

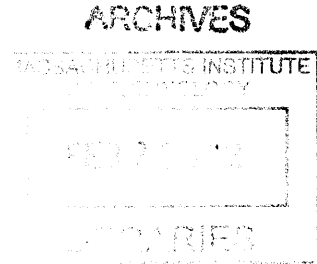
Computational Modeling Techniques for Biological Network Productivity Increases: Optimization and Rate-limiting Reaction Detection

by

Yuanyuan Cui

M.S. Uppsala University (2007)

B.S. Tsinghua University (2005)



Submitted to Computational and Systems Biology Program
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

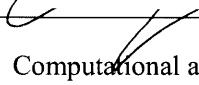
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY


February 2013

©Massachusetts Institute of Technology 2013. All rights reserved.

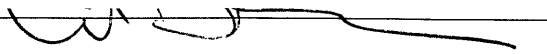
Signature of Author: _____

_____ 
Computational and Systems Biology Program
January 14th, 2013

Certified by: _____

_____ 
Bruce Tidor
Professor of Biological Engineering and Computer Science
Thesis Supervisor

Accepted by: _____

_____ 
Chris B. Burge
Professor of Biology and Biological Engineering
Director, Computational and Systems Biology Graduate Program

Computational Modeling Techniques for Biological Network Productivity Increases: Optimization and Rate-limiting Reaction Detection

by

Yuanyuan Cui

Submitted to Computational and Systems Biology Program
on January 15th, 2013, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The rapid development and applications of high throughput measurement techniques bring the biological sciences into a ‘big data’ era. The vast available data for enzyme and metabolite concentrations, fluxes, and kinetics under normal or perturbed conditions in biological networks provide unprecedented opportunities to understand the cell functions. On the other hand, it brings new challenges of handling, integrating, and interpreting the large amount of data to acquire novel biological knowledge. In this thesis, we address this problem with a new ordinary differential equation (ODE) model based on the mass-action rate law (MRL) of the biochemical reactions. It describes the detailed biochemical mechanisms of the enzyme reactions, and therefore reflects closely of how the enzymes work in the systems. Because the MRL models are constructed with elementary enzyme reaction steps, it is also much more flexible than the aggregated rate law (ARL) model to incorporate new enzyme interactions and regulations. Two versions of the MRL model ensembles for the central carbon metabolic network, which generates most of the precursors for the secondary metabolite, were constructed. The *E. coli* version contains the basic reactions in this network and was applied to optimize the aromatic amino acid production which requires fine-tuned flux partition between glycolysis pathway and the pentose phosphate pathway. The *S. cerevisiae* version is more sophisticated with the incorporated dynamics of the NAD/NADH and NADP/NADPH, as well as the automatic switch from aerobic to anaerobic condition. It was applied to maximize the ethanol production yield, for which the NAD/NADH ratio is a crucial regulating factor. In order to develop methodologies to understand the intrinsic network properties and optimize the network behavior, we further explored approaches for the identification of pathway bottlenecks. Four computational assays were studied, including metabolite accumulation, conditional V_{\max} , increased glucose input, and decreased E_0 , which were applied to the ethanol model ensemble to discover their effectiveness in bottleneck identification in this network. The *TDH* reaction was detected as a major bottleneck restricting

carbon flow towards the ethanol pathway and affecting NADH availability. To manipulate the network for desired production rates of target metabolites, we developed an optimization technique for mass-action rate law ODE models that allows parallel or sequential combinations of enzyme knock-out and over-/under-expression strategies to be conducted on the model. Many strategies were suggested to improve the aromatic amino acid production and help identify the two-direction flux feature of the pentose phosphate pathway. Strategies were also found to enhance the ethanol production yield above 95% of the theoretical yield. Although the two applications studied here are both in the field of metabolic engineering, it is anticipated that the mass-action rate law models for the central carbon metabolism can be extended to study the cancer metabolism. Preliminary studies show promising results for designing cancer clinical trial simulations with a combined model incorporating high level cancer progression and detailed cancer biochemical metabolism.

Thesis Supervisor: Bruce Tidor

Title: Professor of Biological Engineering and Computer Science

*This thesis is dedicated to
Siying Dong, Fengping Li, and Jingjun Cui.
Without your love, support, and sacrifice,
none of this would have been possible.*

Acknowledgements

I would like to first give my sincere appreciation to Bruce Tidor who opened the door for me to the Computational and Systems Biology program at MIT and later on to his lab. It is from his mentoring, I developed the strong standard and skills for scientific research and critical thinking. Equally important, he taught me how to present research precisely and concisely from every correction he made to my presentation during the one-on-one meetings and the group meetings. This will surely benefit my career development and become my life-time treasure.

I would also like to thank all members of the Tidor lab for your generous support and company. In particular, I like to thank David Hagen for his endless support on KroneckerBio and lab cluster, and his very valuable discussions on mathematics and probabilities. I also thank Nirmala Paudel, Tina Toni, Ishan Patel, Andrew Horning, Brian Bonk, Raja Srinivas, Filipe Gracio, Yang Shen, and Nate Silver for your useful comments and inspiring discussions on my projects, as well as your incredible support to me outside the lab. I thank Joshua Apgar who passed me the initial KroneckerBio model of the *E. coli* central carbon metabolic network.

I want to pass my appreciation to my committee members. Jacob White provided critical help on numerical difficulties during the development of the optimization method for KroneckBio models. Collin Stultz and Narendra Maheshri made crucial and constructive suggestions to the projects during the committee meetings.

I'm also very fortunate to have many great collaborators from the industry. They provide insights from a more practical prospective, which makes my projects more realistic and applicable.

Finally, my appreciation goes to my family. Without the sacrifice my husband and my parents took to live long-distance with me for five and half year, none of these works would become possible. My husband, Siying, is the all-time supporter for my scientific research and always carries me through when I encounter hurdles in life. I'm grateful to have his continuous encouragement and love.

Contents

1	Introduction	17
1.1	Computer aided biological network analysis and optimization	18
1.2	Mechanistic modeling of metabolic networks	20
1.3	Techniques for rate-limiting reaction detection and release	24
1.4	Anticipated impact on cancer therapy and clinical trial improvement	26
2	Mass-action model ensemble applied to <i>E. coli</i> aromatic amino acid overproduction	35
2.1	Introduction	37
2.2	Modeling	45
2.2.1	Mass-Action Rate Law Model	45
2.2.2	Network Topology Augmentation	46
2.2.3	Parameter Fitting	48
2.2.4	Model Ensemble	50
2.2.5	Model Manipulation	51
2.3	Results and Discussion	53
2.3.1	Knock-out strategies reveal the complexity of the pentose phosphate	

pathway	53
2.3.2 Up and down regulation allow new engineering strategies to improve aromatic amino acid production	58
2.3.3 Engineering strategies that rebalance carbon fluxes between glycolysis and pentose phosphate pathway lead to improvement of aromatic amino acid production	61
2.4 Conclusions	64
Acknowledgement	66
References	67
Figures	72
Tables	79
3 Systematic bottleneck identification and release for <i>Saccharomyces cerevisiae</i> ethanol overproduction	83
3.1 Introduction	85
3.2 Method	95
3.2.1 Modeling ethanol production in yeast central carbon metabolic network.	95
3.2.2 Calculation of metabolite accumulation	98
3.2.3 Calculation of conditional V_{\max}	99
3.2.4 Measurement of the effect of increasing glucose input	100
3.2.5 Measurement of the effect of decreasing enzyme E_0	100
3.2.6 Sequential bottleneck release	100

3.3 Results and Discussion	102
3.3.1 Ethanol production bottlenecks detection via the four-test framework. .	102
3.3.2 NAD and NADH balance plays a crucial role in regulating ethanol production	106
3.3.3 Sequential bottleneck release increases ethanol production yield	109
3.4 Conclusions	113
References	115
Figures	121
Tables	134
Abbreviations	136
4 Conclusions and future directions	139
Appendix A	146
Appendix B	151

List of Figures

2-1	Structure of the central carbon metabolic network model	72
2-2	Pareto optimal frontier	73
2-3	Single knock-out results for 129 sub-models	74
2-4	Steady-state flux analysis on important knock-outs	75
2-5	Steady-state flux analysis	76
2-6	Flux direction and steady-state concentrations	77
2-7	Balance PEP and E4P steady-state concentrations	78
3-1	NAD and NADH dynamics in central carbon metabolic network	121
3-2	Bottleneck detection framework	122
3-3	Yeast central carbon metabolic network model	123
3-4	Metabolite accumulation test	124
3-5	Conditional V_{\max} test	125
3-6	Flux imbalance around accumulated metabolites	126
3-7	Glucose input test	127
3-8	Decreased E_0 test	128
3-9	NAD and NADH balance at aerobic and anaerobic conditions	129
3-10	Single enzyme sequential optimization	130
3-11	Bottleneck release	131
3-12	Flux change for <i>TDH</i> and <i>EOLE</i> reactions	132
3-13	Sequential bottleneck release	133

List of Tables

2-1	The enzyme reaction modification	79
2-2	The list of best strategies for improving aromatic amino acid productions	80
3-1	Reaction mechanisms used to update NAD/NADH and NADP/NADPH balances	134
3-2	Best enzyme strategies from the sequential single enzyme over- and under-expression .	135

Chapter 1

Introduction

High-throughput technologies in biological and medical studies have improved significantly in quality while reducing their cost in the past decade, which has led to the generation of large data sets describing biological network topology, enzyme kinetic behavior, biochemical reaction dynamics, and species concentrations. Free access to these data sets makes it possible to create quantitative models for large-scale biological networks that describe their behavior at the mechanistic level. High-quality mechanistic models are expected to be capable of providing critical insights into the behavior of biological networks and thus lead to the development of new biological knowledge and guide experiments in network re-engineering. With proper control of uncertainty, computational modeling can save researchers in academia and industry both time and money by providing valuable predictions, prototyping experiments, and directing design. Modeling can also help to discover important but sometimes hidden network properties, such as alternative network states or intrinsic bottlenecks. New techniques are needed to manipulate these detailed and often complex models in order to develop strategies that optimize network behavior to suit biomedical or biotechnological goals. More than just the construction of computational models, the development of strategies for studying, analyzing, and designing with

these models provides tremendous potential for impact and growth of capabilities, particularly to advance metabolic engineering and medical applications.

1.1 Computer-aided biological network analysis and optimization

With biology increasingly becoming a data-rich field, an emerging challenge is how to organize, sort, interrelate, and contextualize all of the high-throughput data sets becoming available. Most of these data sets are large enough to exceed the human ability to directly read out deep biological knowledge. Traditional biological techniques are no longer enough to efficiently interpret these large data sets. This challenge has motivated the field of computational and systems biology, wherein computational and statistical analyses of high-throughput data are used to infer biochemical network structure, function, and response to stimulation and intervention. During the past few decades, most of efforts in computational and systems biology have been focused on solving three major problems: reconstruction of biological networks, simulation of network responses, and optimization of network behaviors. Two classes of biological systems, gene expression regulatory networks and metabolic networks, have received the most attention and are used as case studies to develop novel computational methodologies.

Many modeling techniques have been developed to infer or ‘reverse-engineer’ gene networks (Bansal et al., 2007), which is defined as the process of identifying gene interactions from experimental data through computational analysis, including clustering (Amato et al., 2006; Eisen et al., 1998), Bayesian analysis (Friedman et al., 2000; Yu et al., 2004), information theoretic approaches (Margolin et al., 2006; Steuer et al., 2002), and ordinary differential equation modeling (Bansal et al., 2006; Gardner et al., 2003; di Bernardo et al., 2005). These methodologies generally utilize steady-state gene expression data or short time-series data to

infer gene interactions and have different performance depending on the quality of the data and the network properties (Bansal et al., 2007). With the vast availability of genome sequence data, the genome-scale metabolic reconstructions have also exploded during the past decade (Oberhardt et al., 2009). Since the publication of the first genome-scale metabolic reconstruction of *Haemophilus influenza* (Edwards & Palsson, 1999), the field of genome-scale metabolic network analysis has expanded rapidly, and more than 50 genome-scale metabolic reconstructions have been published.

With the expanded use of computational techniques, many models have been constructed for biological networks, and network topologies as well as molecular mechanisms through which regulation is achieved have been identified and reported in several major databases (e.g., KEGG (Kanehisa & Goto, 2000), BRENDA (Scheer et al., 2010)). However, our understanding of the functioning of the regulatory systems is still insufficient for many applications. Nevertheless, techniques to simulate models and develop novel knowledge are important research frontiers. Simulation techniques have been developed for different formalisms, including directed graphs, Bayesian networks, Boolean networks and their generalizations, ordinary and partial differential equations, qualitative differential equations, stochastic master equations, and rule-based formalisms (de Jong, 2002). The biological results for gene regulation networks obtained through these applications have been the subject of several reviews (Endy & Brent, 2001; Hasty et al., 2001; McAdams & Arkin, 1998; Smolen et al., 2000).

The desire to engineer cellular features to produce desired molecules and other cellular functions has led to the rapid development of metabolic engineering and synthetic biology, both of which aim to manipulate microorganisms or other cells in order to optimize desired cell behavior. Metabolic engineering studies the directed improvement of cellular properties through

modification of specific biochemical reactions or introduction of new ones with the use of recombinant DNA technology. There are numerous applications of metabolic engineering published in the scientific and patent literature, a major effort of which has been on the improved fermentation production of chemicals of commercial and industrial importance, such as amino acids, polymers, lipids, and biofuels (Alper & Stephanopoulos, 2009; Atsumi & Liao, 2008; Bailey, 1991; Barkovich & Liao, 2001; Bongaerts et al., 2001; Cameron & Tong, 1993; Cameron & Chaplen, 1997; Keasling, 1999; Li & Vederas, 2009; Stephanopoulos & Sinskey, 1993; Tyo et al., 2007). Synthetic biology aims to modify cellular behavior to perform new tasks and construct complex networks in single-cell and multicellular systems. Recent achievements include the development of sophisticated non-native behaviors such as bi-stability, oscillations, proteins customized for biosensing, optimized drug synthesis, and programmed spatial pattern formation (Andrianantoandro et al., 2006; Benner & Sismour, 2005; Khalil & Collins, 2010; McDaniel & Weiss, 2005). For both metabolic engineering and synthetic biology, computational modeling and optimization play crucial roles. Research can be done utilizing the strengths of both (Lee et al., 2008).

1.2 Mechanistic modeling of metabolic networks

Metabolic networks, especially the central carbon metabolic network, are popular targets for developing modeling techniques, because they are among the best studied networks with well studied network topology, regulation, and measurement data. *Escherichia coli* has gained the most attention as a model organism given the mature techniques for DNA manipulation of its genome (Oberhardt et al., 2009). *Saccharomyces cerevisiae* is also a frequently used model organism due to its popularity in producing industrial alcohol. Many of the desired metabolic

products are end or intermediate compounds of the central carbon metabolic network, which usually includes the glycolysis pathway and the pentose phosphate pathway. Twelve well-known precursor metabolites serve as branch points from the central carbon metabolic network to generate biomass, and nine of them sit in glycolysis and the pentose phosphate pathway (Neidhardt et al., 1990). Glucose is the main carbon input of the central carbon metabolic network. It is oxidized via either glycolysis to generate ATP and metabolic intermediates, or the pentose phosphate pathway to yield ribose 5-phosphate for nucleic acid synthesis and NADPH for reductive biosynthetic processes. For most eukaryotic cells and many bacteria living under aerobic conditions, pyruvate produced by glycolysis is further oxidized to H₂O and CO₂ via the citric acid cycle (TCA cycle), in the process generating significant energy in the form of ATP (Nelson & Cox, 2008). In order to maximize the production yield of the desired metabolites by designing and predicting productive modifications to this complex network, considerable effort has been put into developing a quantitative understanding and mathematical description of central carbon metabolism. Several mathematical models of various types have been developed and applied over the past few decades, including flux balance analysis (FBA) and ordinary differential equation (ODE) models (Burgard et al., 2003; Chassagnole et al., 2002; Edwards & Palsson, 2000; Pramanik & Keasling, 1998; Schmid et al., 2004; Usuda et al., 2010; Vital-Lopez et al., 2006). The FBA model is one of the most widely used model types in the field of metabolic engineering. It requires relatively modest information regarding biological mechanism, which usually includes a list of chemical reactions with their stoichiometry, flux constraints, and specification of feeds and metabolic demands (Kauffman et al., 2003; Stephanopoulos et al., 1998; Varma & Palsson, 1994). Most of the information can be readily acquired from existing literature and databases (e.g., KEGG (Kanehisa & Goto, 2000) and

BRENDA (Scheer et al., 2010)). A set of linear equations is constructed from the stoichiometry of the reactions so that the fluxes going in and coming out from a node (metabolite) of the network are the same. The fluxes of each branch of the network at steady state can then be learned by solving this set of equations. Therefore, FBA models have advantages when modeling large-scale (e.g., whole genome) networks. However, the steady-state condition is assumed for FBA models, which usually leads to an underdetermined set of equations with a continuous space of acceptable solutions. In practice, maximizing biomass production on the assumption that evolution would favor such solutions or minimizing metabolic adjustment (MOMA), in which it is assumed that metabolic fluxes in a knock-out strain undergo minimal redistribution with respect to the flux configuration of the wild type (Segrè et al., 2002), is used in order to obtain a unique solution. The risk is if the assumptions are not valid, the optimal FBA solution may not correspond to the observed flux distribution in the cell (see review by Edwards et al., 2002). Yet, this condition becomes problematic for mutant strains, in which evolution may not have achieved optimality. Therefore, the predictive ability of FBA can be limited, especially for mutant strains with gene knock-outs.

In contrast to the steady-state nature of FBA models, ODE models, including the aggregated rate law (ARL) and the mass-action rate law (MRL) forms, incorporate network dynamics and are considered to represent the actual enzyme mechanisms of the network (Chassagnole et al., 2002; Lee et al., 2006; Liao et al., 1996; Tzafiriri, 2003). ARL modeling simplifies the description of a single enzymatic step by aggregating the elementary steps associated with a specific mechanism into a single reaction, where the rate becomes a sometimes complex and very non-linear function of the species concentrations involved (Lee et al., 2006; Liao et al., 1996; Tzafiriri, 2003). The rate formulae are usually derived from mass-action laws with certain

assumptions (e.g., quasi-steady state) or acquired as empirical equations from the literature. MRL modeling does not simplify the elementary enzymatic reaction steps and includes all intermediate metabolites as tractable variables, which makes it possible to detect which intermediate step of the enzymatic reaction causes problem when a certain enzyme becomes bottleneck of the network. ODE models generally have more parameters than the corresponding FBA models and require more experimental data to fully determine these parameters. At the same time, the higher dimensional parameter space makes ODE models more flexible to incorporate complicated network topologies and regulations. If unconstrained, the space of steady states reachable by both FBA and ODE models is the same, but ODE models can readily map the parameter constraints into the kinetically feasible regions of the solution space, whereas it is not easily transferable to FBA models (Machado et al., 2012).

In this thesis, I present two versions of mass-action rate law model ensembles for the central carbon metabolic network, one for *E. coli* and one for *S. cerevisiae*. The *E. coli* version contains the basic reactions in this network and is suitable for analyzing basic network behavior and optimizing most amino acid productions. The *S. cerevisiae* version incorporates the dynamics of the NAD/NADH and NADP/NADPH, as well as the switch from aerobic to anaerobic conditions. It is thus a more comprehensive tool and can be used for analyzing effects from a wide variety of experimental conditions, such as different glucose and oxygen conditions. The implemented oxygen consumption mechanism by oxidative phosphorylation enables the model to convert to anaerobic condition when oxygen is depleted, which makes it possible to analyze the steady-state changes for all metabolites in the network after this crucial condition switch. It can also provide deep insight into the regulatory function of the NAD to NADH ratio on the dynamics of this network, which is demonstrated as an important regulating factor for the network state. The *E.*

coli model was applied to optimize the aromatic amino acid production (Chapter 2) and the *S. cerevisiae* model was applied to maximize the ethanol production yield (Chapter 3). They can be easily extended to study production of other chemicals branching from the central carbon metabolic network (e.g., high carbon alcohol).

1.3 Techniques for rate-limiting reaction detection and release

Increasing the productivity of target chemicals is the main goal of metabolic engineering. Although much effort has been made to determine efficient strategies that improve the production rate of target chemicals, there is less research on how to systematically discover the bottlenecks in the system--that is, to identify targets for rational genetic engineering. In order to increase the productivity and yield of metabolite production, researchers have focused almost exclusively on enzyme amplification or other modifications of the product pathway (Stephanopoulos & Vallino, 1991). In those studies, an enzymatic reaction is often labeled a bottleneck if overexpressing that enzyme improves the production of target chemicals (Dai et al., 2002; Lütke-Eversloh & Stephanopoulos, 2008). However, increased production of many metabolites requires significant redirection of flux distributions in primary metabolism, such as glycolysis, the pentose phosphate pathway, and the citric acid (TCA) cycle. It can be especially challenging to identify bottlenecks in primary metabolism, because of the complexity of network topology and regulation. Approaches such as metabolic control analysis (Heinrich & Rapoport, 1974; Kacser & Burns, 1973) have received considerable attention. However, their value in guiding metabolic engineering efforts remains uncertain, given the significant drawback that this method is only valid in the local neighborhood of the operating point evaluated (Stephanopoulos & Vallino, 1991). Dynamic sensitivity analysis is another effort to detect bottlenecks in primary

metabolism, in which relative changes of target metabolite concentration caused by an infinitesimal percentage change in any enzyme activity are calculated and used to predict bottlenecks (Shiraishi & Suzuki, 2009). However, this method provides limited insight into the intrinsic properties of the network that cause those reactions to be bottlenecks. The network rigidity and principal nodes theory has been developed by Stephanopoulos and Vallino (1991), which can identify nodes in the network that have inherent resistance to flux partitioning alterations. This theory is useful to identify branching nodes and enzyme reactions that are crucial and potentially harder to modify in order to increase the yield of target chemicals.

In this thesis, I present a framework to systematically identify bottlenecks in a biological network and to study detect their relevance for target chemical production. Four computational tests, including metabolite accumulation, conditional V_{\max} , glucose input, and decreased E_0 , are developed, which can be easily and efficiently calculated based on mass-action kinetical models or ensembles of models. In particular, the conditional V_{\max} test is able to provide critical insight into the intrinsic cause of network bottlenecks, which is valuable to guide design strategy development for production improvements. A detailed description of the bottleneck detection framework is presented in Chapter 3.

Due to the complex interactions of the central carbon metabolic network, it is usually not obvious how to manipulate the enzymes in the network for the optimized production of the target chemicals, even if we may already know where the bottlenecks are in the network. In Chapter 2, I reported an optimization framework for mass-action kinetic models and their ensembles that can efficiently identify strategies leading to enhanced aromatic amino acid production. This optimization method allows enzyme knock-outs, in which an enzyme activity is completely removed from an organism through genetic disruption, as well as enzyme over- and under-

expression spanning a range from ten times to one-tenth the unperturbed concentration. All combinations of single, double, and triple enzyme knock-outs as well as all combinations of one- and two-enzyme expression changes were constructed and studied for aromatic amino acid overproduction. Efficient strategies with high confidence even in the presence of parameter uncertainties were identified. In Chapter 3, this optimization method was further developed into a sequential single-enzyme over- and under-expression optimization framework. Compared to the exhaustive multiple-enzyme optimization reported in Chapter 2, the single-enzyme optimization is much more computationally efficient, as the optimization number required on the same order of magnitude as the enzyme number in the network; whereas that of the multiple-enzyme optimization could rapidly increase due to the large number of possibilities of enzyme combinations. Therefore, the single-enzyme sequential optimization method is a useful technique to identify efficient enzyme strategies for a large-scale network. The bottleneck identification and release methodologies developed in this thesis are of general value, and are applicable to other metabolic products beyond those studied here.

1.4 Anticipated impact on cancer therapy and clinical trial improvement

In recently years, central carbon metabolism has been increasingly linked to cancer progression and possible therapies. Described decades ago, the Warburg effect of aerobic glycolysis is a key metabolic hallmark of cancer; however, its significance remains unclear (Hsu & Sabatini, 2008; Kroemer & Pouyssegur, 2008). Research has been carried out attempting to decode the causal relation between enhanced glycolysis and cancer development (Hsu & Sabatini,

2008), but cancer mechanism and central carbon metabolism are both sufficiently complex that no simple answer has emerged. The NAD/NADH ratio in cells, however, was reported as being related to tumor development (Koukourakis et al., 2006). With the capability to model the NAD/NADH dynamics in the network, it may now be possible to use the central carbon metabolism model I establish in this thesis to understand cancer mechanisms and predict possible therapeutic approaches. Some preliminary work has been conducted in this direction, which I briefly describe here.

To link detailed mechanistic models of central carbon metabolism with higher-level cancer progression, a cancer progression model can be used with a mutation or growth rate that depends on the mechanistic model's output. A commonly used cancer progression model from Frank et al. (2004), which describe the transition from wild-type cells to transformed cells using four mutation steps with different mutation and reproduction rates, was chosen for this preliminary test. A much simpler mechanistic model that describes selenium metabolism was used to test this concept. Dietary selenium supplementation was reported to possibly reduce the risk for prostate cancer; however, too high selenium levels in cells could potentially increase DNA damage (Gromer & Gross, 2002; Köhrle et al., 2000; Schrauzer, 2000). We therefore built a simplified mass-action model of selenium metabolism in which the selenium input can be metabolized into selenide or methylselenol. The selenide directly can damage DNA and methylselenol can oxidize H_2O_2 in the system to reduce DNA damage. We further define the mutation rate in the cancer progression model to be proportional to the amount of the damaged DNA in the mechanistic one. By simulating this combined model with different selenium input levels, we are able to observe different corresponding cancer progression rates, which can be explained by the relative fluxes of the selenide (DNA damaging) and methylselenol (H_2O_2 neutralizing). The results indicate that

this type of model is capable of predicting how intracellular metabolism can affect cancer progression. Cancer clinical trial simulation can also be established based on this combined model. The kinetic parameters of the reactions in the selenium mechanistic model can be sampled from a Gaussian distribution to simulate enzyme activity variations in populations. The selenium level can also be sampled from a distribution to represent the variation in baseline selenium in different human individuals. Cancer progression can be efficiently simulated for at least tens of thousands of individuals. Statistical tests can then be conducted on the cancer progression simulations to predict the outcome of clinical trials. Preliminary results show that the outcomes of clinical trials are highly dependent on selenium response curves in human individuals and pre-screening of patients may be needed in order to appropriately identify patients who would benefit from dietary selenium supplementation and increase the chance of positive clinical trial outcomes.

The preliminary results we observed suggest this type of combined model and clinical trial simulation potential to be extended to other mechanistic metabolism models and can be used to yield insights into clinical trial design. Further work will be required but there is in principle no barrier to incorporating the entire central carbon metabolic network into cancer progression models. Valuable insights about the relation between glycolysis metabolism and cancer development are anticipated from the analysis of these models.

References

- Alper, H., & Stephanopoulos, G. (2009). Engineering for biofuels: exploiting innate microbial capacity or importing biosynthetic potential? *Nature Reviews. Microbiology*, 7(10), 715–23. doi:10.1038/nrmicro2186
- Amato, R., Ciaramella, a, Deniskina, N., Del Mondo, C., di Bernardo, D., Donalek, C., Longo, G., et al. (2006). A multi-step approach to time series analysis and gene expression clustering. *Bioinformatics (Oxford, England)*, 22(5), 589–96. doi:10.1093/bioinformatics/btk026
- Andrianantoandro, E., Basu, S., Karig, D. K., & Weiss, R. (2006). Synthetic biology: new engineering rules for an emerging discipline. *Molecular Systems Biology*, 2, 2006.0028. doi:10.1038/msb4100073
- Atsumi, S., & Liao, J. C. (2008). Metabolic engineering for advanced biofuels production from *Escherichia coli*. *Current Opinion in Biotechnology*, 19(5), 414–9. doi:10.1016/j.copbio.2008.08.008
- Bailey, J. E. (1991). Toward a science of metabolic engineering. *Science (New York, N.Y.)*, 252(5013), 1668–75.
- Bansal, M., Belcastro, V., Ambesi-Impiombato, A., & di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Molecular Systems Biology*, 3(78), 78. doi:10.1038/msb4100120
- Bansal, M., Della Gatta, G., & di Bernardo, D. (2006). Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics (Oxford, England)*, 22(7), 815–22. doi:10.1093/bioinformatics/btl003
- Barkovich, R., & Liao, J. C. (2001). Metabolic engineering of isoprenoids. *Metabolic Engineering*, 3(1), 27–39. doi:10.1006/mben.2000.0168
- Benner, S. a, & Sismour, a M. (2005). Synthetic biology. *Nature reviews. Genetics*, 6(7), 533–43. doi:10.1038/nrg1637
- Bongaerts, J., Krämer, M., Müller, U., Raeven, L., & Wubbolts, M. (2001). Metabolic engineering for microbial production of aromatic amino acids and derived compounds. *Metabolic Engineering*, 3(4), 289–300. doi:10.1006/mben.2001.0196
- Burgard, A. P., Pharkya, P., & Maranas, C. D. (2003). Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering*, 84(6), 647–57. doi:10.1002/bit.10803
- Cameron, D C, & Tong, I. T. (1993). Cellular and metabolic engineering. An overview. *Applied Biochemistry and Biotechnology*, 38(1-2), 105–40.

- Cameron, D.C., & Chaplen, F. W. R. (1997). Developments in metabolic engineering. *Current Opinion in Biotechnology*, 8(2), 175–180.
- Chassagnole, C., Noisommit-Rizzi, N., Schmid, J. W., Mauch, K., & Reuss, M. (2002). Dynamic modeling of the central carbon metabolism of *Escherichia coli*. *Biotechnology and Bioengineering*, 79(1), 53–73. doi:10.1002/bit.10288
- Dai, S., Vaillancourt, F. H., Maaroufi, H., Drouin, N. M., Neau, D. B., Snieckus, V., Bolin, J. T., et al. (2002). Identification and analysis of a bottleneck in PCB biodegradation. *Nature Structural biology*, 9(12), 934–9. doi:10.1038/nsb866
- Edwards, J S, & Palsson, B. O. (1999). Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *The Journal of Biological Chemistry*, 274(25), 17410–6.
- Edwards, J S, & Palsson, B. O. (2000). Robustness analysis of the *Escherichia coli* metabolic network. *Biotechnology Progress*, 16(6), 927–39. doi:10.1021/bp0000712
- Edwards, Jeremy S., Covert, M., & Palsson, B. (2002). Metabolic modelling of microbes: the flux-balance approach. *Environmental Microbiology*, 4(3), 133–140. doi:10.1046/j.1462-2920.2002.00282.x
- Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25), 14863–8.
- Endy, D., & Brent, R. (2001). Modelling cellular behaviour. *Nature*, 409(6818), 391–5. doi:10.1038/35053181
- Frank, S. a. (2004). Age-specific acceleration of cancer. *Current Biology : CB*, 14(3), 242–6. doi:10.1016/j.cub.2003.12.026
- Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, 7(3-4), 601–20. doi:10.1089/106652700750050961
- Gardner, T. S., di Bernardo, D., Lorenz, D., & Collins, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science (New York, N.Y.)*, 301(5629), 102–5. doi:10.1126/science.1081900
- Gromer, S., & Gross, J. H. (2002). Methylseleninate is a substrate rather than an inhibitor of mammalian thioredoxin reductase. Implications for the antitumor effects of selenium. *The Journal of Biological Chemistry*, 277(12), 9701–6. doi:10.1074/jbc.M109234200
- Hasty, J., McMillen, D., Isaacs, F., & Collins, J. J. (2001). Computational studies of gene regulatory networks: in numero molecular biology. *Nature Reviews. Genetics*, 2(4), 268–79. doi:10.1038/35066056

- Heinrich, R., & Rapoport, T. (1974). A Linear Steady- State Treatment of Enzymatic Chains. *European Journal of Biochemistry*, 42, 89–95. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1111/j.1432-1033.1974.tb03319.x/abstract>
- Hsu, P. P., & Sabatini, D. M. (2008). Cancer cell metabolism: Warburg and beyond. *Cell*, 134(5), 703–7. doi:10.1016/j.cell.2008.08.021
- Kacser, H., & Burns, J. A. (1973). The control of flux. *Symp. Soc. Exp. Biol.*, 27, 65–104.
- Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1), 27–30.
- Kauffman, K. J., Prakash, P., & Edwards, J. S. (2003). Advances in flux balance analysis. *Current Opinion in Biotechnology*, 14(5), 491–496. doi:10.1016/j.copbio.2003.08.001
- Keasling, J. D. (1999). Gene-expression tools for the metabolic engineering of bacteria. *Trends in Biotechnology*, 17(11), 452–460.
- Khalil, A. S., & Collins, J. J. (2010). Synthetic biology: applications come of age. *Nature Reviews. Genetics*, 11(5), 367–79. doi:10.1038/nrg2775
- Koukourakis, M. I., Giatromanolaki, A., Harris, A. L., & Sivridis, E. (2006). Comparison of metabolic pathways between cancer cells and stromal cells in colorectal carcinomas: a metabolic survival role for tumor-associated stroma. *Cancer Research*, 66(2), 632–7. doi:10.1158/0008-5472.CAN-05-3260
- Kroemer, G., & Pouyssegur, J. (2008). Tumor cell metabolism: cancer's Achilles' heel. *Cancer Cell*, 13(6), 472–82. doi:10.1016/j.ccr.2008.05.005
- Köhrle, J., Brigelius-Flohé, R., & Böck, A. (2000). Selenium in biology: facts and medical perspectives. *Biological ...*, 381(October), 849–864.
- Lee, J. M., Gianchandani, E. P., & Papin, J. a. (2006). Flux balance analysis in the era of metabolomics. *Briefings in Bioinformatics*, 7(2), 140–50. doi:10.1093/bib/bbl007
- Lee, S. K., Chou, H., Ham, T. S., Lee, T. S., & Keasling, J. D. (2008). Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. *Current Opinion in Biotechnology*, 19(6), 556–63. doi:10.1016/j.copbio.2008.10.014
- Li, J. W.-H., & Vederas, J. C. (2009). Drug discovery and natural products: end of an era or an endless frontier? *Science (New York, N.Y.)*, 325(5937), 161–5. doi:10.1126/science.1168243
- Liao, J. C., Hou, S. Y., & Chao, Y. P. (1996). Pathway analysis, engineering, and physiological considerations for redirecting central metabolism. *Biotechnology and Bioengineering*, 52(1), 129–40. doi:10.1002/(SICI)1097-0290(19961005)52:1<129::AID-BIT13>3.0.CO;2-J

- Lütke-Eversloh, T., & Stephanopoulos, G. (2008). Combinatorial pathway analysis for improved L-tyrosine production in *Escherichia coli*: identification of enzymatic bottlenecks by systematic gene overexpression. *Metabolic Engineering*, *10*(2), 69–77. doi:10.1016/j.ymben.2007.12.001
- Machado, D., Costa, R. S., Ferreira, E. C., Rocha, I., & Tidor, B. (2012). Exploring the gap between dynamic and constraint-based models of metabolism. *Metabolic Engineering*, *14*(2), 112–9. doi:10.1016/j.ymben.2012.01.003
- Margolin, A. a, Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., & Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, *7 Suppl 1*, S7. doi:10.1186/1471-2105-7-S1-S7
- McAdams, H. H., & Arkin, a. (1998). Simulation of prokaryotic genetic circuits. *Annual Review of Biophysics and Biomolecular Structure*, *27*, 199–224. doi:10.1146/annurev.biophys.27.1.199
- McDaniel, R., & Weiss, R. (2005). Advances in synthetic biology: on the path from prototypes to applications. *Current Opinion in Biotechnology*, *16*(4), 476–83. doi:10.1016/j.copbio.2005.07.002
- Neidhardt, F., Ingraham, J., & Schaechter, M. (1990). *Physiology of the Bacterial Cell: A Molecular Approach*. Sinauer Associates Inc.
- Nelson, D. L., & Cox, M. M. (2008). *Lehninger Principles of Biochemistry* (5th ed., pp. 527–646, 868). W. H. Freeman.
- Oberhardt, M. a, Palsson, B. Ø., & Papin, J. a. (2009). Applications of genome-scale metabolic reconstructions. *Molecular Systems Biology*, *5*(320), 320. doi:10.1038/msb.2009.77
- Pramanik, J., & Keasling, J. D. (1998). Effect of *Escherichia coli* biomass composition on central metabolic fluxes predicted by a stoichiometric model. *Biotechnology and Bioengineering*, *60*(2), 230–8.
- Scheer, M., Grote, a., Chang, a., Schomburg, I., Munaretto, C., Rother, M., Sohngen, C., et al. (2010). BRENDA, the enzyme information system in 2011. *Nucleic Acids Research*, *39*(Database), D670–D676. doi:10.1093/nar/gkq1089
- Schmid, J. W., Mauch, K., Reuss, M., Gilles, E. D., & Kremling, A. (2004). Metabolic design based on a coupled gene expression-metabolic network model of tryptophan production in *Escherichia coli*. *Metabolic Engineering*, *6*(4), 364–77. doi:10.1016/j.ymben.2004.06.003
- Schrauzer, G. (2000). Selenomethionine: a review of its nutritional significance, metabolism and toxicity. *The Journal of Nutrition*, *130*(7), 1653–6.

