation of the ecotoxicity of model nanoparticles. *Chemosphere* **2009**, 75, (7), 850-857.
75. Thompson, W. H.; Leege, P. B.; Millner, P. D.; Watson, M. E., Test methods for the examination of composting and compost. In USCC; USDA, Eds. 2001.
76. Hamilton, M. A.; Russo, R. C.; Thurston, R. V., Trimmed Spearman-Kärber method for estimating median lethal concentrations in toxicity bioassays. *Environ. Sci. Technol.* **1977**, 11, (7), 714-719.
77. Lu, C. M.; Zhang, C. Y.; Wen, J. Q.; Wu, G. R.; Tao, M. X., Research of the effect of nanometer materials on germination and growth enhancement of *Glycine max* and its mechanism. *Soybean Sci.* **2002**, 21, (3), 168-172.
78. Khodakovskaya, M.; Dervishi, E.; Mahmood, M.; Xu, Y.; Li, Z.; Watanabe, F.; Biris, A. S., Carbon Nanotubes Are Able To Penetrate Plant Seed Coat and Dramatically Affect Seed Germination and Plant Growth. *ACS Nano* **2009**, 3, (10), 3221-3227.
79. Nair, R.; Varghese, S. H.; Nair, B. G.; Maekawa, T.; Yoshida, Y.; Kumar, D. S., Nanoparticulate material delivery to plants. *Plant Sci.* **2010**, 179, (3), 154-163.
80. Navarro, E.; Piccapietra, F.; Wagner, B.; Marconi, F.; Kaegi, R.; Odzak, N.; Sigg, L.; Behra, R., Toxicity of Silver Nanoparticles to *Chlamydomonas reinhardtii*. *Environ. Sci. Technol.* **2008**, 42, (23), 8959-8964.
81. Krug, H. F.; Wick, P., Nanotoxicology: An Interdisciplinary Challenge. *Angew. Chem., Int. Ed.* **2011**, 50, (6), 1260-1278.
82. Zhu, H. N.; Lu, Q. X.; Abdollahi, K., Seed Coat Structure of *Pinus Koraiensis*. *Microsc Microanal* **2005**, 11, (SupplementS02), 1158-1159.
83. Zeng, L. W.; Cocks, P. S.; Kailis, S. G.; Kuo, J., The role of fractures and lipids in the seed coat in the loss of hardseededness of six Mediterranean legume species. *J. Agric. Sci.* **2005**, 143, (01), 43-55.



84. Hu, X.; Daun, J.; Scarth, R., Proportions of C18 : 1n-7 and C18 : 1n-9 fatty acids in canola seedcoat surface and internal lipids. *J. Am. Oil Chem. Soc.* **1994**, 71, (2), 221-222.
85. Stark, W. J., Nanoparticles in Biological Systems. *Angew. Chem., Int. Ed.* **2011**, 50, (6), 1242-1258.
86. Yoo, S. I.; Yang, M.; Brender, J. R.; Subramanian, V.; Sun, K.; Joo, N. E.; Jeong, S.-H.; Ramamoorthy, A.; Kotov, N. A., Inhibition of Amyloid Peptide Fibrillation by Inorganic Nanoparticles: Functional Similarities with Proteins. *Angew. Chem., Int. Ed.* **2011**, 50, (22), 5110-5115.
87. Guo, S.; Dong, S., Graphene nanosheet: synthesis, molecular engineering, thin film, hybrids, and energy and analytical applications. *Chem. Soc. Rev.* **2011**, 40, (5), 2644-2672.
88. Wei, H.; Li, B.; Li, J.; Wang, E.; Dong, S., Simple and sensitive aptamer-based colorimetric sensing of protein using unmodified gold nanoparticle probes. *Chem Commun* **2007**, 0, (36), 3735-3737.
89. Nel, A.; Xia, T.; Mädler, L.; Li, N., Toxic Potential of Materials at the Nanolevel. *Science* **2006**, 311, (5761), 622-627.
90. Ma, X.; Geiser-Lee, J.; Deng, Y.; Kolmakov, A., Interactions between engineered nanoparticles (ENPs) and plants: Phytotoxicity, uptake and accumulation. *Sci. Total Environ.* **2010**, 408, (16), 3053-3061.
91. Torney, F.; Trewyn, B. G.; Lin, V. S. Y.; Wang, K., Mesoporous silica nanoparticles deliver DNA and chemicals into plants. *Nat. Nano.* **2007**, 2, (5), 295-300.
92. González-Melendi, P.; Fernández-Pacheco, R.; Coronado, M. J.; Corredor, E.; Testillano, P. S.; Risueño, M. C.; Marquina, C.; Ibarra, M. R.; Rubiales, D.; Pérez-de-Luque, A., Nanoparticles as Smart Treatment-delivery Systems in Plants: Assessment of Different Techniques of Microscopy for their Visualization in Plant Tissues. *Ann Bot-london* **2008**, 101, (1), 187-195.
93. Liu, Q.; Chen, B.; Wang, Q.; Shi, X.; Xiao, Z.; Lin, J.; Fang, X., Carbon Nanotubes as Molecular Transporters for Walled Plant Cells. *Nano Lett.* **2009**, 9, (3), 1007-1010.
94. Martin-Ortigosa, S.; Valenstein, J. S.; Sun, W.; Moeller, L.; Fang, N.; Trewyn, B. G.; Lin, V. S. Y.; Wang, K., Parameters Affecting the Efficient Delivery of Mesoporous Silica Nanoparticle Materials and Gold Nanorods into Plant Tissues by the Biolistic Method. *Small* **2012**, 8, (3), 413-422.
95. Zheng, L.; Hong, F.; Lu, S.; Liu, C., Effect of nano-TiO<sub>2</sub> on strength of naturally aged seeds and growth of spinach. *Biol. Trace Elem. Res.* **2005**, 104, (1), 83-91.
96. Prasad, T. N. V. K. V.; Sudhakar, P.; Sreenivasulu, Y.; Latha, P.; Munaswamy, V.; Reddy, K. R.; Sreeprasad, T. S.; Sajanlal, P. R.; Pradeep, T., Effect of nanoscale zinc oxide particles on the germination, growth and yield of peanut. *J. Plant Nutr.* **2012**, 35, (6), 905-927.
97. Khodakovskaya, M. V.; de Silva, K.; Nedosekin, D. A.; Dervishi, E.; Biris, A. S.; Shashkov, E. V.; Galanzha, E. I.; Zharov, V. P., Complex genetic, photothermal, and photoacoustic analysis of nanoparticle-plant interactions. *Proc. Natl. Acad. Sci.* **2011**, 108, (3), 1028-1033.
98. Kim, S. C.; Chen, D.-R.; Qi, C.; Gelein, R. M.; Finkelstein, J. N.; Elder, A.; Bentley, K.; Oberdörster, G.; Pui, D. Y. H., A nanoparticle dispersion method for in vitro and in vivo nanotoxicity study. *Nanotoxicology* **2010**, 4, (1), 42-51.
99. Chen, D.-R.; Pui, D. Y. H.; Kaufman, S. L., Electro spraying of conducting liquids for monodisperse aerosol generation in the 4 nm to 1.8 μm diameter range. *J. Aerosol Sci.* **1995**, 26, (6), 963-977.

100. Wu, B.; Wang, Y.; Lee, Y.-H.; Horst, A.; Wang, Z.; Chen, D.-R.; Sureshkumar, R.; Tang, Y. J., Comparative Eco-Toxicities of Nano-ZnO Particles under Aquatic and Aerosol Exposure Modes. *Environ. Sci. Technol.* **2010**, 44, (4), 1484-1489.
101. Jaworek, A., Micro- and nanoparticle production by electrospraying. *Powder Technol.* **2007**, 176, (1), 18-35.
102. Pui, D. Y. H. P., MN); Chen, D.-r. L., MN) Electrospraying apparatus and method for introducing material into cells. US 6,399,362 B1, Jun. 4, 2002, 2000.
103. Welbaum, G. E.; Bradford, K. J.; Yim, K.-O.; Booth, D. T.; Oluoch, M. O., Biophysical, physiological and biochemical processes regulating seed germination. *Seed Sci. Res.* **1998**, 8, (02), 161-172.
104. Gordon-Kamm, W. J.; Spencer, T. M.; Mangano, M. L.; Adams, T. R.; Daines, R. J.; Start, W. G.; O'Brien, J. V.; Chambers, S. A.; Adams, W. R.; Willetts, N. G.; Rice, T. B.; Mackey, C. J.; Krueger, R. W.; Kausch, A. P.; Lemaux, P. G., Transformation of Maize Cells and Regeneration of Fertile Transgenic Plants. *Plant Cell* **1990**, 2, (7), 603-618.
105. Chen, D.-R.; Wendt, C. H.; Pui, D. Y. H., A Novel Approach for Introducing Bio-Materials Into Cells. *J. Nanoparticle Res.* **2000**, 2, (2), 133-139.
106. Gu, Z.; Biswas, A.; Zhao, M.; Tang, Y., Tailoring nanocarriers for intracellular protein delivery. *Chem. Soc. Rev.* **2011**, 40, (7), 3638-3655.
107. Gleiser, G.; Picher, M. C.; Veintimilla, P.; Martinez, J.; VerdÚ, M., Seed dormancy in relation to seed storage behaviour in Acer. *Bot. J. Linn. Soc.* **2004**, 145, (2), 203-208.
108. Hund-Rinke, K.; Simon, M., Ecotoxic Effect of Photocatalytic Active Nanoparticles (TiO<sub>2</sub>) on Algae and Daphnids (8 pp). *Environ. Sci. Pollut. R* **2006**, 13, (4), 225-232.
109. Baek, Y.-W.; An, Y.-J., Microbial toxicity of metal oxide nanoparticles (CuO, NiO, ZnO, and Sb<sub>2</sub>O<sub>3</sub>) to Escherichia coli, Bacillus subtilis, and Streptococcus aureus. *Sci. Total Environ.* **2011**, 409, (8), 1603-1608.
110. Karlsson, H. L.; Cronholm, P.; Gustafsson, J.; Möller, L., Copper Oxide Nanoparticles Are Highly Toxic: A Comparison between Metal Oxide Nanoparticles and Carbon Nanotubes. *Chem. Res. Toxicol.* **2008**, 21, (9), 1726-1732.
111. Schwember, A. R.; Bradford, K. J., Quantitative trait loci associated with longevity of lettuce seeds under conventional and controlled deterioration storage conditions. *J. Exp. Bot.* **2010**, 61, (15), 4423-4436.
112. Finch-Savage, W. E.; Leubner-Metzger, G., Seed dormancy and the control of germination. *New Phytol.* **2006**, 171, (3), 501-523.
113. Nadjafi, F.; Bannayan, M.; Tabrizi, L.; Rastgoo, M., Seed germination and dormancy breaking techniques for Ferula gummosa and Teucrium polium. *J. Arid Environ.* **2006**, 64, (3), 542-547.
114. Rieter, W. J.; Pott, K. M.; Taylor, K. M. L.; Lin, W., Nanoscale Coordination Polymers for Platinum-Based Anticancer Drug Delivery. *J. Am. Chem. Soc.* **2008**, 130, (35), 11584-11585.
115. Yan, M.; Du, J.; Gu, Z.; Liang, M.; Hu, Y.; Zhang, W.; Priceman, S.; Wu, L.; Zhou, Z. H.; Liu, Z.; Segura, T.; Tang, Y.; Lu, Y., A novel intracellular protein delivery platform based on single-protein nanocapsules. *Nat. Nano.* **2010**, 5, (1), 48-53.
116. Walser, T.; Limbach, L. K.; Brogioli, R.; Erismann, E.; Flamigni, L.; Hattendorf, B.; Juchli, M.; Krumeich, F.; Ludwig, C.; Prikopsky, K.; Rossier, M.; Saner, D.; Sigg, A.; Hellweg, S.; Gunther, D.; Stark, W. J., Persistence of engineered nanoparticles in a municipal solid-waste incineration plant. *Nat. Nano.* **2012**, 7, (8), 520-524.

