Exploring optimal objective functions and additional constraints for flux prediction in genome-scale models

Nguyen Hoang Son

Thesis to obtain the Master of Science Degree in

Biotechnology

Supervisors:  Dr. Rafael Costa
Prof. Isabel Maria de Sá Correia Leite de Almeida

Examination Committee
Chairperson: Prof. Duarte Miguel de França Teixera dos Prazeres
Supervisor: Dr. Rafael Costa
Members of the Committee: Dr. Daniel Machado

October 2014
Acknowledgments

Apart from the efforts of myself, the success of any project depends largely on the encouragement and guidelines of many others. First and foremost, I would like to thank to my supervisor Dr. Rafael Costa for the valuable guidance and advice. This research project would not have been finished successfully without his continuously support and assistance. His enthusiastic supervision helped me in all the time of research and writing of this thesis.

I want take this opportunity to express my sincere gratitude to Prof. Erik Aurell and Prof. Isabel Sá Correia, the coordinators of euSYSBIO program, for their dedicated and constant support throughout 2 years of my studying. Prof. Isabel Sá Correia is also the co-supervisor of this thesis who helped me so much to keep the work on progress and I really appreciate that.

I would like to gratefully acknowledge the tremendous encouragement and insightful comments of Prof. Susana Vinga and her group during the time of research of this thesis. In addition, the author truly convey thanks to the officers, lab-mates at INESC-ID for providing the useful preferences and laboratory facilities.

I also wish to express love and gratitude to my beloved families; for their understanding and endless love, through the duration of my studies.

Finally, my thanks and appreciations go to my friends in Lisbon, to the city and people here for their warm hearts and all the tremendous memories with them for my meaningful time in this beautiful country that would never be forgotten.
Abstract

With the technological developments in the post-genomic era there has been an increasing focus of metabolic reconstructions for a large number of model organisms, such as, the prokaryotic *Escherichia coli*, the eukaryote *Saccharomyces cerevisiae* and human cells. Genome-scale reconstructions are usually stoichiometric and analyzed under steady-state assumption using constraint-based modelling with Flux Balance Analysis (FBA).

Several applications that use constraint-based models have been particularly successful in computationally assessing metabolic networks, as well as for predicting maximum yield of a certain desired product and analysis of gene essentially, under environmental and/or genetic perturbations. These various applications of FBA require not only the stoichiometric information of the network, but also an appropriate cellular objective function and possible additional physico-chemical constraints to compute the unique/multiple resulting flux distributions of an organism. One approach usually used to address this optimization problem and to explore the metabolic capabilities of the organism is the linear programming (LP) framework.

To compute metabolic flux distributions in microbes, the most common objective assumption is to consider the optimization of the growth rate (“biomass equation”). However, other objectives may be more accurate in predicting phenotypes. Since objective function selection seems to be, in general, highly dependent on the growth conditions, quality of the constraints and comparison datasets specific, more investigations are required for better understanding the universality of the objective function.

In this work, we explore the validity of different cases of optimization criteria and the effect of single (or combinations) cellular constraints in order to improve the predictive power of intracellular flux distribution. These comparisons were evaluated to compare predicted fluxes to published experimental $^{13}$C-labelling fluxomic datasets using three metabolic models with different conditions and comparison datasets.

It can be observed that by using different conditions and metabolic models, the fidelity patterns of FBA can differ considerably. However, despite of the observed variations, several conclusions could be drawn. First, the maximization of biomass yield achieves one of the best objective function under all conditions studied. For the batch growth condition the most consistent optimal criteria appears to be described by maximization of the biomass yield per flux or by the objective of maximization ATP yield per flux unit. Moreover, under N-limited continuous cultures the criteria minimization of the flux distribution across the network was determined as the most significant. Secondly, the predictions
obtained by flux balance analysis using additional combined standard constraints are not necessarily better than those obtained using only the single constraint. Finally, the multi-objective optimization in FBA illustrate a potential improvement of metabolic flux distribution predictions.

The experimental datasets, metabolic models and ready-to-run MATLAB scripts are freely available at [https://github.com/hsnguyen/ObjComparison](https://github.com/hsnguyen/ObjComparison).

**Keywords**

Metabolic networks, constraint-based models, flux balance analysis, objective functions, constraints, flux distribution prediction, multi-objective optimization.
Resumo

Com a evolução tecnológica na era pós-genómica tem havido um foco em crescendo de reconstruções metabólicas para um grande número de organismos, tais como a Escherichia coli, a Sacharomyces cerevisae e células humanas. As reconstruções à escala genómica são geralmente estequiométricas e analisadas assumindo estado-estacionário, utilizando modelação baseado em restrição pela análise de balanço de fluxos (flux balance analysis, FBA).

Várias aplicações que usam modelos baseados em restrições tem sido particularmente bem sucedidas na computação de redes metabólicas, bem como na previsão do rendimento máximo de um determinado produto e análise de genes essenciais por perturbações genéticas e ambientais. Estas várias aplicações de FBA requer não apenas informação estequiométrica da rede mas também uma função objetivo objectivo celular e possíveis restrições físico-químicas adicionais para computar a distribuição de fluxo única/multipla de um organismo. Um método geralmente utilizada para resolver este problema de optimização e explorar as capacidades metabólicas do organismo é a programação linear.

Para computar a distribuição de fluxos metabólicos a função objectivo mais comum é considerar a optimização da taxa de crescimento “equação da biomass”). Contudo, outros objectivos podem ser mais precisos na previsão de fenótipos. Uma vez que a função objetivo parece ser, em geral, altamente dependente das condições de crescimento, qualidade de restrições e conjunto de dados de comparação, mais investigações são necessárias para uma compreensão da universalidade da função objetivo.

Neste trabalho, foi explorado a validade de diferentes funções de optimização e o efeito de restrições celulares singulares (ou combinações) para melhorar o poder preditivo da distribuição de fluxos intracelulares. Estas comparações foram avaliadas para comparar os fluxos previstos com um conjunto de dados fluxomicos de carbono marcado ($^{13}C$) em diferentes condições, usando três modelos metabólicos. Foi possível observar que usando diferentes condições experimentais e modelos metabólicos, os perfis de fidelidade em FBA podem divergir consideravelmente. No entanto, apesar das variações observadas, várias conclusões podem ser retiradas. Em primeiro lugar, a maximização do rendimento da biomassa alcança uma das melhores funções objetivo em todas as condições estudadas. Em condições de crescimento “batch” o critério óptimo aparenta ser descrito pela maximização do rendimento da biomassa ou pelo maximização do rendimento de ATP por unidade de fluxo. Além disso, em condições de quimiostato limitado na fonte de amônia o critério de minimização a distribuição de fluxo através da rede foi determinada como mais signi-
ficativa. Em segundo lugar, as previsões obtidas por FBA usando restrições standard combinadas não são necessariamente melhores do que as restrições singulares. Finalmente, a optimização multi-objectivo em FBA ilustra uma potencial melhoria na previsão da distribuição metabólica de fluxos. O conjunto de dados experimentais, modelos metabólicos e os scripts em MATLAB estão disponíveis e de acesso livre em https://github.com/hanguyen/ObjComparison.