- Segrè, D., Vitkup, D., & Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(23), 15112–7. doi:10.1073/pnas.232349399
- Shiraishi, F., & Suzuki, Y. (2009). KINETICS , CATALYSIS , AND REACTION ENGINEERING Method for Determination of the Main Bottleneck Enzyme in a Metabolic, 415–423.
- Smolen, P., Baxter, D. a., & Byrne, J. H. (2000). Modeling transcriptional control in gene networks--methods, recent results, and future directions. *Bulletin of Mathematical Biology*, 62(2), 247–92. doi:10.1006/bulm.1999.0155
- Stephanopoulos, G., & Sinskey, a J. (1993). Metabolic engineering--methodologies and future prospects. *Trends in Biotechnology*, 11(9), 392–6.
- Stephanopoulos, G., & Vallino, J. J. (1991). Network rigidity and metabolic engineering in metabolite overproduction. *Science (New York, N.Y.)*, 252(5013), 1675–81.
- Stephanopoulos, Gregory, Aristidou, A., & Nielsen, J. (1998). *Metabolic Engineering: Principles and Methodologies* (1st ed.). Academic Press.
- Steuer, R., Kurths, J., Daub, C. O., Weise, J., & Selbig, J. (2002). The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics (Oxford, England)*, 18 Suppl 2, S231–40.
- Tyo, K. E., Alper, H. S., & Stephanopoulos, G. N. (2007). Expanding the metabolic engineering toolbox: more options to engineer cells. *Trends in Biotechnology*, 25(3), 132–7. doi:10.1016/j.tibtech.2007.01.003
- Tzafriri, a R. (2003). Michaelis-Menten kinetics at high enzyme concentrations. *Bulletin of Mathematical Biology*, 65(6), 1111–29. doi:10.1016/S0092-8240(03)00059-4
- Usuda, Y., Nishio, Y., Iwatani, S., Van Dien, S. J., Imaizumi, A., Shimbo, K., Kageyama, N., et al. (2010). Dynamic modeling of Escherichia coli metabolic and regulatory systems for amino-acid production. *Journal of Biotechnology*, 147(1), 17–30. doi:10.1016/j.jbiotec.2010.02.018
- Varma, A., & Palsson, B. O. (1994). Metabolic flux balancing: basic concepts, scientific and practical use. *Nature Biotechnology*, 12(10), 994–998.
- Vital-Lopez, F. G., Armaou, A., Nikolaev, E. V., & Maranas, C. D. (2006). A computational procedure for optimal engineering interventions using kinetic models of metabolism. *Biotechnology Progress*, 22(6), 1507–17. doi:10.1021/bp060156o
- Yu, J., Smith, V. A., Wang, P. P., Hartemink, A. J., & Jarvis, E. D. (2004). Advances to Bayesian network inference for generating causal networks from observational biological

data. *Bioinformatics (Oxford, England)*, 20(18), 3594–603.
doi:10.1093/bioinformatics/bth448

de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, 9(1), 67–103. doi:10.1089/10665270252833208

di Bernardo, D., Thompson, M. J., Gardner, T. S., Chobot, S. E., Eastwood, E. L., Wojtovich, A. P., Elliott, S. J., et al. (2005). Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nature Biotechnology*, 23(3), 377–83.
doi:10.1038/nbt1075

Chapter 2

A Mass-Action Rate Law Ensemble Model of *E. coli* Central Carbon Metabolism Applied to Amino Acid Production

Abstract

Two types of computational models that dominate the field of metabolic engineering are flux balance and aggregated rate law models, both of which have strengths and shortcomings. In this report, we present a model from a third class, a mass-action rate law (MRL) model, for the *E. coli* central carbon metabolic network. This mechanistic model does not require assuming optimal behavior of the metabolic network as for flux balance modeling and also involves fewer assumptions than aggregated rate law modeling. To estimate the uncertainty of model predictions due to parameter uncertainty, an ensemble of sub-models was built using latin hypercube sampling. The ensemble model was used to identify enzyme expression change strategies for overproducing aromatic amino acids. The predicted strategies revealed implications of complexity in the pentose phosphate pathway (PPP) and suggested that fine-tuning both the direction and volume of the PPP flux can play an important role in improving aromatic amino acid production. The ensemble model presented here for central carbon metabolism provides

new opportunities for applying metabolic engineering to the production of commercially and industrially important chemicals.

2.1 Introduction

Metabolic engineering studies the directed improvement of cellular properties through modification of specific biochemical reactions or introduction of new ones with the use of recombinant DNA technology. There are numerous applications of metabolic engineering published in the scientific and patent literature, a major effort of which has been on the improved fermentation production of chemicals of commercial and industrial importance, e.g. amino acids, polymers, lipids, and biofuels (Alper and Stephanopoulos, 2009; Atsumi and Liao, 2008; Bailey, 1991; Barkovich and Liao, 2001; Bongaerts et al., 2001; Cameron and Chaplen, 1997; Cameron and Tong, 1993; Keasling, 1999; Li and Vederas, 2009; Tyo et al., 2007; Stephanopoulos and Sinskey, 1993). One of the primary host organisms for this purpose has been *Escherichia coli*, because of its wide range of growth substrates and the powerful molecular biological tools available for its manipulation (Cameron and Tong, 1993; Feist et al., 2010; Leuchtenberger et al., 2005). Because many of the desired metabolic products are end or intermediate compounds of the central carbon metabolic network, it has become one of the most studied and well understood of the biochemical pathways. Glucose, the main network input, is not only an excellent fuel but also a remarkably versatile precursor, capable of supplying a vast array of metabolic intermediates for biosynthetic reactions. Glucose is oxidized via two major paths — (i) through glycolysis to generate ATP and metabolic intermediates, or (ii) through the pentose phosphate pathway to yield ribose 5-phosphate for nucleic acid synthesis and NADPH for reductive biosynthetic processes. For most eukaryotic cells and many bacteria living under aerobic conditions, pyruvate produced by glycolysis is further oxidized to H₂O and CO₂ via the tricarboxylic acid (TCA) cycle, in the process generating significant energy in the form of ATP (Nelson and Cox, 2008). Twelve well-known precursor metabolites serve as branch points from

the central carbon metabolic network to generate biomass, and nine of them sit in the glycolysis and pentose phosphate pathway (Neidhardt et al., 1990). In order to design and predict productive modifications to this complex network, considerable effort has been put into developing a quantitative understanding and mathematical description of central carbon metabolism. Several mathematical models of various types have been developed and applied over the past few decades, including flux balance analysis (FBA) and ordinary differential equation (ODE) models (Burgard et al., 2003; Chassagnole et al., 2002; Edwards and Palsson, 2000; Pramanik and Keasling, 1998; Schmid et al., 2004; Usuda et al., 2010; Vital-Lopez et al., 2006).

The flux balance form is one of the most widely used model types in the field of metabolic engineering. It is based on the assumption that metabolic transients are more rapid than both cellular growth rates and dynamic changes in the organism's environment (Stephanopoulos et al., 1998; Varma and Palsson, 1994). In this view, the metabolic fluxes are considered in quasi-steady state relative to growth. A useful feature of metabolic flux models is the relatively modest requirements necessary in terms of specific information regarding biological mechanism. Only three types of information are required: a list of chemical reactions with their stoichiometry; flux constraints, such as a maximum rate for each reaction (V_{\max}); and specification of feeds and metabolic demands (Kauffman et al., 2003; Stephanopoulos et al., 1998; Varma and Palsson, 1994). For the *E. coli* metabolic network, this information can be readily acquired from existing literature and databases, which makes it relatively straightforward to build, although the V_{\max} information is less certain. However, the steady-state condition usually leads to an underdetermined set of equations indicating a continuous space of acceptable solutions. In general, a unique solution is obtained by optimizing a metabolic objective function, such as

maximizing biomass production on the assumption that evolution would favor such solutions. Other approaches, like minimization of metabolic adjustment (MOMA), have also been suggested, where it is assumed that metabolic fluxes in a knock-out strain undergo minimal redistribution with respect to the flux configuration of the wild type (Segrè et al., 2002). Thus, FBA can be defined as a linear programming problem with a set of constraints. The constraints are typically upper and lower flux bounds that define the space of allowable distributions, and can include bounds obtained from different carbon sources and limited oxygen supply to simulate environmental conditions (Orth et al., 2010). This construction allows FBA calculations to proceed very quickly even for large networks and makes FBA models attractive for genome-wide modeling. FBA has been widely used to model large-scale *E. coli* metabolic networks, where several successes have been reported in predicting experimental growth rates under different environmental and growth conditions (Edwards and Palsson, 2000; Edwards et al., 2002; Segrè et al., 2002). However, failure to incorporate gene regulatory events and to account for toxic intermediate build-up has led to discrepancies between model results and experimental observations. In addition, FBA assumes consistency between the mathematical objective function and the evolutionary objective. If untrue, the optimal FBA solution may not correspond to the observed flux distribution in the cell (see review: Edwards et al., 2002). Yet, this condition becomes problematic for mutant strains, in which evolution may not have achieved optimality. Therefore, the predictive ability of FBA can be limited, especially for mutant strains with gene knock-outs. Moreover, the detailed dynamic behavior of the network and the metabolite concentrations are also not captured by FBA models, although they may not be necessary for successful metabolic engineering.

In contrast to the steady-state nature of FBA models, in recent years ordinary differential equation (ODE) models have started to attract attention as dynamic models of metabolic networks. Two variants of ODE models have been reported that differ in the types of rate laws used — the aggregated rate law (ARL) model and the mass-action rate law (MRL) model. ARL modeling simplifies the description of a single enzymatic step by aggregating the elementary steps associated with a specific mechanism into a single reaction, where the rate becomes a sometimes complex and very non-linear function of the species concentrations involved (Lee et al., 2006; Liao et al., 1996; Tzafiriri 2003). The rate formulae are usually derived from mass-action laws with certain assumptions (e.g., quasi-steady state) or acquired as empirical equations from the literature. A particularly useful ARL model for the *E. coli* central carbon network was presented by Chassagnole et al. (2002). The model covers both glycolysis and the pentose-phosphate pathway, and it is composed of 30 enzymes and 17 metabolites (Figure 1). The aggregated reactions and their rate parameters were acquired from published literature and databases. The model was validated with measured metabolite concentrations under transient conditions and captures experimentally observed dynamics of metabolite concentrations.

Here we present a new mathematical model for the *E. coli* central carbon metabolic network. The model is based on the Chassagnole et al. (2002) ARL model, but we recast it as an MRL model and re-fit the parameters to additional data. In contrast to ARL, MRL models represent enzyme reactions with a series of elementary reactions and express the reaction rate with rate laws consisting of only second-, first-, and zeroth-order reactions. The MRL model can be mechanistically more accurate, does not require a quasi-steady state assumption, and could be valid over a wider range of concentrations and conditions. The general formulae of the MRL model can be expressed in Kronecker form as

$$\frac{d\vec{x}}{dt} = A^{(1)}\vec{x} + A^{(2)}\vec{x}\otimes\vec{x} + B^{(1)}\vec{u} + B^{(2)}\vec{u}\otimes\vec{x} + \vec{k}$$

where \vec{x} is a vector of dynamic species concentrations in the network as a function of time t ; \vec{u} is a vector of network input concentrations that are externally controlled and not evolved by the model; $\vec{x}\otimes\vec{x}$ represents a vector Kronecker product, which is a column vector containing all possible pair-wise combinations of the species concentrations; $\vec{u}\otimes\vec{x}$ is a column vector containing all pair-wise combinations of species and network input concentrations; $A^{(1)}$, $A^{(2)}$, $B^{(1)}$, and $B^{(2)}$ are the corresponding coefficient matrices, the entries of which are the first- and second-order rate constants, separately; and \vec{k} contains zeroth-order rate constants. As all the reactions in the MRL model are elementary reactions, the species vector contains not only the metabolites and enzymes, but also the different intermediate complexes between metabolites and enzymes. Each species participates in only a small set of elementary reactions; thus, the coefficient matrices are sparse. This ODE-type model can be integrated numerically by a variety of means to acquire the concentrations of all species as a function of time; long-time solutions approach the steady state of the system. In addition to its mechanistic realism, which could make it applicable across broad sets of conditions, the simple and standard Kronecker mathematical representation provides an opportunity to develop standard and general software to study this type of model.

Both the ARL and MRL forms of ODE models can simulate network dynamics and provide both transient and steady-state information on metabolite concentrations, in contrast to FBA models. ODE models tend to have one or a very small number of steady states, eliminating the need for biomass maximization or MOMA assumptions as with FBA to reach a unique steady-

state solution for a given glucose input. It was reported that if unconstrained, the space of steady states by both FBA and ODE models is the same, but constraints of parameter range can be readily mapped into kinetically feasible regions of the solution space of ODE models that is not easily transferable to FBA models (Machado, et al. 2012). However, ODE models require significant knowledge in the form of kinetic parameters, enzyme levels, and mechanistic formulae for enzyme reactions. MRL formulations avoid the quasi-steady state assumption of ARL ones. Moreover, ARL models require knowing the lumped enzyme reaction rate formulae, which in theory can be obtained based on physical mechanism but are often acquired as empirical equations. This can complicate incorporation of new enzyme reactions or regulation, such as inhibitors or activators, which may require new lumped formulae to reactions affected. MRL formulations, however, because of their mechanistic nature, are generally straightforward to augment with inhibition and regulation when known.

We built out initial MRL model for *E. coli* central carbon metabolism from the Chassagnole et al. (2002) model by expanding each aggregated enzyme reaction to expose the intermediate elementary steps with their associated rate constants. The overall MRL model was then fit to the Chassagnole et al. model so that the two models produced essentially identical rates for each enzyme as a function of substrates. This initial model was successful at matching trajectories from the Chassagnole et al. model. The two models showed similar behavior for the wild-type *E. coli* network; interestingly, though, they started to diverge in behavior for enzyme knock-outs. Next, this initial MRL model was improved by refitting to a combination of synthetic data produced from the ARL model and newer experimental data from knock-out strains (Ishii et al., 2007). In this refitting, rather than produce a single model, we produced a collection of models that are similarly good fits to the data. This ensemble provides an estimate of uncertainty in the

parameters fit and in the predictions made. MRL models tend to have a greater number of adjustable parameters than ARL models, which provides the potential for underdetermination and overfitting. Here that possibility was minimized by using the ARL model to produce a substantial number of data points such that the MRL parameters were fit to an overabundance of data, but non-identifiability is not ruled out.

One of the most appealing and challenging goals of metabolic engineering is to design more efficient biological systems for industrial use. The mathematical model ensemble built here provides a foundation for such rational design studies. In order to propose metabolic engineering changes to the *E. coli* central metabolic network to produce valuable metabolites, we introduced an optimization framework, optModulation, for the MRL model. The approach identifies and qualifies metabolic improvements. The proposed optModulation scheme was tested by determining optimal genetic manipulation strategies to maximize a pre-defined reaction flux, namely, aromatic amino acid production.

Over the past 5 years the global market for fermentation amino acid products has increased more than 40%, but the efficiency of aromatic amino acid production, particularly tryptophan, by fermentation remains low (Ikeda, 2003; Leuchtenberger et al., 2005). One difficulty in tryptophan overproduction is to properly balance the supply of its three precursors. In *E. coli*, the production of 1 mole of any aromatic amino acid starts with combining 1 mole of erythrose 4-phosphate (E4P) with 1 mole of phosphoenolpyruvate (PEP) to form a common precursor chorismate. In addition, 1 mole of PEP and 1 mole of serine are further consumed in the pathway from chorismate to tryptophan (Ikeda, 2006; Nelson and Cox, 2008). Different strategies have been attempted to modify either the common pathways or the tryptophan branch by metabolic engineers (see review: Ikeda, 2006). Using the model ensemble, we identified complexities in the

pentose phosphate pathway such that the direction of flux through this bi-direction loop strongly affects the choice of manipulation strategies for overproducing aromatic amino acids. The results of this study also predict that balancing the precursors for tryptophan could be beneficial for its overproduction.

2.2 Modeling

2.2.1 Mass-Action Rate Law Model

The ARL model built by Chassagnole et al. (2002) provides a good reference model from which to generate a more comprehensive MRL model for the central carbon metabolic network (Figure 1). Foundational work by King and Altman (1956) and Cleland (1963) provides a bridge between the elementary rate constants of MRL models and the more abstract parameters of ARL models. The authors of those studies describe a graphical method to derive the steady-state rate law from a system of elementary reactions. This method has since been developed into a formal algorithm (Cornish-Bowden, 1977) and is available as a web tool (Kuzmic, 2008). To convert the Chassagnole et al. ARL model into an MRL model, we constructed elementary reactions for each enzyme in the ARL model based on the enzyme mechanisms reported in Chassagnole et al. Next, a steady-state rate law was derived using the King–Altman method. Then, we optimized a preliminary set of MRL parameters by fitting the rate versus substrate concentration curves calculated from the King–Altman steady-state rate law for MRL model to those simulated from the ARL model. This process was carried out for each enzyme in the ARL model. The objective function for the fitting, $G_{chassagnole}$, was similar to a chi-square metric, being a sum of squared differences weighted by the inverse variance.

$$G_{chassagnole} = \sum_{r_{data,chassagnole}} \left(\frac{r_{pred,chassagnole} - r_{data,chassagnole}}{\sigma_{data,chassagnole}} \right)^2$$

In this objective function, $r_{data,chassagnole}$ is the steady-state reaction rate at given species concentrations based on the ARL model, and $r_{pred,chassagnole}$ is the steady-state reaction rate at the

same species concentrations based on the King–Altman method calculated from the MRL model. The range for the species concentrations used in the fitting varied for different metabolites and different enzyme reactions, but generally spanned from 0.01 mM to 100 mM. Typically 1,000–27,000 points were used to fit each enzyme; the points were equally spaced in the logarithm of the concentration. Grids of species concentrations were created and used for reactions involving more than one species in the rate law. The optimization was done using the `fmincon` function in MATLAB (version 2008b; The MathWorks, Inc.; Natick, MA).

2.2.2 Network Topology Augmentation

The initial MRL model converted from the Chassagnole et al. model was updated to include several important enzymes and metabolites in the glycolysis and pentose phosphate pathways. To make the network model more comprehensive, the KEGG database (8/6/2009; Kanehisa and Goto, 2000) was used to select model additions. The complete list of changes is given in Table 1, and some are described in the following paragraph.

A number of enzymatic conversions in *E. coli* are carried out by multiple isomers with the same general activity but with potentially different rate parameters and cellular concentrations. The original Chassagnole et al. model used a single enzyme to represent such instances. Because of our interest in designing meaningful genetic variants for metabolic engineering, we augmented the topology to include ten separate pairs of isomers where previously ten individual enzymes represented the biochemistry. This expansion led to locally parallel routes, instead of a single path, which makes it possible to design knock-out mutant strains that only partially shut down a pathway. In addition, three new enzyme reactions (due to the enzyme products of the genes *pgl*, *edd*, and *eda*) and 2 new metabolites (gluconolactone-6P and KDPG [2-keto-3-deoxy-

6-phosphogluconate]) were also added to the network. The *edd* and *eda* reactions represent the Entner-Doudoroff pathway, which is an alternate route that catabolizes glucose to pyruvate. It has been shown that the accumulation of KDPG in bacteria is correlated with an immediate and significant decrease in growth. In fact, the gene product of *eda* has been considered a target for the development of new bacteriostatic or bactericidal drugs (Braga et al., 2004). In the Chassagnole et al. model, two constant fluxes are used to model the production of G3P and PYR from tryptophan synthesis. The cellular concentration of tryptophan and its precursors could be dramatically affected by metabolic engineering manipulations, which would not be properly reflected by this constant flux treatment. We thus elected to make these fluxes dependent on the available concentrations of precursors. Specifically, we assumed that all fluxes that produce the aromatic amino acid precursor chorismate are converted to tryptophan, and ignored the production of phenylalanine and tyrosine. Two enzyme reactions based on TrpSynth1 and TrpSynth2 were added to model this process of generating tryptophan and thus directly linked to G3P and PYR dynamically.

The enzyme reaction mechanisms were preserved for enzymes in the Chassagnole et al. model; whereas those of added enzymes were determined based on the BRENDA database (8/6/2009) (Scheer et al., 2010). Most of the isomers share a similar mechanism with their parallel enzymes, but some isomers (e.g., PFKB) do not show the same cooperative interaction with their ligand or are not affected by inhibitors. Additionally, some isomers do not equally share activity; instead, one of the parallel isomers accounts for most of the enzyme activity, either due to rate constants or concentration.

The main input of the model is the supply of glucose. Chassagnole et al. modeled supply as a constant extracellular concentration in the starting model. Because most experimental data about

metabolic reactions are based on chemostats in which glucose flows in and is removed from the reaction vessel as a flux with a specific flow rate, we changed the constant glucose feed into dynamical flux reactions in the model. After all of these changes were made, the new model consists of 43 enzyme reactions, 211 species, and 263 free kinetic parameters. Figure 1 depicts the overall topology of the model. The model is available as supplementary material.

2.2.3 Parameter Fitting

We adopted a dual strategy for parameter estimation. One aim was to select parameters for our MRL model so that the dose-response curve of the enzyme reaction rates as a function of metabolite concentrations matched those from the original model of Chassagnole et al. (2002). Simultaneously our second aim was to select parameters that produced results matching a more recent experimental data set published by Ishii et al. (2007), containing normalized steady-state concentrations for 12 of the 18 metabolites in the model for wild-type *E. coli* K-12 strain BW25113 as well as for 22 variant strains each with one enzyme knocked out. The actual steady-state concentrations were recreated from these normalized data by re-scaling them so that the wild-type metabolite concentrations match those calculated from the Chassagnole et al. model with the dilution rate and glucose feed the same as in the Ishii et al. experiments.

Candidate parameter sets that fit both the Ishii et al. measurement data and the Chassagnole et al. enzyme reaction rates were generated by fitting to a weighted component objective $G(w)$ containing two terms, one representing deviation from the Chassagnole et al. rates and the other deviation from the Ishii et al. data.

$$G(w) = w \cdot \sum_{r_{data,chassagnole}} \left(\frac{r_{pred,chassagnole} - r_{data,chassagnole}}{\sigma_{data,chassagnole}} \right)^2 + \sum_{x_{data,ishii}} \left(\frac{x_{pred,ishii} - x_{data,ishii}}{\sigma_{data,ishii}} \right)^2$$

The first summation was the same as described for Chassagnole et al. model fitting. The second summation represented sum-of-squares fitting of the predicted chemical concentrations from the model to the measured Ishii et al. data, normalized by the variance of the measurement data. To generate initial parameters for this dual fitting, 1000 parameter sets were generated based on the optimized parameters from only fitting the Chassagnole et al. model. For each parameter set, a random number of parameters up to 50% of all parameters were selected and replaced by a random number based on a Gaussian distribution with mean as the value from Chassagnole et al. fitting and the standard variation as one third of the log of the largest parameter value allowed (i.e., 10^{12}). The weighting factor w played an important role in the optimization. Small values caused the optimization to focus only on optimizing the Ishii sub-objective with little control on the Chassagnole sub-objective; whereas for large values the fitting of the Chassagnole sub-objective would dominate the Ishii sub-objective. Rather than attempting to determine an ideal balance between the fitting of Chassagnole and Ishii sub-objectives, we carried out a series of optimizations with a range of values for w chosen that spanned 0.05 to 20.

This collection of optimization with various weights swept out a pareto optimal frontier (Figure 2A). Each point represented the result of a single optimization for a particular value of w . Moving along the frontier improved the fit to one part of the objective and degraded the fit to the

other. Thus, the frontier represented different tradeoffs achievable. Naturally, a higher weight on the Chassagnole part of the objective led to a better fit to that sub-objective and a worse fit to the Ishii sub-objective. Examination of the fitting behavior for each of the 44 points on the frontier demonstrated rather good fits to both sub-objectives. It is a somewhat arbitrary choice to select the best fitted model from the pareto optimal frontier. Here we selected the blue point in Figure 2A as our fitted model and the basis for local sampling in parameter space; the red points were also used but no further sampling around them was performed.

2.2.4 Model Ensemble

A potential problem for detailed mechanistic models is the combination of a high dimensional parameter space and limited data available for training; this is especially an issue for mass-action models due to their large number of parameters. Although there may be a set of optimal parameters, wide ranges of parameters surrounding the optimum may fit the available data almost as well. Such is the case here for our model fit to the Chassagnole et al. model and Ishii et al. data. Moreover, due to measurement error, a parameter set with a somewhat disadvantaged objective value may be closer to the true parameter values. To improve the reliability of model predictions, we generated a model ensemble that represents the parameter uncertainty, and used the ensemble to make predictions.

Latin hypercube sampling (LHS) (McKay et al., 1979; Stein, 1987) was used to collect candidate parameter sets in order to distribute a reasonable number of samples over the parameter space. In LHS all the parameters were assumed to follow a multivariate Gaussian distribution, with mean values as the best fitted parameter values from the previous session and a covariance matrix as the inverse of a modified Fisher information matrix calculated from the best

fitted values. The Fisher information matrix was modified by replacing the eigenvalues smaller than an arbitrary cutoff of 58.1 with this cutoff, which corresponds to approximately a 30% change in the parameter values along the eigendirections. This choice retained 61 original eigenvalues out of the 263 and removed the remaining flat eigendirections that could have caused oversampling of invalid regions. 20,000 parameter set samples were drawn using LHS and 85 of them fell within the tolerance level as fitting well to both the Chassagnole et al. model and the Ishii et al. data. In order to reduce the bias effect of the arbitrarily chosen best fitted model, all 44 frontier samples were also included into the final ensemble, giving 129 sub-models in the ensemble.

2.2.5 Model Manipulation

An optimization framework was applied to the model ensemble to identify strategies leading to enhanced aromatic amino acid production. The objective function was the aromatic amino acid production rate from the reaction catalyzed by the enzyme DAHPS complex from the two substrates PEP and E4P. The set of strategies considered consisted of enzyme knock-outs, in which an enzyme activity was completely removed from the model, as well as enzyme over- and under-expression (termed “expression change” here) spanning a range from ten times to one-tenth the unperturbed concentration. All combinations of single, double, and triple enzyme knock-outs as well as all combinations of one and two enzyme expression changes were constructed and studied. If one enzyme concentration in a two enzyme expression-change strategy was suggested to be the value of the lower bound (one-tenth the unperturbed concentration), a knock-out of that enzyme plus an expression change of the other enzyme was also studied.

The effect of pure knock-outs, whether single or multiples, was studied by simulating until a steady state was reached and evaluating the objective function in the steady state. The effect of perturbations that included one or more expression-changed enzymes was studied by optimizing to find the combination of levels of modified enzymes leading to the maximum objective. The optimization was done using the `fmincon` function in MATLAB (version 2008a; The MathWorks, Inc.; Natick, MA). These evaluations were carried out individually on each sub-model in the ensemble. For the case of enzyme expression changes, the level of modification was taken as the median across all sub-models in the ensemble, and each sub-model was re-simulated with this value to re-evaluate the objective in order to reflect the production improvement for the ensemble. For both knock-out-only and enzyme expression-change perturbations, the score of a strategy was taken as the average production improvement ratio of perturbed strain against wild-type strain over all sub-models in the ensemble, and the support rate of a strategy was defined as the percentage of sub-models showing any improvement (greater than 0.1%) over the objective in the unperturbed network.

2.3 Results and Discussion

The *E. coli* central carbon metabolism model of Chassagnole et al. (2002) was converted to mass-action form, augmented with additional enzymes, and reparameterized using both the rate behavior from the starting model and experimental data from Ishii et al. (2007; see MODELING). The resulting model provided an excellent fit to the training data (Figure 2). The value of the 210 computed steady-state metabolite concentrations was within 2 fold of the values obtained by Ishii et al. (2007), with an average difference of roughly 50% (Figure 2B). Simultaneously, most of the enzyme reactions had a perfect or near perfect fit to the corresponding Chassagnole et al. reactions (Figure 2D), with only a few reactions fitting worse but still reasonably (the worst fits are shown in Figure 2C). From this single parameter set, an ensemble of 129 models was constructed that fit the data essentially equally well and that differed from each other only in their parameters (see MODELING). Each member of the ensemble was an excellent fit to the training data and together the ensemble represented the parameter uncertainty inherent in underdetermined biochemical models; in the work here, the ensemble was used to compute a representation of prediction uncertainty.

2.3.1 Knock-out strategies reveal the complexity of the pentose phosphate pathway

Each of the 129 models in the ensemble was explored using combinations of one, two, or three gene knock-outs to identify variants with increased production of aromatic amino acids. The results for single knock-outs are examined here first and the model ensemble shows strong consensus for nearly all of the knock-outs (Figure 2). Each knock-out was classified by computing a productivity factor, which is the ratio of the steady-state aromatic amino acid production rate with and without the knock-out. The categories used were *reduced* (<0.8 fold,

blue), *neutral* (0.8–1.2 fold, yellow), *marginally increased* (1.2–1.5 fold, orange), and *increased* (>1.5 fold, red). Most of the knock-outs were predicted to lead to reduced or neutral production in each of the 129 models. A few gene knock-outs were predicted to lead to increased production in all (*ppc* and *pykF*) or most (*glpat*, *pgm*, and *rpe*) models. Even when most but not all models predicted increased production, the dissenting models usually predict marginally increased or neutral production. Interestingly, *rpe* is the exception; most models expected increased production, but thirteen predicted reduced production. Moreover, the *talB* knock-out has complementary behavior; most models predicted reduced production, except the same thirteen that predicted an enhancement. Finally, a minority of models predicted enhanced production for *rpiA* and *rppk* knock-outs, with the remaining models predicting neutral or marginally increased production in all but two cases. Interestingly, the two most strongly predicted knock-outs for improved yield (*ppc* and *pykF*) have been tested and proved effective experimentally by several research groups (Backman, 1992; Gosset et al., 1996). In particular, a *pykF* and *pykA* double knock-out in *E. coli* PB103 strain resulted in a 3.4-fold increase in carbon flux to aromatic biosynthesis (Gosset et al., 1996), which is very similar to the predicted 4.03-fold improved aromatic amino acid production by our ensemble model.

Analysis of steady-state fluxes in all 129 models of the ensemble was carried out to understand the source of computed improvements in aromatic amino acid synthesis (see Figure 4). The gene products of *ppc* and *pykF* are both responsible for fluxes away from the metabolite phosphoenolpyruvate (PEP), which is one of the two precursors of aromatic amino acid synthesis. Steady-state flux results indicated that knocking out either gene singly increased the steady-state amount of PEP significantly (average 14.0 and 2.1 fold, respectively) as well as the flux into aromatic amino acid synthesis (average 55.9 and 4.0 fold, respectively). On the other hand, the

genes *rpe*, *rpiA*, and *talB* are all located in the pentose phosphate pathway and thus may be responsible for increasing the steady-state concentration of erythrose-4-phosphate (E4P), the other precursor of aromatic amino acid synthesis, and the corresponding flux. Knocking out the gene *rpe*, *rpiA*, or *talB* did not affect steady-state PEP concentration significantly, but did increase E4P concentration for all but thirteen models, 114 models, and thirteen models, respectively. The thirteen models showing decreased E4P concentrations with *rpe* knock-out or increased E4P concentrations with *talB* knock-out were the same thirteen models mentioned above that showed improved aromatic amino acid production. The other three single knock-outs predicted to increase aromatic amino acid synthesis in some or most of the models (*pgm*, *glpat*, and *rppk*) terminate carbon fluxes flowing away from the central pathway toward biosynthesis; knocking-out each one individually should also direct more carbon into PEP and E4P. The results indeed show that these three knock-outs moderately increase the steady-state concentrations of both PEP and E4P, as well as the flux toward aromatic amino acid synthesis.

Results above show that knocking out the pentose phosphate pathway genes *rpe*, *rpiA*, and *talB* can increase steady-state E4P concentration and thus aromatic amino acid synthesis rates. However, it is still not clear *why* these knock-outs increase E4P concentration, partially due to the complex topology of the pentose phosphate pathway. It is particularly surprising to see that knocking out the gene *talB*, the product of which is the immediate enzyme that generates E4P, is able to increase E4P concentration and aromatic amino acid synthesis in some models. A more detailed steady-state flux analysis was thus carried out to understand the carbon flow in the pentose phosphate pathway. Interestingly, the results indicate two distinct wild-type steady-state flux patterns for the pentose phosphate pathway among the models — a clockwise flux and a counter-clockwise flux (Figure 5). In the wild-type case (black lines), most models have a carbon

flow direction from metabolite Ru5P to X5P/R5P, to S7P/G3P, to E4P/F6P, and to F6P/G3P (clockwise flux); whereas thirteen models have a reversed flux direction for the enzymes RPE, TKTa/TKTb, and TALa/TALb (counter-clockwise flux). Previously we noted that the *rpe* knock-out was predicted to increase aromatic amino acid synthesis in all but thirteen models; whereas the *talB* knock-out had the complementary behavior. Further investigation revealed that the models in which the *rpe* knock-out showed improvement corresponded to the clockwise flux models, and those for which the *talB* knock-out showed improvement corresponded to the counter-clockwise flux models. The steady-state flux analysis in Figure 5 showed that in clockwise flux models the *rpe* knock-out (red lines) reduced the carbon flux from metabolite Ru5P to E4P (*tktA/tktB*-S7P and *talA/talB* flux) and induced a reversal of the *tktA/tktB*-F6P flux, which corresponds to a switch from a clockwise pentose phosphate pathway flux pattern to a counter-clockwise pattern. In the counter-clockwise flux models, the *talB* knock-out (green lines) terminates the flux flowing away from the metabolite E4P and thus directs more flux toward aromatic amino acid synthesis. Putting the knock-out strategies into the “wrong” flux pattern model results in a decrease of aromatic amino acid synthesis. Interestingly, Ikeda (2003) suggested a possible theory of two-way flux for the effect of manipulating the pentose phosphate pathway in his review of amino acid production. He suggested that a clockwise flux from glucose-6-P to ribose-5-P and then to E4P helps the accumulation of ribose-5-P and thus increases histidine production, whereas a counter-clockwise flux increases E4P concentration and improves aromatic amino acid production. This theory is consistent with our results from ensemble modeling, namely that there could be two distinct flux directions in the pentose phosphate pathway and the conversion from clockwise to counter-clockwise flux increases aromatic amino acid production. The results also indicate that the choice of knock-out strategies

for aromatic amino acid overproduction may depend on the particular wild-type flux pattern of the pentose phosphate pathway under the particular growth conditions. On the other hand, experimental tests on *talB* and *rpe* knock-outs will provide insights for eliminating inappropriate models from the ensemble.

For a defined system, the steady-state concentrations are usually determined by the parameters. However, our model topology only covers a subset of the whole cell system, and there is a possibility of having a different steady state for the whole system. In order to determine whether the two-way flux pattern was caused by intrinsic variance of the parameter sets or extrinsic differences of the metabolite steady state, the *tkt* flux generating F6P and G3P was recalculated for each model using the wild-type steady-state concentrations from all models. Only the results of the 44 frontier models are shown in Figure 6, as the LHS models have similar behaviors. Results show that there is one model that always has a clockwise pattern for all steady-state concentrations tested, and there is another model that always has the counter-clockwise pattern for all steady states. The rest of the models can have either the clockwise or the counter-clockwise pattern depending on which steady state they experience, although some models may have a bias. This indicates that both the intrinsic parameters of the model and the environmental species concentrations matter for determining the flux direction in the pentose phosphate pathway.