## Figure Legend

**Figure 1. Flow chart of experimental procedures**

**Figure 2. Effects of NPs on seed germination and elongation; Red line: relative germination rate; Blue dashed line: germination index.**

**Figure 3. SEM images for NPs/lettuce seeds. In the aqueous phase, the SEM image shows that metal oxide NPs (TiO<sub>2</sub> NPs 1000 mg/L) (a) and (CuO NPs 1000 mg/L) were adsorbed on the seed surface (b).**

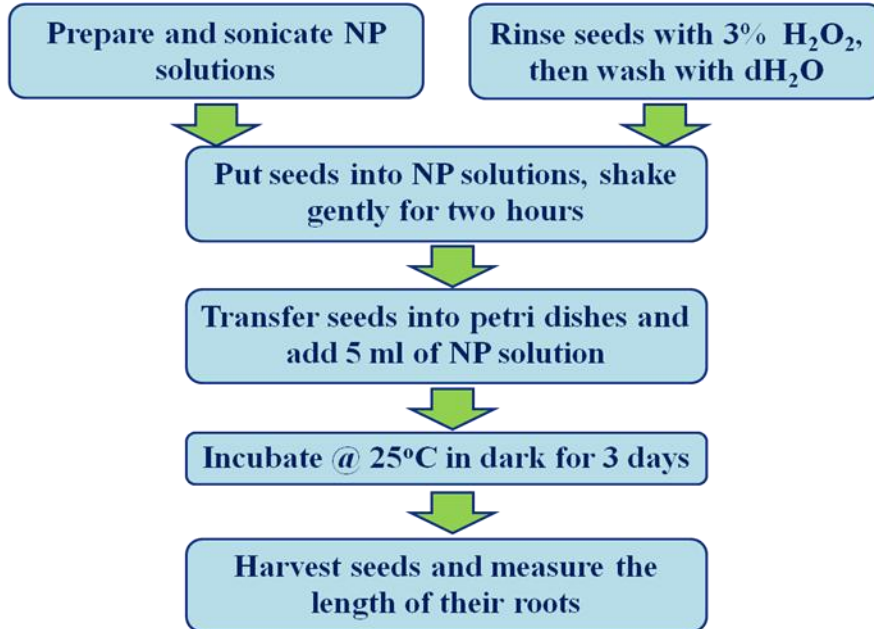
**Figure 4. Effects of CuO NPs on seed germination and root elongation (incubation at 25 °C in dark for 3 days, NPs could be observed on the seed surface.)**

Figure 4-1. Lettuce seeds (a) Incubated in dH<sub>2</sub>O; (b) Incubated in 500 mg/L of CuO NPs.

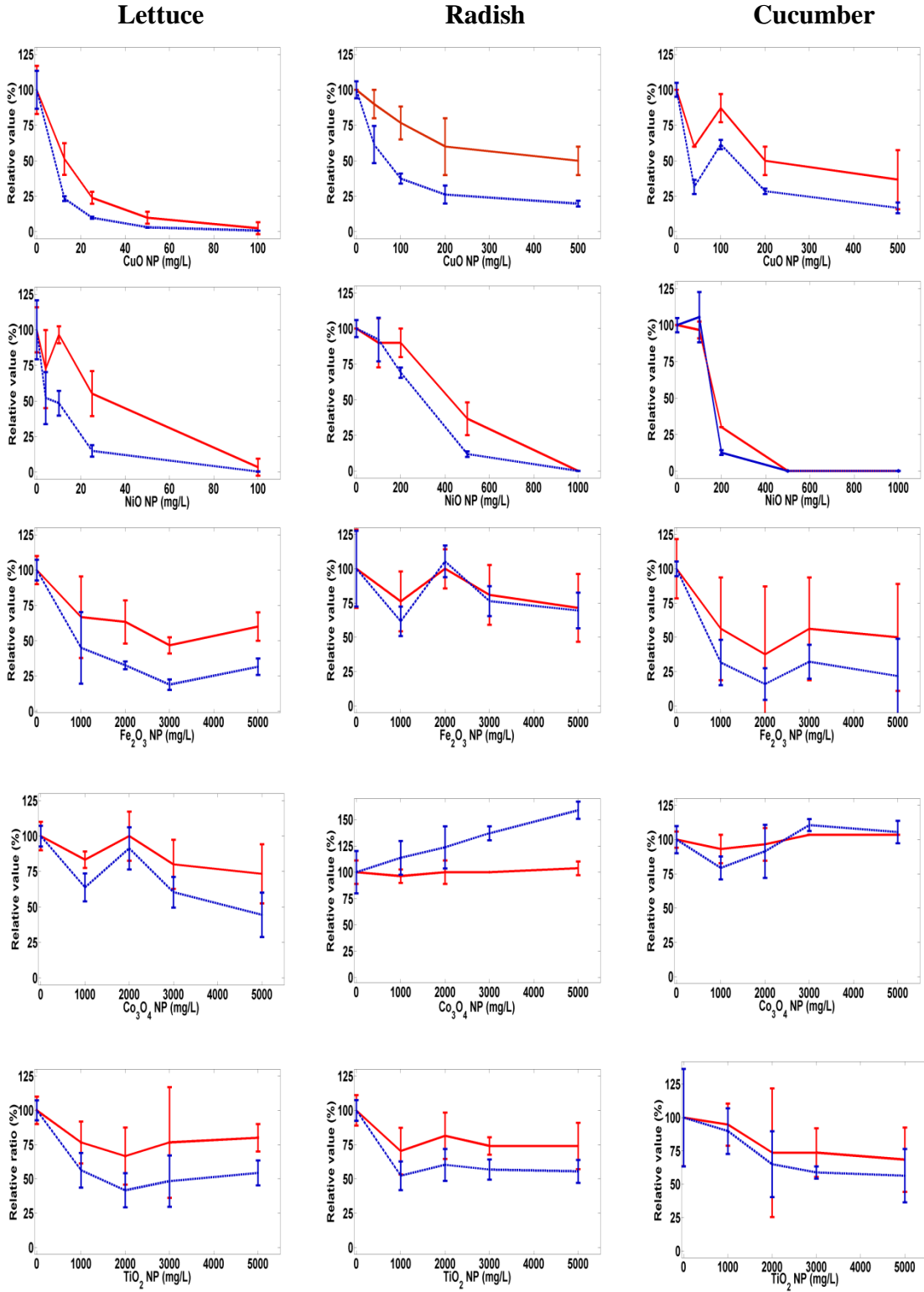
Figure 4-2. Radish seeds (a) Incubated in dH<sub>2</sub>O; (b) Incubated in 500 mg/L of CuO NPs.

Figure 4-3. Cucumber seeds (a) Incubated in dH<sub>2</sub>O; (b) Incubated in 500 mg/L of CuO NPs.

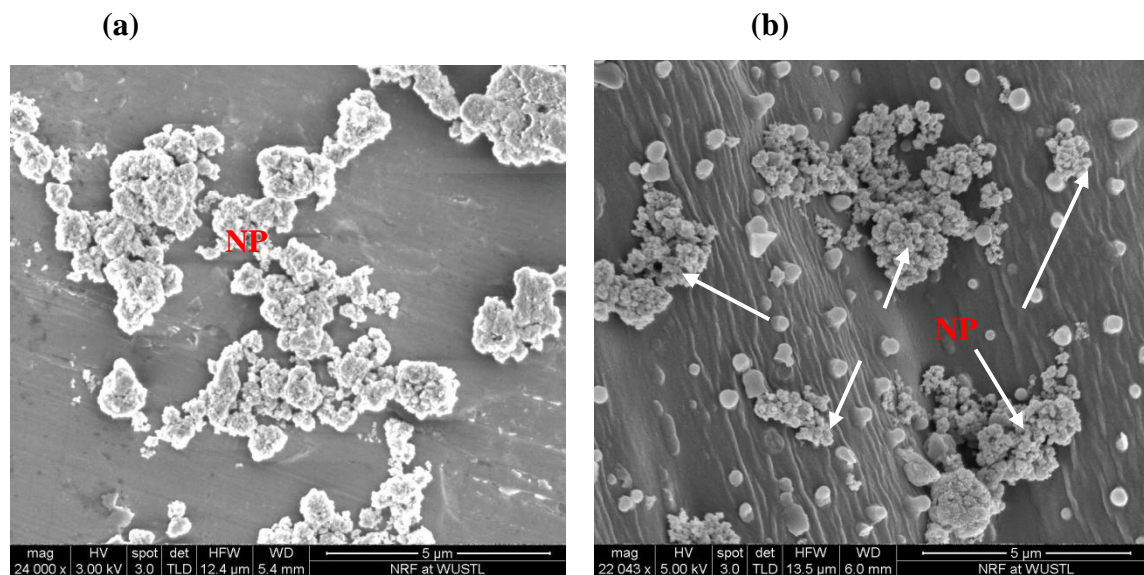
**Figure 1. Flow chart of experimental procedures**



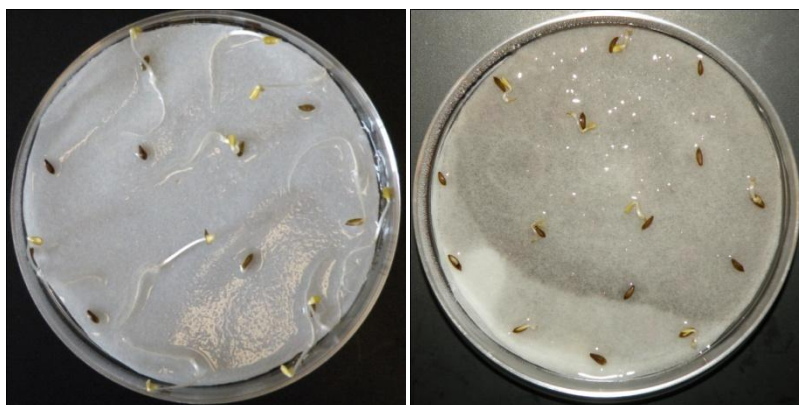
**Figure 2. Effects of NPs on seed germination and elongation. Red line: relative germination rate; Blue dash line: germination index.**