**Palavras Chave**

Redes metabólicas, modelos baseado em restrições, análise de balanco de fluxos, funções objective, restrições, previsão distribuição fluxo, optimização multi-objectivo.
Contents

1 Introduction 1
  1.1 Background ............................................................ 2
    1.1.1 Systems biology .................................................. 2
    1.1.2 Metabolic networks .............................................. 3
    1.1.3 Genome-scale models and metabolic engineering applications . 4
    1.1.4 Cell cultures .................................................... 5
  1.2 Fluxomics .............................................................. 5
    1.2.1 Metabolic Flux Analysis ........................................ 6
    1.2.2 Constraint-based analysis ...................................... 7
  1.3 Motivation and Objectives .......................................... 8
  1.4 Structure of the thesis ............................................. 8

2 Flux Balance Analysis 11
  2.1 General definitions .................................................. 12
  2.2 Variations of Flux Balance Analysis (FBA) .......................... 13
  2.3 Review of recent works examining objective functions ............. 17
  2.4 The COBRA Toolbox .................................................. 19

3 Methods 21
  3.1 Metabolic models and experimental data ............................. 22
  3.2 Cellular constraints .................................................. 22
  3.3 Objective functions .................................................. 23
    3.3.1 Single objective functions ..................................... 23
  3.4 Multi-objective formulations ....................................... 24
    3.4.1 Pareto optimization .............................................. 25
    3.4.2 \( \epsilon \)-constraint algorithm ................................ 25
  3.5 Implementation ....................................................... 26
  3.6 Calculating distance between predicted and experimental data .... 27

4 Results and Discussion 29
  4.1 The impact of objective functions by FBA in different conditions .... 30
    4.1.1 Carbon (C)-limited chemostat cultures ......................... 30
4.1.2 Ammonium (N)-limited chemostat cultures ........................................ 32
4.1.3 Batch cultures ............................................................................. 33
4.2 The impact of objective functions by FBA in different metabolic models ........ 33
4.3 Evaluation of pairwise constraints ....................................................... 36
4.4 Evaluation of multi-objective functions ................................................. 36

5 Conclusions and Future Work ................................................................. 39
5.1 Main conclusions and work summary ................................................... 40
5.2 Ongoing and future work ................................................................ 40

Bibliography ....................................................................................... 43

Appendix A Experimental Datasets and Matches between experimental and model reactions ......................................................... 49
Appendix B FBA simulation results ............................................................. 55
Appendix C Published abstract and article .................................................. 75
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Application of systems biology in metabolic engineering approaches</td>
<td>2</td>
</tr>
<tr>
<td>1.2</td>
<td>A typical metabolic network of <em>E. coli</em> central carbon metabolism with well-known pathways indicated. Solid arrows represent fluxes to biomass building blocks. Figure from [1].</td>
<td>3</td>
</tr>
<tr>
<td>1.3</td>
<td>Applications and number of studies of <em>E. coli</em> genome-scale models. This Figure is taken from [2].</td>
<td>4</td>
</tr>
<tr>
<td>1.4</td>
<td>An example of isotope tracing mechanism in glycolysis pathway. Isotopic carbon atoms $^{13}$C appeared in glycolytic intermediates clearly or being measured by MS or NMR spectroscopy (source [3]).</td>
<td>6</td>
</tr>
<tr>
<td>1.5</td>
<td>Six typical steps of a standard $^{13}$C-Metabolic Flux Analysis (MFA) study (source [4]).</td>
<td>7</td>
</tr>
<tr>
<td>2.1</td>
<td>An example of a network with its corresponding basic Flux Balance Analysis (taken from [5]).</td>
<td>12</td>
</tr>
<tr>
<td>2.2</td>
<td>The solution space is decreased when introducing additional criteria in the formulation of a typical [FBA] model (source [6]). Note that the optimal solution usually is non-unique.</td>
<td>14</td>
</tr>
<tr>
<td>2.3</td>
<td>An example for working mechanism of Parsimonious enzyme usage [FBA] (source [7]).</td>
<td>17</td>
</tr>
<tr>
<td>2.4</td>
<td>COBRA toolbox features overview (source [8]).</td>
<td>20</td>
</tr>
<tr>
<td>3.1</td>
<td>Map from parameter space into objective function space. The red piece of curve is noninferior solutions in this case.</td>
<td>25</td>
</tr>
<tr>
<td>3.2</td>
<td>A graphic illustration for $\epsilon$-constraint algorithm in case of bi-objective problem. Figure 3.2a presents a noninferior solution found by using $\epsilon$-constraint algorithm on a single point. Figure 3.2b shows the Pareto front of the whole problem represented by the red-highlighted region on the objective function space.</td>
<td>26</td>
</tr>
<tr>
<td>3.3</td>
<td>Workflow presents the summary performed in the thesis.</td>
<td>28</td>
</tr>
<tr>
<td>4.1</td>
<td>Predictive fidelities for objective functions using different constraints (a,c,e,g) and corresponding individual flux predictions (b,d,f,h) between predicted and experimental fluxes for all the objective functions evaluated without additional constraint of four <em>E. coli</em> (C)-limited chemostat cultures on the Genome-scale model. Missing lines represent infeasible solutions.</td>
<td>31</td>
</tr>
</tbody>
</table>
4.2 Predictive fidelities for objective functions with constraints (a,c) and corresponding individual flux predictions (b,d) between predicted and experimental fluxes for all the objective functions evaluated without additional constraint of E. coli N-limited chemostat culture conditions on the Core and Genome-scale model. Missing lines represent infeasible solutions in the simulation. 32

4.3 Predictive fidelities for objective functions with constraints (a,c) and corresponding individual flux predictions (b,d) between predicted and experimental fluxes for all the objective functions evaluated without additional constraint of E. coli batch culture conditions on the Genome-scale model. 34

4.4 Predictive fidelities of Parsimonious enzyme usage FBA (pFBA) simulation for each objective function in several datasets with three different metabolic models. Values above 1 are out of range. 35