When double and triple knock-outs were included, many more options were identified that improved aromatic amino acid production. Multiple knock-outs also showed greater increases in the production rate; the model, unaware of limitations on metabolism outside of the central carbon pathway, claimed to have found a 245-fold improvement with the best performing double knock-out and a 410-fold improvement with the best performing triple knock-out, compared to

67 fold for the best single knock-out. Large variation was observed in the support rate across different strategies, where a strategy can receive between 100% and 0.8% support from the sub-models. (The support is defined as the fraction of models in the ensemble that predict an improvement in production, and thus serves as a proxy for level of consensus in the result.) This variation reflects differences and parameter uncertainty among sub-models. Interestingly, a combination of the *talB* knock-out discussed above and one of the *zwf*, *pgl*, or *gnd* knock-outs increases the support rate of the model ensemble to 100%; whereas a single *talB* knock-out alone received a positive vote from 13 of the 129 sub-models. Further investigation showed that an additional knock-out of *zwf*, *pgl*, or *gnd* helped reverse all clockwise models to counter-clockwise models, so that the *talB* knock-out, which blocked carbon from leaving E4P in counter-clockwise models, worked for all the sub-models and reached an average of 6.7-fold increase for aromatic amino acid production. This result indicates that multiple knock-outs may identify more reliable strategies despite the parameter uncertainty, and those strategies may lead to higher success rates in experimental tests.

2.3.2 Up and down regulation allow new engineering strategies to improve aromatic amino acid production

Rather than using only knock-outs, next we considered combinations of different genetic modifications. Specifically, we allowed up or down regulation or knock-out of up to two enzymes, and we optimized for aromatic amino acid production (see Methods). The upper and lower bound on the enzyme gene expression changes were selected based on examples from the literature and common laboratory practice. Some research groups report enhancing the activity of the glycolytic enzymes PPS and aldolase by 10–15 fold (Babul et al., 1993; Patnaik and Liao, 1994). The general practice for enzyme regulation in research laboratories is approximately up to

a 10-fold enhancement or reduction (Alper et al., 2005). For the current study we thus chose expression changes of up to 10 fold in either direction. To increase the reliability of candidate strategies identified, we further required consensus among sub-models — at least 80% for single modifications and 90% for double modifications. Additionally, double modifications were required to have a higher production rate than both of the corresponding single changes. Strategies identified based on these criteria are listed in Table 2. Results of single and double knock-outs, one knock-out plus one up or down regulation, and two up or down regulations are incorporated together to make this table comprehensive.

Compared to knock-out-only strategies, the inclusion of enzyme up or down regulation allows for new modes for increasing aromatic amino acid production. A total of 37 strategies were identified that involved up regulation of at least one enzyme. Many of these strategies involved overexpression of *dahps* alone or in combination with a change to another enzyme, leading to modeled production increases of 9 fold to 220 fold. The gene product of *dahps* in the model represents a lumped enzyme that converts a molecule of erythrose-4-phosphate (E4P) and one of phosphoenolpyruvate (PEP) to shikimate-3-phosphate (S3P), which combines with another molecule of PEP to form, in two enzymatic steps, chorismate, the common precursor to aromatic amino acids (Figure 1). While it is intuitive that *dahps* is on the synthetic pathway and thus *could* be limiting, there are many enzymes (e.g., *trpSynth1*, *trpSynth2*) on the direct synthetic pathway, and the model has done something significant and non-obvious in identifying this enzyme group as limiting. To what extent *dahps* plays as a limiting factor depends on its kinetic parameters. If we increase the k_{cat} and k_{on} of *dahps* binding E4P and PEP by 100 fold, any further increase of *dahps* concentration will not increase the production rate of aromatic amino acids. In other words, *dahps* is no longer the rate limiting factor under this arbitrary condition. In fact, several research

groups have overexpressed the collection of individual enzymes represented by *dahps* and observed production improvements that form the basis for industrial production strains (Azuma *et al.*, 1993; Berry, 1996; Chan *et al.*, 1993). Other enzymes whose overexpression increases aromatic amino acid synthesis were distributed throughout glycolysis (*pts*, *fbaB*, *tis*, *pgk*, *gpmB*, *pykF*, and *pfkA*) and the pentose-phosphate pathway (*rpiB*, *tktB*, and *talB*), indicating other limitations as well. Interestingly, many of the overexpressed enzymes did not optimize to the full 10-fold overexpression bound, but rather converged at an intermediate value. Most of these cases involve enzymes in the pentose phosphate pathway (e.g., *tktB*, *talB*, and *rpiB*) or the triangle region before G3P (e.g., *fbaB* and *tis*), where further overexpression of the enzymes beyond the optimal value starts to decrease the production rate. This indicates the complexity of the two regions in the network and the importance of the fine-tuned enzyme optimization strategies.

The knock-down results provide an interesting comparison to the knock-out results. In some cases, a complete knock-out further enhances productivity compared to partial knock-down, including *ppc* (10-fold knock-down increases production 45.1 fold; knock-out increases production 66.6 fold). However, in a number of cases, a partial knock-down results in greater productivity than a complete knock-out, including *gapA* (10-fold knock-down increases production 35.9 fold; knock-out has zero production due to complete depletion of PEP, which makes sense given the topology of Figure 1). The existence of multiple isomers with the same activity provides a convenient way to partially knock down a pathway by knocking out only one of the parallel isomers.

2.3.3 Engineering strategies that rebalance carbon fluxes between glycolysis and pentose phosphate pathway lead to improvement of aromatic amino acid production

A group of high-performance strategies involving down-regulating enzymes or knocking out isomers located downstream of G3P in the glycolysis pathway (*gapA*, *pgk*, *gpmA*, and *eno*) were observed, which are not intuitive as these enzymes directly lead to the generation of the precursor PEP for aromatic amino acid production. A further examination shows that the steady-state concentration of PEP is approximately 13 fold higher than that of E4P in wild-type models (Figure 7A). A 10-fold knock-down of *gapA* reduced the steady-state concentration of PEP to 91.9% while increasing that of E4P to 3.9 fold (PEP to E4P ratio is 3.2); the corresponding aromatic amino acid production rate was increased to 13.6 fold. A combination of a 10-fold knock-down of *gapA* and a 10-fold knock-down of *gpmA* decreased the concentration of PEP to 28.0% while increasing the concentration of E4P to 7.0 fold (PEP to E4P ratio is 1.4); and the corresponding aromatic amino acid production rate was further increased to 33.4 fold (Figure 7A). On the other hand, a combination of *tktB* and *ppc* knock-outs corresponding to an increase of PEP concentration and decrease of E4P concentration (PEP to E4P ratio is 1243) resulted in a significantly lower aromatic amino acid production rate (12.1% of the wild-type level). The production of 1 molecule of an aromatic amino acid requires 1 molecule of PEP and 1 molecule of E4P — an equal amount of the two precursors. The results above indicate that strategies that increased the level of E4P to match that of PEP led to significant improvement in aromatic amino acid production, whereas strategies that increased the discrepancy in the amounts of the two precursors reduced productivity. This effect can also be seen in Figures 7B and 7C, showing that a relative ratio of PEP to E4P closer to one resulted in higher aromatic amino acid productivity.

Further investigation was conducted to understand the relative sensitivity of aromatic amino acid production to the concentrations of E4P and PEP. The *rpe + rppk + pgm* triple knock-out model and *tktB* knock-out model had approximately the same level for PEP steady-state concentrations, but increased E4P steady-state concentration to 7.6 fold and decreased E4P concentration to 24.0% of the wild-type value, respectively. The corresponding aromatic amino acid production rate increased to 55.1 fold and decreased to 5.9% of the wild-type level, respectively (Figure 7A). On the other hand, the triple knock-outs *rpe + ppc + talB* and *pdh + rpiA + pgl* kept E4P levels approximately the same, but either increase or decrease the PEP steady-state concentration. Interestingly, a 20.5-fold increase in PEP concentration resulted in only a 1.9-fold increase in the aromatic amino acid production rate (Figure 7A). This indicates that the E4P concentration has a limiting role under the conditions of this study, and aromatic amino acid production is much more sensitive to the concentration of E4P than to that of PEP. The *gapA* and *gapA + gpmA* knock-down strategies discussed above reduced carbon flow through the glycolysis pathway toward PEP and redistributed the carbon flux toward E4P. They increased the overall production rate significantly by sacrificing some amount of PEP in order to increase the E4P concentration. Conventionally, metabolic engineers focus more on increasing one or both of the precursors of aromatic amino acid synthesis, but little attention has been focused on creating a balance between the two, partially because of the difficulty in identifying proper experiment strategies. Our results show that the mathematical models we presented here are capable of identifying the limiting precursor and providing strategies that rebalance the precursors and efficiently improve productivity.

Unlike the strategies such as *ppc* knock-out, *synth1* knock-out, and *glpat* knock-out, which work by removing a carbon sink and thus directing more carbon toward aromatic amino acid

production, the rebalancing strategies re-distribute redundant carbon from PEP to E4P and thus do not require increased carbon supply to enhance productivity. A combination of these two types of strategies resulted in the largest computed improvement of aromatic amino acid production.

The synthesis of tryptophan, an aromatic amino acid, also requires serine as an additional precursor in the path that branches past chorismate. Therefore, strategies that reduce serine production, e.g. knock-out of the serine synthesis reaction, would not be good candidates for tryptophan production. We simulated the level of serine production for each strategy in Table 2 and calculated the serine production improvement ratio (see Table 2). If tryptophan overproduction were the target, the serine production improvement ratio would also need to be taken into account. All of the strategies reported in Table 2, except for the serine synthesis reaction knock-out, have a small effect on or negligible reduction of (approximately 10% or less, if any) the serine production rate.

2.4 Conclusions

The central carbon metabolic network is one of the most important biological networks in metabolic engineering, as most of the precursors for primary and secondary metabolites that have industrial interests are coming from this network. Given the complicated enzyme interactions and regulations in this network, it is often not straightforward to identify efficient strategies to optimize the production rates of the target metabolites. Here we presented a mass-action model ensemble, which incorporated the most up-to-date knowledge about the topology and enzyme isomers, for the *E. coli* central carbon metabolic network. The model ensemble includes 129 individual models which have equally good fit to the steady-state data from Ishii et al. and the reaction rate data from Chassagnole et al. The variations of parameters among different models thus provide a useful measurement for the impact of parameter uncertainty on model predictions. An exhaustive optimization search, including single, double, and triple enzyme knock-outs as well as single and double enzyme over- and under-expressions, was applied to the model ensemble in order to identify enzyme strategies that can maximize the aromatic amino acid production rate. The *rpe* and *talB* single knock-outs identified through the optimization reveal the complexity of the pentose phosphate pathway and suggest there could be two natural flux direction of the pentose phosphate pathway, clockwise direction and counter-clockwise direction. More importantly, different strategies may be needed to optimize the aromatic amino acid production, depending on the natural direction of the system. As two precursors are required for the aromatic amino acid production, it is believed that a balanced precursor level could benefit the production the most. Non-obvious strategies that help re-partition the carbon fluxes towards the PEP and E4P precursors were identified by the optimizations. It improves the aromatic amino acid production without the need to increase the network carbon input. Without the

computational model ensemble we built, it would be difficult to identify these strategies given the complexity of the central carbon metabolism. The mathematical model ensemble and the manipulation tool we built here can be easily extended to study other chemicals of interests that have precursors from the central carbon metabolic network. With limited data, it is anticipated to have the capability to provide insightful guidance of experimental designs that optimize the production yields of the desired chemicals.

Acknowledgement

We thank Eugénio Ferreira, Daniel O’Keefe, Mark J. Nelson, Kristala Jones Prather, and Isabel Rocha for helpful and insightful discussions. This work was partially supported by the DuPont MIT Alliance and the MIT Portugal Program.

References

- Alper, H., Fischer, C., Nevoigt, E., Stephanopoulos, Gregory., 2005. Tuning genetic control through promoter engineering. *Proceedings of the National Academy of Sciences of the United States of America* 102, 12678–83.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1200280&tool=pmcentrez&rendertype=abstract>.
- Alper, H., Stephanopoulos, Gregory., 2009. Engineering for biofuels: exploiting innate microbial capacity or importing biosynthetic potential? *Nature reviews. Microbiology* 7, 715–23.
<http://www.ncbi.nlm.nih.gov/pubmed/19756010>.
- Atsumi, S., Liao, James C., 2008. Metabolic engineering for advanced biofuels production from *Escherichia coli*. *Current opinion in biotechnology* 19, 414–9.
<http://www.ncbi.nlm.nih.gov/pubmed/18761088>.
- Azuma, S., Tsunekawa, H., Okabe, M., Okamoto, R., Aiba, S., 1993. Hyper-production of l-tryptophan via fermentation with crystallization. *Applied Microbiology and Biotechnology* 39, 471–476. <http://www.springerlink.com/index/T2P26XUG17N3521R.pdf> (Accessed March 8, 2011).
- Babul, J., Clifton, D., Kretschmer, M., Fraenkel, D.G., 1993. Glucose metabolism in *Escherichia coli* and the effect of increased amount of aldolase. *Biochemistry* 32, 4685–92.
<http://www.ncbi.nlm.nih.gov/pubmed/8485146>.
- Backman, K.C., 1992. Method of biosynthesis of phenylalanine. 5, 169, 768.
- Bailey, J.E., 1991. Toward a science of metabolic engineering. *Science (New York, N.Y.)* 252, 1668–75. <http://www.ncbi.nlm.nih.gov/pubmed/2047876>.
- Barkovich, R., Liao, J C., 2001. Metabolic engineering of isoprenoids. *Metabolic engineering* 3, 27–39. <http://www.ncbi.nlm.nih.gov/pubmed/11162230>.
- Berry, A., 1996. Improving production of aromatic compounds in *Escherichia coli* by metabolic engineering. *Trends in biotechnology* 14, 250–6. <http://www.ncbi.nlm.nih.gov/pubmed/8771798>.
- Bongaerts, J., Krämer, M., Müller, U., Raeven, L., Wubbolts, M., 2001. Metabolic engineering for microbial production of aromatic amino acids and derived compounds. *Metabolic engineering* 3, 289–300. <http://www.ncbi.nlm.nih.gov/pubmed/11676565> (Accessed August 3, 2010).
- Braga, R., Hecquet, L., Blonski, C., 2004. Slow-binding inhibition of 2-keto-3-deoxy-6-phosphogluconate (KDPG) aldolase. *Bioorganic & medicinal chemistry* 12, 2965–72.
<http://www.ncbi.nlm.nih.gov/pubmed/15142555> (Accessed February 22, 2011).

- Burgard, A.P., Pharkya, P., Maranas, C.D., 2003. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering* 84, 647–57. <http://www.ncbi.nlm.nih.gov/pubmed/14595777>.
- Cameron, D C, Tong, I.T., 1993. Cellular and metabolic engineering. An overview. *Applied biochemistry and biotechnology* 38, 105–40. <http://www.ncbi.nlm.nih.gov/pubmed/8346901>.
- Cameron, D.C., Chaplen, F.W.R., 1997. Developments in metabolic engineering. *Current opinion in biotechnology* 8, 175–180. <http://linkinghub.elsevier.com/retrieve/pii/S0958166997800985> (Accessed February 28, 2011).
- Chan, E.-C., Tsai, H.-L., Chen, S.-L., Mou, D.-G., 1993. Amplification of the tryptophan operon gene in *Escherichia coli* chromosome to increase l-tryptophan biosynthesis. *Applied Microbiology and Biotechnology* 40, 301–305. <http://springerlink.metapress.com/openurl.asp?genre=article&id=doi:10.1007/BF00170384>.
- Chassagnole, C., Noisommit-Rizzi, N., Schmid, J.W., Mauch, K., Reuss, M., 2002. Dynamic modeling of the central carbon metabolism of *Escherichia coli*. *Biotechnology and Bioengineering* 79, 53–73. <http://doi.wiley.com/10.1002/bit.10288>.
- Cleland., 1963. The kinetics of enzyme-catalyzed reactions with two or more substrates or products. I. Nomenclature and rate equations. *Biochimica et Biophysica Acta* 67:104. <http://www.ncbi.nlm.nih.gov/pubmed/20875501>.
- Cornish-Bowden, A., 1977. An automatic method for deriving steady-state rate equations. *Biochemical Journal* 165, 55.
- Edwards, J S, Palsson, B O., 2000. Robustness analysis of the *Escherichia coli* metabolic network. *Biotechnology progress* 16, 927–39. <http://www.ncbi.nlm.nih.gov/pubmed/11101318>.
- Edwards, Jeremy S., Covert, M., Palsson, B., 2002. Metabolic modelling of microbes: the flux-balance approach. *Environmental Microbiology* 4, 133–140. <http://doi.wiley.com/10.1046/j.1462-2920.2002.00282.x>.
- Feist, A.M., Zielinski, D.C., Orth, J.D., Schellenberger, J., Herrgard, M.J., Palsson, B.Ø., 2010. Model-driven evaluation of the production potential for growth-coupled products of *Escherichia coli*. *Metabolic engineering* 12, 173–86. <http://www.ncbi.nlm.nih.gov/pubmed/19840862> (Accessed December 15, 2010).
- Gosset, G., Yong-Xiao, J., Berry, a., 1996. A direct comparison of approaches for increasing carbon flow to aromatic biosynthesis in *Escherichia coli*. *Journal of industrial microbiology* 17, 47–52. <http://www.ncbi.nlm.nih.gov/pubmed/8987689>.
- Ikeda, M., 2003. Amino acid production processes. *Advances in biochemical engineering/biotechnology* 79, 1–35. <http://www.ncbi.nlm.nih.gov/pubmed/12523387>.

- Ikeda, M., 2006. Towards bacterial strains overproducing L-tryptophan and other aromatics by metabolic engineering. *Applied microbiology and biotechnology* 69, 615–26.
<http://www.ncbi.nlm.nih.gov/pubmed/16374633> (Accessed September 5, 2010).
- Ishii, N. et al., 2007. Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science (New York, N.Y.)* 316, 593–7.
<http://www.ncbi.nlm.nih.gov/pubmed/17379776>.
- Kanehisa, M., Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* 28, 27–30.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102409&tool=pmcentrez&rendertype=abstract>.
- Kauffman, K.J., Prakash, P., Edwards, Jeremy S., 2003. Advances in flux balance analysis. *Current Opinion in Biotechnology* 14, 491–496.
<http://linkinghub.elsevier.com/retrieve/pii/S0958166903001174> (Accessed July 16, 2010).
- Keasling, J.D., 1999. Gene-expression tools for the metabolic engineering of bacteria. *Trends in biotechnology* 17, 452–460. <http://linkinghub.elsevier.com/retrieve/pii/S0167779999013761> (Accessed February 28, 2011).
- King, E.L., Altman, C., 1956. A schematic method of deriving the rate laws for enzyme-catalyzed reactions. *The Journal of Physical Chemistry* 60, 1375–1378.
- Kuzmic, P., 2008. The king-altman method. <http://www.biokin.com/king-altman/index.html>.
- Lee, J.M., Gianchandani, E.P., Papin, J. a., 2006. Flux balance analysis in the era of metabolomics. *Briefings in bioinformatics* 7, 140–50.
<http://www.ncbi.nlm.nih.gov/pubmed/16772264>.
- Leuchtenberger, W., Huthmacher, K., Drauz, K., 2005. Biotechnological production of amino acids and derivatives: current status and prospects. *Applied microbiology and biotechnology* 69, 1–8.
- Li, J.W.-H., Vederas, J.C., 2009. Drug discovery and natural products: end of an era or an endless frontier? *Science (New York, N.Y.)* 325, 161–5.
<http://www.ncbi.nlm.nih.gov/pubmed/19589993> (Accessed July 16, 2012).
- Liao, J C, Hou, S.Y., Chao, Y.P., 1996. Pathway analysis, engineering, and physiological considerations for redirecting central metabolism. *Biotechnology and bioengineering* 52, 129–40.
<http://www.ncbi.nlm.nih.gov/pubmed/18629859>.
- Machado, D., Costa, R.S., Ferreira, E.C., Rocha, I., Tidor, B., 2012. Exploring the gap between dynamic and constraint-based models of metabolism. *Metabolic engineering* 14, 112–9.
<http://www.ncbi.nlm.nih.gov/pubmed/22306209> (Accessed October 27, 2012).

McKay, M.D., Beckman, R., Conover, W., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21, 239–245. <http://www.jstor.org/stable/1271432> (Accessed January 31, 2011).

Neidhardt, F., Ingraham, J., Schaechter, M., 1990. *Physiology of the Bacterial Cell: A Molecular Approach*. Sinauer Associates Inc.

Nelson, D.L., Cox, M.M., 2008. *Lehninger principles of biochemistry*. 5th ed. W. H. Freeman.

Orth, J.D., Thiele, I., Palsson, B.Ø., 2010. What is flux balance analysis? *Nature biotechnology* 28, 245–8. <http://www.ncbi.nlm.nih.gov/pubmed/20212490>.

Patnaik, R., Liao, J C., 1994. Engineering of *Escherichia coli* central metabolism for aromatic metabolite production with near theoretical yield. *Applied and environmental microbiology* 60, 3903–8. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=201913&tool=pmcentrez&rendertype=abstract>.

Pramanik, J., Keasling, J D., 1998. Effect of *Escherichia coli* biomass composition on central metabolic fluxes predicted by a stoichiometric model. *Biotechnology and bioengineering* 60, 230–8. <http://www.ncbi.nlm.nih.gov/pubmed/10099424>.

Scheer, M. et al., 2010. BRENDA, the enzyme information system in 2011. *Nucleic Acids Research* 39, D670–D676. <http://www.nar.oxfordjournals.org/cgi/doi/10.1093/nar/gkq1089> (Accessed February 11, 2011).

Schmid, J.W., Mauch, K., Reuss, M., Gilles, E.D., Kremling, A., 2004. Metabolic design based on a coupled gene expression-metabolic network model of tryptophan production in *Escherichia coli*. *Metabolic engineering* 6, 364–77. <http://www.ncbi.nlm.nih.gov/pubmed/15491865>.

Segrè, D., Vitkup, D., Church, G.M., 2002. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America* 99, 15112–7. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=137552&tool=pmcentrez&rendertype=abstract>.

Stein, M., 1987. Large sample properties of simulations using Latin hypercube sampling. *Technometrics* 29, 143–151. <http://www.jstor.org/stable/1269769> (Accessed January 31, 2011).

Stephanopoulos, G, Sinskey, a J., 1993. Metabolic engineering--methodologies and future prospects. *Trends in biotechnology* 11, 392–6. <http://www.ncbi.nlm.nih.gov/pubmed/7764086>.

Stephanopoulos, Gregory, Aristidou, A., Nielsen, J., 1998. *Metabolic Engineering: Principles and Methodologies*. 1st ed. Academic Press.

Tyo, K.E., Alper, H.S., Stephanopoulos, G.N., 2007. Expanding the metabolic engineering toolbox: more options to engineer cells. *Trends in biotechnology* 25, 132–7. <http://www.ncbi.nlm.nih.gov/pubmed/17254656>.

Tzafiri, a R., 2003. Michaelis-Menten kinetics at high enzyme concentrations. *Bulletin of mathematical biology* 65, 1111–29. <http://www.ncbi.nlm.nih.gov/pubmed/14607291> (Accessed February 16, 2011).

Usuda, Y. et al., 2010. Dynamic modeling of *Escherichia coli* metabolic and regulatory systems for amino-acid production. *Journal of biotechnology* 147, 17–30. <http://www.ncbi.nlm.nih.gov/pubmed/20219606> (Accessed February 19, 2011).

Varma, A., Palsson, B.O., 1994. Metabolic flux balancing: basic concepts, scientific and practical use. *Nature Biotechnology* 12, 994–998. <http://www.nature.com/nbt/journal/v12/n10/abs/nbt1094-994.html> (Accessed February 17, 2011).

Vital-Lopez, F.G., Armaou, A., Nikolaev, E.V., Maranas, C.D., 2006. A computational procedure for optimal engineering interventions using kinetic models of metabolism. *Biotechnology progress* 22, 1507–17. <http://www.ncbi.nlm.nih.gov/pubmed/17137295>.

Figures

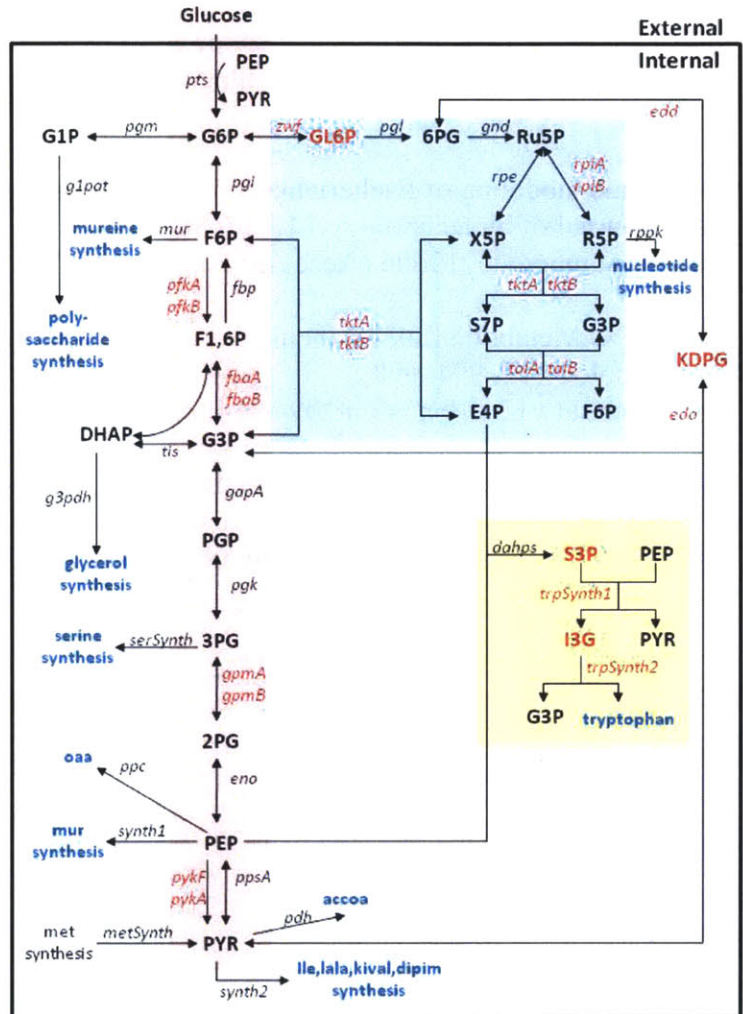


Figure 1. Structure of the central carbon metabolic network model. The red box indicates the glycolysis pathway; the blue box indicates the pentose phosphate pathway; and the yellow box indicates the aromatic amino acid synthesis pathway. The blue letters are the terminal products as output of the network. The enzymes and metabolites in red are the modifications made to the Chassagnole *et al.* model. The aromatic amino acid synthesis pathway (yellow box) was added to the Chassagnole *et al.* model to replace the constant carbon flux to G3P and PYR due to tryptophan synthesis.

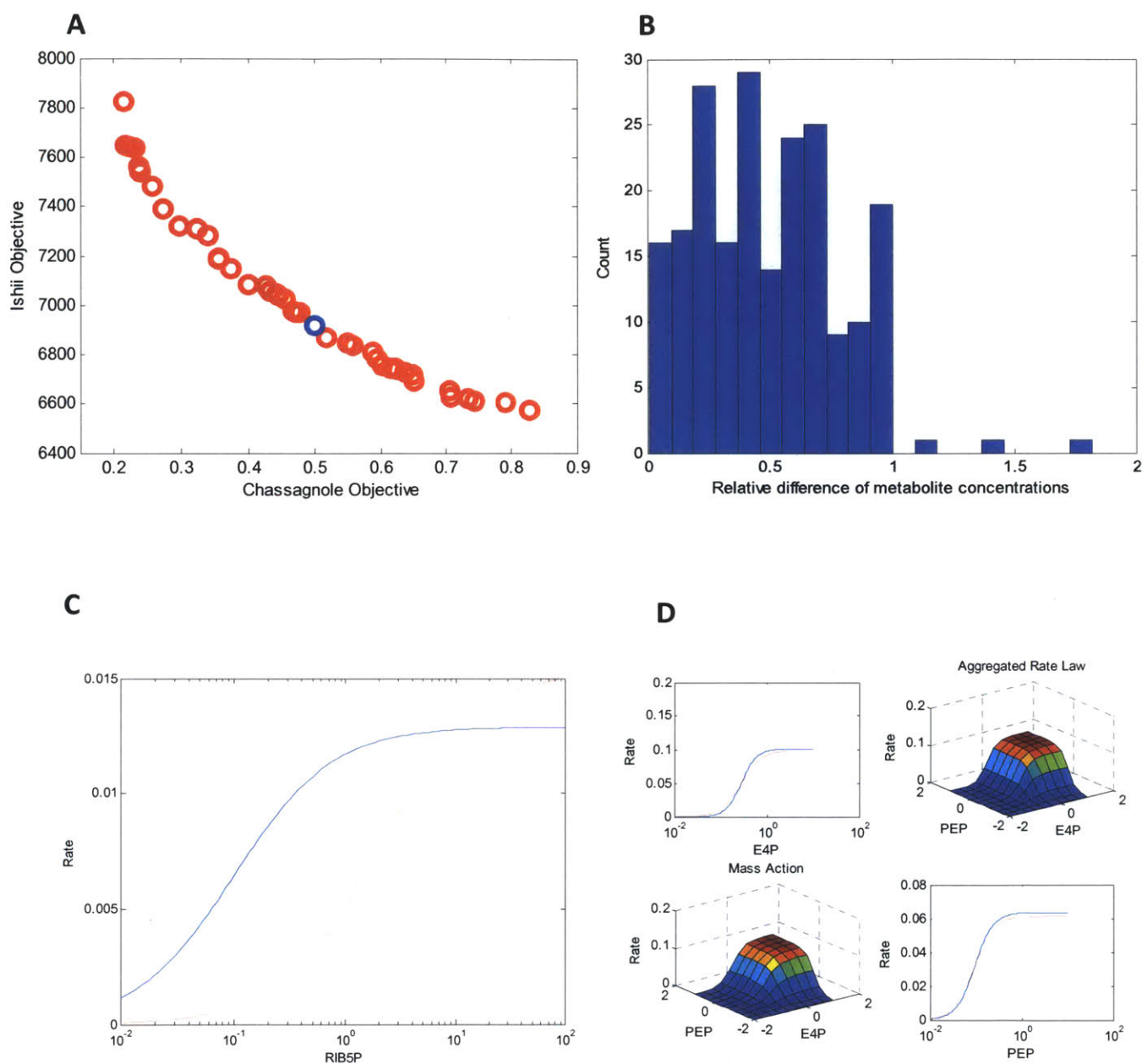


Figure 2. **A.** Pareto optimal frontier. The x-axis is the chi-square objective value of fitting to Chassagnole *et al.*(2002) rate formulae, and the y-axis is the chi-square objective value of fitting to Ishii *et al.* (2007) measurement data. 44 frontier points were obtained from the optimization by systematically varying the weighting factor w . The blue point indicates the parameter set that was chosen as the basis for local sampling in parameter space. **B.** The relative difference of the calculated steady-state metabolite concentrations to the 210 Ishii *et al.*(2007) measured steady-state values for the optimal (blue in panel A) model. **C,** **D.** Representative fitting results for the optimal model (red dotted lines) to the Chassagnole *et al.* model (blue lines); C. (RPPK reaction) represents the worst fitting reaction and D. (DAHPS reaction) represents a typical fitting reaction.

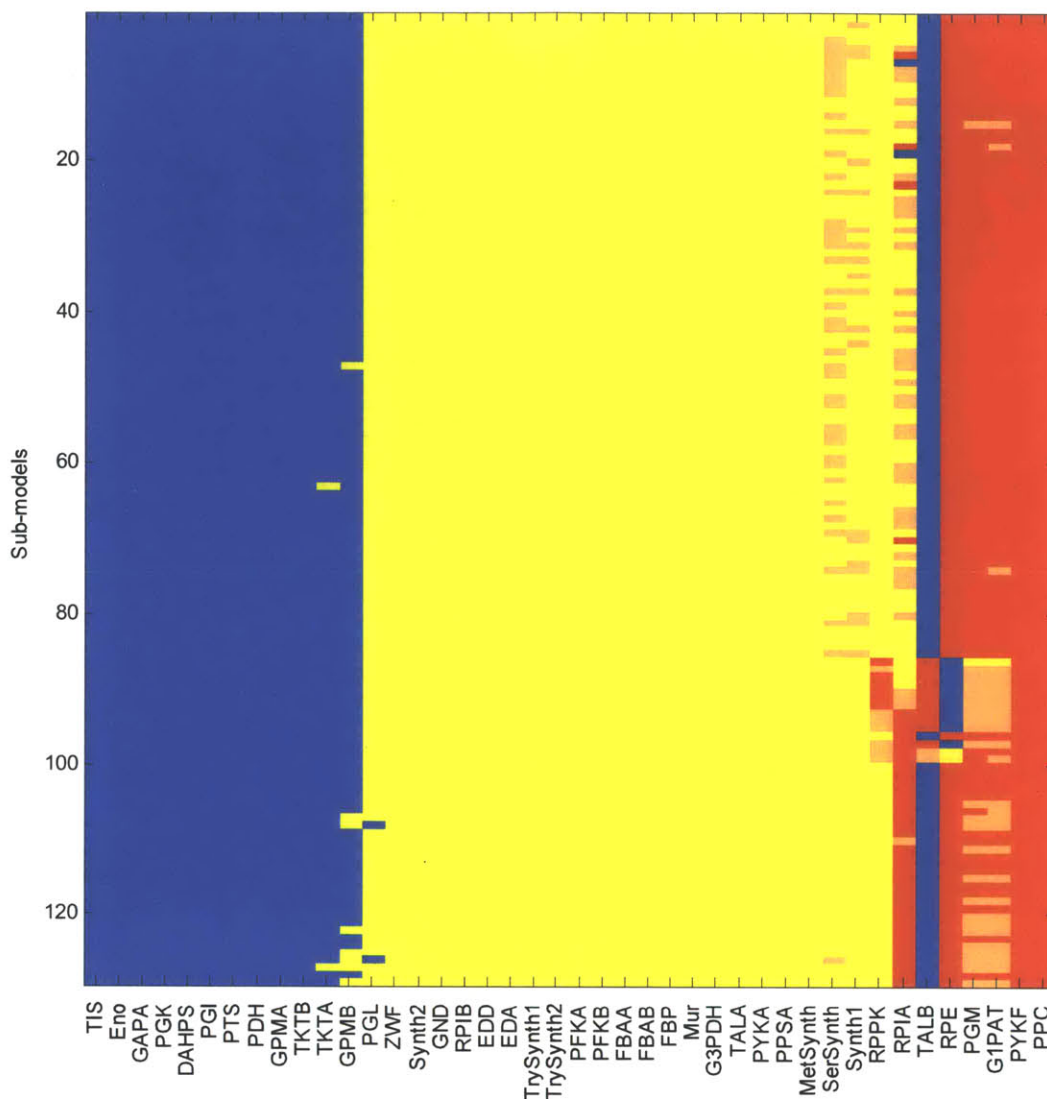


Figure 3. Single knock-out results for 129 sub-models. The columns correspond to the 42 single knockouts and the rows correspond to the 129 sub-models. The data plotted are the ratio of aromatic amino acid production rate for knockouts to that for wildtype. Blue corresponds to a ratio smaller than 0.8 (strategies diminishing aromatic amino acid production); yellow corresponds to a ratio between 0.8 and 1.2 (strategies may or may not increase aromatic amino acid production); orange corresponds to a ratio between 1.2 and 1.5 (strategies improving aromatic amino acid production); and red corresponds to a ratio larger than 1.5 (strategies significantly improving aromatic amino acid production).