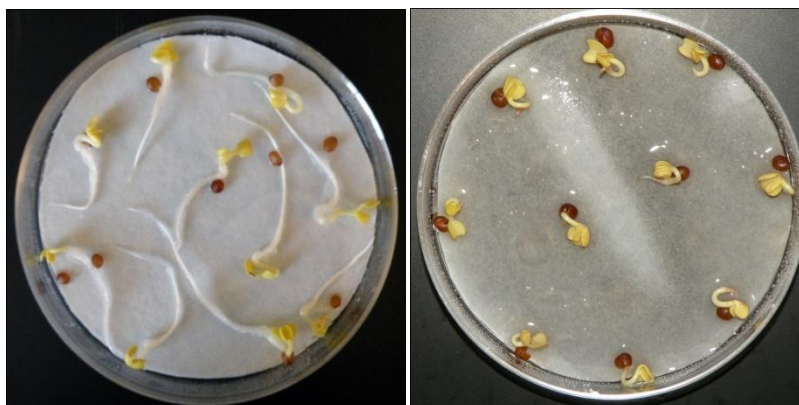
**Figure 3. SEM images for NPs/lettuce seeds. In the aqueous phase, the SEM image shows that metal oxide NPs ( $\text{TiO}_2$  NPs 1000 mg/L) (a) and (CuO NPs 1000 mg/L) were adsorbed on the seed surface (b).**



**Figure 4. Effects of CuO NPs on seed germination and root elongation (incubation at 25 °C in dark for 3 days, NPs could be observed on the seed surface.)**



1. Lettuce seeds (a) Incubated in dH<sub>2</sub>O; (b) Incubated in 500 mg/L of CuO NPs.



2. Radish seeds (a) Incubated in dH<sub>2</sub>O; (b) Incubated in 500 mg/L of CuO NPs.



3. Cucumber seeds (a) Incubated in dH<sub>2</sub>O; (b) Incubated in 500 mg/L of CuO NPs.

**Table 1.** Effects of NPs on seeds activities

NP	Lettuce		Radish		Cucumber	
	EC <sub>50</sub> (mg/L)	GI affected by 1000 mg/L NP	EC <sub>50</sub> (mg/L)	GI affected by 1000 mg/L NP	EC <sub>50</sub> (mg/L)	GI affected by 1000 mg/L NP
<b>CuO</b>	12.9	-100%*	397.6	-100%*	175.4	-100%*
<b>NiO</b>	27.9	-100%*	400.7	-100%*	228.2	-100%*
<b>Fe<sub>2</sub>O<sub>3</sub></b>	> 5000	-55.0%*	> 5000	-38.4%	1682	-68.4%*
<b>TiO<sub>2</sub></b>	> 5000	-36.2%	> 5000	-47.6%*	> 5000	-10.2%
<b>Co<sub>3</sub>O<sub>4</sub></b>	> 5000	-43.6%	> 5000	+13.7%	> 5000	-20.7%

GI – Germination Index; ‘+’ - enhancement, ‘-’ – inhibition, ‘\*’ – significant difference

**Table 2.** EC<sub>50</sub> values of Cu<sup>2+</sup>/Ni<sup>2+</sup> vs. released ions from NPs at their EC<sub>50</sub> concentration

Seeds		Lettuce	Radish	Cucumber
EC <sub>50</sub> for ions (mg/L) *	Cu <sup>2+</sup>	4.9 [3.9, 6.0]	8.0 [5.8, 11.0]	4.8 [3.5, 6.6]
	Ni <sup>2+</sup>	8.8 [6.5, 11.9]	18.7 [15.9, 22.0]	15.7 [12.6, 19.6]
Released ions in solution from CuO and NiO NPs (mg/L) **	Cu <sup>2+</sup>	0.20 ± 0.16 (13)	1.75 ± 0.45 (400)	0.47 ± 0.28 (230)
	Ni <sup>2+</sup>	0.26 ± 0.19 (28)	1.97 ± 0.64 (400)	1.32 ± 0.11 (175)

\* Values of 95% confidence interval of free metal ions were in the bracket [].

\*\* Concentrations of released metal ions from NP solutions incubated with different seeds: NPs in the experiments were at the concentrations of their approximate respective EC<sub>50</sub> values (in parentheses). Data were averaged based on duplicated samples.

**Table 3.** Size distribution of typical metal oxide NP solutions

Total *NP in solution (mg/L)	CuO	NiO	Fe <sub>2</sub> O <sub>3</sub>	Co <sub>3</sub> O <sub>4</sub>	TiO <sub>2</sub>
	100	100	1000	1000	1000
**Average size (nm)	984	576	246	440	562

\* NPs suspension were prepared in dH<sub>2</sub>O (pH = 7)

\*\* Average sizes (Z average) were determined by DLS after NPs incubated at room temperature for 30 min.



# Electrospray Facilitates the Germination of Plant Seeds

Stephen G. Wu<sup>\*</sup>, Li Huang<sup>\*</sup>, Jennifer Head, Madelyn Ball, Yinjie J. Tang, Da-Ren Chen<sup>#</sup>

Department of Energy, Environmental and Chemical Engineering, Washington University in St. Louis, St. Louis, Missouri 63130, USA

## *Abstract*

We proposed a new approach to enhance the plant seed germination via the electrospray of nanoparticles (NPs). A single-capillary electrospray system with a particle deposition stage (where seeds are placed) was set up for this investigation. For demonstration, lettuce (*Lactuca sativa*) seeds were bombarded by TiO<sub>2</sub> NPs via the electrospray for 2~4 minutes in order to promote their germination. Based on our study, the enhancement on germination was significant in cases with aged seeds or seeds placed in an unfavorable growth condition (e.g., low pH medium). The electrospray of other NPs (i.e., Au and CuO) were as also shown to be effective in enhancing the germination of aged lettuce seeds. TEM (Transmission Electron Microscopy) and SEM-EDX (Scanning Electron Microscopes and Energy-Dispersive X-ray Spectroscopy)

analyses suggested that sprayed NPs penetrate the seed coat via the frequent bombardment of NPs at high speeds, thus breaking the coat-imposed seed dormancy. The enhancement on the germination of grass seeds was also observed in this study. The proposed seed treatment may have the potential to improve the germination of various recalcitrant crop seeds.

***Keywords:*** aerosolized TiO<sub>2</sub> NP, crop seeds, lettuce, shelf life, TEM, seed dormancy

## 1. Introduction

In recent years, Nanotechnology and Nanoparticle-related research have undergone rapid growth in various fields, including nanomedicines, drug delivery, biomedical imaging and sensing, and solar energy conversion (Yoo *et al.* 2011; Guo and Dong 2011; Wei *et al.* 2007). Nanoparticles are defined as objects with at least two dimensions less than 100 nm. The unique surface area and solubility of nanoparticles distinguish them from their molecular and bulk counterparts, contributing to various biological effects (Menard *et al.* 2011; Nel *et al.* 2006). In work related to plants, several studies have focused on the translocation of NPs in plants (Lin and Xing 2008; Khodakovskaya *et al.* 2009; Ma *et al.* 2010b). For instance, nanomaterials can be used as a vector to deliver DNAs or other chemicals into plant cells and tissues (Nair *et al.* 2010; Torney *et al.* 2007; González-Melendi *et al.* 2008; Liu *et al.* 2009; Martin-Ortigosa *et al.* 2012). There are also extensive studies on nanotoxicity on plants in the hopes of advancing nanobiotechnology into agricultural applications while relieving the public concern of its potential risk (Rico *et al.* 2011). Zheng *et al.* (2005) has showed that a low dosage of TiO<sub>2</sub> NPs had no harmful effects on spinach plants, but rather promoted photosynthesis and nitrogen metabolism that benefited the plant's growth (Zheng *et al.* 2005). Multi-walled carbon nanotubes and zinc oxide NPs were found to be able to stimulate the seed germination, thus enhancing the plant's growth in the aqueous culture (Khodakovskaya *et al.* 2009; Prasad *et al.* 2012). More, a series of approaches (i.e., genetic, photo-thermal and photo-acoustic methods) were combined to characterize the interactions between multiwalled carbon nanotubes and tomato tissues, providing new insights into the gene transcription regulations of plants under the influence of nanomaterials (Khodakovskaya *et al.* 2011). However, colloidal suspensions of NPs were used in the previous literature. It is known that NPs in suspended solutions tend to agglomerate in general, resulting

in the reduction of their nanoscale effect (Nel *et al.* 2006). It is thus more desirable to have the NPs in their singlet form to study the effect of NPs. Electro spray has been demonstrated to accomplish the above task (Kim *et al.* 2010).

In this study, we employed a single-capillary electro spray setup to simultaneously disperse and deliver individual NPs onto the surface of plant seeds. TiO<sub>2</sub> NPs were applied in this investigation because of their low cost and low toxicity. *Lactuca sativa* (an edible lettuce) was selected as an example plant to demonstrate the feasibility of the proposed treatment for crop seeds. Tests on other NPs (Au and CuO) and grass seeds were also performed to support our idea.

## **2. Experiment**

### **2.1 Electro spray setup**

The detail of the experimental setup has been described in the work of Wu *et al.* (2010). A brief description is provided herein for the reference. The schematic diagram of the experimental setup is shown in Figure 1. The single-capillary electro spray setup consisted of four components: a spray head, a particle deposition stage, a high voltage power supply (Bertan Model 230), and an optical monitor system. The electro spray head was a single capillary, connected to a syringe driven by a programmable Harvard syringe pump (PHD 2000). Sprayed TiO<sub>2</sub> suspension was fed through the spray head at 2  $\mu\text{L}/\text{min}$ . A non-uniform electrical field was established between the spray head and deposition stage by applying a positive high voltage on the spray nozzle and electrically grounding the stage. The deposition stage provided a platform on which the lettuce seeds (typically 30 lettuce seeds for each run) were exposed to the NP electro spray. The distance between the capillary tip and the seed platform was kept at 2.0 cm. In

this study, a typical working voltage was about 7 kV ~ 10 kV for operating the electrospray at the so-called cone-jet mode (Chen *et al.* 1995). The optical monitor system, which included a microscopic lens (InfiniGage, Infinity Photo-Optical Co. Japan), CCD camera (XC-ST 70, Infinity Photo-Optical Co. Japan) and a LCD screen, magnified the liquid meniscus at the capillary exit in order to monitor the cone-jet operation.

All spray suspensions were freshly prepared by dispersing the NPs (via 1min of sonication, 15 W, Misonix XL-2000 Ultrasonic Liquid Processors, NY, US) in 1 mM of a sodium acetate buffer (pH = 7) to ensure that the conductivity of sprayed suspensions was in the range of 150-200  $\mu$ S/cm.

## **2.2 Chemicals and Materials.**

TiO<sub>2</sub> NPs (30-50 nm, rutile) and CuO NPs (30-50 nm) were obtained from Nanostructured & Amorphous Materials, Inc. (Houston, TX, US), while Au NPs (50 nm) was obtained from BBInternational Inc. (Cardiff, UK). The lettuce seeds (*Lactuca sativa*, Black Seeded Simpson, #2846) were purchased from Ferry-Morse Seed Co. (Fulton, KY, US), while the yarrow (*Achillea millefolium*) and common ragweed (*Ambrosia artemisiifolia*) seeds were purchased from Herbiseed Company (Twyford, UK).

To investigate the effect of the electrospray treatment on seed germination in an acidic or basic environment (e.g., to mimic the growth conditions in acidic & alkaline soil), we prepared the following buffer solutions for the seed culture: 2 mM of citric acid buffer for a pH of 3; 2 mM of MES (2-(N-morpholino) ethanesulfonic acid) buffer for a pH of 5; 2 mM of HEPES (4-2-hydroxyethyl-1-piperazineethanesulfonic acid) buffer for a pH of 7; 2 mM of Tricine (N-tris

(hydroxymethyl) methylglycine) buffer for a pH of 9; and 2 mM of N-cyclohexyl-3-aminopropanesulfonic acid buffer for a pH of 11 (Reddy and Singh 1992).