4.5 Comparison between pFBA predictive fidelities using single (4.5a) and pairwise (4.5b) constraints for the dataset from Ishii under the dilution rate of 0.7 h⁻¹. 36

B.1 Predictive fidelities for objective functions with single constraints for E. coli Core model. 56
B.2 Predictive fidelities for objective functions with single constraints for E. coli Schuetz model. 58
B.3 Predictive fidelities for objective functions with single constraints for E. coli iAF1260 Genome-scale model. 60
B.4 Predictive fidelities for objective functions with pairwise constraints for E. coli Core model. 62
B.5 Predictive fidelities for objective functions with pairwise constraints for E. coli Schuetz model. 64
B.6 Predictive fidelities for objective functions with pairwise constraints for E. coli iAF1260 Genome-scale model. 66
B.7 Individual flux predictions between predicted and experimental fluxes for all the objective functions evaluated without additional constraint of the Core model. 68
B.8 Individual flux predictions between predicted and experimental fluxes for all the objective functions evaluated without additional constraint of the Schuetz model. 70
B.9 Individual flux predictions between predicted and experimental fluxes for all the objective functions evaluated without additional constraint of the Genome-scale model. 72
B.10 Predictive fidelities of pFBA simulation for each objective function in six remaining datasets with three different metabolic models. Values above 1 are out of range. 74
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Single objective functions tested in FBA simulation from [9]</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>Experimental datasets for <em>E. coli</em> from different literature references.</td>
<td></td>
</tr>
<tr>
<td>3.2</td>
<td>Cellular constraints used as additional information for FBA in this work. Most of them are adapted from [9].</td>
<td></td>
</tr>
<tr>
<td>3.3</td>
<td>List of single objective functions tested and related references.</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Predictive fidelities of pFBA simulation using single and multi-objective functions in the Core model.</td>
<td></td>
</tr>
<tr>
<td>4.2</td>
<td>Predictive fidelities of pFBA simulation using single and multi-objective functions in the Schuetz model.</td>
<td></td>
</tr>
<tr>
<td>4.3</td>
<td>Predictive fidelities of pFBA simulation using single and multi-objective functions in the Genome-scale model.</td>
<td></td>
</tr>
<tr>
<td>A.1</td>
<td>Experimental flux data for growth of <em>E. coli</em> in different conditions from different sources (Ishii and Emmerling).</td>
<td></td>
</tr>
<tr>
<td>A.2</td>
<td>(continue) from Nanchen, Yang, Perrenoud and Holm.</td>
<td></td>
</tr>
<tr>
<td>A.3</td>
<td>Mapping from experimental reactions to the corresponding ones in model reconstructions. The &quot;-&quot; symbol indicates opposite direction of the reaction. Note that beside this common mapping, some dataset may provide information about additional specific reactions in the model reconstruction also.</td>
<td></td>
</tr>
<tr>
<td>A.4</td>
<td>Specific reaction fluxes used in each objective function. Reactions with 'f' and 'b' suffixes mean forward and backward irreversible reactions respectively which are originated from a reversible reaction.</td>
<td></td>
</tr>
</tbody>
</table>
Abbreviations

COBRA  Constraint-Based Reconstruction and Analysis
LP  Linear Programming
MILP  Mix-Integer Linear Programming
MFA  Metabolic Flux Analysis
NLP  Non-Linear Programming
QP  Quadratic Programming
MOMA  Minimization Of Metabolic Adjustment
FCA  Flux Coupling Analysis
FBA  Flux Balance Analysis
FVA  Flux Variability Analysis
FSA  Flux Sensitivity Analysis
OMNI  Optimal Metabolic Network Identification
ROOM  Regulatory On/Off Minimization
pFBA  Parsimonious enzyme usage FBA
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Background</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Fluxomics</td>
<td>5</td>
</tr>
<tr>
<td>1.3 Motivation and Objectives</td>
<td>8</td>
</tr>
<tr>
<td>1.4 Structure of the thesis</td>
<td>8</td>
</tr>
</tbody>
</table>
Chapter 1. Introduction

1.1 Background

In this Chapter some background knowledge and general definitions related to the thesis topic is presented.

1.1.1 Systems biology

System biology is an interdisciplinary field that studies mainly the complex biological interactions systemically. The ‘systems thinking’, in contrast with the traditional ‘partitions thinking’ of reductionists, requires incredible workload of analysis and integrations that could only be accomplished thanks to our innovative computational technologies. The history of the orientation shift from regular molecular biology to systems biology has been well described in [10]. In which, the authors proposed that the contemporary holistic methodology is not only evolved from the development of reductive mainstream molecular biology to deal with bursting informations harnessed by high-throughput technologies. It results also from the lesser well-known line of works involving nonequilibrium thermodynamics that initially study novel functional states of multiple molecules interacted simultaneously.

Figure 1.1: Application of systems biology in metabolic engineering approaches.

Systems biologists usually utilize mathematical modelling methods to analyse biological interactions represented by different types of networks such as metabolic pathways, transcriptional regulations, signal transductions, etc to understand behaviour of cell as a whole. Together with the development of high-throughput technologies in molecular biology [11], the extension of computational and systems biology is becoming increasingly significant. Among cellular activities, metabolism has direct relations with phenotyping mechanisms thus very helpful in diagnosis and treatment of diseases. Metabolic engineering, genetic therapies and various other biotechnologies also benefit immensely from systematic researches in metabolic networks [12].
1.1.2 Metabolic networks

A metabolic network could be briefly depicted as the relationships between nodes of biochemical molecules (metabolites) represented by corresponding hyper edges (reactions) in a biological system. An edge in the graph is usually free of self-loops and report an irreversible flux (rate) of a reaction. Thus, a reversible one constitutes two opposite fluxes and usually being indicated by a bi-directional arrow. Internal reactions are edges that connected two groups of nodes (reactants and productions) while exchange fluxes contain only one node with the other side lies on the outer environment. An example for metabolic network of central carbon flow in *E. coli* is given in Figure 1.2.

![Figure 1.2: A typical metabolic network of *E. coli* central carbon metabolism with well-known pathways indicated. Solid arrows represent fluxes to biomass building blocks. Figure from [1].](image)

Metabolic networks always play critical roles in numerous studies about organisms since this type of network theoretically provides us the very underlying reactions controlling all the physico-chemical aspects of a cell. Although still at starting points, the bursting development in this field keeps bringing significant achievements to our knowledge about this biological world.