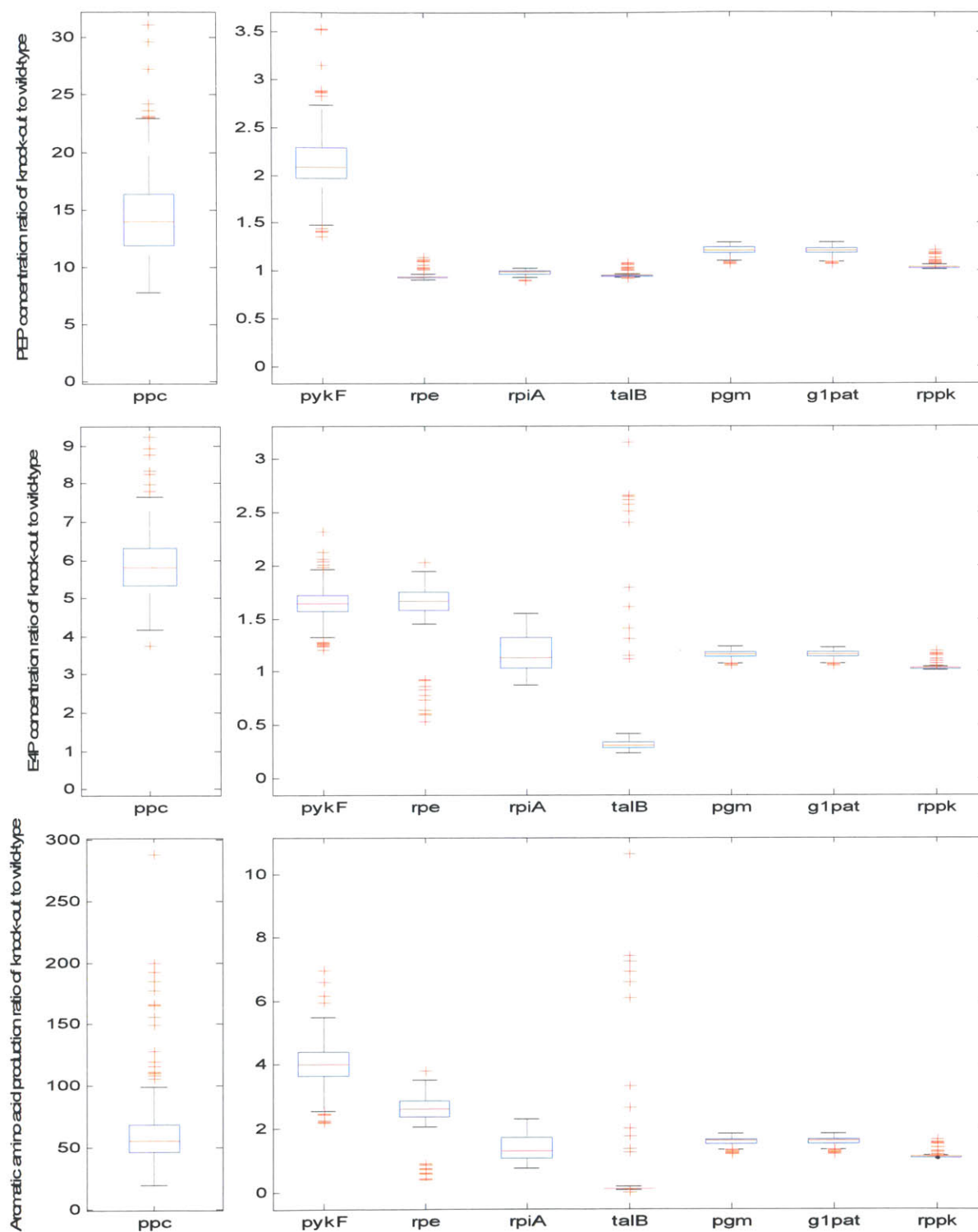


Figure 4. Steady-state flux analysis on important knock-outs. The top panels (PEP) and middle panels (E4P) show the relative steady-state concentrations of the knock-out strains to the wild-type strain; the bottom panels show the relative steady-state aromatic amino acid fluxes of the knock-out strains to the wild-type strain. The box plot indicates the results of the 129 models in the ensemble, with the middle red line as the median, the edges of the box as the 25th and 75th percentiles, the whiskers extended to the most extreme data points not considered outliers, and the outliers plotted individually.

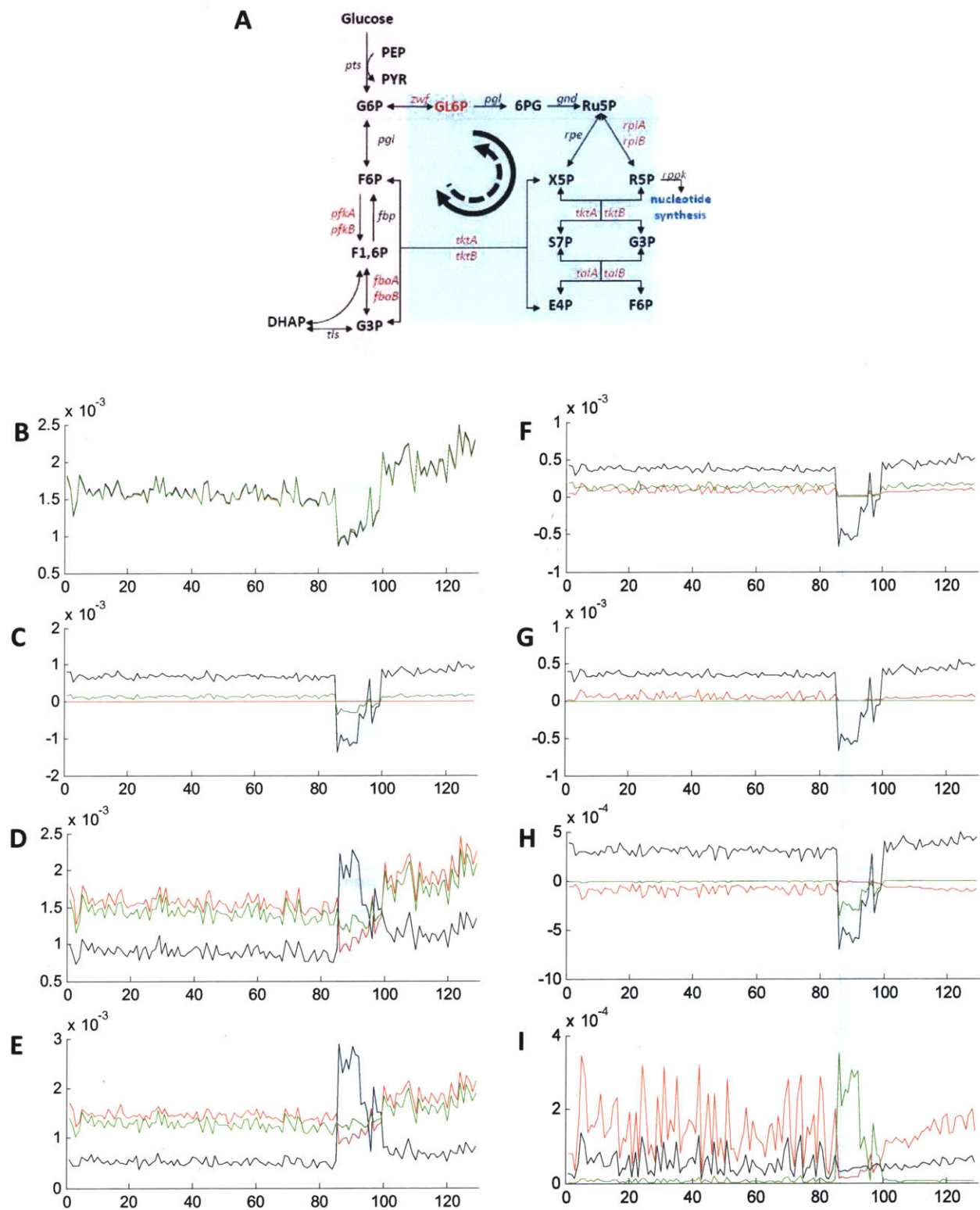


Figure 5. Steady-state flux analysis. **A.** the pentose phosphate pathway. The solid arrow shows the direction of clockwise flux and dotted arrow shows the direction of counter-clockwise flux. **B-I.** The black lines are the wild-type cases; the red lines are the *rpe* knock-out cases; and the green lines are the *talB* knock-out cases. The x-axis corresponds to each of the 129 models and the y-axis corresponds to the fluxes (mM/s) for each of the eight enzyme reactions. The 86 – 98 models (shaded) are the thirteen models that could have reversed fluxes. **B:** *gnd* flux; **C:** *rpe* flux; **D:** *rpi* flux; **E:** *rpk* flux; **F:** *tkt* (to S7P) flux; **G:** *tal* flux; **H:** *tkt* (to F6P) flux; **I:** *dahps* flux.

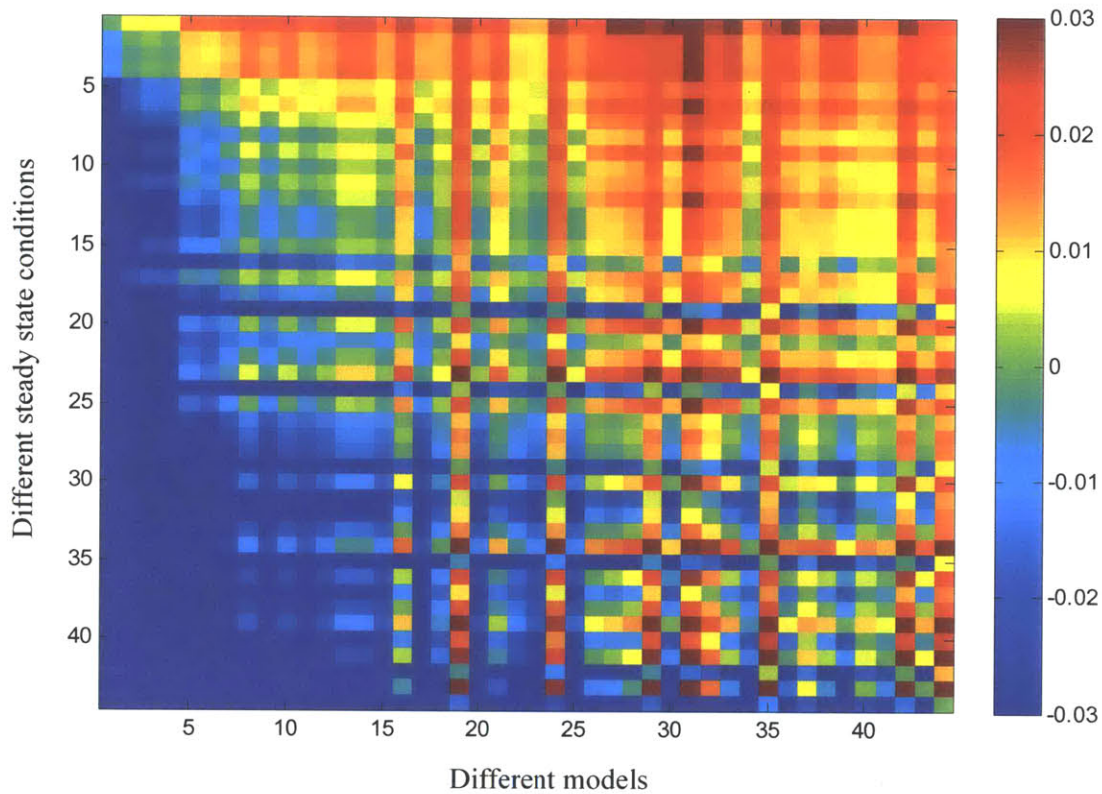


Figure 6. Flux direction and steady-state concentrations. The x-axis represents different frontier models and the y-axis represents different steady-state conditions the models experience. The steady-state conditions were collected from the wild-type steady-state concentrations of the 44 frontier models. The color indicates the flux value, with a unit of mM/s.

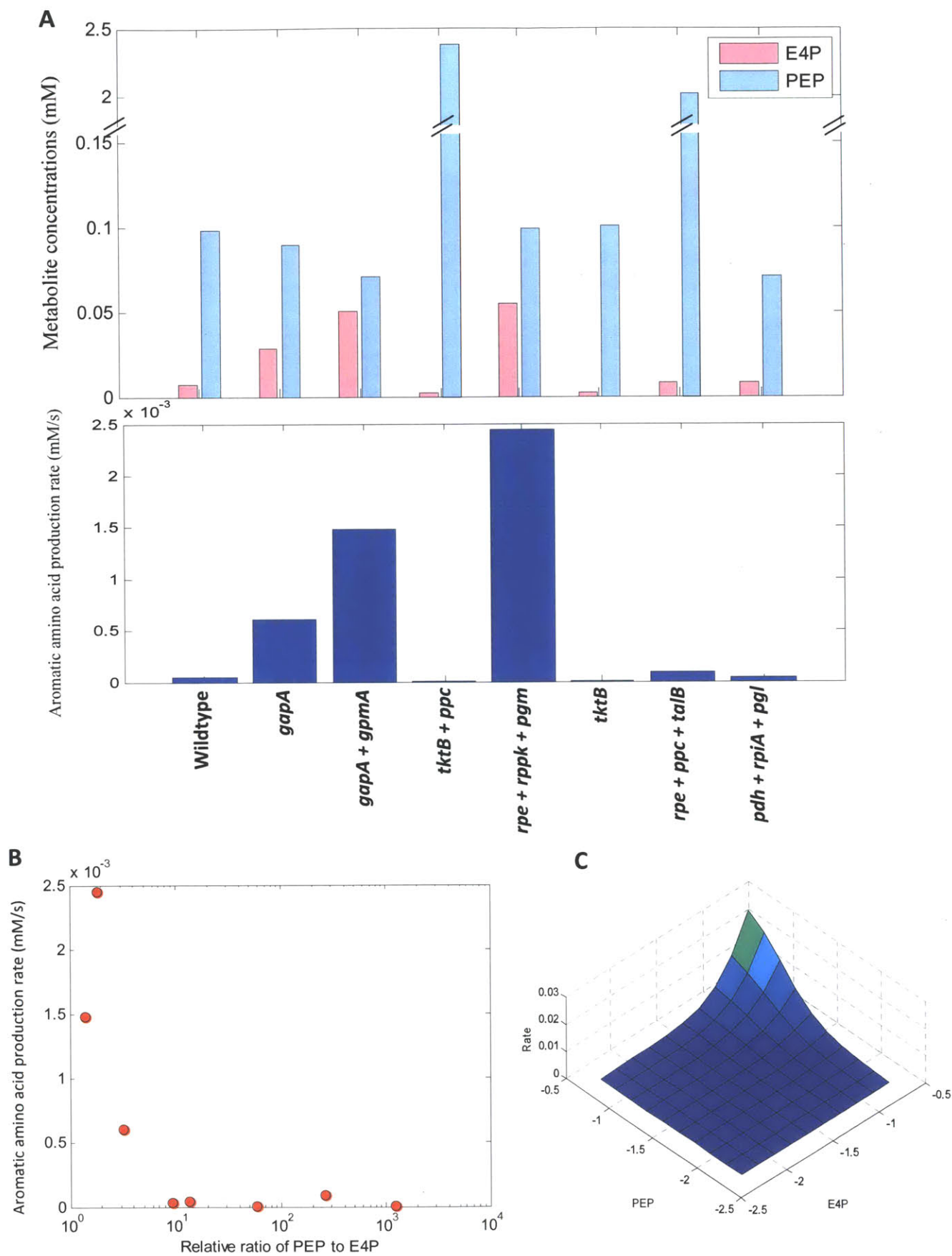


Figure 7. Balance PEP and E4P steady-state concentrations. **A.** the upper panel shows the steady-state concentrations of PEP and E4P for different enzyme manipulation strategies; the lower panel shows the corresponding aromatic amino acid production rates. **B.** the relative ratio of PEP to E4P for each of the strategies in **A** were calculated and plotted against the corresponding aromatic amino acid production rate. **C.** the reaction rate of *dahps* enzyme for different PEP and E4P concentrations.

Tables

Old enzyme	New enzymes	Mechanisms
PGLuMu	GPMA/GPMB	Uni-Uni reversible
R5PI	RPIA/RPIB	Uni-Uni reversible
Aldolase	FBAA/FBAB	Uni-Bi ordered reversible
PFK	PFKA/PFKB/FBP	Uni-Uni irreversible with and without Hill Coef
TA	TALA/TALB	Bi-Bi ordered reversible
Tka	Tka/TKb	Bi-Bi ordered reversible
TKb	Tka/TKb	Bi-Bi ordered reversible
PK	PYKF/PYKA/PPSA	Uni-Uni irreversible with Hill Coef
G6PDH	ZWF/PGL	Uni-Uni reversible/Uni-Uni irreversible
	EDD/EDA	Uni-Uni reversible/Uni-Bi ordered reversible
TrpSynth	TrpSynth1/TrpSynth2	Bi-Bi ordered irreversible/Uni-Uni irreversible

Table 1. The enzyme reaction modification.

Strategies	Support	E/E ₀	Score	Serine	Strategies	Support	E/E ₀	Score	Serine	Strategies	Support	E/E ₀	Score	Serine
EDD + PYKA	96.1%	KO/KO	1.001	1.00	RPIA	91.5%	KO/KO	1.40	0.94	PFKA	100%	KO/KO	1.24	1.01
EDD + FBAA	92.2%	KO/KO	1.004	1.00	Ru5P	93.0%	KO/KO	2.84	0.89	PFKB	100%	KO/KO	1.26	1.02
PYKA + FBAA	92.2%	KO/KO	1.004	1.00	GPMB	99.2%	10/10	2.93	2.52	RPPK	100%	KO/KO	1.30	1.10
PGK + PTS	92.2%	0.1/0.1	1.29	0.97	PFKB	100%	KO	1.08	0.96	RPIA	96.9%	KO/KO	1.59	1.04
RPIA + GND	99.2%	KO/KO	1.78	0.93	TALA	100%	KO/KO	1.09	0.96	Synth1	100%	KO	1.16	1.07
RPIA + PGL	99.2%	KO/KO	1.86	0.94	MetSynth	100%	KO/KO	1.09	0.96	G3PDH	100%	KO/KO	1.17	1.07
GPMA + PTS	98.4%	0.1/0.1	3.09	2.43	PPSA	100%	KO/KO	1.09	0.96	ZWF	100%	KO/KO	1.21	1.09
GPMA + PGM	100%	0.1/KO	5.15	2.82	PYKA	100%	KO/KO	1.09	0.96	PFKA	100%	KO/KO	1.23	1.02
GPMA + Ru5P	99.2%	0.1/KO	5.62	2.43	EDD	100%	KO/KO	1.09	0.96	PFKB	100%	KO/KO	1.26	1.02
TALB + GND	100%	KO/KO	6.24	0.97	FBAB	100%	KO/KO	1.09	0.97	RPPK	100%	KO/KO	1.31	1.11
TALB + PGL	100%	KO/KO	6.39	0.99	FBAA	100%	KO/KO	1.09	0.97	Mur	100%	KO/KO	1.35	1.13
TALB + ZWF	100%	KO/KO	6.67	1.02	G3PDH	100%	KO/KO	1.09	0.97	RPIA	96.9%	KO/KO	1.59	1.04
G3PDH	100%	KO	1.01	1.00	ZWF	100%	KO/KO	1.14	0.99	SerSynth	100%	KO	1.20	0
TALA	100%	KO/KO	1.01	1.00	RPIA	91.5%	KO/KO	1.42	0.95	ZWF	100%	KO/KO	1.25	0
PPSA	100%	KO/KO	1.01	1.00	RPPK	100%	KO	1.12	1.04	PFKA	100%	KO/KO	1.27	0
EDD	100%	KO/KO	1.01	1.00	TrpSynth2	100%	KO/KO	1.12	1.04	PFKB	100%	KO/KO	1.29	0
PYKA	100%	KO/KO	1.01	1.00	TALA	100%	KO/KO	1.12	1.04	RPPK	100%	KO/KO	1.35	0
FBAB	100%	KO/KO	1.01	1.00	PPSA	100%	KO/KO	1.12	1.04	Mur	100%	KO/KO	1.39	0
FBAA	100%	KO/KO	1.01	1.00	PYKA	100%	KO/KO	1.12	1.04	Synth1	100%	KO/KO	1.42	0
ZWF	100%	KO	1.04	1.02	EDD	100%	KO/KO	1.12	1.04	RPIA	98.4%	KO/KO	1.64	0
MetSynth	100%	KO/KO	1.04	1.02	FBAB	100%	KO/KO	1.13	1.04	Ru5P	90.7%	KO/KO	2.83	0
TALA	100%	KO/KO	1.04	1.02	FBAA	100%	KO/KO	1.13	1.04	G1PAT	100%	KO	1.56	1.19
PPSA	100%	KO/KO	1.04	1.02	G3PDH	100%	KO/KO	1.13	1.04	G3PDH	100%	KO/KO	1.57	1.19
PYKA	100%	KO/KO	1.04	1.02	ZWF	100%	KO/KO	1.17	1.06	FBAB	100%	KO/KO	1.58	1.19
FBAA	100%	KO/KO	1.05	1.02	PFKA	100%	KO/KO	1.23	1.00	FBAA	100%	KO/KO	1.58	1.19
FBAB	100%	KO/KO	1.05	1.02	PFKB	100%	KO/KO	1.24	1.01	ZWF	100%	KO/KO	1.64	1.21
G3PDH	100%	KO/KO	1.05	1.02	RPIA	100%	KO/KO	6.70	1.02	RPPK	100%	KO/KO	1.75	1.24
RPIA	99.2%	KO/KO	2.05	0.98	Ru5P	100%	KO/KO	50.60	0.86	PFKA	100%	KO/KO	1.77	1.16
PFKA	98.4%	KO	1.07	0.96	Mur	100%	KO	1.16	1.06	PFKB	100%	KO/KO	1.78	1.17
TALA	98.4%	KO/KO	1.07	0.96	FBP	100%	KO/KO	1.16	1.06	Mur	100%	KO/KO	1.80	1.26
MetSynth	98.4%	KO/KO	1.07	0.96	MetSynth	100%	KO/KO	1.16	1.06	SerSynth	100%	KO/KO	1.92	0
PPSA	98.4%	KO/KO	1.07	0.96	TALA	100%	KO/KO	1.16	1.06	Synth1	100%	KO/KO	1.85	1.28
PYKA	98.4%	KO/KO	1.07	0.96	PPSA	100%	KO/KO	1.16	1.06	RPIA	100%	KO/KO	2.17	1.16
EDD	98.4%	KO/KO	1.07	0.96	EDD	100%	KO/KO	1.16	1.06	Ru5P	92.2%	KO/KO	3.64	1.13
FBAB	98.4%	KO/KO	1.07	0.96	PYKA	100%	KO/KO	1.16	1.06	PGM	100%	KO	1.57	1.19
FBAA	98.4%	KO/KO	1.07	0.96	FBAA	100%	KO/KO	1.16	1.06	G3PDH	100%	KO/KO	1.58	1.19
G3PDH	98.4%	KO/KO	1.07	0.96	FBAB	100%	KO/KO	1.16	1.06	FBAB	100%	KO/KO	1.58	1.19
Tkb	93.0%	2.00/1.90	1.09	1.00	G3PDH	100%	KO/KO	1.16	1.06	FBAA	100%	KO/KO	1.58	1.19
ZWF	100%	KO/KO	1.13	0.98	ZWF	100%	KO/KO	1.20	1.08	ZWF	100%	KO/KO	1.64	1.21

Table 2. The list of best strategies for improving aromatic amino acid productions

Strategies	Support	E/E ₀	Score	Serine	Strategies	Support	E/E ₀	Score	Serine	Strategies	Support	E/E ₀	Score	Serine
RPPK	100%	KO/KO	1.76	1.24	PGK	98.4%	10/0.1	12.32	0.96	G3PDH	100%	KO/KO	68.82	6.94
PFKA	100%	KO/KO	1.78	1.17	PGM	100%	10/KO	13.63	1.13	ZWF	100%	KO/KO	69.20	6.94
PFKB	100%	KO/KO	1.79	1.17	Ru5P	100%	10/KO	16.62	0.92	RPIB	100%	KO/10	69.93	6.78
Mur	100%	KO/KO	1.80	1.26	GPMB	99.2%	10/10	25.76	2.28	TKa	100%	KO/0.55	70.18	6.68
Synth1	100%	KO/KO	1.86	1.28	GPMA	98.4%	10/0.1	26.37	2.33	TKb	100%	KO/2	70.68	6.65
SerSynth	100%	KO/KO	1.92	0	PYKF	100%	10/KO	28.41	1.66	Mur	100%	KO/KO	73.19	7.05
RPIA	100%	KO/KO	2.18	1.16	GAPA	99.2%	0.1	13.08	0.92	PFKB	100%	KO/KO	73.22	6.37
Ru5P	92.2%	KO/KO	3.65	1.13	PYKA	100%	0.1/KO	13.09	0.92	RPIA	100%	KO/KO	77.60	6.54
PYKF	100%	KO	4.03	1.93	TIS	99.2%	0.1/1.61	13.09	0.92	RPPK	100%	KO/KO	78.02	7.18
G3PDH	100%	KO/KO	4.08	1.94	TALB	99.2%	0.1/1.90	13.09	0.92	G1PAT	100%	KO/KO	94.90	7.53
PFKA	100%	KO/KO	4.14	1.84	G3PDH	100%	0.1/KO	13.44	0.92	PGM	100%	KO/KO	95.12	7.53
ZWF	100%	KO/KO	4.28	1.98	RPIB	99.2%	0.1/1.27	13.11	0.92	GPMB	100%	KO/10	109.47	8.16
PFKB	100%	KO/KO	4.29	1.80	FBAB	99.2%	0.1/2.34	13.11	0.92	GPMA	100%	KO/0.1	110.09	8.18
Mur	100%	KO/KO	4.79	2.09	ENO	99.2%	0.1/0.66	13.12	0.92	Synth1	100%	KO/KO	120.10	8.23
RPPK	100%	KO/KO	4.84	2.09	RPPK	99.2%	0.1/0.97	13.14	0.92	SerSynth	100%	KO/KO	138.41	0
RPIA	100%	KO/KO	5.50	1.84	G1PAT	99.2%	0.1/0.98	13.17	0.92	GAPA	100%	KO/0.1	202.32	3.68
PGM	100%	0.1/KO	5.70	2.23	ZWF	100%	0.1/KO	13.69	0.93	DAHPS	100%	KO/10	220.30	4.66
Synth1	100%	KO/KO	6.17	2.38	TKa	99.2%	0.1/0.49	14.17	0.89	PYKF	100%	KO/KO	244.80	9.80
SerSynth	100%	KO/KO	6.80	0	TKb	93.8%	0.1/2.40	14.23	0.88					
G1PAT	100%	KO/KO	6.94	2.47	Ru5P	100%	0.1/KO	14.62	0.92					
PGM	100%	KO/KO	6.96	2.47	Synth1	100%	0.1/KO	14.72	0.96					
Ru5P	100%	KO/KO	7.14	1.82	Mur	100%	0.1/KO	14.95	0.96					
GPMB	100%	10/KO	9.94	3.62	SerSynth	100%	0.1/KO	15.07	0					
GPMA	100%	0.1/KO	10.09	3.66	PGK	99.2%	0.1/0.1	17.58	0.89					
DAHPS	98.4%	10	9.02	0.97	PGM	100%	0.1/KO	19.81	1.06					
TIS	98.4%	10/4.44	9.03	0.97	GPMA	99.2%	0.1/0.1	31.95	2.13					
G3PDH	100%	10/KO	9.06	0.97	PYKF	100%	0.1/KO	35.52	1.44					
TALB	98.4%	10/5.72	9.11	0.97	DAHPS	99.2%	0.1/10	74.74	0.75					
ZWF	100%	10/KO	9.36	0.99	PPC	100%	KO	66.63	6.88					
ENO	98.4%	10/0.1	9.50	1.01	TALA	100%	KO/KO	66.64	6.88					
RPIB	98.4%	10/8.64	9.63	0.96	PPSA	100%	KO/KO	66.64	6.88					
RPPK	98.4%	10/0.17	9.84	1.00	PTS	100%	KO/2	66.65	6.88					
TKa	98.4%	10/0.39	10.11	0.94	PYKA	100%	KO/KO	66.66	6.88					
TKb	98.4%	10/2.58	10.25	0.94	EDD	100%	KO/KO	66.66	6.88					
Synth1	100%	10/KO	10.28	1.03	TALB	100%	KO/2	66.73	6.87					
Mur	100%	10/KO	10.34	1.02	TIS	100%	KO/2	66.74	6.88					
SerSynth	100%	10/KO	10.55	0	PFKA	100%	KO/KO	67.06	6.80					
RPIA	100%	10/KO	10.72	0.95	FBAA	100%	KO/KO	67.54	6.90					
G1PAT	98.4%	10/0.52	10.98	1.04	FBAB	100%	KO/8	68.42	6.92					

Table 2 (continued). The list of best strategies for improving aromatic amino acid productions

Chapter 3

Systematic bottleneck identification and release for *Saccharomyces cerevisiae* ethanol production

Abstract

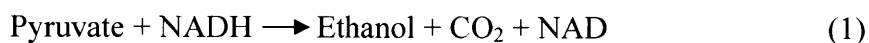
Increased concern for the cost and supply of oil and its negative impact on the environment has led to recent interest in renewable fuel alternatives. Most biofuels share similar precursors in the central carbon metabolic network, and ethanol is the most common renewable fuel today. The study of ethanol production from the central carbon metabolic network is becoming a classic case for the development of techniques to understand and manipulate *Saccharomyces cerevisiae* for optimized biofuel production. Here we present and apply a mass-action model ensemble for the *Saccharomyces cerevisiae* central carbon metabolic network that incorporates the ethanol synthesis pathway as well as the dynamics of NAD and NADH interconversion. This model ensemble, which samples over parameter uncertainties, was used to design and then to analyze strategies that improve ethanol production. We further explored approaches for the identification of pathway bottlenecks. Four computational assays were studied, including metabolite accumulation, conditional V_{\max} , increased input, and decreased enzyme, which were applied to the ethanol model ensemble to study bottleneck identification in this network. The *TDH* reaction

was detected as a major bottleneck restricting carbon flow towards the ethanol pathway and affecting NADH availability. A computational process of greedy sequential single enzyme over- and under-expression optimization was then conducted and several strategies were identified that together improve the ethanol yield of a majority of models beyond 95% of the theoretical yield.

3.1 Introduction

Metabolic engineering has been widely used as an efficient tool to optimize the production of industrial and commercial chemicals. It usually involves directed improvement of cellular properties, using recombinant DNA technologies, through modification of specific biochemical reactions or introduction of new reactions. Many successful applications have been reported to elevate chemical production by re-engineering the genomes of microorganisms, in which popular synthetic chemicals include amino acids, polymers, lipids, and biofuels (Alper & Stephanopoulos, 2009; Atsumi & Liao, 2008; Bailey, 1991; Barkovich & Liao, 2001; Bongaerts et al., 2001; Cameron & Tong, 1993; Cameron & Chaplen, 1997; Keasling, 1999; Li & Vederas, 2009; Stephanopoulos & Sinskey, 1993; Tyo et al., 2007). Because many of the desired metabolic products are terminal or intermediate compounds of central carbon metabolism, composed of glycolysis and the pentose phosphate pathway, it has become one of the most intensively studied biochemical systems. Recent concerns about greenhouse gas emissions, as well as the supply and cost of oil, have led to interest in alternative liquid transportation fuels, an often-touted version of which is the cellular conversion of biomass into ethanol and other alternative fuels produced from similar metabolic intermediates that are found in the glycolysis pathway (Alper & Stephanopoulos, 2009). Regardless of which molecule or mix of molecules becomes the dominant biomass-based fuel of the future, microbial conversion is at present the main avenue for alternative fuel production. Ethanol production through the central carbon metabolic network has become a useful case study to develop computational techniques to help improve the fermentation production of renewable biofuels (Bro et al., 2006; Lee et al., 2008; Matsushika et al., 2009).

The most commonly used microbe for ethanol production has been yeast; *Saccharomyces cerevisiae* which can produce ethanol to concentrations as high as 18% of the fermentation broth, is the preferred species for most ethanol fermentation (Lin & Tanaka, 2006). This yeast can grow both on simple sugars, like glucose, and on the disaccharide sucrose. As with many microorganisms, *Saccharomyces cerevisiae* metabolizes glucose by the Embden-Meyerhof (EM) pathway (Lin & Tanaka, 2006). Under aerobic conditions, the pyruvate formed in the final step of glycolysis is oxidized to acetyl-CoA, which enters the citric acid cycle and is oxidized to CO₂ and H₂O. Under anaerobic conditions, there is no O₂ to accept the electrons of NADH and thus to reoxidize it to NAD, which threatens the function of the glycolysis pathway. *Saccharomyces cerevisiae* evolved to continually regenerate NAD during anaerobic glycolysis by transferring electrons from NADH to ethanol, a reduced end product (Nelson & Cox, 2008; Pronk et al., 1996). Pyruvate is converted to ethanol and CO₂ in a two-step process: pyruvate is decarboxylated in an irreversible reaction catalyzed by pyruvate decarboxylase; the acetaldehyde generated is then reduced to ethanol through the action of alcohol dehydrogenase with the reducing power furnished by NADH. These two steps can be combined into an overall reaction:



As can be seen in this pathway, the NAD/NADH balance plays a crucial role in regulating the flux distribution in the network. NADH and the related coenzyme NADPH serve different functions in *Saccharomyces cerevisiae* metabolism. NADPH is mostly used as a reductant in biosynthetic reactions; whereas the NADH/NAD ratio mainly determines the intracellular redox potential (Bakker et al., 2001; Vemuri et al., 2007). The total amount of NADH and NAD can be considered as a conserved quantity (Bakker et al., 2001). Reduction of NAD must be matched by continuous reoxidation of NADH. There are several major ‘sources and sinks’ of NADH in the

central carbon metabolic network (Bakker et al., 2001; Hou & Vemuri, 2010; Vemuri et al., 2007) as shown in Figure 1. NADH is mainly produced by the *gapA* reaction in the glycolysis pathway, where it participates in the conversion of glyceraldehyde 3-phosphate to 1,3-bisphosphoglycerate. Under aerobic conditions, additional NADH is produced through the citric acid cycle. The major pathway to reoxidize NADH to NAD when oxygen is available is the respiration pathway in mitochondria. In anaerobic conditions oxygen is not present to accept electrons from NADH. Instead, ethanol is synthesized from pyruvate to reoxidize the NADH. In addition, the glycerol synthesis pathway, which uses dihydroxyacetone phosphate in the glycolysis pathway as a reactant, is also used to reoxidize the NADH that cannot be fully consumed by ethanol synthesis in order to maintain the NAD/NADH balance. The dynamics of the NAD/NADH interconversion is controlled by this complex network under aerobic and anaerobic conditions. It determines the relative fluxes among different pathways, and thus is crucial to understand for the purpose of optimizing ethanol production in *Saccharomyces cerevisiae*.

Due to the importance of ethanol in industrial use, much effort has gone into maximizing its production. Improvements include pretreatment of feedstocks and introducing enzymes to digest xylose (Hahn-Hägerdal et al., 2006; Lee et al., 2008; Lynd, 1996; Matsushika et al., 2009; Nevoigt, 2008; Sánchez & Cardona, 2008). Less attention has been focused on regulation of the overall central carbon metabolic network and on understanding bottlenecks in this network. Because of the complexity of the NAD/NADH dynamics in this network, computational and mathematical models may be especially useful in understanding and optimizing ethanol production.

Recent research has led to computational models in order to efficiently analyze and optimize metabolic networks. Two of the major model types that have been developed and applied are

flux balance analysis (FBA) and ordinary differential equation (ODE) models (Burgard et al., 2003; Chassagnole et al., 2002; Edwards & Palsson, 2000; Pramanik & Keasling, 1998; Schmid et al., 2004; Usuda et al., 2010; Vital-Lopez et al., 2006). FBA models require relatively modest information regarding biological mechanism, including a list of chemical reactions with their stoichiometry, flux constraints, and specification of feeds and metabolic demands (Kauffman et al., 2003; Stephanopoulos et al., 1998; Varma & Palsson, 1994), most of which can be readily acquired from existing literature and databases. Therefore, FBA models have the advantage of feasibility and simplicity when modeling large-scale (e.g., whole genome) networks. In contrast to the steady-state nature of FBA models, ODE models, including aggregated rate law (ARL) and mass-action rate law (MRL) forms, incorporate network dynamics and attempt to represent detailed enzyme behavior (Chassagnole et al., 2002; Lee et al., 2006; Liao et al., 1996; Tzafiriri, 2003). If unconstrained, the space of steady states from both FBA and ODE models are the same, but ODE models can readily map parameter constraints into the kinetically feasible regions of the solution space, whereas this information is not easily transferable to FBA models (Machado et al., 2012).