### **2.3 Determination of size distribution of Electro sprayed NPs.**

The size distribution of TiO<sub>2</sub> NPs after electro spray was characterized by a scanning mobility particle sizer (SMPS, TSI, USA) via the electro spray setup described by Chen *et al* (1995). The size distribution of the TiO<sub>2</sub> NPs in suspension was determined by dynamic light scattering (DLS) (Zetasizer, Malvern Instruments, Worcestershire, UK).

### **2.4 Protocol for SEM and TEM Images of Treated Seeds**

To observe the NPs' bombardment on the seeds, the sprayed lettuce seeds were air dried and subjected to scanning electron microscopy (SEM, Nova 2300 FEI, OR, US). When necessary, samples were coated with gold by a low vacuum sputter coater (SPI supplies, PA, US) to increase the surface conductivity before imaging. To validate the penetration of nanoparticles through the seed coat, a TEM (Transmission Electron Microscope, JEOL 1200 EX, MA, US) was used to image the cross sections of treated seeds. Seeds were first fixed overnight at 4 °C in a phosphate buffer solution containing 2% paraformaldehyde and 2.5% glutaraldehyde (pH 7.2), followed by post-fixing with 1% osmium tetroxide for 2 hours after a phosphate buffer wash. Subsequently, these samples were stained with 1% aqueous uranyl acetate overnight at 4 °C. After dehydration with sequential ethanol concentrations ranging from 50 to 100%, sections of each sample were cut with a Leica Ultracut UCT ultramicrotome (Leica Microsystems Inc., Bannockburn, IL, US) and placed on grids for TEM imaging. To further confirm the existence of nanoparticles in the seed sections, the sections were characterized by an energy dispersive X-ray spectroscopy, coupled with SEM (SEM-EDX) for elemental analysis of titanium.

## 2.5 Seed germination.

Thirty lettuce seeds were placed on the electrically-grounded deposition stage to undergo NP electrospray treatment for time periods of approximately 5 minutes. After the spray, all seeds were transferred into petri dishes (15 mm x 100 mm) containing 5 mL of medium solution. Fifteen lettuce seeds were incubated in each petri dish to ensure adequate space to germinate and grow. All dishes were sealed and incubated at 25 °C in the dark (Reddy and Singh 1992). After the end of the germination period (usually three days), the seedling's length were measured (note: majority of lettuce seeds were germinated in the first three days of incubation). The lettuce seeds with seedling lengths longer than 1 cm were considered to be well-germinated seeds. A total of 60 seeds were used for each treatment condition for analysis. Similar protocols were performed for the other seeds with slight variances in seed number per petri dish (i.e., Yarrow seed: about 120 per petri dish/2 weeks; Common ragweed seed: 50 per petri dish/2 weeks.).

The impact of NP electrospray was evaluated based on the seed germination percentage, calculated by the following equation:

$$\text{Germination percentage} = \frac{\text{Number of seeds well germinated}}{\text{Total number of seeds}}$$

Note that we used the germination percentage as the sole parameter to evaluate the effect of the NP electrospray treatment. The Statistics Toolbox of MATLAB was employed to conduct the data analyses based on the Z test for germination percentage, where statistically significant was defined as  $P < 0.05$ .

## 3. Results and Discussion

### 3.1 The electrospray of TiO<sub>2</sub> NP onto plant seeds

Figure 2 shows the size distribution measurement of TiO<sub>2</sub> NPs in aqueous solutions. It is evidenced that TiO<sub>2</sub> NPs in aqueous solution formed agglomerates up to several μm in size. The agglomeration of NPs reduces the particle size effect, particularly on cellular function (Wu *et al.* 2010). By adjusting the electrical conductivity and the feeding flow rate of TiO<sub>2</sub> NP suspensions, the single-capillary electrospray operated at the cone-jet mode ensured the production of monodisperse droplets in sub-micrometer and nanometer size ranges (Chen *et al.* 1995; Jaworek 2007). When the concentration of TiO<sub>2</sub> NPs in the solutions was diluted, the solvent in the droplets evaporated during electrospray so that singlet TiO<sub>2</sub> NPs could be dispersed onto seed surface.

Figure 2 shows the size distributions measured by the SMPS when freshly prepared TiO<sub>2</sub> NP suspensions of 1 g/L are electrosprayed. The measured particle size distribution exhibited a narrow peak at 36.0 nm, suggesting that most particles were singly dispersed after the electrospray process. Once electrosprayed, gas-borne NPs were accelerated by the presence of a DC electric field. The terminal velocity of sprayed NPs was estimated to be in the range of 100 to 500 m/s prior to bombarding the seed coat (Pui and Chen 2000). The penetration of NPs through the seed coat under the proposed treatment was verified by the SEM images (shown in Figure 3). To our knowledge, this is the first report demonstrating the application of using aerosolized NPs to penetrate plant seed coats.

The seed coat, consisting of layers of the testa and endosperm envelope (Welbaum *et al.* 1998), provides protection against the entry of parasites and mechanical injury. However, the coat may impose the seed dormancy (Zeng *et al.* 2005). To improve seed germination, we



employed NPs with high electrical charges to facilitate the break of seed coat by accelerating them in a DC electrical field. Different from the delivery of particles into plant cells via high-pressure gas (Torney *et al.* 2007; Gordon-Kamm *et al.* 1990), the electrospray accelerates particles primarily through the space charge effect (due to the presence of highly charged particles in high concentration) (Chen *et al.* 2000). The NP's entry into seeds via electrospray demonstrates the potential of this proposed method for delivering various materials (DNAs or plant hormones) into embryos (Gu *et al.* 2011). By tuning the electrical field strength and controlling the charges on the NPs, a broad range of particle velocities can be achieved for the bombardment of targeted organelles. The optimal speeds needed for the successful delivery of nanomaterials (varying in size, density, and shape) into various types of seeds would be an interesting topic to explore in the near future.

### **3.2 Seed germination enhancement**

Two factors may affect the germination of seeds. First factor, if the seeds were stored for a long period of time, their germination are typically reduced (i.e., become more recalcitrant). Second factor is the soil pH, an important factor influencing seed germination, and plants have to adapt in acidic or alkaline environments. Thereby, we tested the effectiveness of NP treatment for seed germination under unfavorable conditions. Figure 4 summarizes the experimental results of seed germination of lettuce seeds when treated with TiO<sub>2</sub> NPs electrospray and incubated in buffer solutions. Figure 4a is the germination percentage of aged lettuce (stored for 10 months) treated by the NP electrospray as a function of NP concentration and spray time. After planting aged lettuce seeds in an unfavorable pH condition for germination (pH 5), treated lettuce seeds showed clear germination enhancement. Aged lettuce seeds (stored for over 10 months) in a MES buffer (pH = 5, a typical pH in acidic soils) had a natural germination percentage of ~ 40%.

The germination of the same lettuce seeds reached its peak value (65%) when seeds were pretreated by electro spraying  $\text{TiO}_2$  NP suspension (at the concentration of 1 g/L) for about 4 min. Prolonged NP electro spraying displayed less effectiveness on the seeds' germination.

Figure 4b shows the germination of lettuce seeds when they were incubated in the buffer solutions of various pH values. Three sets of fresh lettuce seeds were used in each test incubation condition: untreated seeds (black bars), pre-treated seeds by electro spraying the solvent only (grey bars), and pre-treated seeds by electro spraying with  $\text{TiO}_2$  NPs (white bars). When lettuce seeds were fresh (i.e., recently purchase from the seed company), the control seeds (seeds without pretreatment) usually had high germination percentages (~ 80%) even without NP electro spray treatment under a wide range of pH growth conditions (i.e., pH = 5~9, normal soil range). When the lettuce seeds were placed in an extreme harsh pH conditions (i.e., pH = 3), the control seeds showed minimal germination (~ 0%) while the NP-treated seeds recovered to a germination percentage of 20%. This result indicates that NP-treated plant seeds may be useful in to vegetate the polluted lands (e.g., phytoremediation of acid contaminated soil).

### **3.3 Effect of various NPs on the seed germination enhancement**

To verify the feasibility of this proposed treatment, CuO and Au NPs were also applied in this study to pre-treat different sets of aged lettuce seeds. Figure 5a shows the seed germination results when aged lettuce seeds (stored for 16 months) were electro sprayed by CuO NP suspensions. Figure 4b gives the germination percentage when the same aged lettuce seeds were pre-treated by electro spraying Au NP suspensions at various concentrations. The germination percentages for the controls were also included in the figure as the reference. The results indicate that the electro spray of either CuO NP (1 g/L) or Au NP suspensions had significant enhancement effects on aged lettuce seeds. Similar to  $\text{TiO}_2$  NP electro spray, the best seed

germination percentage for CuO NP electro spray were obtained in the case of a short spray time of 4 min. For the 4-min spray period, the best germination percentage occurred at the concentration of  $10^7$  NPs/cm<sup>3</sup> for Au NP suspensions. Based on the above observation, nano-materials of different compositions are possible to use for enhancing the seed germination.

We further performed the comparison test to characterize the effectiveness of various NPs on the germination of aged lettuce seeds in a prolonged incubation period (shown in Figure 6). Figure 6a shows the germination percentage under the electro spray treatment by suspensions of TiO<sub>2</sub>, CuO and Au NPs. The results indicate that the treatment by electro spraying either TiO<sub>2</sub> or Au NP suspensions gave the most improved germination percentage. These pretreated seeds also germinated faster, while the control samples (seeds without pretreatment) showed a longer germination window (over seven days). Although the electro spray of the buffer solution (1 mM NaAc) accelerated the seed germination, we did not observe an improvement in the overall germination percentages in a seven-day incubation period. For agricultural and horticultural crops, delayed and sporadic germination is undesired because it reduces the harvest efficiency. NP-electro spray treated seeds may have the advantage of better crop productivity because of their early and homogeneous germination behavior.

No significant difference on the shoot length of germinated lettuce seeds among all treatment conditions (only the case with CuO NPs showing somewhat negative impact on the shoot length) was observed in this study (shown in Figure 6b). This observation suggests that the proposed treatment has no adverse effect on the early state of shoot development and could probably improve overall seed germination.

### **3.4 Mechanism for lettuce seed germination enhancement by TiO<sub>2</sub> NP electro spray**

Plant seeds are incapable of germination at 100% even under the favorable conditions of temperature and hydration. In this study, we have demonstrated that the electrospray of NP suspensions can enhance the germination of lettuce seeds while immersing seeds in the buffer solution only (without spray) showed no significant impact on the seed germination percentage. Although the electrospray of buffer solutions without NPs showed minor positive enhancement on the seed germination, it was not as effective as the electrospray of NP suspensions. The possible explanation to the observed enhancement on seed germination due to NP electrospray treatment is the breaking of the coat-imposed seed dormancy. Electrosprayed NPs bombarding the seed coat may weaken the structure of the seed coat (Figure 3), which is considered as an influential factor in controlling seed dormancy (Gleiser *et al.* 2004).