Recently, *omics* methods have been developing to cover all biological activities in cells. They range step-by-step from levels of nucleotide sequences (genomics) to transcriptional matters (transcriptomics), then interactions of amino acids (proteomics) and metabolites’ properties (metabolomics) [13]. However, most of them only offer qualitative *in vitro* aspects of molecular biology in the cell. An emerging application of systems biology in metabolism, known as fluxomics, is expected to quantitatively predict the flux values of metabolic reactions *in vivo* which provide the critical link between genes, proteins and the observable metabolic phenotypes. To do that, a number of modelling methods...
have been employed to tackle the problem of complicated kinetic parameters which are used to measure the absolute values of metabolic concentrations in a direct way [14–16]. These methods usually require the combination of an experimental metabolic flux analysis (Metabolic Flux Analysis (MFA)) with stoichiometric model of the network. Details for MFA and current technology will be described in the next section.

The metabolism reconstruction of prokaryotic *Escherichia coli* has been one of the first but most concerned problem and still being updated day-by-day. This is reasoned by various peer-reviewed studies related to genome-scale metabolic network reconstruction of this microbial organism in different applications [2]. On this thesis work, only investigations on this bacteria's metabolism will be established under various living conditions by different model configurations.

### 1.1.3 Genome-scale models and metabolic engineering applications

The holistic approach of Systems Biology encourages the integration of more detailed biological components to achieve comprehensive systems of living organism. The development of genome-scale metabolic reconstructions is an example for that phenomenon. From the earlier studying that utilize networks with only central carbohydrate metabolism [17, 18], successive versions with larger scales emerged and perpetually being completed [19, 20]. The expansion has been done by adding

---

**Figure 1.3:** Applications and number of studies of *E. coli* genome-scale models. This Figure is taken from [2].

---

4
novel subsystems such as fatty acid, alternate carbon metabolism or cell wall synthesis \cite{2}. The promotion of metabolic reconstructions to the genome scale stems from the boosting availability of biological information in this post-genomic era and efforts to incorporate all these literature into single systemic model. A genome-scale model, as being defined in \cite{21}, is a mathematical representation of reconstructed genome-scale metabolism which contains information about metabolites and fluxes, their stoichiometry, also the genes encoding enzymes of the metabolic network.

Figure 1.3 demonstrates six major applications of E. coli genome-scale models and the amount of researches in each of those fields. Among them, metabolic engineering is one of the most topic being concerned with the genome scale models. This field serves as a step required to design newly cell families with expected characteristics by using mathematical and experimental tools in metabolic analysis and modification \cite{22}. Therefore, a systemic modelling would shed light to the problem of complex nature of cellular metabolism and regulation in metabolic engineering. Recently, unlike traditional methods for genetic modifications such as random mutagenesis and screening, the contemporary engineering strategy is increasingly applying genome-scale metabolic reconstructions to predict cellular phenotype from a systems level before in vivo implementation \cite{2}.

1.1.4 Cell cultures

Culture in this case is considered as the environmental condition where the cells grow. Studying biological systems under various cultures is important because organisms living in different conditions utilize appropriate mechanisms to demonstrate diversified phenotypes, thus give us knowledge about adaptations and/or evolutions in biological world. There are two typical classes of cultures in research: batch and chemostat cultures. The former stands for an ideal circumstance in which living things could grow with a surplus amount of resources. The later represents more interesting and noteworthy cases that usually happen in real world: certain nutrient is restrained. This could be achieved by operating a bioreactor which consists of a continuous flow of the feed (fresh medium) coming in and an outflow of the effluent (culture liquid) being removed to keep the culture volume constant \cite{23, 24}. The addition rate of the limiting nutrient is well calibrated in order to control the growth rate of the micro-organism. When the steady state is reach, the specific growth rate ($\mu$) of the cells is the same as the dilution rate of the bioreactor ($D$).

1.2 Fluxomics

Fluxomics is a quantitative representation of metabolic state by estimating the reaction rates of genome-scale metabolism. As one of ‘omic’ studying field, it has been growing fast with numerous of approaches to be implemented. Basically, there appeared two classes of modelling methods for all techniques in fluxomics: stoichiometric (static) and kinetic (dynamic). While kinetic modelling could be applied to study the potential space of dynamic metabolic fluxes under pertubations away from steady state, this approach requires the use of complex biochemical reactions with large set of kinetic parameters which are technically not easy to determine \cite{25}. On the other hand, stoichiometric
models utilize the mass balance and other constraints in form of relative simple linear equations system which is more convenient for further analysis.

Metabolic flux analysis (MFA) and constraint-based analysis are two prominent techniques in fluxomics that take advantage of stoichiometric information of the metabolic network. While MFA allows the experimental measurement of reaction fluxes in vivo by isotope labeling, the constraint-based analysis perform predictions on reaction rates in silico by using computational modelling. Both of the analyses have their own pros and cons and integration of those two is becoming the current trend in fluxomics approaches. MFA technologies are just on the first steps to be applied to large-scale metabolic system in a high-throughput fashion. However, the experiment on living cells usually reflect more accurate values and usually used as references for the results from computer simulations.

1.2.1 Metabolic Flux Analysis

Metabolic Flux Analysis (MFA) is usually known as an experimental method used in fluxomics to measure incoming and outgoing rates of metabolites in a biological system. This definition is applied in the realm of this thesis, although not ubiquitous in the whole system biology community. In some circumstances, the term is used interchangeable with 'fluxomics' definition.

In isotopic labelling, the reactant undergone the interested reaction is marked by atom with a variation (isotope). The isotopes' location and pattern in the products (under steady state) are detected by MS (mass spectrometry) or NMR (nuclear magnetic resonance) technology thanks to their different characteristics compared to the original atoms. This information could be used to determine the passage the isotopic atom followed in the metabolic pathway and the magnitude of the flux through each reaction. Figure 1.4 demonstrates a simple example of this technique applied to study the glycolysis pathway.

13C-MFA is one of the most popular experimental MFA techniques which functions by using 13C isotope to measure intracellular metabolic fluxes. Together with the improvement of its accurateness and reliability on determining in vivo fluxes of metabolic pathways, the studies utilized 13C-MFA has
been increasing and reached a high level of maturity [4]. Recently, there are numerous of datasets on E. coli central carbon fluxes published and ready for used in modelling analysis.

**1.2. Fluxomics**

**1.2.2 Constraint-based analysis**

Metabolite mass balancing and stoichiometric information could provide the formulation of an linear equation system as a constraint. Theoretically, if this equation could be solved completely then all reaction rates in the metabolic network are determined values. However, in fact, this ideal condition is hardly happened because the number of linear equations, or constraints, is usually less than the number of variables (or fluxes). Mathematically, this linear equation systems occurring in the analysis is called under-determined or there is a degree of freedom in those flux variables space.