A mass-action rate law model for the *E. coli* central carbon metabolic network has been reported in Chapter 2, which includes glycolysis and the pentose phosphate pathway. As the ethanol synthesis pathway is an extension at the end of the glycolysis pathway, the model in Chapter 2 serves as a convenient initial model for this study. To convert the *E. coli* model into a *Saccharomyces cerevisiae* model, the topology of the model in Chapter 2 was updated based on the KEGG database (Kanehisa & Goto, 2000). To keep the conversion simple and focused on the important features of NAD and NADH dynamics, we kept the same number of isomers for the *S. cerevisiae* model as for the *E. coli* model. For reactions that have isomers in *S. cerevisiae* but not

in *E. coli*, the general name of the *S. cerevisiae* gene was used (e.g., *PGM* was used for *S. cerevisiae* instead of *PGM1*, *PGM2*, and *PGM3*). For reactions have isomers in *E. coli* but not *S. cerevisiae*, the weaker isomer reaction in *E. coli* was removed. The gene names were updated to *S. cerevisiae* names based on the KGEE database. The parameters were re-fitted against the steady-state flux data of the key branching reactions in the glycolysis pathway, pentose phosphate pathway, and TCA cycle measured for *S. cerevisiae* (Jouhten et al., 2008). More importantly, given NADH as a reactant for the ethanol synthesis reaction, the NAD and NADH concentrations were added as explicit state variables into the model in order to analyze their impact on ethanol production. The steady-state concentrations for NAD and NADH under aerobic conditions (1.47mM and 0.1mM, respectively) were borrowed from their *E. coli* measured data (Chassagnole et al., 2002) as no *S. cerevisiae* measurements reported; however, the total amount of NAD and NADH (around 1mM) reported for *S. cerevisiae* (Bakker et al., 2001) is similar to the total amount used in our models. The ethanol synthesis pathway was also added to the model, with pyruvate and NADH as the two reactants. In the refitting, a collection of models, instead of a single model, was produced, all of which have similarly good fits to the flux and concentration data mentioned above. This ensemble provides an estimate of uncertainty in the parameters fit and in the predictions made.

Although much effort in the field of metabolic engineering has gone into identifying efficient strategies to improve the production rate of desired chemicals, there is less research on how to directly and systematically discover the bottlenecks in the system -- that is, to identify the slow steps that would presumably be high-priority targets for rational genetic engineering. In order to increase productivity and metabolite yield, researchers have focused on enzyme amplification or other modifications of the pathway that produce increased yield (Stephanopoulos & Vallino,

1991). Retrospectively, an enzyme reaction can then be called a bottleneck because its overexpression has improved productivity (Dai et al., 2002; Lütke-Eversloh & Stephanopoulos, 2008). Here we take a complementary approach of independently studying bottlenecks through a variety of approaches and then examining the effect of releasing them through overexpression in simulation.

Alternative approaches such as metabolic control analysis (MCA) (Heinrich & Rapoport, 1974; Kacser & Burns, 1973) have been developed; however, their value in directing metabolic engineering efforts remains uncertain. MCA, for example, is only valid in the local neighborhood of the operating point (e.g., steady state) evaluated (Stephanopoulos & Vallino, 1991). This can limit its applicability. Dynamic sensitivity analysis was developed and applied to primary metabolism, in which the relative change of target metabolite concentration caused by an infinitesimal percentage change in enzyme activity is calculated for each enzyme and used to predict bottlenecks (Shiraishi & Suzuki, 2009). One drawback of this method is that the sensitivity is calculated through differential equations which evaluate the production improvement effect when changing the enzyme activity for only an infinitesimal percentage, which is not realistic in experiments and may not hold valid when changing the enzyme activity for a finite level. While useful, this approach provides only one perspective on bottlenecks (discussed below). Network rigidity and principal nodes theory, developed by Stephanopoulos and Vallino (1991), can identify nodes in the network that have inherent resistance to flux partitioning alterations. The relative flux going down a certain branch of those nodes may not be changed by simply modifying the corresponding enzyme activities. Those nodes, therefore, should be given more attention when designing engineering strategies to improve the production

rate of metabolic products on the branching pathways and can be considered as a different type of “bottlenecks” for this concern.

Here we present a framework to systematically identify bottlenecks in the system and study their relevance for production of a desired output. Four computational tests involved in the framework are illustrated in Figure 2. They are metabolite accumulation, conditional V_{\max} , increased glucose input, and decreased E_0 . Based on the model ensemble we built for *Saccharomyces cerevisiae*, we were able to identify bottlenecks in the central carbon metabolic network through their framework. Analysis of these bottlenecks has led to insights about the production of ethanol by the network.

Here we describe the four tests used for bottleneck identification. Detecting metabolite accumulation is a method that is often used by experimentalists to identify network bottlenecks. Significant (and possibly increasing) accumulation of certain metabolites when others appear to have reached a constant low level can indicate a flux imbalance adjacent to the accumulated metabolite. That is, a slower consuming flux cannot keep up with a faster generating flux. The slow consuming flux usually indicates pathway bottleneck. Several studies have achieved improved production rates of target chemicals by releasing these bottlenecks (Martin et al., 2003; Simonsen et al., 2012). Adopting the same logic, our metabolite accumulation test works by computationally detecting metabolite accumulations in the network. We simulate the model ensemble of the central carbon metabolic network until most of the metabolites in the system have reached a constant level, consistent with the experimental fermentation time (Lin & Tanaka, 2006). Upon further simulation, some metabolites retain their constant concentrations, but others increase significantly. The first time point defines a pseudo-steady state of the system and serves as a reference state, and the later one is used to detect metabolite accumulation. The flux

immediately downstream of these accumulated metabolites is bottleneck candidate to be examined further with the other three tests. Interestingly, rather than identifying a single rate-determining step, here this approach found multiple slow steps simultaneously.

Based on our calculations (see Results and Discussion), many of the slow fluxes identified through metabolite accumulation operate below their maximum V_{\max} . These fluxes don't use the extra capacity because they tend to be reactions with multiple (two) substrates with one of the substrates (usually the non-accumulated of two) limiting. Increasing the concentration of the limiting substrate (a metabolite) generally could involve a shift to a new steady state which leads to a faster flux. We thus introduce the concept of a conditional V_{\max} of an enzyme with respect to a substrate to describe a system property of the enzyme where the actual flux capacity is limited by the given steady state (or pseudo-steady state) of the substrates (occasionally also products) and is smaller than the overall V_{\max} . The overall V_{\max} is a local property of the enzymes determined by the enzyme kinetic parameters (e.g., k_{cat}) and the enzyme availability (e.g., total enzyme concentration, E_0). It does not depend on the system or network the enzyme locates and describes a fixed upper limit for enzyme flux capacity. In contrast, the conditional V_{\max} describes a realizable enzyme flux capacity at a given (pseudo-) steady state. For example, if one substrate (usually a co-factor) of a two-substrate enzymatic reaction remains at a significant low steady-state concentration determined by the system, no matter how much the other substrate concentration is increased, the enzyme cannot run at its V_{\max} . We call the enzyme is running at its conditional V_{\max} respecting to the first substrate and the first substrate is the limiting factor. Under situation like this, the overall V_{\max} is useless, but it is the system property, the conditional V_{\max} which depends on the current network concentration states, that provides valuable

information of the available capacity of an enzymatic reaction. For single substrate reactions, the conditional V_{\max} , when holding product level at zero, is the same as the overall V_{\max} .

The metabolite accumulation test and the conditional V_{\max} test are methods to identify bottleneck candidates. They make no reference to the desired output compound however and so can suggest bottlenecks not relevant to production. The glucose input test and decreased E_0 test are additional tests that measure the relevance of candidates to production of the desired output. As the ethanol production pathway is located near the end of glycolysis, whether the carbon resources provided by glucose can reach this pathway efficiently is important for optimizing ethanol yield. The glucose input test increases the glucose input flux by 2 to 1000 fold. If there is no bottleneck for carbon flux in the network, all fluxes should be elevated with increased glucose input. On the other hand, the bottleneck flux (and those downstream) will not increase when increasing glucose input, as maximum capacity has been reached. By observing flux changes for each enzyme in central carbon metabolism for increased glucose input, we were able to identify the enzymes that constrain the carbon flux.

The fourth test, decreased E_0 , is based on the assertion that if an enzyme reaction is a bottleneck, it is already running at capacity; reducing the capacity by decreasing E_0 should further reduce downstream fluxes to the output, but not for non-bottlenecks. It is tempting to propose an increased E_0 test, as this corresponds to experimentally implemented enzyme overexpression. Similar conclusions cannot as clearly be drawn for an increased E_0 test, however, because the flux might still be limited by something else (for instance, a second bottleneck); release of either alone would not increase production, but both together would.

Based on this four-test framework, we were able to identify bottleneck candidates in the central carbon metabolic network and validate their relevance for ethanol production. The *TDH* reaction, which is several steps upstream of the ethanol synthesis reaction, is identified as an important bottleneck that regulates the carbon flow towards ethanol production. The NAD to NADH ratio emerges as a crucial regulator for the *TDH* reaction and ethanol production. It controls the balance and relative abundance of the carbon precursor, pyruvate, and the NADH for ethanol synthesis. As the NAD and NADH molecules are involved in several enzyme reactions, the fluxes of which are interdependent due to the network topology, it is not obvious how to manipulate the enzymes in the network for optimum balance between NAD and NADH in order to maximize ethanol yield. Here we use a single-enzyme, greedy sequential optimization method to identify efficient strategies to enhance ethanol production. In each round, single-enzyme over- or under-expression optimization was conducted, using similar procedures to those reported in Chapter 2, for each enzyme in the network and for each model in the ensemble. Compared to the exhaustive multiple-enzyme optimization reported in Chapter 2, the single-enzyme greedy optimization is computationally more efficient (scales as the number of enzymes in the network as opposed to combinatorially). Two rounds of sequential optimization are computed to lead to raising ethanol yields from 75-85% to over 95% of the theoretical maximum for most of the models in the ensemble. Further studies show that the strategies identified through this method helped release the bottlenecks identified by the four-test framework. The model ensemble technique and the bottleneck identification and release methods reported here can be readily applied to extended pathways of central carbon metabolism and other networks to help improve the production of chemicals of academic or commercial interest.

3.2 Method

3.2.1 Modeling ethanol production in yeast central carbon metabolic network

A mass-action rate law model for *E. coli* central carbon metabolic network, which includes glycolysis pathway, pentose phosphate pathway, and Entner-Doudoroff pathway, has been reported in Chapter 2. As the central carbon metabolic network is highly conservative among many microorganisms (Nelson & Cox, 2008), this model provides a good reference for generating a comprehensive mass-action rate law model for *Saccharomyces cerevisiae*. In order to update the model topologies from *E. coli* network to *Saccharomyces cerevisiae* network, the KEGG database (Kanehisa & Goto, 2000) was used to select model additions and deletions. The glucose uptake in *Saccharomyces cerevisiae* involves glucose membrane transporters (*HXT*) and hexose phosphorylation enzymes (*HXK1*, *HXK2*, and *GLK*) (Barnett, 2008; Boles & Hollenberg, 1997; Fernandez, Herrero, 1985; Gancedo, 2008; Leandro, Fonseca, & Gonçalves, 2009; Ozcan & Johnston, 1999; Rintala, Wiebe, Tamminen, Ruohonen, & Penttilä, 2008; Rolland, Winderickx, & Thevelein, 2002), the mechanisms of which are different from the glucose uptake of *E. coli*. The *pts* reaction (*E. coli* glucose uptake and phosphorylation reaction) in the *E. coli* model was replaced by an artificial *GT* reaction for glucose transporter and *HXK* reaction for glucose phosphorylation to simulate the glucose uptake for *Saccharomyces cerevisiae*. The chemostat experimental setting in the *E. coli* model was replaced by a batch experimental setting, by removing the constant in-flux and out-flux of glucose for the system and replacing with a fixed initial extracellular glucose supply of 36 g/L. The elementary parameters for reactions *GT* and *HXK* were estimated based on the literature K_m and k_{cat} value (Fernandez, Herrero, 1985; Gao & Leary, 2003), and the enzyme concentrations of them were adjusted so that the system

remains a similar glucose uptake rate as the model in Chapter 2. The Entner-Doudoroff pathway (*edd* and *eda*) was removed from the model, as it does not exist in *Saccharomyces cerevisiae* (Blank, Lehmbeck, & Sauer, 2005). The anaplerotic reaction *pepc* in the *E. coli* model was replaced by *PYRD* reaction for the *Saccharomyces cerevisiae* model, as the pyruvate carboxylase, instead of the phosphoenolpyruvate carboxylase, is the major enzyme to replenish the citric acid cycle for oxaloacetate which is consumed during biosynthesis of amino acids. The *PYRD* reaction was implemented to be sensitive to the concentration of acetyl-CoA to simulate the natural behavior of this enzyme (Pronk et al., 1996). The ethanol production pathway was added as a new branch taking away pyruvate and converting it into ethanol (Pronk et al., 1996). The detailed reaction mechanisms can be found in the Appendix B.

The critical modifications of the Cui *et al.* model were the introduction of NAD/NADH and NADP/NADPH balance into the model as well as the capability to model the transition from aerobic to anaerobic conditions. The mechanisms of the major reactions responsible for the NAD/NADH and NADP/NADPH balance were updated to include NAD, NADH, NADP, and NADPH as reactants or products. In particular, the *TDH* enzyme reaction was updated from a uni-uni reaction to a bi-bi reaction to include NAD as reactant and NADH as product. A simplified three-reaction citric acid cycle (TCA cycle), which only takes into account of the three steps involving NADH generation, was also added into the model to simulate the NADH generation from NAD. The glycerol synthesis reaction and the ethanol synthesis reaction were updated to include NADH as reactant and NAD as product, as the alcohol generation steps consume NADH for redox balance. A simplified oxidative phosphorylation reaction was added to simulate the conversion of NADH to NAD by oxygen under aerobic condition. Oxygen, which was added as a tractable variable in the model, starts at a constant level and linearly decreases

while consumed by the oxidative phosphorylation reaction. Oxygen concentration hitting a zero marks the system converts from aerobic condition to anaerobic condition. The *ZWF1* and *GND* reactions were updated to include NADP as reactant and NADPH as product; a first-order reaction was added to simulate the consumption of NADPH to NADP by biosynthesis reactions in yeast. The detailed reaction mechanisms and reference used for the reactions mentioned above are listed in Table 1. After the modifications, the new yeast model consists of 46 enzyme reactions, 235 species, and 423 free kinetic parameters. Figure 3 depicts the overall topology of the model.

The carbon flux distribution measurement data of *Saccharomyces cerevisiae* CEN.PK1 13-1A in different oxygenation conditions reported by Jouhten et al. (Jouhten et al., 2008) were used to train the wild-type model behavior. In particular, the fluxes of the crucial branching reactions, *ZWF1*, *GND*, glycerol synthesis, *TDH*, ethanol synthesis, *FBA1*, *PYK*, *PYRD*, *PGII*, *PDB1*, and the first TCA reaction, in both aerobic and anaerobic conditions were included in the objective function for the parameter fitting. In addition, the concentrations of NAD and NADH under aerobic condition were also included in the objective function of the fitting, so that the system has a reasonable NAD/NADH balance at the initial stage of the experiment. The overall objective function for the fitting is given as below:

$$G = \sum_{r_{data,aerobic}} \left(\frac{r_{pred,aerobic} - r_{data,aerobic}}{\sigma_{data,aerobic}} \right)^2 + \sum_{r_{data,anaerobic}} \left(\frac{r_{pred,anaerobic} - r_{data,anaerobic}}{\sigma_{data,anaerobic}} \right)^2 + \sum_{x_{data,aerobic}} \left(\frac{x_{pred,aerobic} - x_{data,aerobic}}{\sigma_{data,aerobic}} \right)^2$$

The first and second sums are for the reaction rate fitting for the 11 branching reaction fluxes under aerobic and anaerobic conditions. The third sum is for the NAD and NADH concentration

fitting under aerobic conditions. The NAD and NADH concentrations under aerobic conditions were taken from Chassagnole et al. (2002). Six parameter sets from the frontier model sets in the Chapter 2 model that represent the different pentose phosphate pathway behaviors were selected as the initial parameters for the fitting. For each parameter set, four additional initial parameter sets were generated by adding 10% random noise to each parameters. An E_0 of 0.0176mM was used for all enzymes except for the glucose transporters (Fraenkel, 2003). To take into account of the uncertainties of average enzyme concentrations in yeast, each parameter set was also paired with one of the five E_0 , 0.001mM, 0.005mM, 0.01mM, 0.025mM, 0.05mM, for the parameter fitting. In total, 180 initial parameter sets were used for the optimization which was done using the `fmincon` function in MATLAB (version 2008a; The MathWorks, Inc.; Natick, MA). The boundaries of parameters were partially estimated based on K_m and k_{cat} reported in BRENDA database (Scheer et al., 2010). 52 parameter sets were collected with good fits. These parameter sets formed a model ensemble with 52 sub-models which share the same topology but have different parameter values.

3.2.2 Calculation of metabolite accumulation

Each model was simulated to $t_1 = 10^5$ s (27.8 hour) and to $t_2 = 1.5 \times 10^5$ s (41.7 hour). The t_1 time point was selected as a long enough time to allow the system reaches a pseudo-steady state. It is consistent with the range of experimentally used measurement time for batch reactors (Cheng & Hasan, 2009; Dombek & Ingram, 1987; Çaylak & Sukan, 1998). The concentrations of all metabolites were collected at both time points. A comparison between the concentrations at the first time point and those at the second time point showed that most of the intermediate metabolites but eight stay at the same level. We then defined the first time point as a “pseudo-steady state”, which refers to a state that all but some “special behaving” metabolites have

reached stable concentrations. The “special behaving” metabolites did not reach stable concentrations when allowed to simulate up to 2×10^5 s (55.6 hour). The ratios of the metabolite concentrations at t_2 over t_1 were reported as indicators of abnormal metabolite accumulations in the system.

3.2.3 Calculation of conditional V_{\max}

To calculate conditional V_{\max} , each model in the ensemble was first simulated to acquire pseudo-steady state concentrations for all species at the given measurement time point $t = 27.8$ hour. For multiple reactants enzyme reaction, the conditional V_{\max} for a particular reactant was calculated as the maximum flux this reaction can achieve when changing the concentration of this reactant from zero to infinity (10^{10} mM was used for calculation) while fixing the concentrations of all other reactants and products at the pseudo-steady state values simulated as described above. The fluxes for each concentration state were calculated based on the steady-state rate law derived using the King–Altman method from the elementary reactions (Cleland, 1963; Cornish-Bowden, 1977; King & Altman, 1956; Kuzmic, 2008). For single reactant enzyme reactions, the conditional V_{\max} was calculated with fixed product pseudo-steady state concentrations while changing the concentration for reactant from zero to infinity (10^{10} mM was used for calculation). The conditional V_{\max} was calculated for every reactant of each enzyme reaction. The actual enzyme flux for each reaction was also calculated based on the pseudo-steady state concentrations of all species. The ratio between the actual flux and the conditional V_{\max} is an indicator of the usage of flux capacity and is reported in Results and Discussion session.

3.2.4 Measurement of the effect of increasing glucose input

The E_0 of *GT* and *HXK* enzymes were increased for 2, 5, 10, 25, 50, 100, and 1000-fold to test increasing the glucose input for the corresponding fold for each model. The extracellular glucose was kept excessive for all the tests. Under each glucose input level, the models were simulated to the pseudo-steady state ($t = t_1$) and the enzyme fluxes were calculated as the time derivative of the product concentration of each reaction at the pseudo-steady state. The pseudo-steady state flux changes for each enzyme of each model when increasing the glucose input level were reported as a measurement of effect of increasing glucose input.

3.2.5 Measurement of the effect of decreasing enzyme E_0

The enzymes detected as bottleneck candidates based on the metabolite accumulation test and the conditional V_{\max} test were measured for their effect on ethanol production rate when decreasing their corresponding E_0 . The E_0 of glycerol synthesis enzyme, *TDH*, *TAL1*, *NQMI*, and the first enzyme in the TCA cycle were decreased to 80%, 50%, 10%, and 1%, respectively. The corresponding pseudo-steady state flux changes for ethanol production were reported as indicators of the relevance of the particular bottleneck candidate to the production rate of interest.

3.2.6 Sequential bottleneck release

A sequential single enzyme optimization framework was applied to the model ensemble to identify strategies leading to enhanced ethanol production. The objective function was the ethanol production rate from the reaction catalyzed by the enzyme E_{ole} from the substrates pyruvate, evaluated at the pseudo-steady state. A single enzyme over- and under-expression (termed “expression change” here) spanning a range from 50 times to 1/50 the unperturbed concentration was tested for each of the enzymes in the system for the model ensemble. The

enzyme expression change strategy that results in the most increase of the ethanol yield compared to that of the wild-type model was selected as the best strategy for that model at round 1. The best strategies for each model in the ensemble were applied to the wild-type models, which generates the new base models. A second round optimization of single enzyme expression change was then conducted on the base models to identify strategies that lead to the best increase of ethanol yield. This process can be repeated several times until a preferred yield is achieved or no further yield improvement can be acquired. A corresponding sequence of single enzyme over- and under-expression for each model in the ensemble can be generated as the roadmap to enhance ethanol production. The optimization was done using the `fmincon` function in MATLAB (version 2008a; The MathWorks, Inc.; Natick, MA).

3.3 Results and Discussion

The central carbon metabolism model for *E. coli* built in Chapter 2 was converted to a model ensemble for *Saccharomyces cerevisiae* with updated network topology. The steady-state concentrations of NAD and NADH under aerobic condition were borrowed from those in *E. coli* (Chassagnole et al., 2002) given the total amount of them is similar in *E. coli* (1.57mM) and in *S. cerevisiae* (around 1mM) described in Introduction. The model ensemble was reparameterized using steady-state data of major carbon fluxes in *S. cerevisiae* through the glycolysis pathway, pentose phosphate pathway, and the TCA cycle under both aerobic and anaerobic conditions (Jouhten et al., 2008). The resulting models fit the training data very well, with 82.1% of the calculated fluxes within 10% of the measured values and a maximum deviation of 24.6%.

3.3.1 Ethanol production bottleneck detection via the four-test framework

Metabolite accumulation test and conditional V_{max} test identify bottleneck candidates for further investigation

Each of the 52 models in the ensemble was examined for metabolite accumulation. Each model was simulated and examined at $t_1 = 27.8$ h and $t_2 = 41.7$ h as described in Methods session. Nearly all metabolites had reach low concentrations that remained constant over time at t_1 , comparison between t_2 and t_1 provided a convenient mechanism to identify those that accumulated. The log ratios of the metabolite concentrations collected at t_2 to that at t_1 are shown as a heat map in Figure 4A, where the x -axis corresponds to different metabolites in the model and y -axis corresponds to different models in the ensemble. The rows are ordered by the model's ethanol yield, with the highest yield model at the top. The panel on the right of the heat map

shows the corresponding yield for each model. Red indicates a significant concentration increase; blue represents a significant decrease. Intermediate colors indicate intermediate changes.

The majority of the 26 metabolites consistently show consistent green color across all or nearly all models, which indicates small or no concentration change. The consistency across models indicates insensitivity to parameter uncertainties between models. One can consider the systems as having reached a pseudo-steady state by t_1 , as most of the metabolites have reached a stable level. Interestingly, several metabolites increase concentration (yellow to red) – namely, F1,6P, G3P, E4P, DHAP, and acetyl-CoA. Moreover, those increases are consistent across most of the models, which again indicates a relative insensitivity to parameters. Figure 4B shows a box plot that summarizes the distributions across models from the heat map for each metabolite. The five accumulated metabolites have the largest variation across models partially because the models vary in the level of accumulation predicted but also because some models predict no accumulation.

The carbon flow through glycolysis starts at glucose and progresses towards pyruvate. The metabolites F1,6P, DHAP, and G3P are located immediately before the enzyme reaction catalyzed by the product of the *TDH* gene. A concentration accumulation at this location indicates an imbalance of fluxes going into and out of these metabolites. In particular, it suggests the *TDH* and glycerol synthesis fluxes are slower than the *PFK* and *FBA1* fluxes and that the *FBA1* fluxes are slower than the *PFK* ones (also confirmed by other results, see below), which identifies *TDH* and the glycerol synthesis reactions as potential bottlenecks. Similarly, the accumulation of acetyl-CoA suggests the first reaction in the TCA cycle, *TCA1*, could be a potential bottleneck of the system. It is more complicated for the accumulation of E4P, as reported in Chapter 2 that the flux in pentose phosphate pathway can flow in either the clockwise

or counter-clockwise direction. Both the *TAL* and *TKL* reactions could be potential bottlenecks depending on the flux direction.

To complement the metabolite accumulation simulations, the conditional V_{\max} for each enzyme reaction was calculated for each model in the ensemble. As conditional V_{\max} measures the maximum flux under current condition, the log ratios of actual fluxes to their condition V_{\max} are given in Figure 5. Red indicates fluxes that are similar to the conditional V_{\max} and thus close to capacity; the deepest blue indicates fluxes than 1% of their conditional V_{\max} and thus far below capacity. Of the five accumulated metabolites identified above (E4P, acetyl-CoA, G3P, DHAP, and F1,6P), there is a corresponding enzyme running at its conditional V_{\max} just downstream (*TAL*, *TCA*, *TDH*, and *GLCE*, respectively). This is as expected as described in Figure 2B: bottleneck candidates predicted by the metabolite accumulation test should also be selected by the conditional V_{\max} test, because metabolite accumulation indicates the consuming flux cannot keep with the generating flux, which suggests the consuming flux is already running at capacity. Interestingly, several new enzyme fluxes (*ZWF1*, *GND*, and the ethanol synthesis reaction) are also identified as running at or close to the conditional V_{\max} , but were not selected by the metabolite accumulation test. Further examination shows that, compared to the candidates selected by both tests, there is no flux imbalance around *ZWF1* and *GND*, (Figure 6), consistent with the lack of metabolite accumulation. A more detailed study (Figure 5) shows that the limiting metabolites for the conditional V_{\max} for *ZWF1* and *GND* are carbon precursors, whereas those for the *TDH*, glycerol synthesis, and *TCA* reactions are the co-factors NAD or NADH. Increasing glucose input (carbon input) increases the carbon precursor levels and thus increases the conditional V_{\max} for *ZWF1* and *GND*. By contrast, the conditional V_{\max} for reactions with NAD or NADH as the limiting cannot be released by simply increasing glucose input level.

These reactions (*TDH*, glycerol synthesis, and TCA reaction) are important bottleneck candidates for further examination.

Glucose input test detected TDH reaction as major bottleneck for carbon flow

Ethanol is produced by converting glucose through glycolysis. For the glucose input test, increasing amounts of glucose from 2 to 1000-fold were input and the resulting flux changes were observed. Figure 7 shows a small a heat map of relative fluxes adjacent to each reaction (color from white to red indicates a flux increase from negligible to large). Two patterns are clear. Fluxes near the top of the network increase with increased input; those near the bottom remain unchanged. In particular, the ethanol production rate does not increase with enhanced glucose input. This phenomenon is consistent across the model ensemble and is thus not sensitive to the parameter uncertainties. The switch of the patterns occurs on the glycolysis pathway at the *TDH* reaction. Thus, *TDH* is identified as a bottleneck. The flux of glycerol synthesis enzyme also does not increase with increased glucose level and appears to be another bottleneck for carbon flow. Therefore, the glucose input test further confirms that the *TDH* and glycerol synthesis reactions suggested by the metabolite accumulation and conditional V_{\max} tests are relevant bottlenecks for ethanol production. Additional bottlenecks could exist downstream of the *TDH* reaction, the effect of which could have been hidden by the effect of *TDH* reaction.

Decreased E_0 test validates TDH, glycerol synthesis reaction, and tca1 as bottlenecks for ethanol production

Sequentially observing the effect of reduced enzyme concentrations (the decreased E_0 test) was conducted to examine further the relevance of bottleneck candidates identified previously. In particular, the E_0 for *TDH*, glycerol synthesis, *TCA1*, *TAL1*, and *NQM1* were each decreased to

80%, 50%, 10%, and 1% of their base concentration. The corresponding log ratios of the ethanol production rates relative to those with the original E_0 are plotted in Figure 8. Decreasing the E_0 (and thus the V_{\max}) of *TDH*, glycerol synthesis, and *TCAI* causes the corresponding ethanol production rates to decrease correspondingly. This matches the intuition that decreasing the flux capacity of a limiting reaction would further reduce ethanol production. By contrast, Figure 8 shows that changes of E_0 for *TALI* and *NQMI* do not affect ethanol production. It indicates that although these enzymes may be bottlenecks for the network predicted by the metabolite accumulation and conditional V_{\max} tests, they are not located on the pathways affecting ethanol production and thus are not directly relevant to the improvement of ethanol production.

Interestingly, increasing the E_0 of *TDH*, glycerol synthesis, and *TCAI* do not provide much benefit for improving ethanol production (data not shown), which suggests that a network solution for maximizing the ethanol yield is not obvious even if the bottleneck enzymes have already been identified. More sophisticated techniques are needed to manipulate the network.

3.3.2 NAD and NADH balance plays a crucial role in regulating ethanol production

1 mole of ethanol is produced by consuming 1 mole of pyruvate and 1 mole of NADH. The total amount of NAD plus NADH is relatively stable and has been reported around 1-1.57mM in cells (Chassagnole et al., 2002; Richard, Teusink, Westerhoff, & van Dam, 1993; de Koning & van Dam, 1992). The total amount is implemented as a constant of 1.57mM in the model based on the Chassagnole et al. model. Consequently, reduction of NAD has to be matched by a continuous reoxidation of NADH. The ratio of NADH over NAD is crucial for ethanol production and higher NADH/NAD ratio favors the ethanol generation. Interestingly, the bottleneck tests described above have identified the *TDH* and glycerol synthesis reactions, two

enzyme reactions much earlier than the ethanol synthesis reaction, as the major bottlenecks that affecting ethanol production rate in the yeast central carbon metabolic network. When we examine the ethanol production reaction alone, there is a significant shift of nicotinamide towards the NADH format after switching from aerobic to anaerobic condition and most of the models have NADH as the dominant form (Figure 9). This major increase of NADH level matches well with the significant increase of ethanol production in the anaerobic condition as shown in Figure 9. Under aerobic condition, the NADH formed by *TDH* reaction is ultimately reoxidized to NAD by passage of its electrons to O₂ in mitochondrial respiration. However, under anaerobic condition, NADH generated by glycolysis cannot be reoxidized by O₂. Failure to regenerate NAD would leave the cell with no electron acceptor and thus stop the glycolysis pathway (Nelson & Cox, 2008). The pyruvate metabolism is thus switched from going towards the TCA cycle which also requires NAD as reactant to going towards the ethanol synthesis pathway which regenerates NAD by accepting the electrons. It is impressive that the models can automatically capture this oxygen condition change and switch the pyruvate consumption pathway accordingly. As mentioned in the conditional V_{max} test, some models have the carbon precursor, pyruvate, instead of the NADH as the limiting precursor for the ethanol production reaction. How much pyruvate the downstream pathway can acquire, as pointed out by the glucose input test, depends on the flux capacity of the *TDH* reaction. The conditional V_{max} test for the *TDH* reaction indicates that the limiting factor for the capacity of this reaction is the NAD level in the system. Based on these analyses, the NAD and NADH balance is crucial for determining the ethanol production rate in the system. A higher NAD to NADH ratio could release the limitation of the *TDH* reaction and thus increase the carbon converted to pyruvate; on the other hand, a lower NAD to NADH ratio would favor the ethanol synthesis reaction goes

towards the ethanol production. The NADH generated by *TDH* and TCA reactions are consumed by ethanol synthesis and glycerol synthesis reactions. The faster the ethanol synthesis reaction runs, the more NAD gets recycled, and thus the faster the *TDH* reaction runs, until the NADH has been consumed so much that it starts to become the limiting precursor instead of pyruvate. At the anaerobic pseudo-steady state observed in the model ensemble, there is a shift towards the NADH format, which reduces the capacity of the *TDH* reaction and limits the pyruvate available for ethanol production. For ethanol production, the availability of NADH and pyruvate is interconnected. A careful network flux control is needed to balance the precursors and maximize the productivity.

The glycerol synthesis reaction also plays an important role in ethanol production. As the generation of glycerol from DHAP converts NADH to NAD, it helps release the *TDH* bottleneck by providing more NAD. However, as indicated by the conditional V_{\max} test, the glycerol synthesis reaction may be already running at the overall V_{\max} . This means the glycerol synthesis reaction already recycles NAD at its maximum capacity, and it cannot further help release the *TDH* bottleneck. On the other hand, glycerol is usually considered as an unwanted side-product for ethanol production. It has been estimated that elimination of glycerol production in industrial yeast fermentations aimed at the production of alcohol might increase the annual worldwide production of ethanol by 1.25 billion liters (Nissen et al., 2000). Therefore limiting the glycerol reaction rate is usually desired in industry productions (Bakker et al., 2001). A flux control that balances the benefit of releasing the *TDH* bottleneck and limiting the glycerol yield is desired to maximize the ethanol industrial yield.

3.3.3 Sequential bottleneck release increases ethanol production yield

The NAD and NADH inter-conversion is involved in multiple reactions in the central carbon metabolic network and controls the yield of ethanol against glucose and glycerol. As discussed in previous session, a careful control of the NAD to NADH balance is desired in order to achieve the maximum ethanol production rate. However, due to the complicated inter-dependency of the NAD/NADH related reactions, it is not obvious how to adjust the enzyme levels of each reaction to optimize the network performance. A sequential single enzyme over- and under-expression optimization was conducted on each enzyme and each model. When applied the first round of single enzyme optimization, multiple optimization strategies have been identified that improve the ethanol production. The maximum ethanol production yields for each model with single enzyme optimization are shown in Figure 10. The models are ordered based on their wild-type ethanol yield against glucose (blue bars). Enzyme strategies that are elected by more than one model are marked out in different colors. 88.5% of the models in the ensemble have a wild-type ethanol yield between 75-80% of the theoretical yield, which is similar to the reported production from *Saccharomyces cerevisiae* (Nevoigt, 2008). The enzyme names and average modulation ratios that lead to the best ethanol production rate for the models are listed in Table 2A.

The over-expression of the ethanol synthesis enzyme (*EOLE*) is the most popular strategy which is elected as the best for 30 of the 52 models in the ensemble. In fact, over-expressing *EOLE* improves the ethanol production in all the 52 models, although the effects are very mild for some models. Recall that the ethanol synthesis enzyme is identified as running close to the overall V_{\max} for many models based on the conditional V_{\max} test. An over-expression of this enzyme for those models could increase the capacity of this reaction and allows higher flux.

Indeed the ethanol synthesis fluxes of the models elected *EOLE* as the best strategy are on average 13.9% closer to their corresponding conditional V_{\max} . Previously, the ethanol synthesis reaction is not defined as a network bottleneck, mainly because it is only identified by the conditional V_{\max} test but no major metabolite accumulation observed. However, we discussed that under the anaerobic condition, there is a shift towards the NADH form compared to the NAD form, and most of the nicotinamide exists in the NADH form for many models. This can be considered as a type of metabolite accumulation of the NADH as well. The constant level of NAD plus NADH of a mild 1.57mM made it hard to observe through the metabolite accumulation test. Here, we would therefore define the *EOLE* reaction as another network bottleneck. In previous discussion we mentioned that a faster ethanol production reaction can lead to more NADH recycled to NAD which, in return, releases the *TDH* bottleneck for more downstream pyruvate. The release of *TDH* bottleneck by *EOLE* over-expression can be indeed observed in Figure 11 which shows that for all the models having metabolite accumulation problem, the accumulations of F1,6P, G3P, and DHAP that represent the *TDH* bottleneck have been released. Figure 12, which plots the reaction rates of *TDH* and ethanol synthesis reaction before and after applied the *EOLE* over-expression strategy, confirms that the increased *EOLE* reaction rate does help increase the *TDH* flux and thus release the bottleneck. Under-expressing glycerol synthesis enzyme (*GLCE*) is also selected by 4 models. The metabolite accumulation test for these 4 models displays no obvious accumulation for any metabolites, which indicates no significant bottleneck in these models. Under-expressing *GLCE* limits the NADH reoxidized by glycerol synthesis reaction and thus could increase the NADH level in the system. The NADH concentration indeed increased by 10% to 12-fold for these 4 models after applied *GLCE* under-expression. The corresponding ethanol reaction and *TDH* reaction are increased for 8.6% and

5.0%, respectively. The glucose transporter (*GT*) over-expression and *CDC19* over-expression are also selected by 4 and 3 models, respectively, which potentially increase the carbon availability for the downstream ethanol pathway. Other strategies have relatively fewer supporting models, which reflect the affects of parameter uncertainties on strategy selection.