In an aqueous suspension of NPs, it has been previously reported that NPs can slowly penetrate into seeds of various types and affect their metabolism *in vivo* (Navarro *et al.* 2008a; Ma *et al.* 2010b). For example, multiwalled carbon nanotubes (MWCNTs) were able to penetrate through the coat of tomato seeds after several days of co-incubation (Khodakovskaya *et al.* 2009). In such bulk solutions, NPs are often observed in the agglomerate form and their interaction with seeds is weak. In this study, the electrospray process effectively disperses most of NPs in singlet form, accelerates the velocity of NPs via the presence of electric field and space charge effect, and bombards the seeds by NP collision at high frequency. The proposed process thus increases the chance for NPs entering into seeds via piecing their coats or through natural pores. In case that nano-sized holes are created on the seed coat, oxygen transfer and water uptake might occur and drive the metabolic process for plant growth (Khodakovskaya *et al.* 2009).

Further, the effect of metal ions on seed germination can be excluded in our study as TiO<sub>2</sub> NPs are considered to be insoluble (< 5 ppb, confirmed by our ICP-MS measurement).

Because the incubation process for seed germination took place in the dark, the production of oxidative  $\text{H}_2\text{O}_2$  through the reaction of light with  $\text{TiO}_2$  NPs was minimized in this study. It may explain why toxicity of  $\text{TiO}_2$  NPs, reported in the works of Menard *et al.* (2011) and Hund-Rinke and Simon (2006), was not observed in our investigation (Menard *et al.* 2011; Hund-Rinke and Simon 2006). For the case of CuO NP spray, the decrease in seed germination effectiveness is presumably because of the known phytotoxicity of CuO NP (e.g., release of  $\text{Cu}^{2+}$  ions to interfere with seed functions) (Wang *et al.* 2010b; Baek and An 2011; Karlsson *et al.* 2008).

### **3.5 Shelf life of treated lettuce seeds**

The shelf life test was performed on the lettuce seeds which had undergone the NP electro spray treatment, in order to evaluate the practical potential of the proposed method (Schwember and Bradford 2010). Aged lettuce seeds treated by NP electro spray were sealed and stored in the dark and at room temperature for various time periods (i.e., 1 day, 1 week, and 1 month) prior to the incubation. Figure 7 shows the germination percentage of aged and  $\text{TiO}_2$  NP treated lettuce seeds after the defined storage periods. Significant enhancement on the germination of NP-treated seeds was observed when compared with the control. The germination percentage of aged seeds slightly dropped after one month storage, indicating that the seed coat of NP-treated seeds remained a sufficient protection for the seed embryo during the storage.

### **3.6 NP electro spray treatment of grass seeds**

Several types of grass seeds (e.g., recalcitrant yarrow seeds and ragweed seeds) which have naturally low germination percentages under favorable growth conditions (typically 1~2% each year) were also used in this study to demonstrate broad applications of the proposed seed treatment. In this part of experiment, recalcitrant yarrow and ragweed seeds were subjected to 5 min electro spray of suspension with 1 g/L  $\text{TiO}_2$  NPs. After culturing in de-ionized  $\text{H}_2\text{O}$  at pH = 7,

the germination percentages of recalcitrant yarrow seeds were enhanced from 1.6 (untreated) % to 6.8% (pre-treated) (shown in Figure 8). The germination percentage of ragweed seeds was improved from 1% (untreated) to 3% (treated). The germination percentage of grass seeds studied herein has the potential to be further improved by optimizing the electrospray operation conditions and NP concentration in spray suspensions. Because of the diversity of structure of seeds & coat of seeds in different species (Finch-Savage and Leubner-Metzger 2006), it may require different electrospray conditions to achieve the best result for various species as compared to that for lettuce seeds.

#### **4. Conclusion**

Electrospray of NP suspension was proposed to treat plant seeds for the enhancement of seed germination. A single-capillary electrospray system having a particle deposition stage was set up for this investigation. By applying a positive high voltage on the spray capillary and the electrically grounded stage, the electrospray was operated at the cone-jet mode for the production of monodisperse droplets. The solvent in droplets evaporated right after the droplet production, resulting in highly charged NPs in singlet form. Charged NPs were then accelerated in the presence of DC electrical field and space charge effect, and bombarded lettuce seeds on the deposition stage. Our study demonstrates that NP electrospray effectively dispersed NPs for breaking the seed coat to enhance seed germination. Our study further shows that the proposed treatment enables seeds to germinate under harsh environments (i.e., low pH soil). The enhancement on seed germination were also observed when electrospraying NP suspension of CuO and Au, independent of particle composition. Such treatment was further investigated and proven to be effective for certain weed seeds.

Our proposed seed treatment method can be further optimized by varying the NP concentrations/sizes, electric fields, and spray time. Such method can be cost effective for scaled-up for industrial applications because dilute NP suspensions were used in the treatment (estimate: 1 g of NPs can spray 3.6 million lettuce seeds). We believe that the proposed NP electrospray has the potential to be applied to various plant seeds (Pui and Chen 2000; Nadjafi *et al.* 2006). Meanwhile, environmental friendly NPs (Rieter *et al.* 2008; Yan *et al.* 2010) can be employed in the future to alleviate the concerns on the cytotoxicity of metal oxide NPs (Walser *et al.* 2012). In addition, the use of engineered NPs carrying DNA, plant hormone or other chemicals in the proposed process for seed treatment may open up new opportunities for broad application of nanotechnology in agricultural industry.

## **Acknowledgements**

We appreciate valuable suggestions from Prof. G. Welbaum at Virginia Tech. We would also thank Prof. D. Ho from the Department of Biology at Washington University in St. Louis and Prof. I. Kong at Yeungnam University for close readings and helpful discussions of this manuscript.



## References

- Baek, Y.-W. and An, Y.-J. (2011). Microbial Toxicity of Metal Oxide Nanoparticles (CuO, NiO, ZnO, and Sb<sub>2</sub>O<sub>3</sub>) to *Escherichia coli*, *Bacillus subtilis*, and *Streptococcus aureus*. *Sci. Total Environ.* 409:1603-1608.
- Roco, M. C. (2005). Environmentally Responsible Development of Nanotechnology. *Environ. Sci. Technol.* 39:106A-112A.
- Lin, D. and Xing, B. (2008). Root Uptake and Phytotoxicity of ZnO Nanoparticles. *Environ. Sci. Technol.* 42:5580-5585.
- Klaine, S. J., Alvarez, P. J. J., Batley, G. E., Fernandes, T. F., Handy, R. D., Lyon, D. Y., Mahendra, S., McLaughlin, M. J. and Lead, J. R. (2008). Nanomaterials in the environment: Behavior, fate, bioavailability, and effects. *Environ. Toxicol. Chem.* 27:1825-1851.
- Marambio-Jones, C. and Hoek, E. (2010). A review of the antibacterial effects of silver nanomaterials and potential implications for human health and the environment. *J. Nanoparticle Res.* 12:1531-1551.
- Wang, Z., Lee, Y.-H., Wu, B., Horst, A., Kang, Y., Tang, Y. J. and Chen, D.-R. (2010b). Anti-microbial activities of aerosolized transition metal oxide nanoparticles. *Chemosphere* 80:525-529.
- Ji, J., Long, Z. and Lin, D. (2011). Toxicity of oxide nanoparticles to the green algae *Chlorella* sp. *Chem. Eng. J.* 170:525-530.
- Ma, Y., Kuang, L., He, X., Bai, W., Ding, Y., Zhang, Z., Zhao, Y. and Chai, Z. (2010a). Effects of rare earth oxide nanoparticles on root elongation of plants. *Chemosphere* 78:273-279.
- Wang, B., Feng, W.-Y., Wang, T.-C., Jia, G., Wang, M., Shi, J.-W., Zhang, F., Zhao, Y.-L. and Chai, Z.-F. (2006b). Acute toxicity of nano- and micro-scale zinc powder in healthy adult mice. *Toxicol. Lett.* 161:115-123.

Navarro, E., Baun, A., Behra, R., Hartmann, N., Filser, J., Miao, A.-J., Quigg, A., Santschi, P. and Sigg, L. (2008a). Environmental behavior and ecotoxicity of engineered nanoparticles to algae, plants, and fungi. *Ecotoxicology* 17:372-386.

Menard, A., Drobne, D. and Jemec, A. (2011). Ecotoxicity of nanosized TiO<sub>2</sub>. Review of in vivo data. *Environ. Pollut.* 159:677-684.

Limbach, L. K., Wick, P., Manser, P., Grass, R. N., Bruinink, A. and Stark, W. J. (2007). Exposure of Engineered Nanoparticles to Human Lung Epithelial Cells: Influence of Chemical Composition and Catalytic Activity on Oxidative Stress. *Environ. Sci. Technol.* 41:4158-4163.

Di Salvatore, M., Carafa, A. M. and Carratù G. (2008). Assessment of heavy metals phytotoxicity using seed germination and root elongation tests: A comparison of two growth substrates. *Chemosphere* 73:1461-1464.

Lin, D. and Xing, B. (2007). Phytotoxicity of nanoparticles: Inhibition of seed germination and root growth. *Environ. Pollut.* 150:243-250.

Yang, L. and Watts, D. J. (2005). Particle surface characteristics may play an important role in phytotoxicity of alumina nanoparticles. *Toxicol. Lett.* 158:122-132.

Mondal, A., Basu, R., Das, S. and Nandy, P. (2011). Beneficial role of carbon nanotubes on mustard plant growth: an agricultural prospect. *J. Nanoparticle Res.* 13:4519-4528.

Rico, C. M., Majumdar, S., Duarte-Gardea, M., Peralta-Videa, J. R. and Gardea-Torresdey, J. L. (2011). Interaction of Nanoparticles with Edible Plants and Their Possible Implications in the Food Chain. *J. Agr. Food Chem.* 59:3485-3498.

Rivetta, A., Negrini, N. and Cocucci, M. (1997). Involvement of Ca<sup>2+</sup>-calmodulin in Cd<sup>2+</sup> toxicity during the early phases of radish (*Raphanus sativus* L.) seed germination. *Plant, Cell Environ.* 20:600-608.

Wang, X., Sun, C., Gao, S., Wang, L. and Shuokui, H. (2001). Validation of germination rate and root elongation as indicator to assess phytotoxicity with *Cucumis sativus*. *Chemosphere* 44:1711-1721.

U.S.EPA (1996). Ecological Effects Test Guidelines (OPPTS 850.4200): Seed Germination/Root Elongation Toxicity Test, E. P. Agency, ed.

Reddy, K. N. and Singh, M. (1992). Germination and Emergence of Hairy Beggarticks (*Bidens pilosa*). *Weed Sci.* 40:195-199.

El-Temsah, Y. S. and Joner, E. J. (2010). Impact of Fe and Ag nanoparticles on seed germination and differences in bioavailability during exposure in aqueous suspension and soil. *Environ Toxicol:n/a-n/a*.

Barrena, R., Casals, E., Colón, J., Font, X., Sánchez, A. and Puentes, V. (2009). Evaluation of the ecotoxicity of model nanoparticles. *Chemosphere* 75:850-857.

Thompson, W. H., Leege, P. B., Millner, P. D. and Watson, M. E. (2001). Test methods for the examination of composting and compost, USCC and USDA, eds.

Hamilton, M. A., Russo, R. C. and Thurston, R. V. (1977). Trimmed Spearman-Kärber method for estimating median lethal concentrations in toxicity bioassays. *Environ. Sci. Technol.* 11:714-719.

Lu, C. M., Zhang, C. Y., Wen, J. Q., Wu, G. R. and Tao, M. X. (2002). Research of the effect of nanometer materials on germination and growth enhancement of *Glycine max* and its mechanism *Soybean Sci.* 21:168-172.