Because of this lack of measurement or constraints, the under-determined network could not be computed. To tackle the problem, several approached has been suggested [32]: (i) utilizing isotopic tracing experiment (ii) adding new metabolic constraints, e.g empirical knowledge about uptake rate, flux boundaries, co-metabolites constraints (NAD(P)H, ATP), etc. (iii) using optimization framework with suitable objective functions. The first approach (i) has been discussed already in the previous section. Though this method give extra measurements with high confidence, the cost for experimental and time setting made it unavailable for large system. Options (ii) and (iii) are highly feasible and still being applied for in silico studying in fluxomics. They were the reason for constraint-based analysis to be invented. This method introduced the application of optimization into metabolic network analysis together with appropriate physico-chemical constraints.

A constraint optimization problem in general, is to optimize a cost function of optimization variables subject to a set of constraints. The constraints could be represented by equalities or inequalities. The
Chapter 1. Introduction

The mathematical formulation of an optimization problem is given as follows.

\[
\min_x f(x) \\
\text{s.t.} \quad h_1(x) = 0 \\
\quad \vdots \\
\quad h_p(x) = 0 \\
\quad g_1(x) \leq 0 \\
\quad \vdots \\
\quad g_m(x) \leq 0
\]

where \( x \in \mathbb{R}^n \) is the optimization variable, \( f : \mathbb{R}^n \to \mathbb{R} \) is the cost function; \( h_1, \ldots, h_p \) are equality constraints while \( g_1, \ldots, g_m \) are inequality constraints. It is worth noting that every maximization problem can be converted to a minimization problem by changing the sign of \( f(x) \).

Usually in constraint-based analysis, the cost function is linear which represents a set of biochemical of interest and also known as objective function. The detail formulation will be given in Chapter 2. In this thesis, a typical constraint-based modeling technique, Flux Balance Analysis, is used as the main tool for predicting the flux distribution of \( E. coli \) metabolic networks.

1.3 Motivation and Objectives

Inspired by the importance of fluxomics in System Biology and the fast-growing constraint-based modeling techniques in this field, this thesis work is conducted based on the supports from previous studies. This work considers identifying objective functions and extra constraints that could be used in order to improve the predictive power of constraint-based models of metabolism, by using different metabolic networks and experimental data than previous works [9, 33, 34]. Although there appeared some studies followed this direction, the universality of the objective function of biomass growth remains an open question. Since objective functions are highly depend on the metabolic models/systems, growth conditions, model parameters and training data in general, more investigation should be established for better understanding the universality of the objective function in various conditions. This need has been recognized also by the system biology community [35, 36].

In line with that, the main goal of the current work is to explore different optimality principles and the effect of single (or combinations) empirical constraints in order to improve the predictive power of intracellular flux distribution in metabolic networks. This idea has been tested with three \( E. coli \) metabolic subsystems in different cultures and novel comparison datasets, expanding a previous evaluation [9].

1.4 Structure of the thesis

This thesis is divided into five chapters. In the present Chapter 1, the developed work is contextualized and the aims are mentioned. Chapter 2 provides a background and relevant knowledge about
Flux balance analysis, a common method in constraint-based analysis. The mathematical formulation of Flux Balance Analysis (FBA) is introduced along with an example and its derived forms in different applications are described. Finally, the recent works related to objective function studying in FBA is described. Chapter 3 starts with the description about the objective functions, cellular constraints, type of model reconstructions examined and the datasets that are used in this work. The MATLAB implementation is also introduced in this chapter. Chapter 4 presents and discusses the obtained results of all simulations. Finally, Chapter 5 provides the summary and the conclusion of the thesis, together with suggestions to extend it in the future. The supporting information are given as Appendix. Appendix A is for more detail about the used experimental flux datasets and the mapping from those to the models’ reactions. Appendix B contains all plots generated from simulation results, divided by three parts: simulations on single constraints, on pairwise constraints and relationship between predicted versus experimental fluxes represented by scatter plots. The publications based on this work are given in Appendix C.
### Contents

2.1 General definitions ................................................................. 12  
2.2 Variations of FBA ................................................................. 13  
2.3 Review of recent works examining objective functions ............... 17  
2.4 The COBRA Toolbox ......................................................... 19
2.1 General definitions

The metabolic networks usually contain only stoichiometric information and being analysed under steady-state assumption using constraint-based modelling with flux balance analysis (FBA) [37]. The constraints here refer to the different limited conditions that a given biological system must satisfy, such as physico-chemical, topological and environmental factors [5]. Steady-state assumption forms the underlined linear constraint that prevent metabolites from dilution due to growth [38].

**Figure 2.1**: An example of a network with its corresponding basic Flux Balance Analysis (taken from [5]).

FBA is a well-known method to deal with modelling and analyzing these sets of flux. In which, the flux balance and other constraints could be manifested as mathematical expressions of a constrained system that is subjected to an optimization criterion later. The reason for further optimizing step is that in fact, normally the available constrains are still not sufficient to make the system fully determined. Thus the desired flux distribution, among subspace of solutions, would be presumed to optimize a certain cellular objective which is also described as a function in this method. For example, the most common objective is to consider the maximization of the growth rate (biomass equation) in microbes to compute metabolic flux distributions [39].

An illustration about the use of FBA to a given system is shown in Fig. 2.1. This example used a toy model of metabolic network which consists of only 3 metabolites (A, B, C) and 7 fluxes (R1, ...
2.2 Variations of FBA

Among them, R1, R2, and R3 are internal fluxes while the remaining four represent drain (or exchange) reactions. At first, based on the reconstruction of the network, the stoichiometric information is extracted and stored in form of a node-arc incidence matrix. This stoichiometric matrix, denoted as $S$, has rows corresponding to metabolites and columns mapping to all fluxes involved. So that $(i,j)$ element of this matrix represents the stoichiometric coefficients of metabolite $i$ taking part in reaction $j$. The sign of this value depend on the role of the metabolite in the corresponding flux:

$$\text{sign}(S_{ij}) = \begin{cases} 
1 & \text{if metabolite } i \text{ is a product in reaction } j \\
-1 & \text{if metabolite } i \text{ is a reactant in reaction } j \\
0 & \text{if metabolite } i \text{ does not take part in reaction } j 
\end{cases}$$

For the large-scale metabolic network, $S$ contains mostly zeros among very few of non-zeros elements and thus it would be treated as sparse matrix with special operations for the sake of computational cost. The FBA problem is then formulated with the first constraint regarding the steady-state assumption, namely mass balance condition:

$$Sv = \frac{dC}{dt} = 0 \quad (2.1)$$

In which, the flux rates are limited by upper and lower bound: $v_{i}^{ub} \geq v_{i} \geq v_{i}^{lb}$ or in short with matrix notation: $v_{i}^{ub} \geq v \geq v_{i}^{lb}$.