After the first round of single enzyme optimization, the ethanol yields of many models are enhanced to 80-90% of the theoretical yield (Figure 10) and all models have found at least one strategy to improve their ethanol yields. A second round of single enzyme over- and under-expression optimization was conducted based on the improved models from the first round. Further benefit for ethanol yield is achieved. 57.7% models now have ethanol yield beyond 90% of the theoretical yield (73.3% of them have ethanol yield beyond 95%) compared to only 15.4% after the first round optimization (Figure 13). All but one model chooses a different enzyme strategy than the first round. It indicates the first enzyme strategy already achieves the best benefit it can lead to and it needs a manipulation of a different enzyme to achieve further improvement. The strategies selected by the second round optimization are listed in Table 2B. The *CDC19* over-expression, which increases the conversion from PEP to PYR, is the most elected strategy for the second round optimization. It may indicate that after the *EOLE* bottleneck release, it now can use more carbon resources from the upstream network. It is interesting that the over-expression of the bottleneck enzyme *TDH* is only elected by one model in the first round and 3 model in the second round, which suggests that a direct over-expression of the bottleneck enzyme may not be the most efficient strategy to improve the network production. It may become significant after more optimization rounds when some other bottlenecks have been released first. This also illustrates the complexity and difficulty when there are multiple bottlenecks in the network that there may be an optimal order to release

different bottlenecks. The order of these strategies may not be obvious and require computational modeling to reveal the best recipe. For the second round optimization, the *EOLE*, *GT*, and some other strategies show up again as popular strategies. Although coming in different orders for different models due to the parameter uncertainties, it seems an exhaustive experimental test for the combinations of four or five enzymes (*EOLE*, *CDC19*, *GT*, *GLCE*) would have a high likelihood to result in the best improvement of the ethanol production. Further round of optimizations can be conducted until it meets the ethanol yield requirement or no further improvement can be acquired.

3.4 Conclusions

There have been many practices using the experimental methods to detect the bottlenecks of the metabolic networks, most of which involves detecting accumulated metabolites in the system or randomly increasing the E_0 of the enzymes on the direct pathway. These approaches are usually time consuming and do not always generate biological insights, as enhanced production of target chemicals caused by increased enzyme E_0 does not directly indicate those enzymes are bottleneck, because there may be other intrinsic bottlenecks get released by the increased E_0 of those enzymes. Many times, there are multiple bottlenecks in the network, without uncover all of which, it is difficult to design strategies to achieve the best production. In this report, we demonstrated a computational framework which can systematically identify bottlenecks in the system with very limited data. In particular, *TDH* reaction, which is much earlier in the pathway than the step of ethanol production, is detected as one of the major bottlenecks for ethanol production in the central carbon metabolic network. The metabolite accumulation test shows significant accumulation of precursors of the *TDH* reaction and glucose input test indicates *TDH* reaction limits the carbon flow towards the downstream pathway (e.g., ethanol production). The conditional V_{\max} test further suggests the NAD is the limiting factor for the *TDH* reaction. Our study further indicates that the NAD and NADH balance is determined by several key reactions, including *TDH*, ethanol synthesis reaction, glycerol synthesis reaction, TCA reactions, etc. and the concentration ratio between NAD and NADH is a crucial regulator of the ethanol production rate. In particular, the ethanol production requires two precursors, the carbon precursor pyruvate and the NADH. Our analysis shows that the NADH and pyruvate concentrations are inter-dependent, as higher NADH level leads to lower NAD level for *TDH* reaction and thus fewer carbon available for downstream pathway. It is thus not obvious of how to manipulate the

enzyme levels in the network in order to maximize the ethanol production. We presented a sequential single enzyme optimization method to identify strategies that gradually release network bottlenecks and increase the ethanol production yield. The over-expression of ethanol synthesis enzyme is elected as one of the most popular strategies for the first round optimization, which is consistent with the result that the ethanol production reaction has reached overall V_{\max} and is also a major bottleneck for many models. Interestingly, *CDC19* is selected as the most efficient strategies for the second round of optimization, which indicates different strategies are usually desired to further improve the production after releasing the first bottleneck. After two rounds of single enzyme over- and under-expression optimization, 57.7% of the models in the ensemble have reached over 90% of the theoretical ethanol production yield and all models have improved ethanol production, compared to 70-80% of the theoretical yield for all wild-type models. The model ensemble we built covers the parameter uncertainties introduced by fitting the experimental data. Although different strategy orders are suggested for different models, a similar set of strategies is elected for the first round and second round optimizations. Therefore, an exhaustive experimental test of a limited set of enzyme strategies would lead to a high likelihood to secure the most efficient strategies to achieve the best ethanol production yield. The yeast central carbon metabolic network model we constructed here can be easily extended to study other academically or commercially interesting chemicals, e.g. high carbon biofuel, etc., which are directly or indirectly linked to the glycolysis or pentose phosphate pathway. The four-test bottleneck detection framework we developed demonstrated as an efficient technique to discover valuable insights for network re-engineering before complicated experiments are required.

References

- Alper, H., & Stephanopoulos, G. (2009). Engineering for biofuels: exploiting innate microbial capacity or importing biosynthetic potential? *Nature reviews. Microbiology*, 7(10), 715–23. doi:10.1038/nrmicro2186
- Atsumi, S., & Liao, J. C. (2008). Metabolic engineering for advanced biofuels production from *Escherichia coli*. *Current opinion in biotechnology*, 19(5), 414–9. doi:10.1016/j.copbio.2008.08.008
- Bailey, J. E. (1991). Toward a science of metabolic engineering. *Science (New York, N.Y.)*, 252(5013), 1668–75. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/2047876>
- Bakker, B. M., Overkamp, K. M., van Maris AJ, Kötter, P., Luttik, M. a, van Dijken JP, & Pronk, J. T. (2001). Stoichiometry and compartmentation of NADH metabolism in *Saccharomyces cerevisiae*. *FEMS microbiology reviews*, 25(1), 15–37. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11152939>
- Barkovich, R., & Liao, J. C. (2001). Metabolic engineering of isoprenoids. *Metabolic engineering*, 3(1), 27–39. doi:10.1006/mben.2000.0168
- Barnett, J. A. (2008). A history of research on yeasts 13 . Active transport and the uptake of various metabolites 1, 689–731. doi:10.1002/yea
- Blank, L. M., Lehmbeck, F., & Sauer, U. (2005). Metabolic-flux and network analysis in fourteen hemiascomycetous yeasts. *FEMS yeast research*, 5(6-7), 545–58. doi:10.1016/j.femsyr.2004.09.008
- Boles, E., & Hollenberg, C. (1997). The molecular genetics of hexose transport in yeasts. *FEMS microbiology reviews*, (21), 85–111. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1111/j.1574-6976.1997.tb00346.x/full>
- Bongaerts, J., Krämer, M., Müller, U., Raeven, L., & Wubbolts, M. (2001). Metabolic engineering for microbial production of aromatic amino acids and derived compounds. *Metabolic engineering*, 3(4), 289–300. doi:10.1006/mben.2001.0196
- Bro, C., Regenber, B., Förster, J., & Nielsen, J. (2006). In silico aided metabolic engineering of *Saccharomyces cerevisiae* for improved bioethanol production. *Metabolic engineering*, 8(2), 102–11. doi:10.1016/j.ymben.2005.09.007
- Burgard, A. P., Pharkya, P., & Maranas, C. D. (2003). Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering*, 84(6), 647–57. doi:10.1002/bit.10803

- Cameron, D C, & Tong, I. T. (1993). Cellular and metabolic engineering. An overview. *Applied biochemistry and biotechnology*, 38(1-2), 105–40. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/8346901>
- Cameron, D.C., & Chaplen, F. W. R. (1997). Developments in metabolic engineering. *Current opinion in biotechnology*, 8(2), 175–180. Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/S0958166997800985>
- Çaylak, B., & Sukan, F. V. (1998). Comparison of different production processes for bioethanol. *Turkish Journal of Chemistry*, 22, 351–359.
- Chassagnole, C., Noisommit-Rizzi, N., Schmid, J. W., Mauch, K., & Reuss, M. (2002). Dynamic modeling of the central carbon metabolism of *Escherichia coli*. *Biotechnology and Bioengineering*, 79(1), 53–73. doi:10.1002/bit.10288
- Cheng, N., & Hasan, M. (2009). Production of ethanol by fed-batch fermentation. *Pertanika J. Sci ...*, 17(May 2008), 399–408.
- Cleland. (1963). The kinetics of enzyme-catalyzed reactions with two or more substrates or products. I. Nomenclature and rate equations. *Biochimica et Biophysica Acta*, 67:104. doi:10.1016/j.ajog.2010.07.025
- Cornish-Bowden, A. (1977). An automatic method for deriving steady-state rate equations. *Biochemical Journal*, 165(1), 55.
- Dombek, K., & Ingram, L. (1987). Ethanol production during batch fermentation with *Saccharomyces cerevisiae*: changes in glycolytic enzymes and internal pH. *Applied and environmental microbiology*.
- Edwards, J. S., & Palsson, B. O. (2000). Robustness analysis of the *Escherichia coli* metabolic network. *Biotechnology progress*, 16(6), 927–39. doi:10.1021/bp0000712
- Fernandez, Herrero, and M. (1985). Inhibition and inactivation of glucose-phosphorylating enzymes from *Saccharomyces cerevisiae* by D-xylose. *Journal of general ...*, 131, 2705–2709. Retrieved from <http://mic.sgmjournals.org/content/131/10/2705.short>
- Fraenkel, D. G. (2003). The top genes: on the distance from transcript to function in yeast glycolysis. *Current Opinion in Microbiology*, 6(2), 198–201. doi:10.1016/S1369-5274(03)00023-7
- Gancedo, J. M. (2008). The early steps of glucose signalling in yeast. *FEMS microbiology reviews*, 32(4), 673–704. doi:10.1111/j.1574-6976.2008.00117.x
- Gao, H., & Leary, J. a. (2003). Multiplex inhibitor screening and kinetic constant determinations for yeast hexokinase using mass spectrometry based assays. *Journal of the American Society for Mass Spectrometry*, 14(3), 173–81. doi:10.1016/S1044-0305(02)00867-X

- Hahn-Hägerdal, B., Galbe, M., Gorwa-Grauslund, M. F., Lidén, G., & Zacchi, G. (2006). Bioethanol--the fuel of tomorrow from the residues of today. *Trends in biotechnology*, 24(12), 549–56. doi:10.1016/j.tibtech.2006.10.004
- Hou, J., & Vemuri, G. N. (2010). Using regulatory information to manipulate glycerol metabolism in *Saccharomyces cerevisiae*. *Applied microbiology and biotechnology*, 85(4), 1123–30. doi:10.1007/s00253-009-2202-6
- Jouhten, P., Rintala, E., Huuskonen, A., Tamminen, A., Toivari, M., Wiebe, M., Ruohonen, L., et al. (2008). Oxygen dependence of metabolic fluxes and energy generation of *Saccharomyces cerevisiae* CEN.PK113-1A. *BMC systems biology*, 2, 60. doi:10.1186/1752-0509-2-60
- Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1), 27–30. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102409&tool=pmcentrez&rendertype=abstract>
- Kauffman, K. J., Prakash, P., & Edwards, J. S. (2003). Advances in flux balance analysis. *Current Opinion in Biotechnology*, 14(5), 491–496. doi:10.1016/j.copbio.2003.08.001
- Keasling, J. D. (1999). Gene-expression tools for the metabolic engineering of bacteria. *Trends in biotechnology*, 17(11), 452–460. Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/S0167779999013761>
- King, E. L., & Altman, C. (1956). A schematic method of deriving the rate laws for enzyme-catalyzed reactions. *The Journal of Physical Chemistry*, 60(10), 1375–1378.
- Kuzmic, P. (2008). The king-altman method. Retrieved from <http://www.biokin.com/king-altman/index.html>
- Leandro, M. J., Fonseca, C., & Gonçalves, P. (2009). Hexose and pentose transport in ascomycetous yeasts: an overview. *FEMS yeast research*, 9(4), 511–25. doi:10.1111/j.1567-1364.2009.00509.x
- Lee, J. M., Gianchandani, E. P., & Papin, J. a. (2006). Flux balance analysis in the era of metabolomics. *Briefings in bioinformatics*, 7(2), 140–50. doi:10.1093/bib/bbl007
- Lee, S. K., Chou, H., Ham, T. S., Lee, T. S., & Keasling, J. D. (2008). Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. *Current opinion in biotechnology*, 19(6), 556–63. doi:10.1016/j.copbio.2008.10.014
- Li, J. W.-H., & Vederas, J. C. (2009). Drug discovery and natural products: end of an era or an endless frontier? *Science (New York, N.Y.)*, 325(5937), 161–5. doi:10.1126/science.1168243

- Liao, J. C., Hou, S. Y., & Chao, Y. P. (1996). Pathway analysis, engineering, and physiological considerations for redirecting central metabolism. *Biotechnology and bioengineering*, 52(1), 129–40. doi:10.1002/(SICI)1097-0290(19961005)52:1<129::AID-BIT13>3.0.CO;2-J
- Lin, Y., & Tanaka, S. (2006). Ethanol fermentation from biomass resources: current state and prospects. *Applied microbiology and biotechnology*, 69(6), 627–42. doi:10.1007/s00253-005-0229-x
- Lynd, L. R. (1996). OVERVIEW AND EVALUATION OF FUEL ETHANOL FROM CELLULOSIC BIOMASS: Technology, Economics, the Environment, and Policy. *Annual Review of Energy and the Environment*, 21(1), 403–465. doi:10.1146/annurev.energy.21.1.403
- Machado, D., Costa, R. S., Ferreira, E. C., Rocha, I., & Tidor, B. (2012). Exploring the gap between dynamic and constraint-based models of metabolism. *Metabolic engineering*, 14(2), 112–9. doi:10.1016/j.ymben.2012.01.003
- Martin, V. J. J., Pitera, D. J., Withers, S. T., Newman, J. D., & Keasling, J. D. (2003). Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nature biotechnology*, 21(7), 796–802. doi:10.1038/nbt833
- Matsushika, A., Inoue, H., Kodaki, T., & Sawayama, S. (2009). Ethanol production from xylose in engineered *Saccharomyces cerevisiae* strains: current state and perspectives. *Applied microbiology and biotechnology*, 84(1), 37–53. doi:10.1007/s00253-009-2101-x
- Nelson, D. L., & Cox, M. M. (2008). *Lehninger principles of biochemistry* (5th ed., pp. 527–646, 868). W. H. Freeman.
- Nevoigt, E. (2008). Progress in metabolic engineering of *Saccharomyces cerevisiae*. *Microbiology and molecular biology reviews : MMBR*, 72(3), 379–412. doi:10.1128/MMBR.00025-07
- Nissen, T. L., Kielland-Brandt, M. C., Nielsen, J., & Villadsen, J. (2000). Optimization of ethanol production in *Saccharomyces cerevisiae* by metabolic engineering of the ammonium assimilation. *Metabolic engineering*, 2(1), 69–77. doi:10.1006/mben.1999.0140
- Ozcan, S., & Johnston, M. (1999). Function and regulation of yeast hexose transporters. *Microbiology and molecular biology reviews : MMBR*, 63(3), 554–69. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=103746&tool=pmcentrez&rendertype=abstract>
- Pramanik, J., & Keasling, J. D. (1998). Effect of *Escherichia coli* biomass composition on central metabolic fluxes predicted by a stoichiometric model. *Biotechnology and bioengineering*, 60(2), 230–8. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10099424>

- Pronk, J., Steensma, H., & Dijken, J. V. (1996). Pyruvate metabolism in *Saccharomyces cerevisiae*. *Yeast*, *12*, 1607–1633. Retrieved from <http://repository.tudelft.nl/assets/uuid:4e855902-cd08-4fb8-b185-a8c5804f79ac/Pronk18.pdf>
- Richard, P., Teusink, B., Westerhoff, H. V., & van Dam, K. (1993). Around the growth phase transition *S. cerevisiae*'s make-up favours sustained oscillations of intracellular metabolites. *FEBS Letters*, *318*(1), 80–82. doi:10.1016/0014-5793(93)81332-T
- Rintala, E., Wiebe, M. G., Tamminen, A., Ruohonen, L., & Penttilä, M. (2008). Transcription of hexose transporters of *Saccharomyces cerevisiae* is affected by change in oxygen provision. *BMC microbiology*, *8*, 53. doi:10.1186/1471-2180-8-53
- Rolland, F., Winderickx, J., & Thevelein, J. M. (2002). Glucose-sensing and -signalling mechanisms in yeast. *FEMS yeast research*, *2*(2), 183–201. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12702307>
- Scheer, M., Grote, a., Chang, a., Schomburg, I., Munaretto, C., Rother, M., Sohngen, C., et al. (2010). BRENDA, the enzyme information system in 2011. *Nucleic Acids Research*, *39*(Database), D670–D676. doi:10.1093/nar/gkq1089
- Schmid, J. W., Mauch, K., Reuss, M., Gilles, E. D., & Kremling, A. (2004). Metabolic design based on a coupled gene expression-metabolic network model of tryptophan production in *Escherichia coli*. *Metabolic engineering*, *6*(4), 364–77. doi:10.1016/j.ymben.2004.06.003
- Simonsen, A., Badawi, N., Anskjær, G. G., Albers, C. N., Sørensen, S. R., Sørensen, J., & Aamand, J. (2012). Intermediate accumulation of metabolites results in a bottleneck for mineralisation of the herbicide metabolite 2,6-dichlorobenzamide (BAM) by *Aminobacter* spp. *Applied microbiology and biotechnology*, *94*(1), 237–45. doi:10.1007/s00253-011-3591-x
- Stephanopoulos, G., & Sinskey, a J. (1993). Metabolic engineering--methodologies and future prospects. *Trends in biotechnology*, *11*(9), 392–6. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/7764086>
- Stephanopoulos, Gregory, Aristidou, A., & Nielsen, J. (1998). *Metabolic Engineering: Principles and Methodologies* (1st ed.). Academic Press.
- Sánchez, O. J., & Cardona, C. a. (2008). Trends in biotechnological production of fuel ethanol from different feedstocks. *Bioresource technology*, *99*(13), 5270–95. doi:10.1016/j.biortech.2007.11.013
- Tyo, K. E., Alper, H. S., & Stephanopoulos, G. N. (2007). Expanding the metabolic engineering toolbox: more options to engineer cells. *Trends in biotechnology*, *25*(3), 132–7. doi:10.1016/j.tibtech.2007.01.003

- Tzafiriri, a R. (2003). Michaelis-Menten kinetics at high enzyme concentrations. *Bulletin of mathematical biology*, 65(6), 1111–29. doi:10.1016/S0092-8240(03)00059-4
- Usuda, Y., Nishio, Y., Iwatani, S., Van Dien, S. J., Imaizumi, A., Shimbo, K., Kageyama, N., et al. (2010). Dynamic modeling of Escherichia coli metabolic and regulatory systems for amino-acid production. *Journal of biotechnology*, 147(1), 17–30. doi:10.1016/j.jbiotec.2010.02.018
- Varma, A., & Palsson, B. O. (1994). Metabolic flux balancing: basic concepts, scientific and practical use. *Nature Biotechnology*, 12(10), 994–998. Retrieved from <http://www.nature.com/nbt/journal/v12/n10/abs/nbt1094-994.html>
- Vemuri, G. N., Eiteman, M. a, McEwen, J. E., Olsson, L., & Nielsen, J. (2007). Increasing NADH oxidation reduces overflow metabolism in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 104(7), 2402–7. doi:10.1073/pnas.0607469104
- Vital-Lopez, F. G., Armaou, A., Nikolaev, E. V., & Maranas, C. D. (2006). A computational procedure for optimal engineering interventions using kinetic models of metabolism. *Biotechnology progress*, 22(6), 1507–17. doi:10.1021/bp060156o
- de Koning, W., & van Dam, K. (1992). A method for the determination of changes of glycolytic metabolites in yeast on a subsecond time scale using extraction at neutral pH. *Analytical biochemistry*, 204(1), 118–23. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/1514678>

Figures

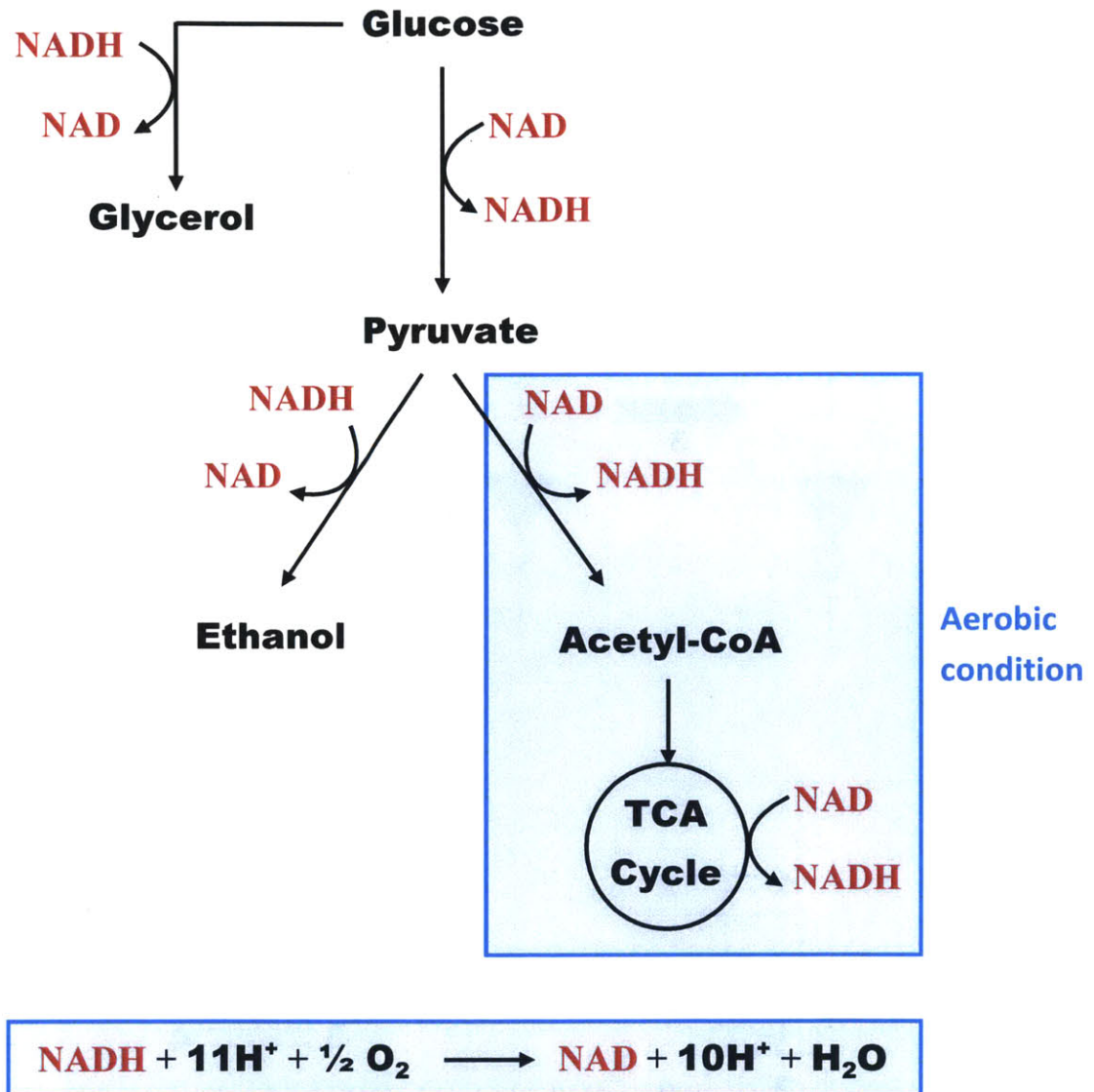
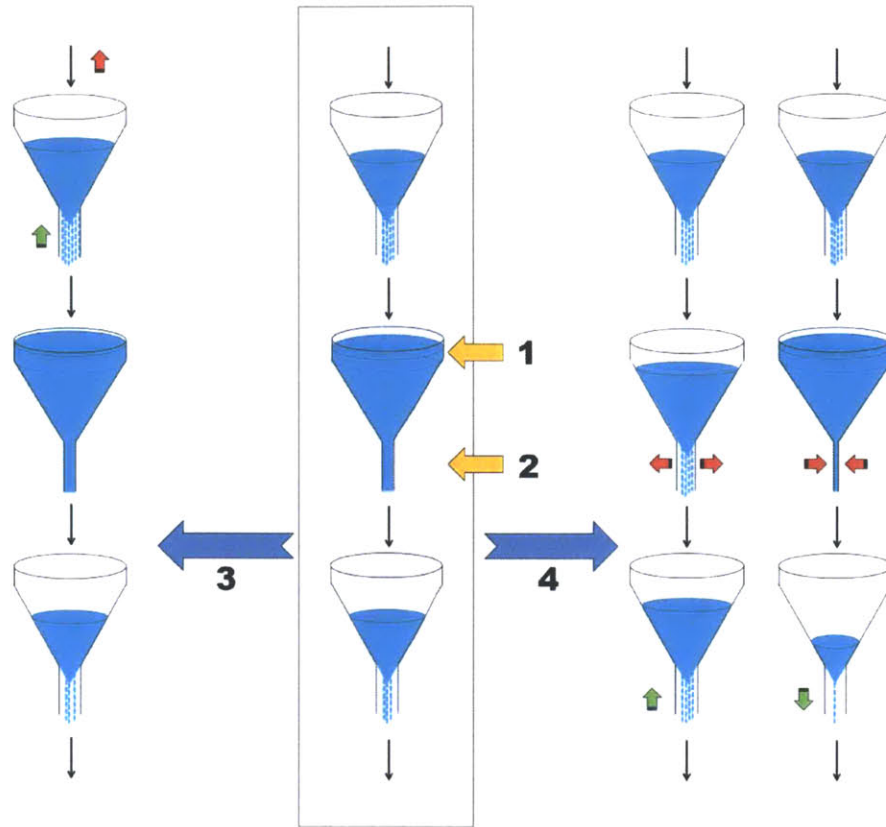
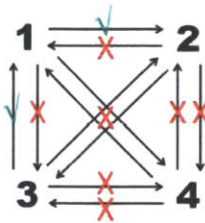
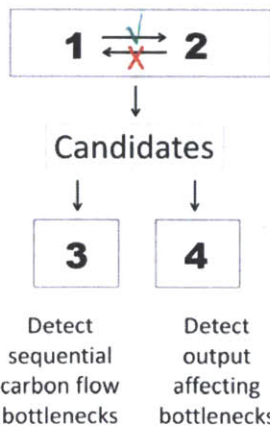


Figure 1. NAD and NADH dynamics in central carbon metabolic network. The blue blocks mark the reactions that only happen under aerobic conditions. The ethanol and glycerol pathways are mainly responsible for reoxidizing NADH under anaerobic condition.

A



B



- 1. Metabolite accumulation**
- 2. Conditional V_{max}**
- 3. Increased input**
- 4. Increased/decreased E_0**

Figure 2. Bottleneck detection framework. A. Bottleneck detection framework. B. Relations among the four tests for detecting bottlenecks

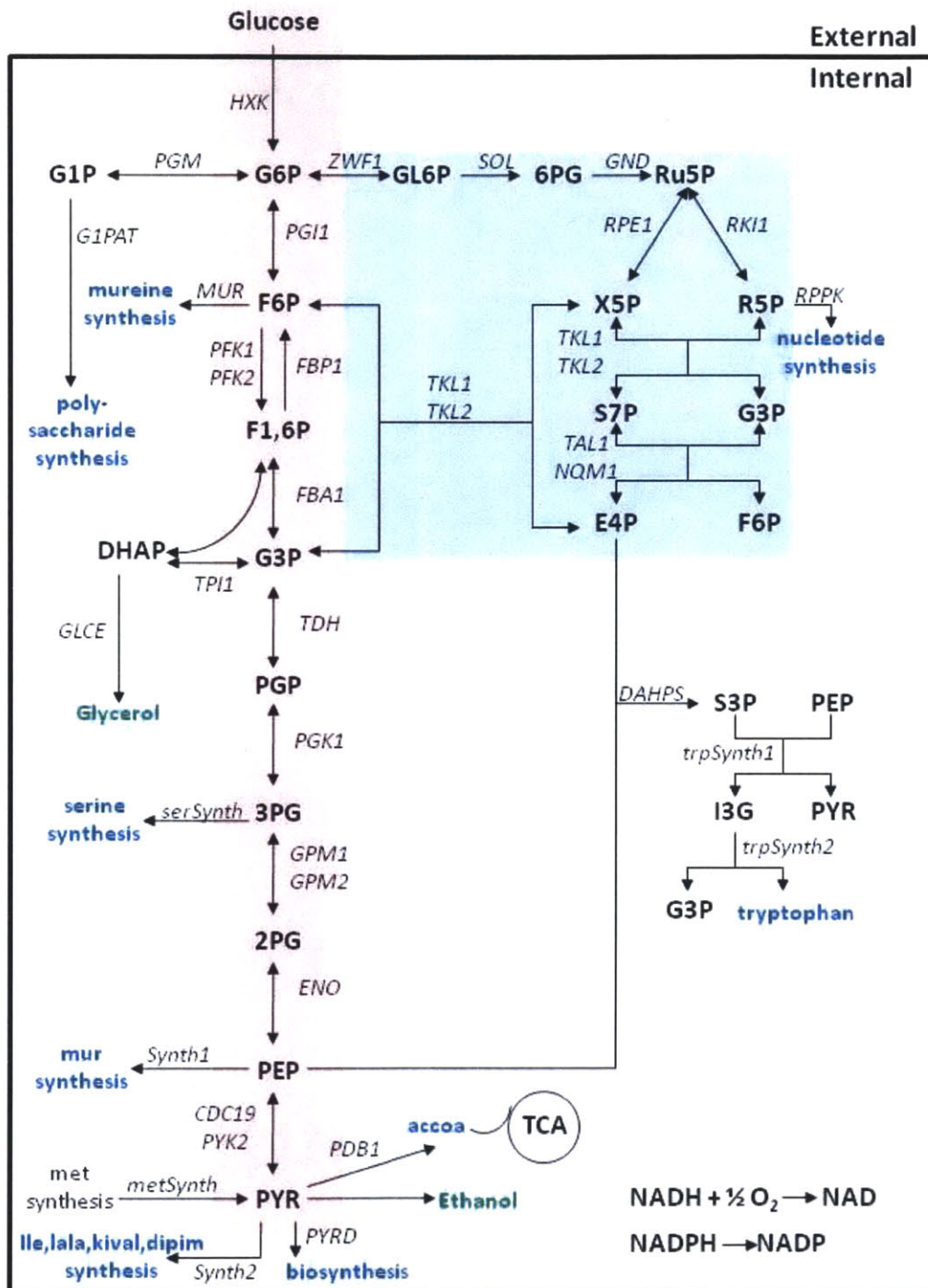


Figure 3. Yeast central carbon metabolic network model. The red block marks the glycolysis pathway. The blue block marks the pentose phosphate pathway.

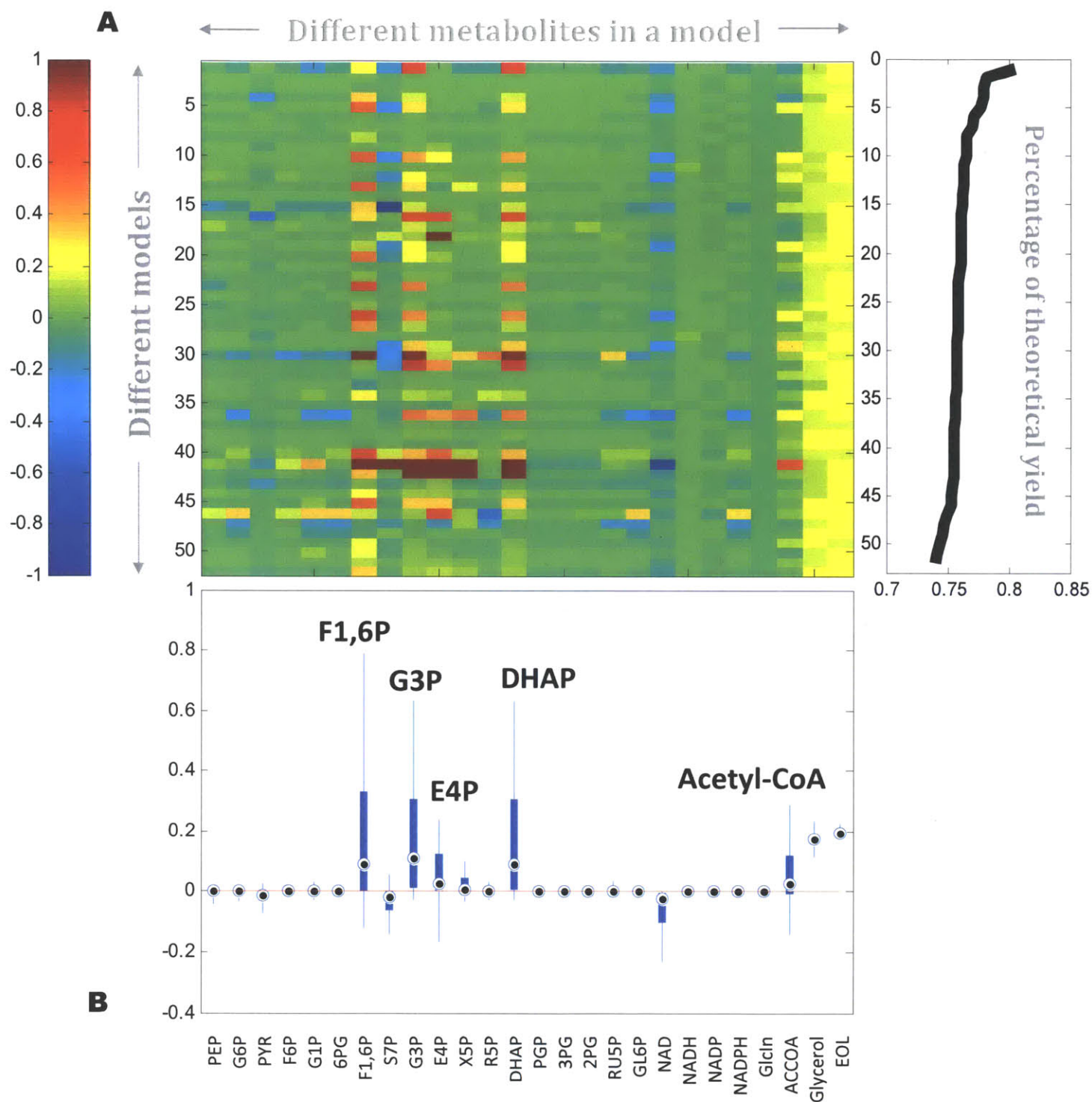


Figure 4. Metabolite accumulation test. **A.** the log ratio of metabolite concentrations at t_2 over t_1 . The models are ranked by their wild-type percentage of theoretical yield. Yellow to red color indicates metabolite accumulations. **B.** The box plot of the data in **A** for each metabolite. Besides ethanol and glycerol, the two end products, F1,6P, G3P, E4P, DHAP, and Acetyl-CoA are also accumulated in the system.