Khodakovskaya, M., Dervishi, E., Mahmood, M., Xu, Y., Li, Z., Watanabe, F. and Biris, A. S. (2009). Carbon Nanotubes Are Able To Penetrate Plant Seed Coat and Dramatically Affect Seed Germination and Plant Growth. *ACS Nano* 3:3221-3227.

- Nair, R., Varghese, S. H., Nair, B. G., Maekawa, T., Yoshida, Y. and Kumar, D. S. (2010). Nanoparticulate material delivery to plants. *Plant Sci.* 179:154-163.
- Navarro, E., Piccapietra, F., Wagner, B., Marconi, F., Kaegi, R., Odzak, N., Sigg, L. and Behra, R. (2008b). Toxicity of Silver Nanoparticles to *Chlamydomonas reinhardtii*. *Environ. Sci. Technol.* 42:8959-8964.
- Krug, H. F. and Wick, P. (2011). Nanotoxicology: An Interdisciplinary Challenge. *Angew. Chem., Int. Ed.* 50:1260-1278.
- Zhu, H. N., Lu, Q. X. and Abdollahi, K. (2005). Seed Coat Structure of *Pinus Koraiensis*. *Microsc Microanal* 11:1158-1159.
- Zeng, L. W., Cocks, P. S., Kailis, S. G. and Kuo, J. (2005). The role of fractures and lipids in the seed coat in the loss of hardseededness of six Mediterranean legume species. *J. Agric. Sci.* 143:43-55.
- Hu, X., Daun, J. and Scarth, R. (1994). Proportions of C18 : 1n-7 and C18 : 1n-9 fatty acids in canola seedcoat surface and internal lipids. *J. Am. Oil Chem. Soc.* 71:221-222.
- Stark, W. J. (2011). Nanoparticles in Biological Systems. *Angew. Chem., Int. Ed.* 50:1242-1258.
- Yoo, S. I., Yang, M., Brender, J. R., Subramanian, V., Sun, K., Joo, N. E., Jeong, S.-H., Ramamoorthy, A. and Kotov, N. A. (2011). Inhibition of Amyloid Peptide Fibrillation by Inorganic Nanoparticles: Functional Similarities with Proteins. *Angew. Chem., Int. Ed.* 50:5110-5115.
- Guo, S. and Dong, S. (2011). Graphene nanosheet: synthesis, molecular engineering, thin film, hybrids, and energy and analytical applications. *Chem. Soc. Rev.* 40:2644-2672.

- Wei, H., Li, B., Li, J., Wang, E. and Dong, S. (2007). Simple and sensitive aptamer-based colorimetric sensing of protein using unmodified gold nanoparticle probes. *Chem Commun* 0:3735-3737.
- Nel, A., Xia, T., Mädler, L. and Li, N. (2006). Toxic Potential of Materials at the Nanolevel. *Science* 311:622-627.
- Ma, X., Geiser-Lee, J., Deng, Y. and Kolmakov, A. (2010b). Interactions between engineered nanoparticles (ENPs) and plants: Phytotoxicity, uptake and accumulation. *Sci. Total Environ.* 408:3053-3061.
- Torney, F., Trewyn, B. G., Lin, V. S. Y. and Wang, K. (2007). Mesoporous silica nanoparticles deliver DNA and chemicals into plants. *Nat. Nano.* 2:295-300.
- González-Melendi, P., Fernández-Pacheco, R., Coronado, M. J., Corredor, E., Testillano, P. S., Risueño, M. C., Marquina, C., Ibarra, M. R., Rubiales, D. and Pérez-de-Luque, A. (2008). Nanoparticles as Smart Treatment-delivery Systems in Plants: Assessment of Different Techniques of Microscopy for their Visualization in Plant Tissues. *Ann Bot-london* 101:187-195.
- Liu, Q., Chen, B., Wang, Q., Shi, X., Xiao, Z., Lin, J. and Fang, X. (2009). Carbon Nanotubes as Molecular Transporters for Walled Plant Cells. *Nano Lett.* 9:1007-1010.
- Martin-Ortigosa, S., Valenstein, J. S., Sun, W., Moeller, L., Fang, N., Trewyn, B. G., Lin, V. S. Y. and Wang, K. (2012). Parameters Affecting the Efficient Delivery of Mesoporous Silica Nanoparticle Materials and Gold Nanorods into Plant Tissues by the Biolistic Method. *Small* 8:413-422.
- Zheng, L., Hong, F., Lu, S. and Liu, C. (2005). Effect of nano-TiO<sub>2</sub> on strength of naturally aged seeds and growth of spinach. *Biol. Trace Elem. Res.* 104:83-91.

Prasad, T. N. V. K. V., Sudhakar, P., Sreenivasulu, Y., Latha, P., Munaswamy, V., Reddy, K. R., Sreeprasad, T. S., Sajanalal, P. R. and Pradeep, T. (2012). Effect of nanoscale zinc oxide particles on the germination, growth and yield of peanut. *J. Plant Nutr.* 35:905-927.

Khodakovskaya, M. V., de Silva, K., Nedosekin, D. A., Dervishi, E., Biris, A. S., Shashkov, E. V., Galanzha, E. I. and Zharov, V. P. (2011). Complex genetic, photothermal, and photoacoustic analysis of nanoparticle-plant interactions. *Proc. Natl. Acad. Sci.* 108:1028-1033.

Kim, S. C., Chen, D.-R., Qi, C., Gelein, R. M., Finkelstein, J. N., Elder, A., Bentley, K., Oberdörster, G. and Pui, D. Y. H. (2010). A nanoparticle dispersion method for in vitro and in vivo nanotoxicity study. *Nanotoxicology* 4:42-51.

Chen, D.-R., Pui, D. Y. H. and Kaufman, S. L. (1995). Electro spraying of conducting liquids for monodisperse aerosol generation in the 4 nm to 1.8  $\mu\text{m}$  diameter range. *J. Aerosol Sci.* 26:963-977.

Wu, B., Wang, Y., Lee, Y.-H., Horst, A., Wang, Z., Chen, D.-R., Sureshkumar, R. and Tang, Y. J. (2010). Comparative Eco-Toxicities of Nano-ZnO Particles under Aquatic and Aerosol Exposure Modes. *Environ. Sci. Technol.* 44:1484-1489.

Jaworek, A. (2007). Micro- and nanoparticle production by electro spraying. *Powder Technol.* 176:18-35.

Pui, D. Y. H. P., MN) and Chen, D.-r. L., MN) (2000). Electro spraying apparatus and method for introducing material into cells, Regents of the University of Minnesota (Minneapolis, MN), United States.

Welbaum, G. E., Bradford, K. J., Yim, K.-O., Booth, D. T. and Oluoch, M. O. (1998). Biophysical, physiological and biochemical processes regulating seed germination. *Seed Sci. Res.* 8:161-172.

Gordon-Kamm, W. J., Spencer, T. M., Mangano, M. L., Adams, T. R., Daines, R. J., Start, W. G., O'Brien, J. V., Chambers, S. A., Adams, W. R., Willetts, N. G., Rice, T. B., Mackey, C. J., Krueger, R. W., Kausch, A. P. and Lemaux, P. G. (1990). Transformation of Maize Cells and Regeneration of Fertile Transgenic Plants. *Plant Cell* 2:603-618.

Chen, D.-R., Wendt, C. H. and Pui, D. Y. H. (2000). A Novel Approach for Introducing Bio-Materials Into Cells. *J. Nanoparticle Res.* 2:133-139.

Gu, Z., Biswas, A., Zhao, M. and Tang, Y. (2011). Tailoring nanocarriers for intracellular protein delivery. *Chem. Soc. Rev.* 40:3638-3655.

Gleiser, G., Picher, M. C., Veintimilla, P., Martinez, J. and VerdÚ, M. (2004). Seed dormancy in relation to seed storage behaviour in Acer. *Bot. J. Linn. Soc.* 145:203-208.

Hund-Rinke, K. and Simon, M. (2006). Ecotoxic Effect of Photocatalytic Active Nanoparticles (TiO<sub>2</sub>) on Algae and Daphnids (8 pp). *Environ. Sci. Pollut. R* 13:225-232.

Baek, Y.-W. and An, Y.-J. (2011). Microbial toxicity of metal oxide nanoparticles (CuO, NiO, ZnO, and Sb<sub>2</sub>O<sub>3</sub>) to Escherichia coli, Bacillus subtilis, and Streptococcus aureus. *Sci. Total Environ.* 409:1603-1608.

Karlsson, H. L., Cronholm, P., Gustafsson, J. and Möller, L. (2008). Copper Oxide Nanoparticles Are Highly Toxic: A Comparison between Metal Oxide Nanoparticles and Carbon Nanotubes. *Chem. Res. Toxicol.* 21:1726-1732.

Schwember, A. R. and Bradford, K. J. (2010). Quantitative trait loci associated with longevity of lettuce seeds under conventional and controlled deterioration storage conditions. *J. Exp. Bot.* 61:4423-4436.

Finch-Savage, W. E. and Leubner-Metzger, G. (2006). Seed dormancy and the control of germination. *New Phytol.* 171:501-523.

Nadjafi, F., Bannayan, M., Tabrizi, L. and Rastgoo, M. (2006). Seed germination and dormancy breaking techniques for *Ferula gummosa* and *Teucrium polium*. *J. Arid Environ.* 64:542-547.

Rieter, W. J., Pott, K. M., Taylor, K. M. L. and Lin, W. (2008). Nanoscale Coordination Polymers for Platinum-Based Anticancer Drug Delivery. *J. Am. Chem. Soc.* 130:11584-11585.

Yan, M., Du, J., Gu, Z., Liang, M., Hu, Y., Zhang, W., Priceman, S., Wu, L., Zhou, Z. H., Liu, Z., Segura, T., Tang, Y. and Lu, Y. (2010). A novel intracellular protein delivery platform based on single-protein nanocapsules. *Nat. Nano.* 5:48-53.

Walser, T., Limbach, L. K., Brogioli, R., Erismann, E., Flamigni, L., Hattendorf, B., Juchli, M., Krumeich, F., Ludwig, C., Prikopsky, K., Rossier, M., Saner, D., Sigg, A., Hellweg, S., Gunther, D. and Stark, W. J. (2012). Persistence of engineered nanoparticles in a municipal solid-waste incineration plant. *Nat. Nano.* 7:520-524.



## Figure Captions

**Figure 1** A schematic diagram of single-capillary electrospray setup with the particle deposition stage (used in this study).

**Figure 2** The measurement of TiO<sub>2</sub> NP size distributions: (a) for a freshly prepared NP suspension (at 1 g/L TiO<sub>2</sub> NPs) measured by dynamic light scattering; (b) for gas-borne TiO<sub>2</sub> NPs after electrospray; (c) SEM image of freshly prepared NP solution (1 g/L) and (d) SEM image of TiO<sub>2</sub> NPs after electrospray.

**Figure 3** SEM (a), TEM (b) and SEM-EDX (c, d) images of lettuce seeds treated by electrospray of TiO<sub>2</sub> NP suspension at a concentration of 1 g/L for 5 min. The SEM image (a) shows that TiO<sub>2</sub> NPs were individually adsorbed onto the seed surface. The TEM image (b) and SEM-EDX images (c, d) are for the cross sections of treated lettuce seeds. The images evidenced that the TiO<sub>2</sub> NPs can penetrate the coat of lettuce seeds and reside in the seeds. The scale bar in (b) indicates the length of 200 nm on the image.