Another constraints for certain flux could be usually included depend on the specific environmental or physico-chemical restrictions in order to characterize certain cellular network function. The mass balance equation together with these constraints form the solution space of all possible flux state points. Then FBA typically solve an optimization problem on this space, with objective functions linearly being expressed as:

$$Z = c \times v$$

where $c$ is coefficient vector that define the weight of each flux in the objective function. The parameter $c$ can take the form of simple sparse vector with only one non-zero element in case of objective involving single reaction, such as biomass yield or ATP yield [40]. If multiple fluxes involved, such as redox potential or ATP producing optimization [33], objective function $Z$ is then a linear combination of $v_{i}$ with varied coefficients $c_{i}$. In the example from Figure 2.1, $Z = v_{5}$ thus we could infer that $c = [0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0]$.

In case of more complex functions, mixed-integer or non-linear programming have to be used to solve the optimization problem [9]. The general mathematical expression for simple FBA problem is:

$$\min/\max \ (Z = f(v)) \quad (2.2)$$

$$s.t. \quad Sv = 0$$

$$v_{i}^{ub} \geq v \geq v_{i}^{lb}$$

2.2 Variations of FBA

Recently, the applications of constraint-based analysis has been expanding perpetually with expect to improve our knowledge about the complicated characteristics of metabolism in living cells.
Chapter 2. Flux Balance Analysis

Among them, numerous modelling techniques were remoulded from the idea of flux balance analysis. FBA has become an important tool for metabolic engineering and systems biology [5, 12, 41]. Next, more details of selected in silico algorithms will be described.

The original FBA is typically applied to estimate the system’s state in which certain biochemical productions and/or growth rate (biomass) are optimized. For further comprehend the property of metabolic network and relationship of reactions, there appeared some other tools such as flux variability analysis (Flux Variability Analysis (FVA)), flux coupling analysis (Flux Coupling Analysis (FCA)) and flux sensitivity analysis (Flux Sensitivity Analysis (FSA)) [28]. While FVA provides information about the variation in flux distribution when the objective is reached its optimized value due to the under-determined of the solution space [42], FCA studies the flux change of objective function in response to the perturbations of other fluxes [43] and FCA is to examine correlation between all pairwise fluxes in a metabolic network [44].

![Figure 2.2:](image)

**Figure 2.2:** The solution space is decreased when introducing additional criteria in the formulation of a typical FBA model (source [6]). Note that the optimal solution usually is non-unique.

On the other hand, several formulations of FBA had been developed to investigate the flux states of mutants created by specific genetic modifications. This stems from the fact that the objective used for wild-type might not accurately represents the behaviour of systems with gene-knockouts. There were two common frameworks usually proposed for this case, namely Minimization Of Metabolic Adjustment (MOMA) (minimization of metabolic adjustment) and Regulatory On/Off Minimization (ROOM) (regulatory on/off minimization). Both of two approaches calculated flux state of a mutant by minimizing corresponding distance metric to adjust the solutions from wild-type to mutant case. In MOMA [15], the adjustment metric is the Euclidean distance between flux state points in mutant solution space with optimal point in wild-type solution found by FBA beforehand. This could be formally expressed as a quadratic programming as follow:

$$
\text{min}(Lv + v^TQv)
$$

s.t.

$$
Sv = 0
$$

$$
v^{ub} \geq v \geq v^{lb}
$$

$$
v_j = 0, j \in A
$$

here $L = -2v^w$ where $v^w$ is the optimal solution found by FBA in wild-type strain; Q in quadratic part is simply unit matrix and indices $j$ represent reactions involved when gene is deleted.

On the other hand, ROOM [16] used the number of 'significant' flux changes from the wild-type
flux distribution as the metric and try to minimize this measure to find the flux state of mutant. The mixed-integer linear programming is used to solve this problem:

\[
\min \left( \sum_{i=1}^{m} y_i \right)
\]

s.t.

\[
Sv = 0
\]

\[
v^{ub} \geq v \geq v^{lb}
\]

\[
v_j = 0, j \in A
\]

\[
v_i - y_i(v^{ub}_i - w^u_i) \leq w^u_i
\]

\[
v_i - y_i(v^{lb}_i - w^l_i) \geq w^l_i
\]

\[
y_i \in \{0, 1\}
\]

\[
w^u_i = w_i + \delta |w_i| + \epsilon
\]

\[
w^l_i = w_i - \delta |w_i| - \epsilon
\]

where \(w^u_i\) and \(w^l_i\) are used for determine the 'significant' of flux changes, which are denoted by Boolean variable \(y_i\). The tolerance parameters \(\delta\) (multiplicative) and \(\epsilon\) (additive) are chosen regarding the trade-off between resultant fluxes and running time.

Another emerged class of constraint-based analysis takes advantage of multi-objective system. The theory is to optimize an objective function while other precursor objectives are still achieved. A typical case study is to enhancing the production rate of a desired biochemical metabolite while the growth rate is still secured by setting an additional constrain about biomass optimization (primal problem) in the system. OptKnock [45] is a well-known representative which functions such that bilevel optimization mechanism:

\[
\max v_{product} \quad (\text{OptKnock})
\]

s.t.

\[
\max v_{biomass}(\text{Primal})
\]

s.t.

\[
Sv = 0
\]

\[
v^{ub} \geq v \geq v^{lb}
\]

\[
v_{biomass} \geq v_{biomass}^{min}
\]

\[
v^y_{j} \leq y_j \leq v^x_{j} y_j
\]

\[
\sum_{j} (1 - y_j) \leq K
\]

\[
y_i \in \{0, 1\}
\]

where \(v_{biomass}^{min}\) is a minimum level of biomass production, \(v^y_j\) and \(v^x_j\) are determined by minimizing and maximizing every reaction flux subject to the constrains in Primal problem , \(K\) is the number of allowable knock-outs [45].

Using the basic concept of OptKnock framework as starting point, several constraint-based modelling methods has been invented for the purpose of network design. OptStrain, OptGene and OptReg were among them. OptStrain [46] utilized a two-step algorithm to determine the maximum yield of the desired biochemical production and at the same time, the number of required non-native reactions for
Chapter 2. Flux Balance Analysis

the system to reach this optimal state had been minimizing. OptGene [47] used genetic programming to do the optimization work to get the phenotypic objective of interest. OptReg [48] was designed to investigate the activation/inhibition and elimination reaction set needed to satisfy the level of desired phenotype.