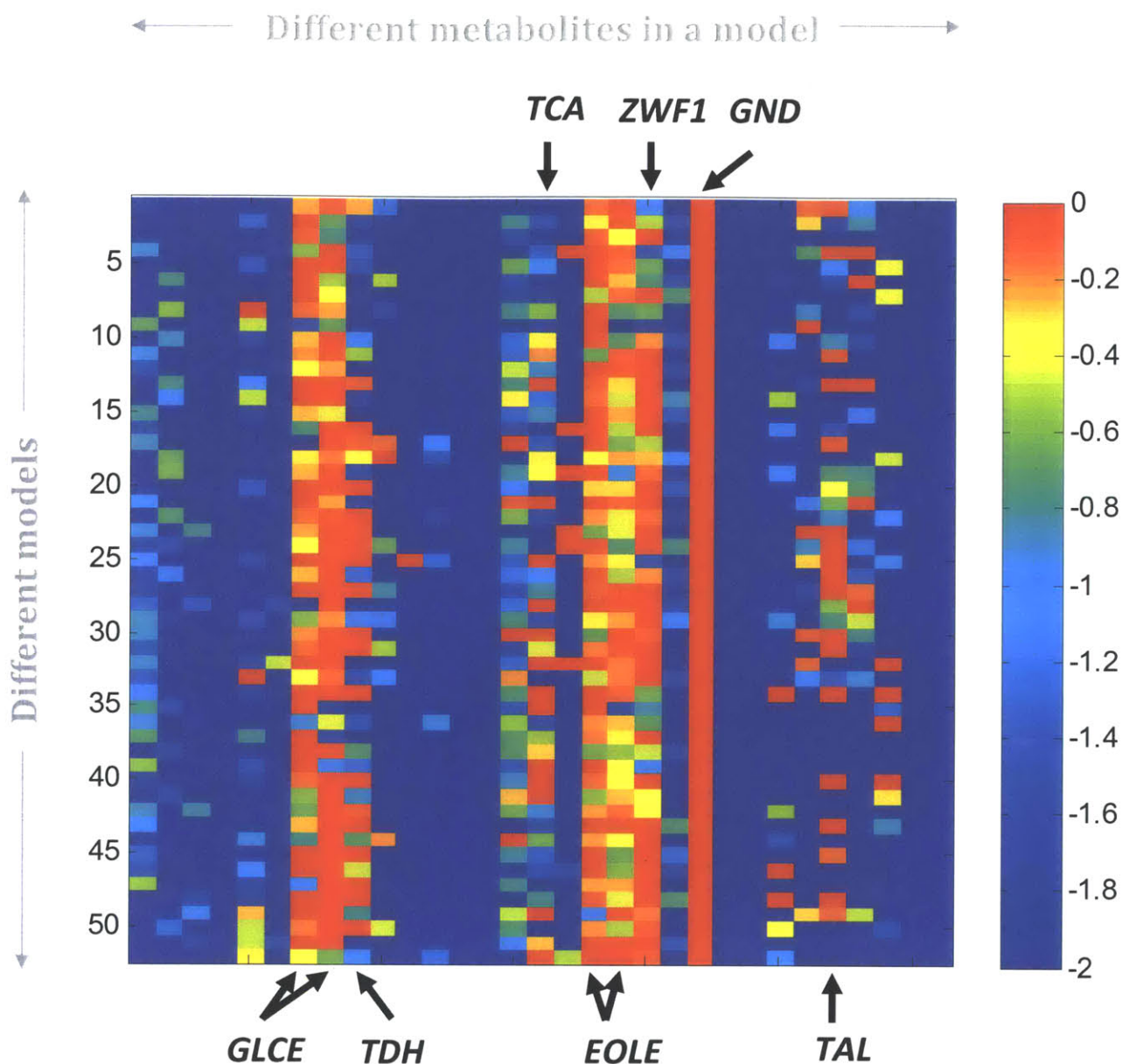


Figure 5. Conditional V_{\max} test. The enzyme fluxes from the leftmost column to the rightmost column are *PGM*, *GIPAT*, *PGII*, *PFK*, *FBA1*, *TPII*, *GLCE* (DHAP), *GLCE* (NADH), *TDH* (NAD), *PGK1*, *serSynth*, *ENO*, *Synth1*, *PYK*, *PDB1* (PYR), *TCA1* (NAD), *Synth2*, *EOLE* (NADH), *EOLE* (PYR), *ZWF1* (G6P), *SOL*, *GND*, *RPE1*, *PKII*, *RPPK*, *TKL* (R5P) (to S7P), *TAL* (GAP), *TKL* (E4P) (to F6P), *DHAPS*, *trpSynth1* (S3P), *trpSynth2*, where the metabolites in the brackets are the limiting metabolites.

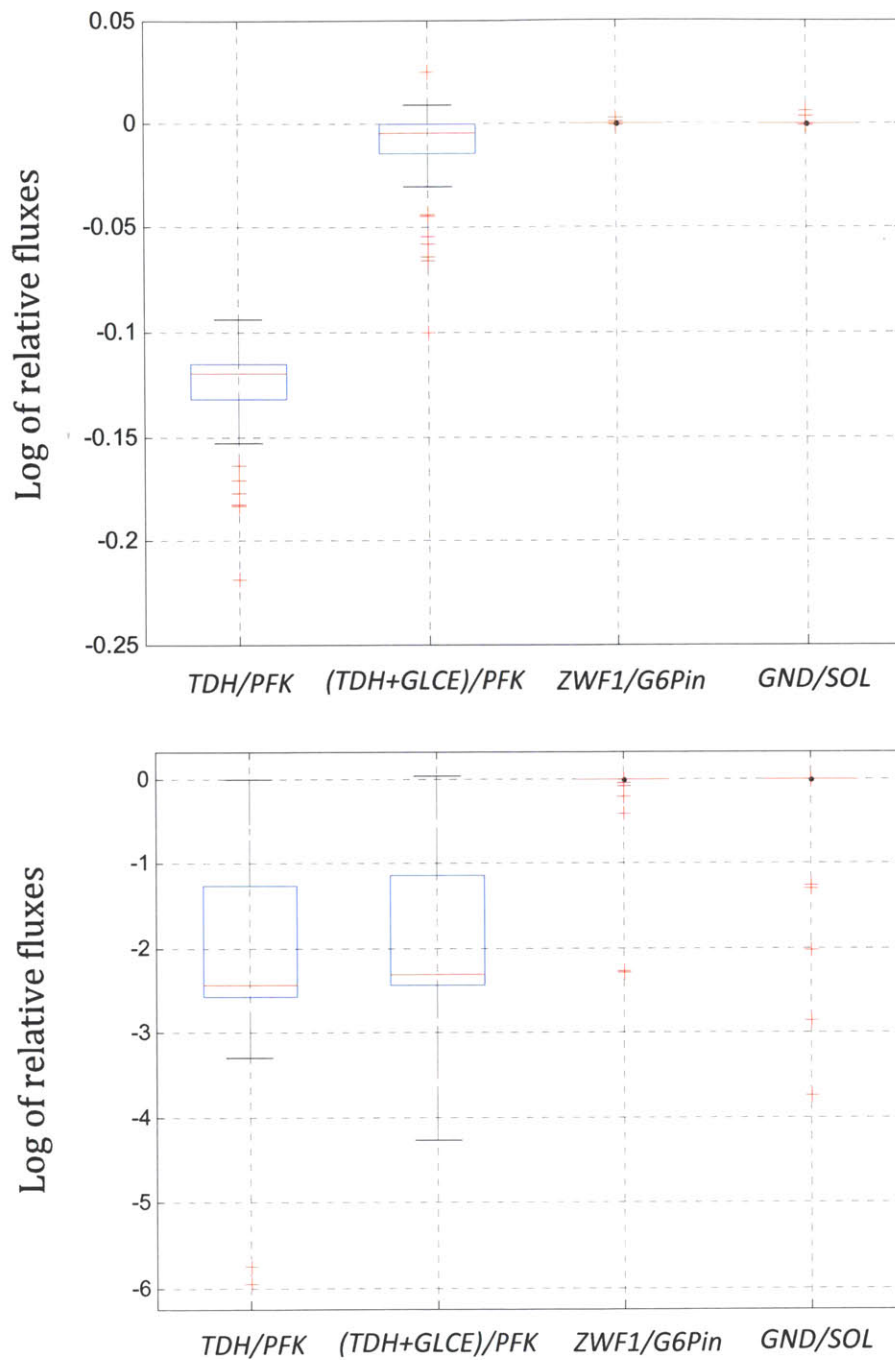


Figure 6. Flux imbalance around accumulated metabolites. The top panel shows under the original glucose input level, the log of the relative fluxes for the corresponding reactions. The bottom panel shows the flux difference under 10-fold of the original glucose input level. The fluxes going out of G3P and DHAP are smaller than the fluxes going in, which indicates the *TDH* and *GLCE* as bottlenecks for the system. The fluxes going out and going in the G6P and 6PG are, however, at the same level, whether or not increase the glucose input. This indicates that the *ZWF1* and *GND* reactions are not critical bottlenecks, as they can be opened up with higher glucose input.

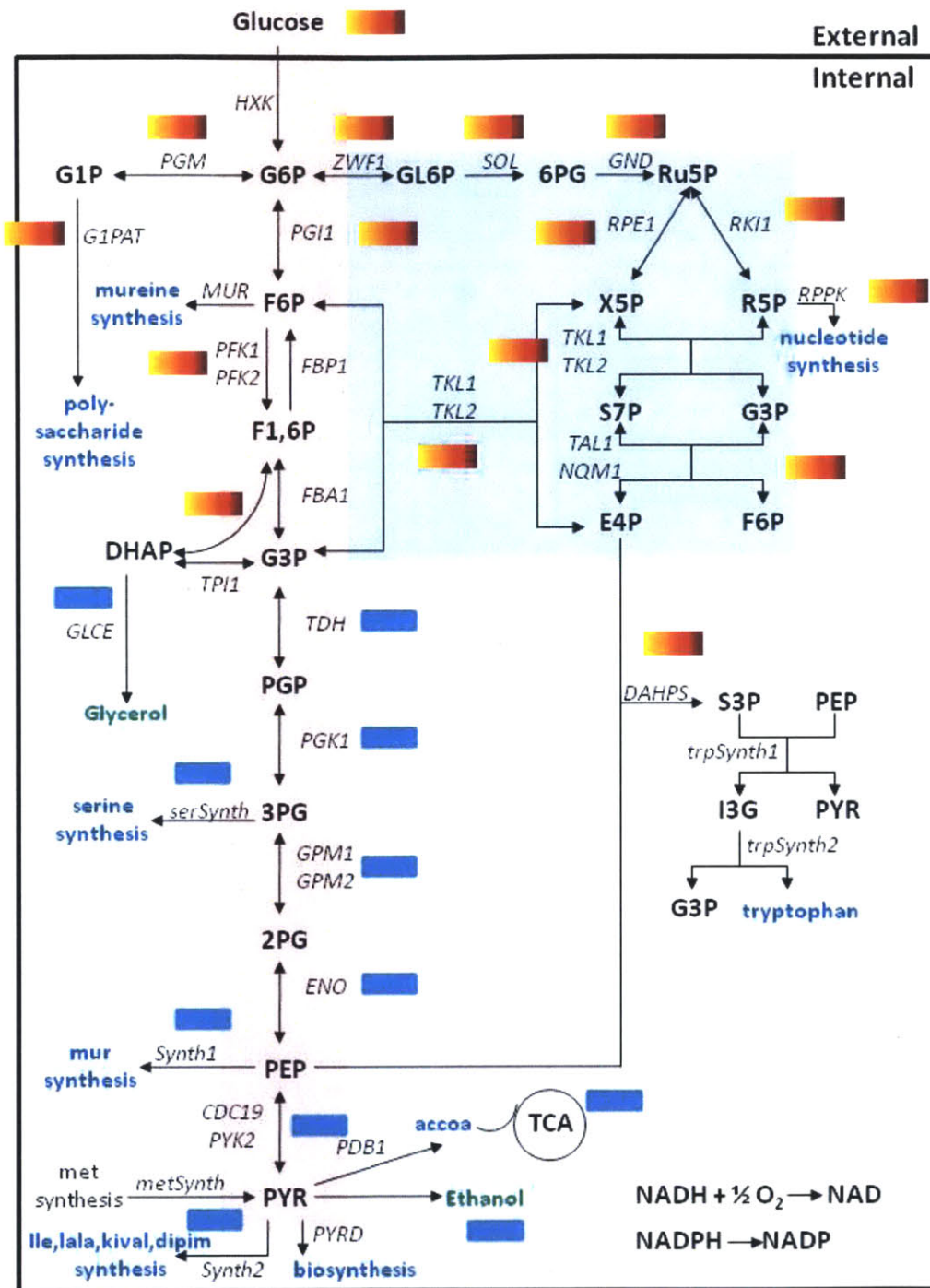


Figure 7. Glucose input test. Glucose input level was increased for 2 to 1000-fold. Yellow to red mark indicates increased fluxes and blue mark indicates no change of fluxes compared to the original glucose input. The *TDH* reaction marks a pattern switch between the upper and lower glycolysis pathway.

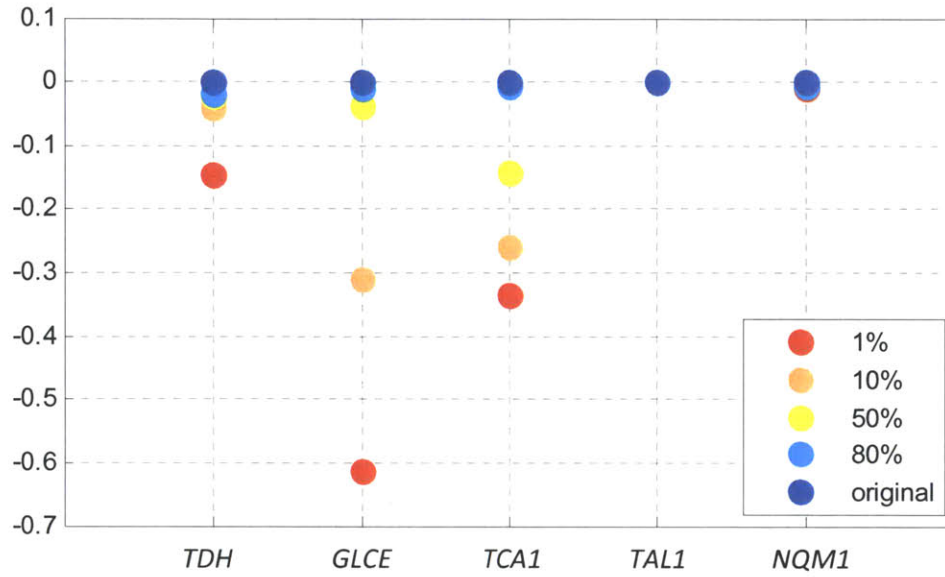


Figure 8. Decreased E_0 test. The change of E_0 of *TDH*, *GLCE*, and *TCA1* reactions affect the ethanol production rate correspondingly, but that of *TAL1* and *NQM1* do not.

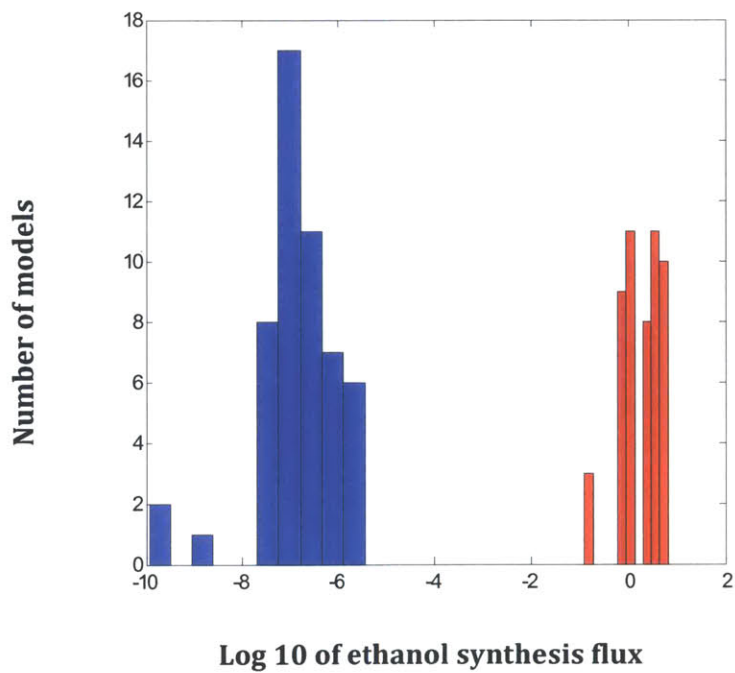
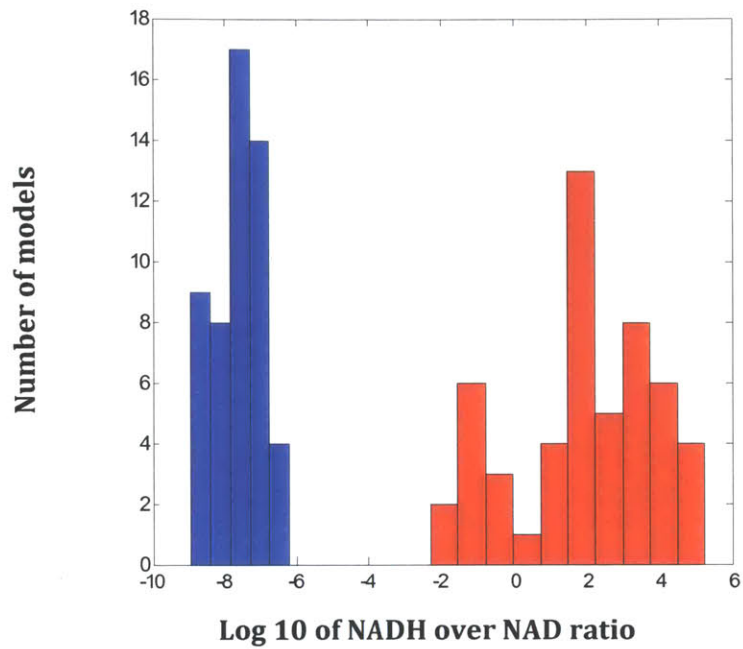


Figure 9. NAD and NADH balance at aerobic and anaerobic conditions. The blue bars are the log₁₀ of the NADH over NAD ratio for model ensemble under aerobic conditions; the red bars are those under anaerobic condition. As the total amount of NADH plus NAD is fixed, this indicates the balance between the two is shifted towards NADH after entering the anaerobic condition.

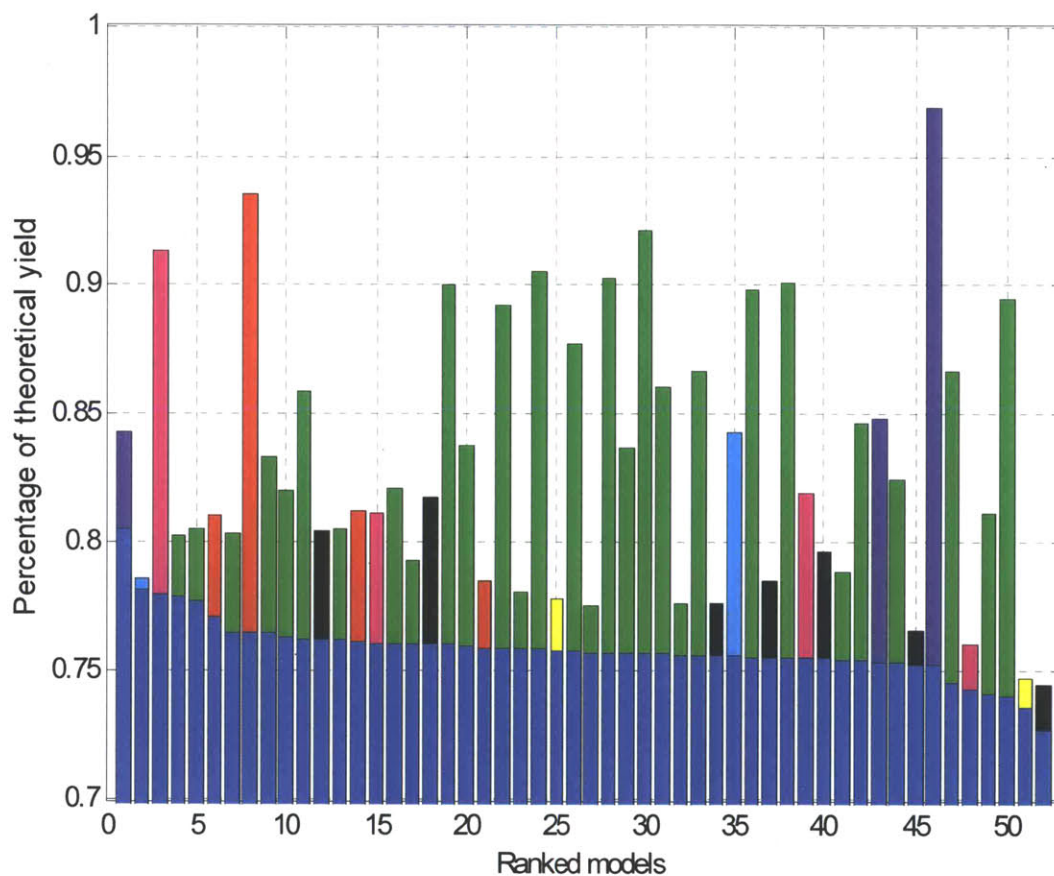


Figure 10. Single enzyme sequential optimization. Blue bars are ethanol yield for wild-type models. Other color bars are ethanol yield after applied the best strategy of each model. The same color indicates the same strategies are applied. The black color marks strategies elected by only one model. The green bar is *EOLE*; the pink bar is *GLCE*; the red bar is *GT*; the purple bar is *CDC19*; the cyan bar is *TPII*; yellow bar is *TCAI*.

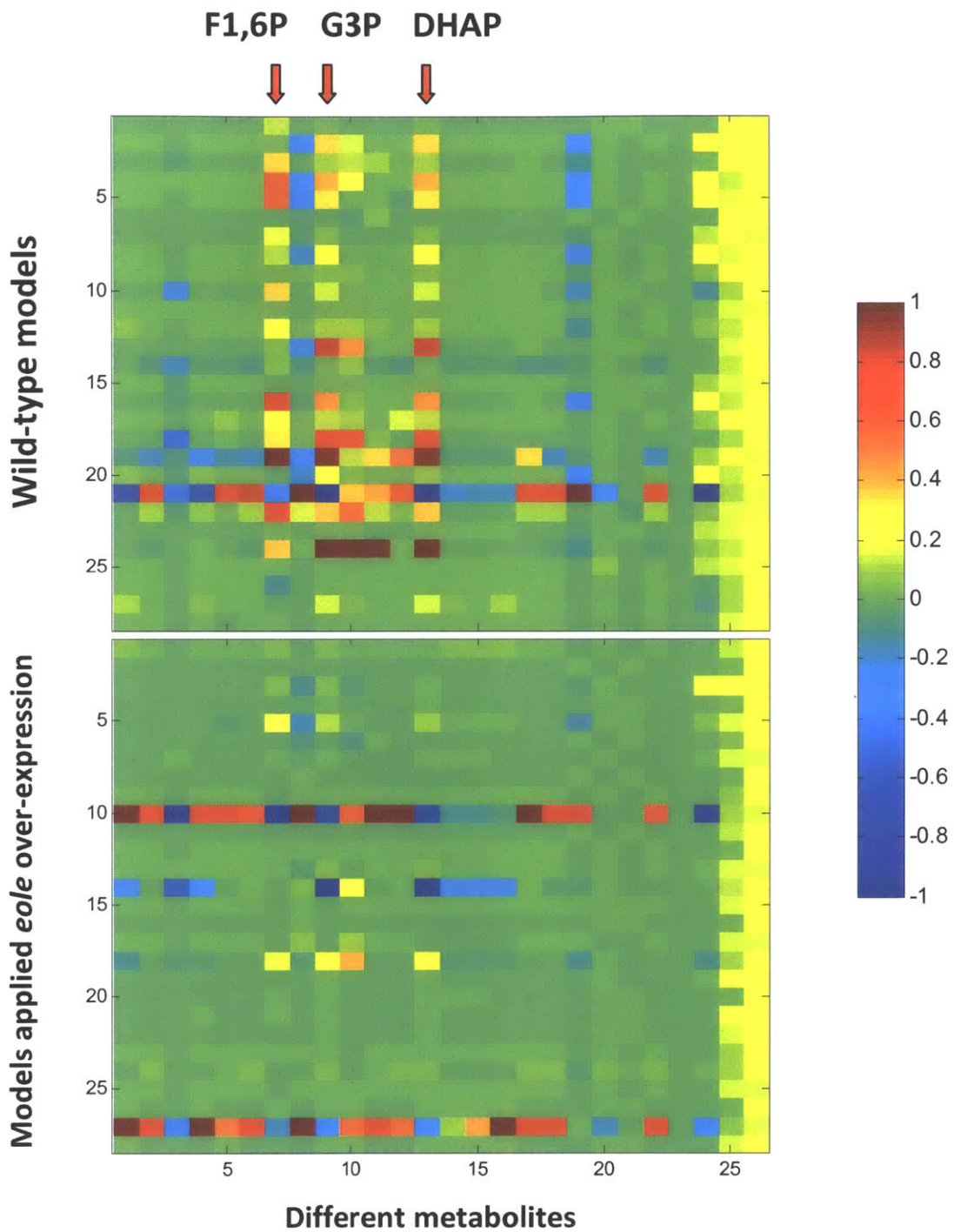


Figure 11. Bottleneck release. After applying the *EOLE* over-expression. The metabolite accumulation for F1,6P, G3P, and DHAP have been released. The rows correspond to models elected *EOLE* as the best first round strategy.

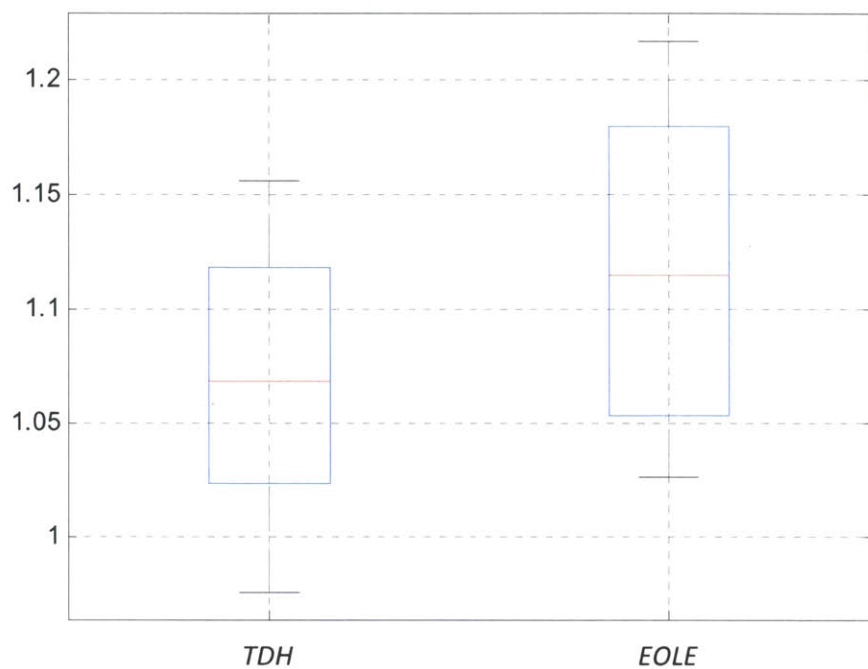


Figure 12. Flux change for *TDH* and *EOLE* reactions. The boxplot shows the ratio of the fluxes after applied the *EOLE* over-expression against those for the wild-type for the models elected *EOLE* as the best strategy. It can be seen all but one model has increased *TDH* and *EOLE* fluxes (ratio above one).

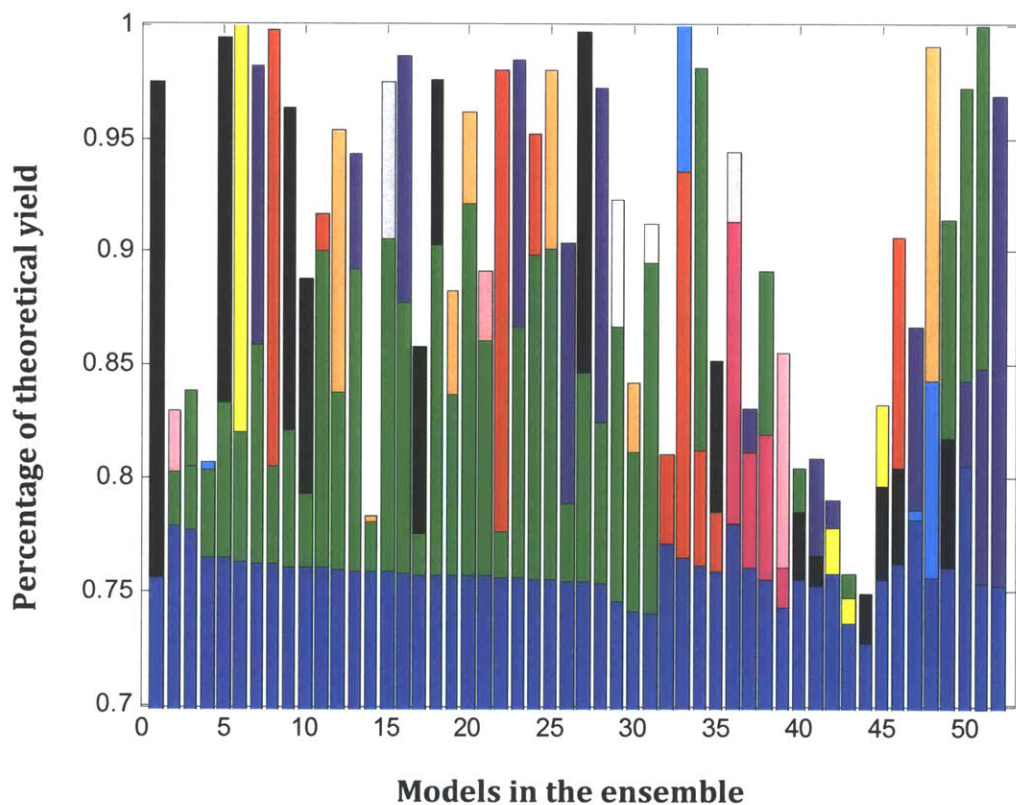


Figure 13. Sequential bottleneck release. Two rounds of single enzyme over- and under-expression optimization were applied to model ensemble. The blue bars are the wild-type model yield. The green bars are *EOLE*; the magenta bars are *GLCE*; the red bars are *GT*; the purple bars are *CDC19*; the cyan bars are *TPH*; yellow bars are *TCAI*; the white bars are *PGII*; the grey bars are *PFKI*; the black bars are strategies only elected by one model.

Tables

Enzyme reactions	Functions	Reaction mechanisms
<i>TDH</i>	NAD to NADH	$G3P + NAD \rightleftharpoons PGP + NADH$
<i>PDB1</i>	NAD to NADH	$PYR + NAD \rightarrow ACCOA + NADH$
<i>TCA1</i>	NAD to NADH	$OXA + ACCOA + NAD \rightleftharpoons KG + NADH$
<i>TCA2</i>	NAD to NADH	$KG + NAD \rightleftharpoons SCOA + NADH$
<i>TCA3</i>	NAD to NADH	$SCOA + NAD \rightleftharpoons OXA + NADH$
<i>OPE</i>	NADH to NAD	$NADH + \frac{1}{2} O_2 \rightarrow NAD + H_2O$
<i>EOLE</i>	NADH to NAD	$PYR + NADH \rightarrow EOL + NAD$
<i>GLCE</i>	NADH to NAD	$DHAP + NADH \rightarrow Glycerol + NAD$

Table 1. Reaction mechanisms used to update NAD/NADH and NADP/NADPH balances. The full names of the enzymes and metabolites can be found in *Abbreviations* session.

A

Strategy name	Number of models	Modulation ratio
<i>EOLE</i>	30	22.7
<i>GLCE, GT</i>	4	0.07, 2.7
<i>CDC19</i>	3	40.3
<i>TPII, TCAI</i>	2	25.0, 29.1
<i>TDH, PGI1, RKII, TCA2, OPE, PYRD</i>	1	12.0, 3.6, 32.2, 41.0, 0.02, 0.02

B

Strategy name	Number of models	Modulation ratio
<i>CDC19</i>	11	45.7
<i>EOLE</i>	9	23.1
<i>GT</i>	5	2.0
<i>TDH</i>	3	24.0
<i>TPII, PGI1, PFK1, TCAI</i>	2	27.0, 45.5, 50, 30.5
<i>ENO, PGM, PDB1, GIPAT, NQMI, TCA2, GLCE, DEG, PYRD</i>	1	50, 0.02, 0.04, 0.02, 38.8, 1.2, 2.4, 0.02, 0.02

Table 2. Best enzyme strategies from the sequential single enzyme over- and under-expression. A. the enzyme strategies from the first round optimization; B. the enzyme strategies from the second round optimization. The first column shows the strategy name. The second column shows the number of models elected the corresponding strategies. The third column shows the average modulation ratio for the corresponding enzymes, with number above 1 for over-expression and number below 1 for under-expression.

Abbreviations

Enzymes

FBA

DAHPS

ENO

GIPAT

GLCE

TDH

metSynth

MUR

PFK

GND

PGII

PGKI

PDBI

PGM

PYK

RKII

RPPK

RPEI

TAL

TPII

TKL

trpSynth

Aldolase

DAHPS synthases

Enolase

Glucose-1-phosphate adenylyltransferase

Glycerol-3-phosphate dehydrogenase

Glyceraldehydes-3-phosphate dehydrogenase

Methionine synthesis

Mureine synthesis

Phosphofructokinase

6-phosphogluconate dehydrogenase

Glucose-6-phosphate isomerase

Phosphoglycerate kinase

Pyruvate dehydrogenase

Phosphoglucomutase

Pyruvate kinase

Ribose-phosphate isomerase

Ribose-phosphate pyrophosphokinase

Ribulose-phosphate epimerase

transaldolase

Triosephosphate isomerase

Transketolase

Tryptophan synthesis

Metabolites

2pg

3pg

6pg

accoa

e4p

f6p

f1,6p

g1p

g6p

g3p

ile

lala

kival

dipim

nad

nadh

nadp

nadph

2-phosphoglycerate

3-phosphoglycerate

6-phosphogluconate

Acetyl-coenzyme A

Erythrose-4-phosphate

Fructose-6-phosphate

Fructose-1,6-bisphosphate

Glucose-1-phosphate

Glucose-6-phosphate

Glyceraldehydes-3-phosphate

Isoleucine

L-alanine

Alpha-ketoisovalerate

Diaminopimelate

Diphosphopyridindinucleotide, oxidized

Diphosphopyridindinucleotide, reduced

Diphosphopyridindinucleotide-phosphate, oxidized

Diphosphopyridindinucleotide-phosphate, reduced

pep
pyr
r5p
ru5p
s7p
x5p

Phosphoenolpyruvate
Pyruvate
Ribose-5-phosphate
Ribulose-5-phosphate
Sedoheptulose-7-phosphate
Xylulose-5-phosphate

Chapter 4

Conclusions and future directions

The rapid development of genome sequencing and high-throughput measurement techniques of enzymes and species concentrations has the potential to bring the biological sciences into the era of so-called ‘big-data’. The large amount of available data for concentrations, fluxes, and kinetics of enzymes under normal or perturbed conditions in biological networks provide unprecedented opportunities to understand the functional mechanisms of cells. On the other hand, it brings new challenges of handling, integrating, and interpreting the large amount of data to acquire novel biological knowledge. With the development of computational and systems biology, it is now commonly believed that system-level modeling may provide unique opportunities to understand cellular function. New techniques for modeling biological networks, which can incorporate the vast amount of available data and describe the underlying biochemical mechanisms, are needed. It is also important to develop methodologies to analyze intrinsic network properties and optimize system behavior to fit desired performance.

In this thesis, I presented new ordinary differential equation (ODE) models of central carbon metabolism for *E. coli* and *S. cerevisiae* based on mass-action rate laws (MRL) of the biochemical reactions. They describe actual biochemical mechanisms of the enzyme reactions, e.g., the binding and releasing of reactants, the conversion from reactant-enzyme complex to product-enzyme complex, and the binding and release of the products, using separate kinetic parameters. Therefore, it reflects closely of how real enzymes work and has the potential to guide protein engineering. If well trained, the models can help predict the most efficient ways to modify single enzymes, e.g. improving reactant binding, improving k_{cat} , or improving product release, in order to achieve a better level of network performance. Because the MRL models are constructed with elementary enzyme reaction steps, it is much easier than in aggregated rate law (ARL) models to incorporate new enzyme interactions and regulation, which makes this model type very flexible to be extended for studying a different aspect of the same network. This point is demonstrated in this thesis by converting the *E. coli* model from Chapter 2 to the *Saccharomyces cerevisiae* model in Chapter 3. The new model includes the important new features such as the oxygen dynamics, the automatic switch from aerobic to anaerobic condition, and NAD/NADH balance, but it only requires minor changes of several elementary reaction mechanisms. On the other hand, the ARL models would require updated aggregated reaction formula incorporating the effect of the new co-factors. The modifications are not straightforward as many reactions use empirical formula based on experiments with little theoretical rationale. With new measurement data becoming available, the model can be easily re-fit to any new concentration, fluxes, or enzyme kinetic data to refine the performance. This point is demonstrated by re-fitting the *E. coli* model with new *S. cerevisiae* data in Chapter 3. The high flexibility and mechanistically realistic features of the mass-action ODE models make it an

attractive technique to model important biological systems, such as the highly conserved central carbon metabolic network and gene expression regulatory networks.