**Figure 4** Germination percentage of lettuce seeds after treated by TiO<sub>2</sub> NP electrospray as a function of spray time and NP mass concentration in spray solutions: (a) for the case with aged seeds (stored for 10 months) and incubated in buffer solution of pH 5; (b) for the cases with fresh seeds and incubated in the buffer solutions of pH = 3~11. Error bars in the figure are adapted from four replicates in each treatment.

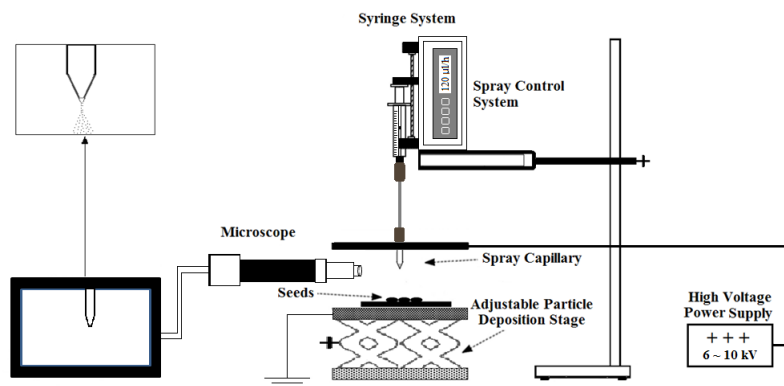
**Figure 5** Germination percentage of lettuce seeds after treated by NP electrospray: (a) for the cases with fresh seeds, using CuO NPs and incubated in the buffer solutions of pH = 7 for three days, and (b) for the cases with aged seeds (stored for 16 months), using Au NPs and incubated in the buffer solutions of pH = 7 for three days. Error bars in the figure are adapted from four replicates in each treatment.

**Figure 6** Comparison of lettuce seed germination after the electrospray treatment using NPs of various materials (i.e., TiO<sub>2</sub>, CuO and Au). Also included in the figure are the data for the control (without the treatment) and the case treated by electrospraying buffer solutions only for the reference. (a) germination percentage of seeds at both Day 3 and Day 7; (b) shoot length of seeds at Day 3 and Day 7; Error bar in (a) and (b) are adapted from four replicates in each treatment with aged seeds (Stored for 14 months and incubated in the buffer of pH = 5).

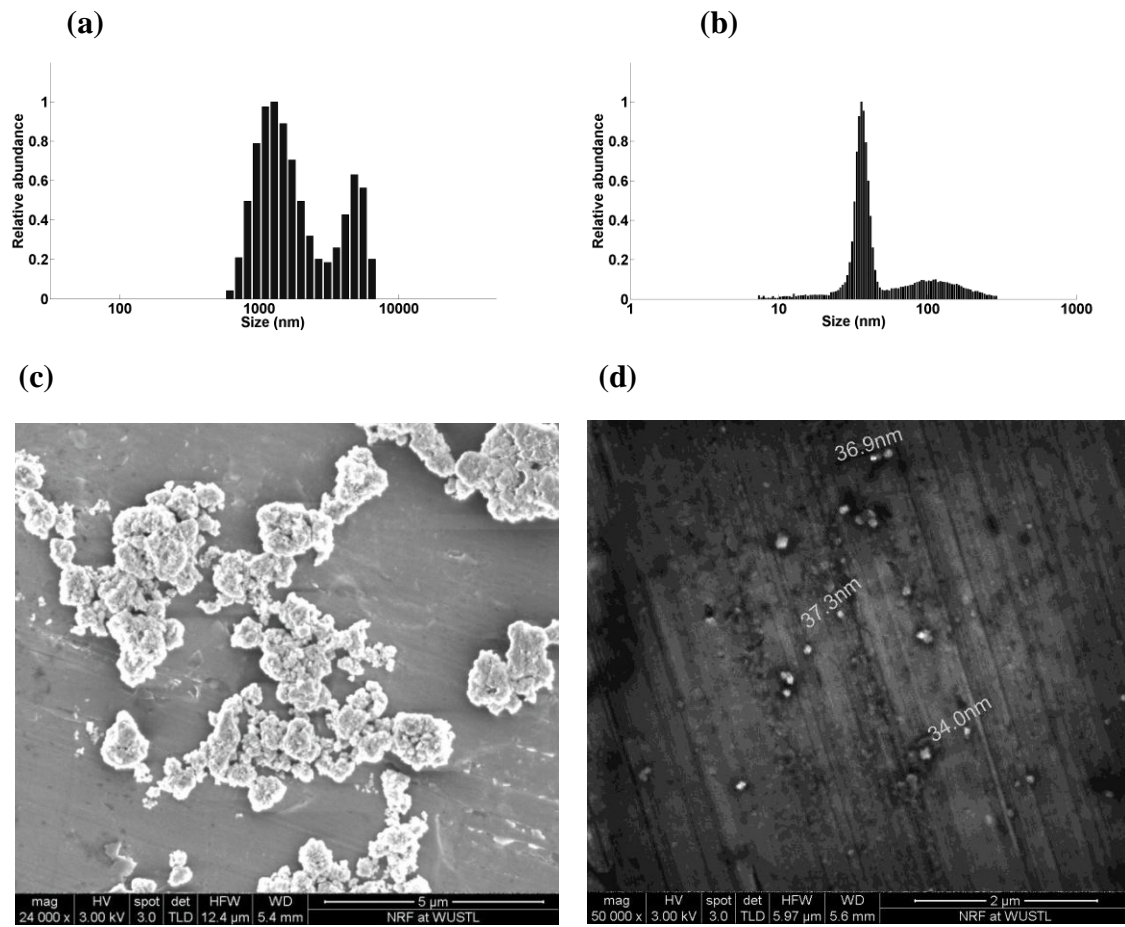
**Figure 7** Germination of aged lettuce seeds (stored for 10 months) treated by TiO<sub>2</sub> NP electro spraying after being placed in the dark for one day, one week and one month prior to the incubation.

**Figure 8** Germination of yarrow seeds after being treated by TiO<sub>2</sub> NP electro spray and incubated for 15 days. Also included in the figure are the germination of untreated seeds and those treated by spraying buffer solutions only (for the reference).

**Figure 1**



**Figure 2**



**Figure 3**

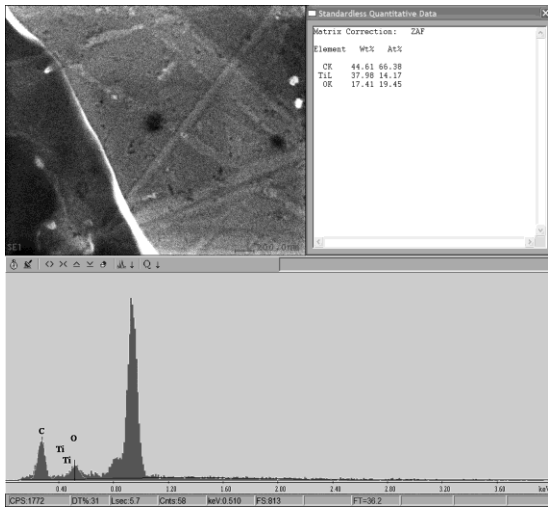
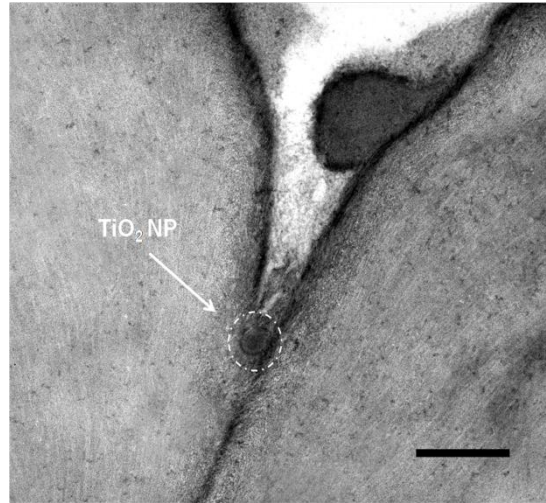
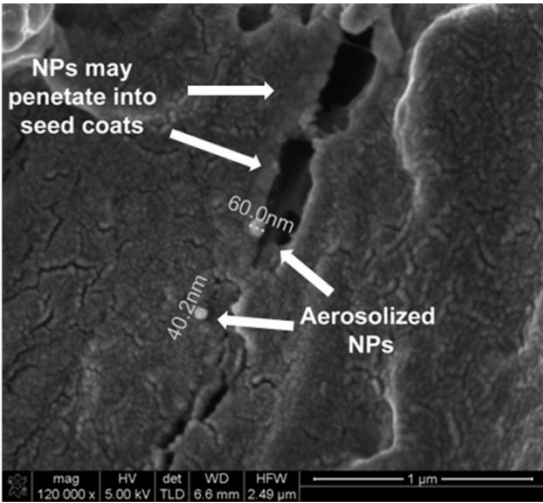
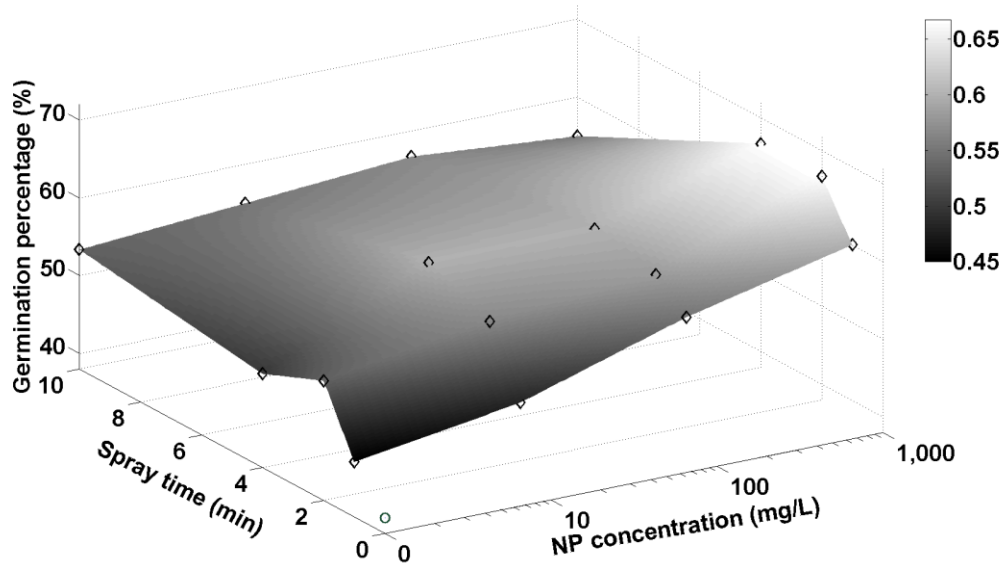
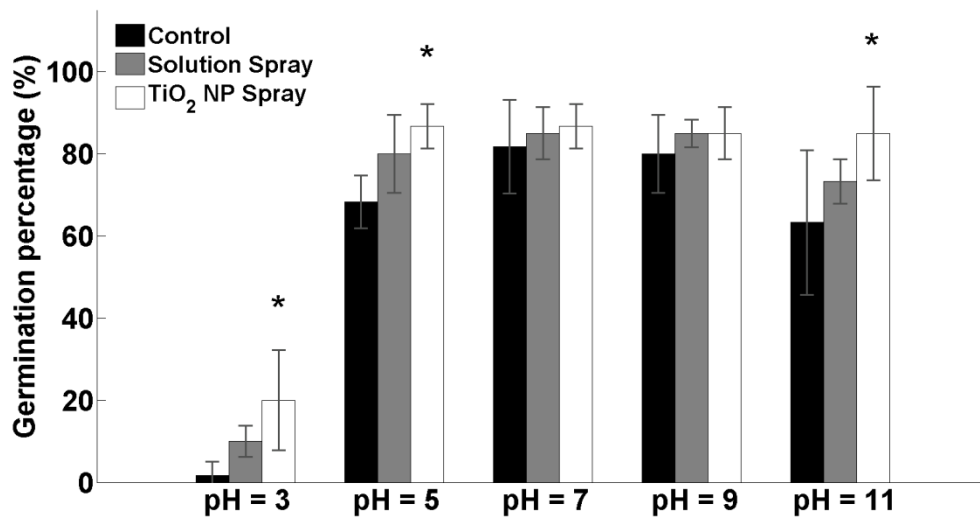