Another noteworthy algorithm is called optimal metabolic network identification (Optimal Metabolic Network Identification (OMNI)) [49]. This tool is for identifying set of reactions in which experimental $^{13}$C-based fluxes would come to agreement with predicted values. A bilevel Mix-Integer Linear Programming (MILP) optimization is employed: the first step is to generate solution space of fluxes by FBA with biomass being optimized, and the second optimization explores the set of ‘agreed’ reactions by minimizing the discrepancy between experimental data and predicted flux distribution.

$$\min \left( \sum_{j} \omega_i \left| v_{j}^{\text{opt}} - v_{j}^{\text{exp}} \right| \right) \quad (2.6)$$

subject to:

- \( v_{\text{opt}} = \max \, v_{\text{biomass}} \)
- \( S_{v} = 0 \)
- \( v_{ub} \geq v \geq v_{lb} \)
- \( v_{d, min} \leq v_{d} \leq v_{d, max} \)
- \( v_{l} = v_{l}^{exp} \)
- \( v_{opt}^{\text{biomass}} \geq v_{min}^{\text{biomass}} \)
- \( \sum_{j} \left( 1 - y_{d} \right) \leq K \)
- \( y_{d} \in \{0, 1\} \)

in which, \( d, l, j \) denoted for indices of deleted reactions due to genetic knock-out, experimentally measured reactions and all of the reactions in the system [28]. The parameter \( \omega \) is weight vector for measured fluxes.

**Parsimonious enzyme usage FBA** is an important derivative of FBA that has been utilizing broadly in fields related to metabolic analysis of genome-scale constraint-based models [7]. This approach is originally not only used to study flux distribution of a network as a typical constraint-based modelling analysis but also be able to classify genes based on condition-specific pathway usage as predicted in silico. As the output of this method, there are five classes of genes associated with reactions that (1) are essential for optimal and suboptimal growth, (2) are inside the Parsimonious enzyme usage FBA (pFBA) optima, (3) are ELE, requiring more enzymatic steps than alternative pathways that meet the same cellular need, (4) are MLE, requiring a reduction in growth rate if used, or (5) cannot carry a flux in the given environmental condition/genotype (pFBA no-flux) [7]. An example is given in Figure 2.3. In which, Gene A, classified as MLE, represents an enzyme that uses a suboptimal co-factor to catalyze a reaction, thereby reducing the growth rate if used. Gene B, classified as pFBA no-flux, cannot carry a flux in this example since it is unable to take up or produce a necessary precursor metabolite. Genes E and F in this example require two different enzymes to catalyze the same transformation which Gene D can do alone; therefore they are classified as ELE. Gene G is
2.3. Review of recent works examining objective functions

essential, since its removal will stop the flux through all pathways. Genes C and D represent the most efficient (topologically and metabolically) pathway and therefore are part of the pFBA optima [7]. Basically this approach finds a flux distribution with minimum absolute values among the alternative optima, assuming that the cell attempts to achieve the selected objective function while allocating the minimum amount of resources (i.e. minimal enzyme usage). Mathematical formulation is given by:

\[
\min \sum_{j=1}^{m} v_{irrev,j} \\
\text{s.t. } \max v_{\text{biomass}} = v_{\text{biomass,lb}} \\
\text{s.t. } S_{irrev}v_{irrev} = 0 \\
0 \leq v_{irrev,j} \leq v_{\text{max}}
\]

where \(m\) is the number of gene-associated irreversible reactions in the network, \(v_{\text{biomass}}\) refers to the growth rate and \(v_{\text{biomass,lb}}\) is the lower bound for the biomass rate which is determined by FBA of irreversible network with \(S_{irrev}\) and non-negative steady-state flux \(v_{irrev}\). In order to achieve irreversible network from a certain one, all reversible reactions are split into two irreversible reactions.

2.3 Review of recent works examining objective functions

In numerous researches, maximizing growth rate has been assumed as the most appropriate objective function for micro-organisms; however it has been found that this may not occur in all cases [39]. Over the last years, a number of studies have been carried out to test the use of objective function optimization by FBA with different networks for predicting metabolic phenotypes. This set of studies can roughly be divided into two ways [35] (i) studies examining hypotheses on presumed cellular objective functions through comparison to measured fluxes generated by \(^{13}\)C-labeling techniques.
Chapter 2. Flux Balance Analysis

[9, 34, 50], and (ii) studies examining optimization techniques to discover or algorithmically predict biological objective functions from experimental data [51, 52].

As part of the former, an evaluation of objective functions for *Escherichia coli* central metabolism model has been established by Schuetz et al [9]. There are in total 11 objective functions, together with 8 physico-chemical constraint forming 99 simulations has been evaluated in attempt of finding the combination that could best predict the $^{13}$C-determined *in vivo* intracellular fluxes in different environment conditions. Table 2.1 shows a summary of the different comparative objective functions for FBA reported in [9].

<table>
<thead>
<tr>
<th>Objective function</th>
<th>Description</th>
<th>Mathematical definition</th>
<th>Reported performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max biomass</td>
<td>Maximization of biomass yield [19, 40, 53]</td>
<td>$\max \frac{v_{\text{biomass}}}{v_{\text{glucose}}}$</td>
<td>Good</td>
</tr>
<tr>
<td>Max ATP</td>
<td>Maximization of ATP yield [40, 54]</td>
<td>$\max \frac{v_{\text{ATP}}}{v_{\text{glucose}}}$</td>
<td>Good</td>
</tr>
<tr>
<td>$\min \sum v_i^2$</td>
<td>Minimization of $l_2$-norm of all fluxes [55, 56]</td>
<td>$\min \sum v_i^2$</td>
<td>+/-</td>
</tr>
<tr>
<td>Max ATP per flux unit</td>
<td>Maximization of ATP yield per flux unit [57]</td>
<td>$\max \frac{v_{\text{ATP}}}{\sum v_i^2}$</td>
<td>Very good</td>
</tr>
<tr>
<td>Max BM per flux unit</td>
<td>Maximization of biomass yield per flux unit.</td>
<td>$\max \frac{v_{\text{biomass}}}{\sum v_i^2}$</td>
<td>Good</td>
</tr>
<tr>
<td>Min glucose</td>
<td>Minimization of glucose consumption.</td>
<td>$\min \frac{v_{\text{glucose}}}{v_{\text{biomass}}}$</td>
<td>Useless</td>
</tr>
<tr>
<td>Min reaction step</td>
<td>Minimization of reaction steps</td>
<td>$\min \sum_{i=1}^{n} y_i^2, y_i \in {0, 1}$</td>
<td>Poor</td>
</tr>
<tr>
<td>Max ATP per reaction step</td>
<td>Maximization of ATP yield per reaction step</td>
<td>$\min \frac{v_{\text{ATP}}}{\sum_{i=1}^{n} y_i}$, $y_i \in {0, 1}$</td>
<td>Useless</td>
</tr>
<tr>
<td>Min redox potential</td>
<td>Minimization of redox potential [33]</td>
<td>$\min \frac{v_{\text{NADH}}}{v_{\text{glucose}}}$</td>
<td>Poor</td>
</tr>
<tr>
<td>Min ATP production</td>
<td>Minimization of ATP producing fluxes [33]</td>
<td>$\min \frac{v_{\text{ATP}}}{v_{\text{glucose}}}$</td>
<td>Poor</td>
</tr>
<tr>
<td>Max ATP production</td>
<td>Maximization of ATP producing fluxes [33, 58, 59]</td>
<td>$\max \frac{v_{\text{ATP}}}{v_{\text{glucose}}}$</td>
<td>Useless</td>
</tr>
</tbody>
</table>