The issue of parameter uncertainties exists for almost all modeling work, due to incomplete data and lack of biological knowledge. In this thesis, we handle the parameter uncertainty problem by introducing model ensembles. Multiple parameter values that fit equally well to the measured data can be found. Instead of using one single model that best fits the available data to make predictions, we collected a group of models based on proper sampling methodologies and use the model ensemble to draw more reliable conclusions. The models in the ensemble all share the same topology but have different parameter values. Without further experimental data, there is no basis for choosing a single parameter set that represents the best biology. It is especially true when it considering mutant networks that a good wild-type parameter set may not necessarily represent the correct behavior after enzyme mutations or expression changes are applied. We observed some inconsistency in predictions from the models in the ensemble from the studies both in Chapter 2 and Chapter 3. It illustrates the significant impact of parameter uncertainty on model predictions and demonstrates the risks of using a single model to draw conclusions with limited data. In Chapter 2, we assign equal weight to each model and let them vote for the best enzyme strategies that optimize the aromatic amino acid productions. If a high percentage of models agree on certain strategies, we have more confidence in them as they are less sensitive to parameter uncertainties existing in the models. It thus provides a way to evaluate the robustness of different enzyme strategies and can potentially improve success rates when applying the predicted strategies to experiments and industrial production. Compared to mathematical methods (e.g., uncertainty propagation) to evaluate parameter uncertainties, the model ensemble method is more straightforward to understand and thus easier to communicate with

experimentalists. It could also be more accurate in many situations given the highly non-linear properties of biological networks. Theoretically, the more models to be included in the ensemble, the better coverage there could be for the parameter space. However, available computational power could limit how many models it is feasible to incorporate into the ensemble. More sophisticated technologies for uncertainty control should be developed in the future to more efficiently handle the predictions from mass-action rate law ODE models.

The central carbon metabolic network is one of the most studied biological networks. It has many good features as a case study to develop computational methodologies (e.g., the well known topology, the known kinetics for many enzymes in this network, and large data set of measured data). On the other hand, the complicated enzyme interactions and regulations in central carbon metabolism also make it a challenging network to analyze and interpret with computational models. Of the different properties we can learn about a network, bottleneck analysis is one of the most important ones, because the increasing use of computational models in metabolic engineering studies that aim to improve production rates of target chemicals.

However, limited research has been done in systematically identify bottlenecks in the central carbon metabolic network. In Chapter 3 we developed a bottleneck identification framework, composed of four computational tests, (i.e., metabolite accumulation, conditional V_{\max} , glucose input, and decreased E_0). This framework is shown to efficiently identify relevant bottlenecks limiting ethanol productions. In particular, the conditional V_{\max} test can directly determine the utilization of available enzyme capacity at a given system state. More importantly, it can provide valuable insights into the intrinsic rationale for the observed rate-limiting steps. Based on these analyses, it is suggested that the balance between NAD and NADH molecules, determined by the relative fluxes of eight enzyme reactions, is the crucial limiting factor for ethanol production.

The *TDH* and ethanol synthesis reactions are two important bottlenecks in the network that constrain the improvement of ethanol yield. Although the bottleneck identification framework was developed and applied to the central carbon metabolic network, the concepts have general value and can be applied to other metabolic networks.

Due to the complexity of central carbon metabolism, it is not obvious of how to manipulate the network for optimized target production, even if we may have gained the knowledge of bottlenecks in the network. In Chapter 2, we showed that aromatic amino acid production requires two precursors, E4P and PEP, one from the pentose phosphate pathway and the other from the glycolysis pathway. There are several branching points between the pentose phosphate pathway and glycolysis. It is not immediately clear how to manipulate the network in order to achieve balanced production of these two metabolites. In Chapter 3 we showed that there is a trade-off between the production of two precursors, pyruvate and NADH, of ethanol synthesis. Due to the constant level of total NAD plus NADH, more NADH means less NAD in the system, which in return, limits the *TDH* reaction rate and thus the availability of pyruvate to the ethanol synthesis pathway. Six other enzyme reactions also contribute to the balance of NAD and NADH in the system. It is not straightforward to identify the right strategies to balance the precursors for best ethanol production. In this thesis, we developed an optimization methodology for mass-action rate law ODE models that allows parallel or sequential combinations of enzyme knock-out and over-/under-expression strategies to the model in order to search for the enzyme strategies that optimize target production rates. The method is shown to be efficient and reliable, with many of the suggested strategies tested to be positive through previous experiments. More strategies that are never considered by previous researches are also predicted by the optimization results, which serve as useful guidance for future experimental design. Results from

experimental tests, whether successful or not, can be used to further refine the parameters of the model or the composition of the ensemble. It can be expected cycles of prediction and testing will lead to improved performance of computational models.

In this thesis, the two applications are both in the field of metabolic engineering, in which optimization of the target chemical production is the major objective. However, the scope of applications in which these types of models could be used is much broader. Recently there is increasing interest in cancer metabolism, which suggests potential links between cancer development and aerobic glycolysis. Decades ago, the Warburg effect described the increased utilization of the glycolysis pathway under aerobic condition for cancer cells compared to normal ones (Warburg, 1956). Since then, significant work has been conducted trying to determine the causal relation between cancer development and abnormal flux in the glycolysis pathway (Fantin et al., 2006; Hsu & Sabatini, 2008; Kroemer & Pouyssegur, 2008). Interestingly, the NAD/NADH ratio has been identified as an important factor connecting central metabolism with cancer cells (Koukourakis et al., 2006). Thus, it is possible that central carbon metabolism models built from those described in this thesis, especially the NAD/NADH involved models described in Chapter 3, can be used to understand cancer mechanism. A bridge between mechanistic models of central carbon metabolism and higher-level cancer progression processes needs to be found for the purpose of this study. As described in Chapter 1, some preliminary research has been conducted where we adopted a commonly used cancer progression model and defined the mutation rates of that model to depend on the output of the mechanistic models that describe the selenium metabolism. The preliminary results of this combined model show the variations of the cancer progress rates for different selenium input, which indicates that we can indeed construct a cancer progression model that depends on detailed mechanistic models. With

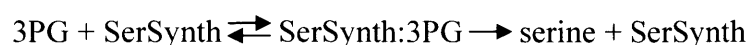
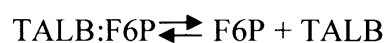
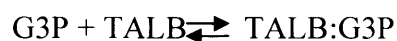
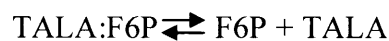
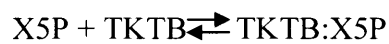
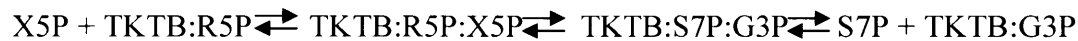
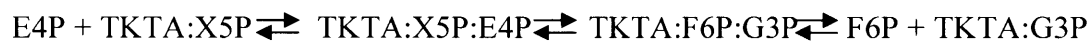
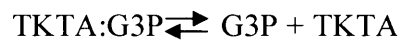
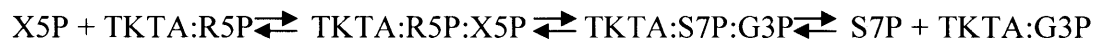
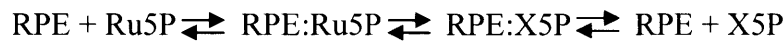
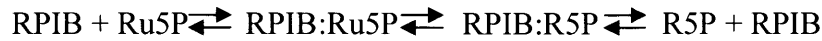
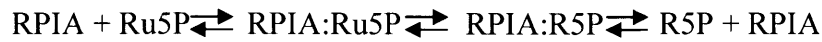
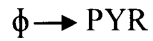
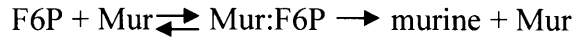
this proof-of-concept test on the simplified selenium metabolism model, it is expected that it would be possible to incorporate the entire central carbon metabolism model with the cancer progression model (for example using NAD/NADH ratio as a linker for the two models). Significant impact can be anticipated for the application of this combined model. It can be used to understand whether changes in central metabolism can speed up cancer progression, and whether the development of cancer requires the enhanced glycolysis pathway to provide the ‘building blocks’. Clinical trial simulations can be built on top of the combined model, so that it is possible to identify possible therapies that reduce the cancer progression rate. With the increasing availability of experimental and clinical data, the mass-action rate law models we built here, and especially the methods developed and applied, will surely make a great impact on biomedicine and metabolic engineering.

References

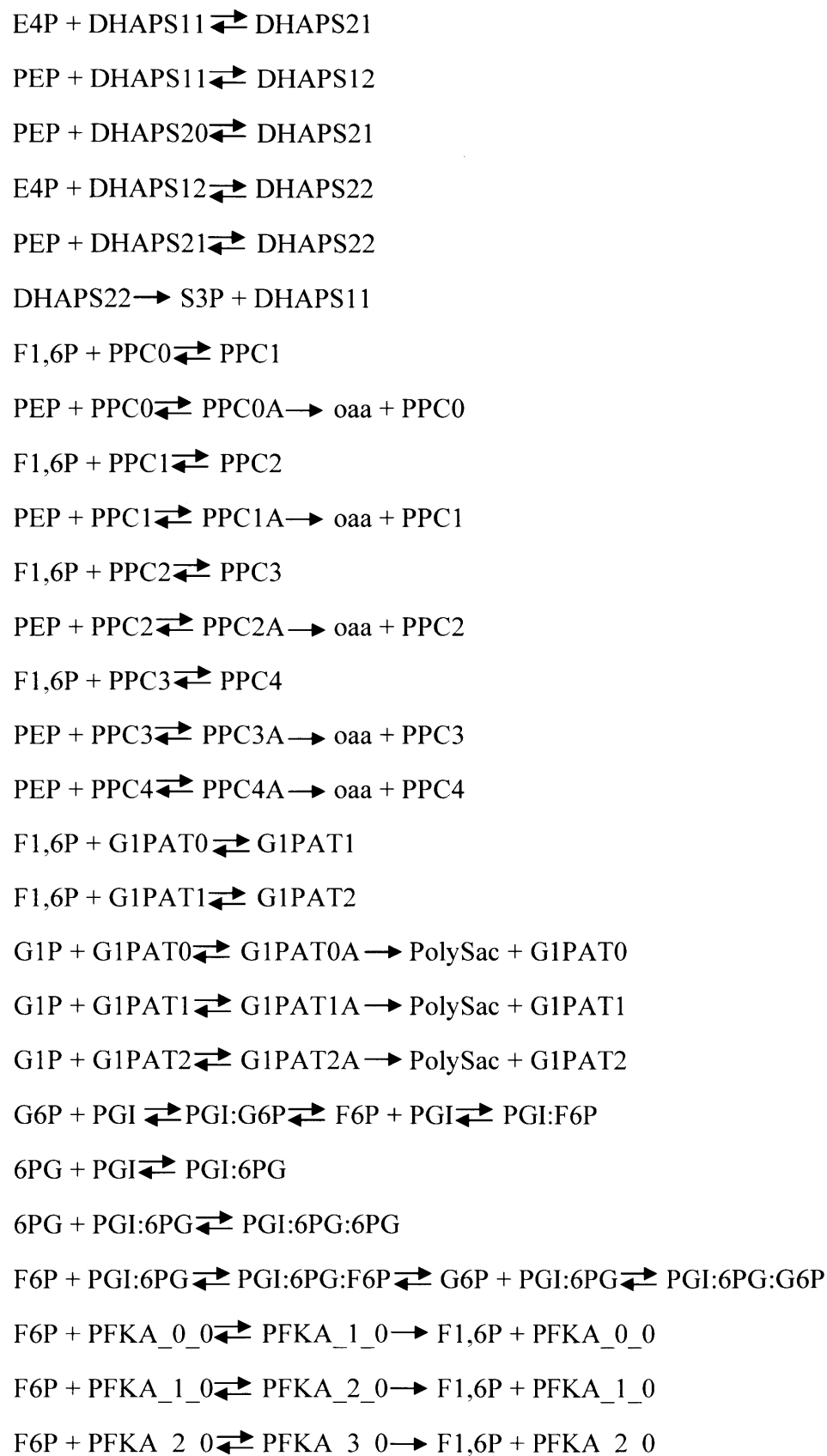
- Fantin, V. R., St-Pierre, J., & Leder, P. (2006). Attenuation of LDH-A expression uncovers a link between glycolysis, mitochondrial physiology, and tumor maintenance. *Cancer Cell*, 9(6), 425–34. doi:10.1016/j.ccr.2006.04.023
- Hsu, P. P., & Sabatini, D. M. (2008). Cancer cell metabolism: Warburg and beyond. *Cell*, 134(5), 703–7. doi:10.1016/j.cell.2008.08.021
- Koukourakis, M. I., Giatromanolaki, A., Harris, A. L., & Sivridis, E. (2006). Comparison of metabolic pathways between cancer cells and stromal cells in colorectal carcinomas: a metabolic survival role for tumor-associated stroma. *Cancer Research*, 66(2), 632–7. doi:10.1158/0008-5472.CAN-05-3260
- Kroemer, G., & Pouyssegur, J. (2008). Tumor cell metabolism: cancer’s Achilles' heel. *Cancer Cell*, 13(6), 472–82. doi:10.1016/j.ccr.2008.05.005
- Warburg, O. (1956). On the origin of cancer cells. *Science*, 123(3191).

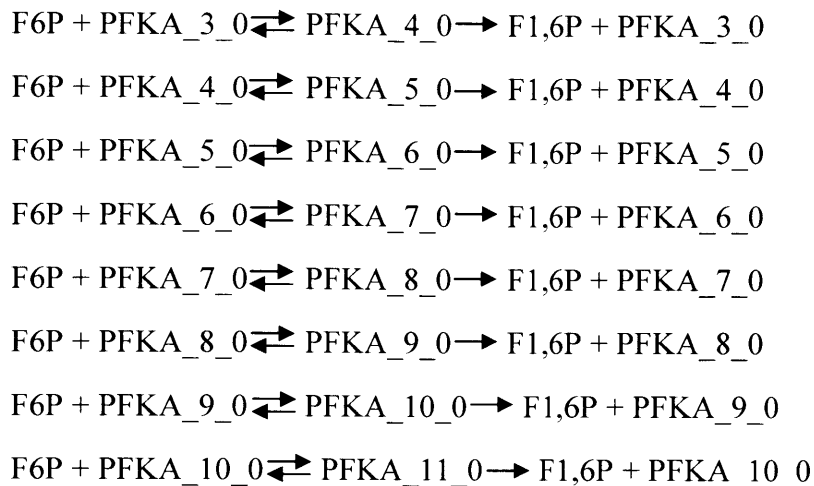
Appendix A

Enzyme reaction mechanisms in *E. coli* mass-action model

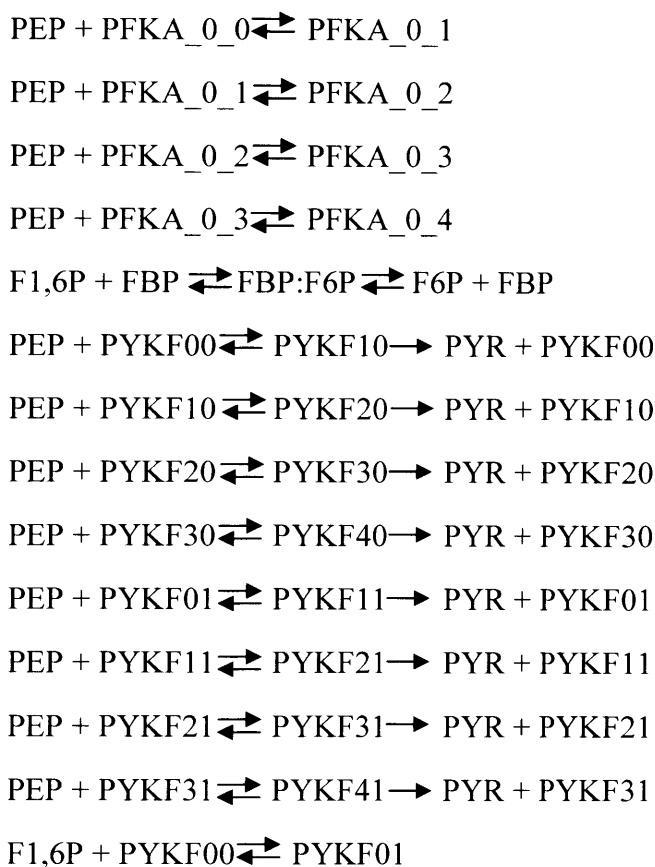


$\text{PEP} + \text{Synth1} \rightleftharpoons \text{Synth1:PEP} \rightarrow \text{cho_mur} + \text{Synth1}$
 $\text{PYR} + \text{Synth2} \rightleftharpoons \text{Synth2:PYR} \rightarrow \text{ile} + \text{Synth2}$
 $\text{G6P} + \text{ZWF} \rightleftharpoons \text{ZWF:G6P} \rightleftharpoons \text{ZWF:GL6P} \rightarrow \text{GL6P} + \text{ZWF}$
 $\text{GL6P} + \text{PGL} \rightleftharpoons \text{PGL:GL6P} \rightarrow \text{PGL} + \text{6PG}$
 $\text{6PG} + \text{GND} \rightleftharpoons \text{GND:6PG} \rightarrow \text{Ru5P} + \text{GND}$
 $\text{DHAP} + \text{TIS} \rightleftharpoons \text{TIS:DHAP} \rightleftharpoons \text{TIS:G3P} \rightleftharpoons \text{G3P} + \text{TIS}$
 $\text{3PG} + \text{GPMA} \rightleftharpoons \text{GPMA:3PG} \rightleftharpoons \text{GPMA:2PG} \rightleftharpoons \text{2PG} + \text{GPMA}$
 $\text{3PG} + \text{GPMB} \rightleftharpoons \text{GPMB:3PG} \rightleftharpoons \text{GPMB:2PG} \rightleftharpoons \text{2PG} + \text{GPMB}$
 $\text{G6P} + \text{PGM} \rightleftharpoons \text{PGM:G6P} \rightleftharpoons \text{PGM:G1P} \rightleftharpoons \text{G1P} + \text{PGM}$
 $\text{2PG} + \text{ENO} \rightleftharpoons \text{ENO:2PG} \rightleftharpoons \text{ENO:PEP} \rightleftharpoons \text{PEP} + \text{ENO}$
 $\text{G3P} + \text{GAPA} \rightleftharpoons \text{GAPA:G3P} \rightleftharpoons \text{GAPA:PGP} \rightleftharpoons \text{PGP} + \text{GAPA}$
 $\text{PGP} + \text{PGK} \rightleftharpoons \text{PGK:PGP} \rightleftharpoons \text{PGK:3PG} \rightleftharpoons \text{3PG} + \text{PGK}$
 $\text{F1,6P} + \text{FBAA} \rightleftharpoons \text{FBAA:F1,6P} \rightleftharpoons \text{FBAA:DHAP:G3P} \rightleftharpoons \text{G3P} + \text{FBAA:DHAP}$
 $\text{FBAA:DHAP} \rightleftharpoons \text{DHAP} + \text{FBAA}$
 $\text{F1,6P} + \text{FBAB} \rightleftharpoons \text{FBAB:F1,6P} \rightleftharpoons \text{FBAB:DHAP:G3P} \rightleftharpoons \text{G3P} + \text{FBAB:DHAP}$
 $\text{FBAB:DHAP} \rightleftharpoons \text{DHAP} + \text{FBAB}$
 $\text{PYR} + \text{PDH} \rightleftharpoons \text{PDH:PYR}_1 \rightarrow \text{accoa} + \text{PDH}$
 $\text{PYR} + \text{PDH:PYR}_1 \rightleftharpoons \text{PDH:PYR}_2 \rightarrow \text{accoa} + \text{PDH:PYR}_1$
 $\text{PYR} + \text{PDH:PYR}_2 \rightleftharpoons \text{PDH:PYR}_3 \rightarrow \text{accoa} + \text{PDH:PYR}_2$
 $\text{PYR} + \text{PDH:PYR}_3 \rightleftharpoons \text{PDH:PYR}_4 \rightarrow \text{accoa} + \text{PDH:PYR}_3$
 $\text{E4P} + \text{DHAPS00} \rightleftharpoons \text{DHAPS10}$
 $\text{PEP} + \text{DHAPS00} \rightleftharpoons \text{DHAPS01}$
 $\text{E4P} + \text{DHAPS01} \rightleftharpoons \text{DHAPS11}$
 $\text{PEP} + \text{DHAPS01} \rightleftharpoons \text{DHAPS02}$
 $\text{E4P} + \text{DHAPS02} \rightleftharpoons \text{DHAPS12}$
 $\text{E4P} + \text{DHAPS10} \rightleftharpoons \text{DHAPS20}$
 $\text{PEP} + \text{DHAPS10} \rightleftharpoons \text{DHAPS11}$

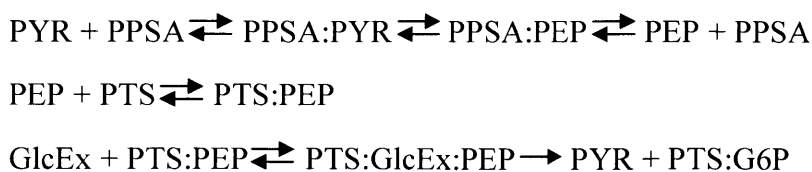




Repeat the above for PFKB



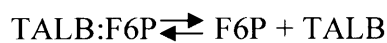
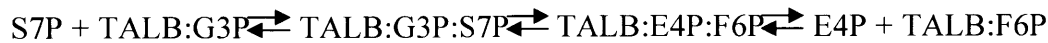
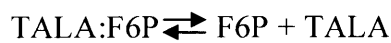
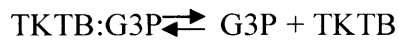
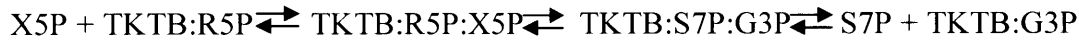
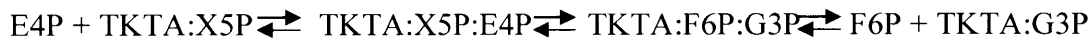
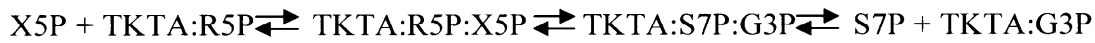
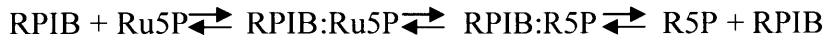
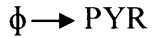
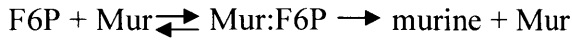
Repeat above for PYKA



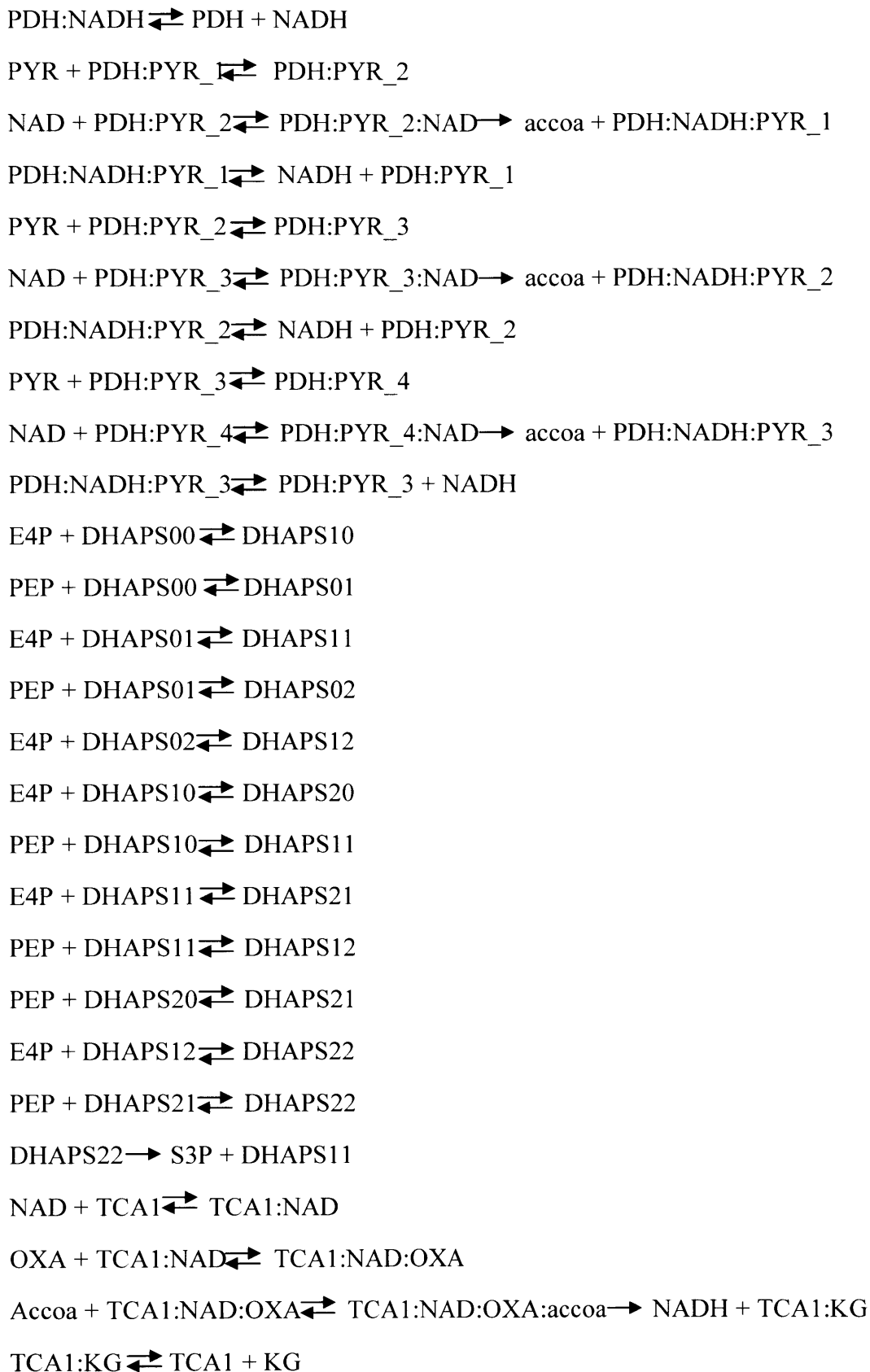
$\text{PTS:G6P} \rightarrow \text{G6P} + \text{PTS}$
 $\text{PYR} + \text{PTS} \rightleftharpoons \text{PTS:PYR}$
 $\text{G6P} + \text{PTS} \rightleftharpoons \text{PTS:I1}$
 $\text{G6P} + \text{PTS:I1} \rightleftharpoons \text{PTS:I2}$
 $\text{G6P} + \text{PTS:I2} \rightleftharpoons \text{PTS:I3}$
 $\text{G6P} + \text{PTS:I3} \rightleftharpoons \text{PTS:I4}$
 $\text{G6P} + \text{PTS:PEP} \rightleftharpoons \text{PTS:PEP:I1}$
 $\text{G6P} + \text{PTS:PEP:I1} \rightleftharpoons \text{PTS:PEP:I2}$
 $\text{G6P} + \text{PTS:PEP:I2} \rightleftharpoons \text{PTS:PEP:I3}$
 $\text{G6P} + \text{PTS:PEP:I3} \rightleftharpoons \text{PTS:PEP:I4}$
 $\text{G6P} + \text{PTS:PYR} \rightleftharpoons \text{PTS:PYR:I1}$
 $\text{G6P} + \text{PTS:PYR:I1} \rightleftharpoons \text{PTS:PYR:I2}$
 $\text{G6P} + \text{PTS:PYR:I2} \rightleftharpoons \text{PTS:PYR:I3}$
 $\text{G6P} + \text{PTS:PYR:I3} \rightleftharpoons \text{PTS:PYR:I4}$
 $\text{G6P} + \text{PTS:GlcEx:PEP} \rightleftharpoons \text{PTS:GlcEx:PEP:I1}$
 $\text{G6P} + \text{PTS:GlcEx:PEP:I1} \rightleftharpoons \text{PTS:GlcEx:PEP:I2}$
 $\text{G6P} + \text{PTS:GlcEx:PEP:I2} \rightleftharpoons \text{PTS:GlcEx:PEP:I3}$
 $\text{G6P} + \text{PTS:GlcEx:PEP:I3} \rightleftharpoons \text{PTS:GlcEx:PEP:I4}$
 $6\text{PG} + \text{EDD} \rightleftharpoons \text{EDD:6PG} \rightleftharpoons \text{EDD:2KDPG} \rightleftharpoons 2\text{KDPG} + \text{EDD}$
 $2\text{KDPG} + \text{EDA} \rightleftharpoons \text{EDA:2KDPG} \rightleftharpoons \text{EDA:PYR:G3P} \rightleftharpoons \text{G3P} + \text{EDA:PYR}$
 $\text{EDA:PYR} \rightleftharpoons \text{PYR} + \text{EDA}$
 $\text{S3P} + \text{E1} \rightleftharpoons \text{E1:S3P}$
 $\text{PEP} + \text{E1:S3P} \rightleftharpoons \text{E1:S3P:PEP} \rightarrow \text{PYR} + \text{E1:I3G}$
 $\text{E1:I3G} \rightarrow \text{I3G} + \text{E1}$
 $\text{I3G} + \text{E2} \rightleftharpoons \text{E2:I3G} \rightarrow \text{G3P} + \text{E2}$

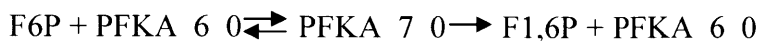
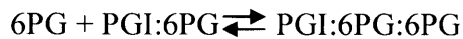
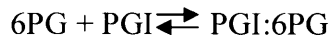
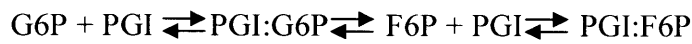
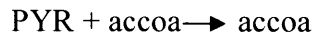
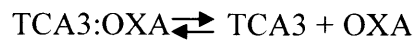
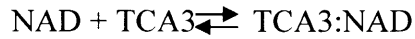
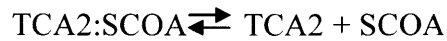
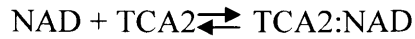
Appendix B

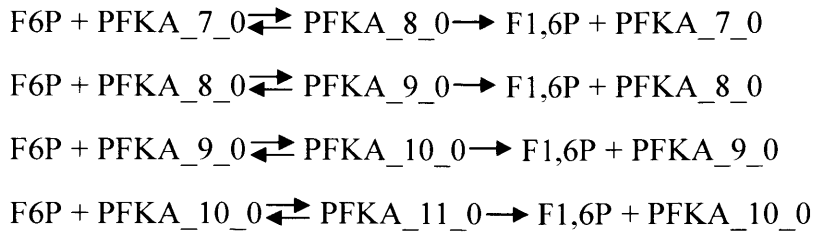
Enzyme reaction mechanisms in *Saccharomyces cerevisiae* mass-action model



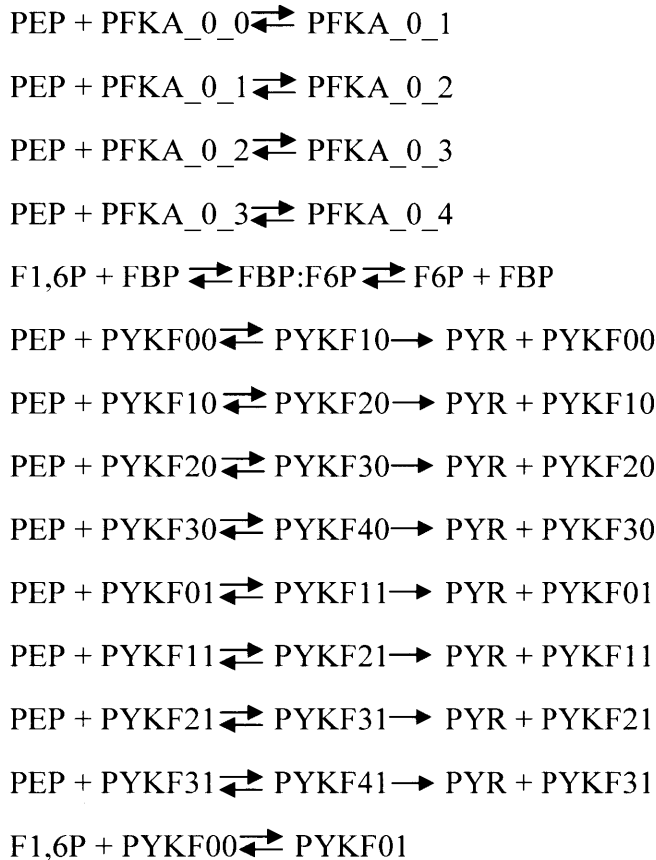
GLCE:Glycerol \rightleftharpoons Glycerol + GLCE
 3PG + SerSynth \rightleftharpoons SerSynth:3PG \rightarrow serine + SerSynth
 PEP + Synth1 \rightleftharpoons Synth1:PEP \rightarrow cho_mur + Synth1
 PYR + Synth2 \rightleftharpoons Synth2:PYR \rightarrow ile + Synth2
 NADP + ZWF \rightleftharpoons ZWF:NADP
 G6P + ZWF:NADP \rightleftharpoons ZWF:NADP:G6P \rightleftharpoons ZWF:NADPH:GL6P \rightleftharpoons NADPH + ZWF:GL6P
 ZWF:GL6P \rightleftharpoons ZWF + GL6P
 GL6P + PGL \rightleftharpoons PGL:GL6P \rightarrow PGL + 6PG
 NADP + GND \rightleftharpoons GND:NADP
 6PG + GND:NADP \rightleftharpoons GND:NADP:6PG \rightarrow NADPH + GND:Ru5P
 GND:Ru5P \rightleftharpoons Ru5P + GND
 DHAP + TIS \rightleftharpoons TIS:DHAP \rightleftharpoons TIS:G3P \rightleftharpoons G3P + TIS
 3PG + GPMA \rightleftharpoons GPMA:3PG \rightleftharpoons GPMA:2PG \rightleftharpoons 2PG + GPMA
 3PG + GPMB \rightleftharpoons GPMB:3PG \rightleftharpoons GPMB:2PG \rightleftharpoons 2PG + GPMB
 G6P + PGM \rightleftharpoons PGM:G6P \rightleftharpoons PGM:G1P \rightleftharpoons G1P + PGM
 2PG + ENO \rightleftharpoons ENO:2PG \rightleftharpoons ENO:PEP \rightleftharpoons PEP + ENO
 NAD + GAPA \rightleftharpoons GAPA:NAD
 G3P + GAPA:NAD \rightleftharpoons GAPA:NAD:G3P \rightleftharpoons GAPA:NADH:PGP \rightleftharpoons NADH + GAPA:PGP
 GAPA:PGP \rightleftharpoons GAPA + PGP
 G3P + GAPA \rightleftharpoons GAPA:G3P \rightleftharpoons GAPA:PGP \rightleftharpoons PGP + GAPA
 PGP + PGK \rightleftharpoons PGK:PGP \rightleftharpoons PGK:3PG \rightleftharpoons 3PG + PGK
 F1,6P + FBAA \rightleftharpoons FBAA:F1,6P \rightleftharpoons FBAA:DHAP:G3P \rightleftharpoons G3P + FBAA:DHAP
 FBAA:DHAP \rightleftharpoons DHAP + FBAA
 F1,6P + FBAB \rightleftharpoons FBAB:F1,6P \rightleftharpoons FBAB:DHAP:G3P \rightleftharpoons G3P + FBAB:DHAP
 FBAB:DHAP \rightleftharpoons DHAP + FBAB
 PYR + PDH \rightleftharpoons PDH:PYR_1
 NAD + PDH:PYR_1 \rightleftharpoons PDH:PYR_1:NAD \rightarrow accoa + PDH:NADH







Repeat the above for PFKB



Repeat above for PYKA