Figure 4

(a)

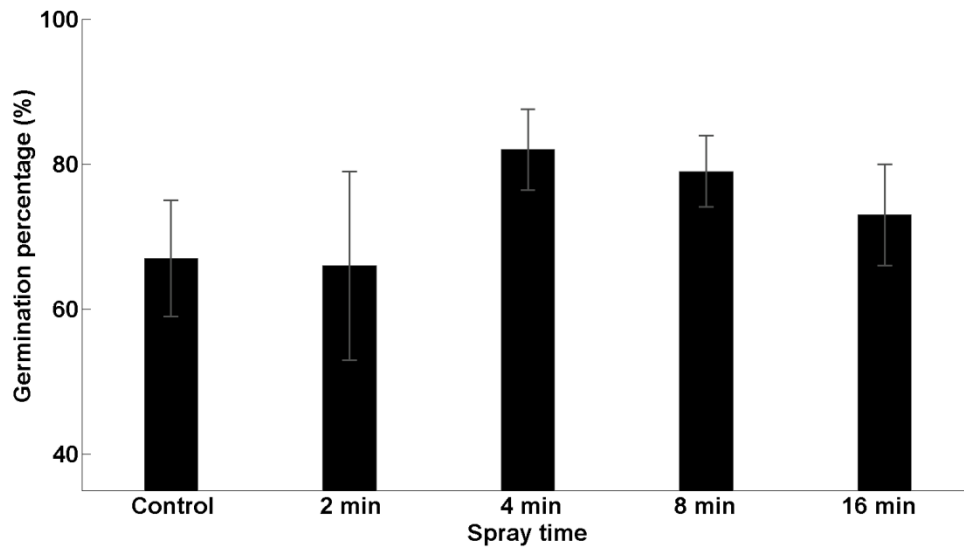


(b)



**Figure 5**

**(a)**



**(b)**

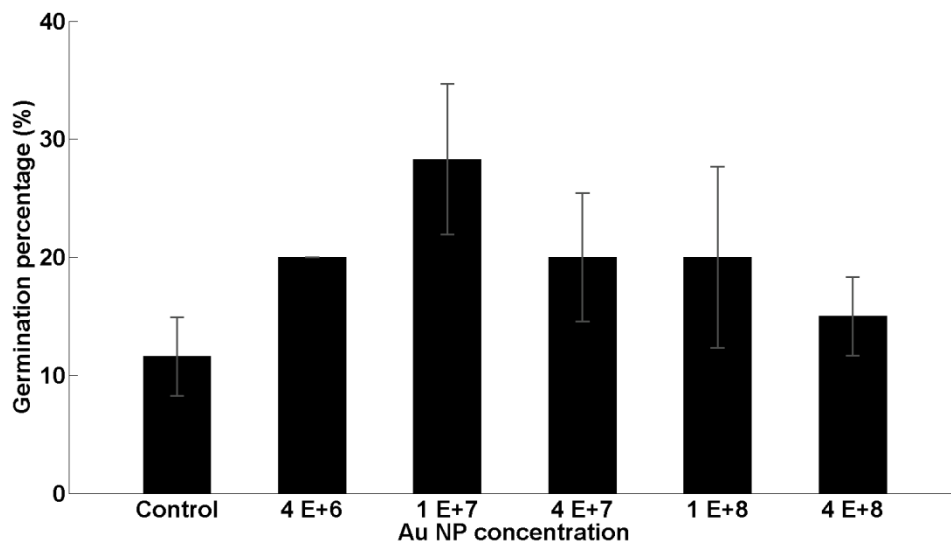
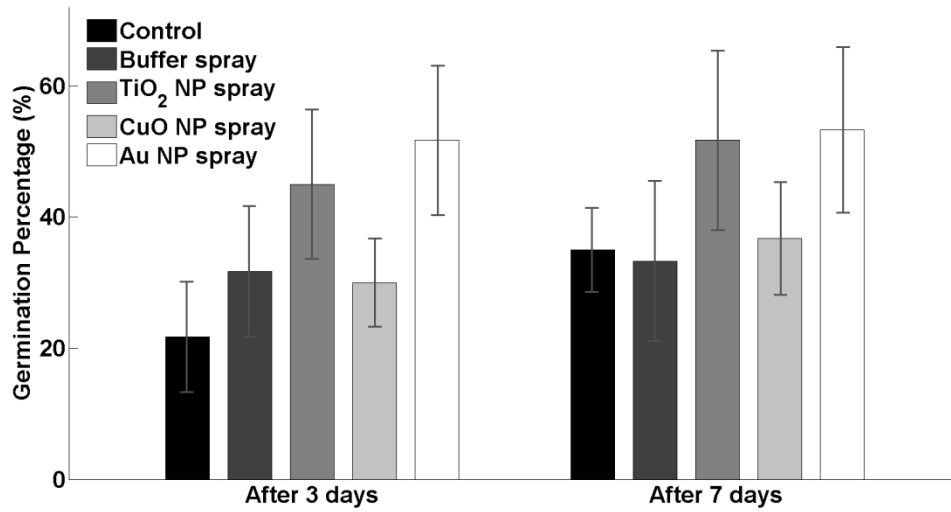
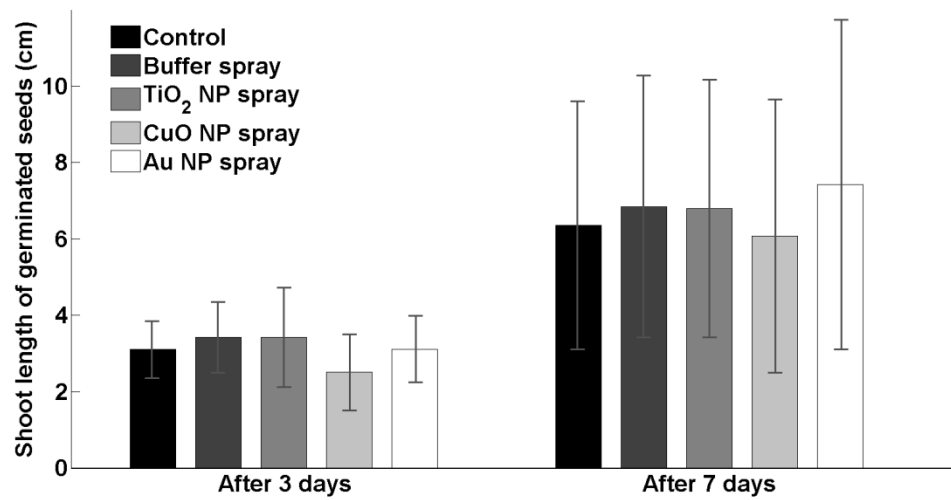


Figure 6



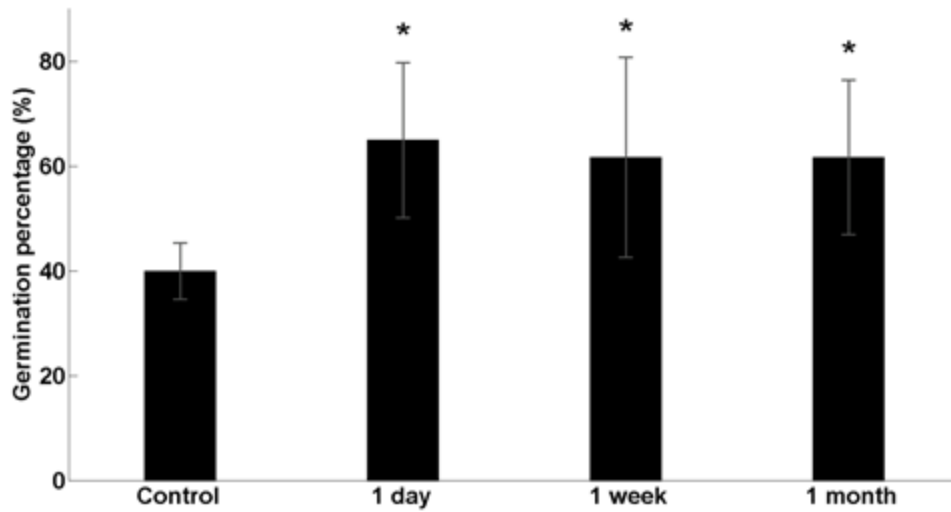
(a)



(b)

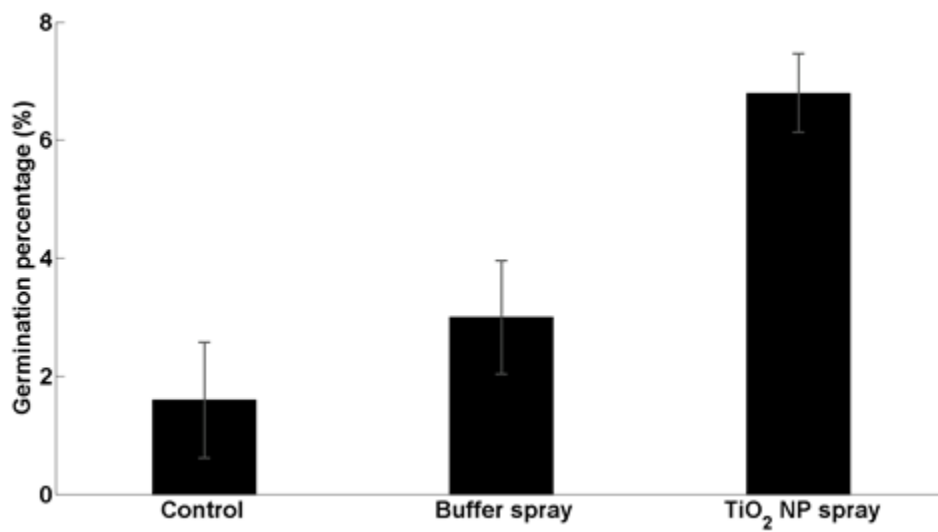


**Figure 7**



Note: \* indicates a significant improvement in seed germination as compared with the control (based on Z-test, SE is based on four replicates).

Figure 8



## **Vita**

Gang Wu

Data of Birth	August 27, 1984
Place of Birth	Taizhou, China
Degrees	B.S. Biotechnology, June 2006
Degrees	M.S. Biological Systems Engineering, December 2010
Degrees	M.S. Energy, Environmental, and Chemical Engineering, May 2015