The result shows that in unlimited resource condition (batch culture), the best objective is maximiza-
tion of ATP yield per unit of flux while in chemostat culture with limited nutrient, the linear maximization of ATP or biomass yield proved to be more suitable objectives; however in the cases tested there was still significant variation between predicted and experimental fluxes. It also worth noting that in this result the artificial constraints become unimportant if appropriated objective function is chosen in given condition. As a continuous work, a study based on multi-objective optimization theory is taken place to understand the principles behind the evolution of flux states [50]. This research pointed out a combination of three efficiency objectives, namely maximum ATP yield and maximum biomass yield (as defined in Table 2.1) with minimum sum of absolute fluxes, produced good approximation to the experimental data. It is worth noting that in this study, the $l_1$ norm (Manhattan norm) of all intracellular fluxes was used instead of $l_2$ norm (also Euclidean norm) as in the premise. The results from these two has offer useful hints to the method used in this thesis which will be revealed in the next sections.

Alternatively, others have developed algorithms to systematically identify or predict a relevant objective function using experimental data. ObjFind [51] is an optimization-based framework that created for that purpose. This in silico procedure attempts to solve the coefficients of importance, which equivalent to vector $c$ in $Z$, on reaction fluxes while keeping the divergence between resultant and experimental flux distribution to be as small as possible [51]. A bi-level optimization problem is used to describe this task: minimize the error between in vivo and in silico fluxes by quadratic programming, subject to the fundamental FBA problem. BOSS is another tool that is claimed to be able to recapitulate the actual objective including even excluded reaction from reconstruction model [52].

Another idea rising recently is that metabolism should be represented by near-optimal flux distributions, rather a single and fixed solution by introducing a novel objective function. This would form a region that regulated cellular fluxes are driven into and considered plausible without any further optimized step. PSEUDO (perturbed solution expected under degenerate optimality) [60] is created as a tool to predict suboptimal flux distribution utilizing a mathematical formulation of this theory. A good review for these objective discovery approaches could be found in [35].

### 2.4 The COBRA Toolbox

The COnstraints Based Reconstruction and Analysis (Constraint-Based Reconstruction and Analysis (COBRA)) toolbox is a set of utilities integrated as form of a software package to provide researchers with easy access to core COBRA methodologies [8]. The openCOBRA project has been developing based on that idea. Starting with tools for MATLAB, it now has grown to include Python modules and still on the way of improving itself to deal with the next complex COBRA models.

In this thesis, COBRA toolbox for MATLAB will be employed to deal with the model reconstructions and optimization framework of FBA. This toolbox offers the processing of SBML files of metabolic network structure and also several functions that are able to deal with flux analysis tasks. Especially, the MATLAB function `optimizeCbModel` could be invoked for FBA simulations and its variant with appropriate additional model settings. It supports optimization on linear programming (Linear Programming (LP), mix integer linear programming (MILP) and quadratic programming (Quadratic
Chapter 2. Flux Balance Analysis

Programming (QP) with suitable solvers integrated. The functionality for non linear programming (Non-Linear Programming (NLP)) is currently still on progress. For more details see [8, 14].

Figure 2.4: COBRA toolbox features overview (source [8]).
3 Methods

Contents

3.1 Metabolic models and experimental data .................................................. 22
3.2 Cellular constraints .................................................................................. 22
3.3 Objective functions .................................................................................. 23
3.4 Multi-objective formulations ..................................................................... 24
3.5 Implementation ......................................................................................... 26
3.6 Calculating distance between predicted and experimental data ............. 27
Chapter 3. Methods

3.1 Metabolic models and experimental data

In present thesis two published condensed and one genome-scale metabolic reconstruction of *E. coli* were selected as the metabolic models for flux analysis. The first condensed genome-scale metabolic model (*Core Model*) [6] contains 72 metabolites forming in total a set of 95 chemical reactions which consist of 75 internal and 20 drain fluxes. The second model (*Schuetz model*) [9] includes 98 reactions and 60 metabolites. The genome-scale model reconstruction iAF1260 [20] was also used in this work (*Genome-scale model*). This model contains in total 1668 metabolites and 2382 reactions.

The 12 experimental datasets used to compare with the model predictions consists of measurement of fluxes and extracellular rates in different cultures extracted from $^{13}$C-MFA analyses on *E. coli*. These datasets are available at KiMoSys repository [61] and come from various literature sources [1, 62–66] to diversify the test environments (chemostat aerobic N- and C-limited, batch aerobic), as described in Table 3.1.

Table 3.1: Experimental datasets for *E. coli* from different literature references.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Dilution rate</th>
<th>Environment</th>
<th>$a$m$_{core/gc}$</th>
<th>$b$m$_{Schuetz}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nanchen et al. [63]</td>
<td>0.09h$^{-1}$</td>
<td>chemostat, C-limited</td>
<td>38</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>0.4h$^{-1}$</td>
<td>chemostat, C-limited</td>
<td>38</td>
<td>66</td>
</tr>
<tr>
<td>Emmerling et al. [1]</td>
<td>0.09h$^{-1}$</td>
<td>chemostat, N-limited</td>
<td>33</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>0.09h$^{-1}$</td>
<td>chemostat, C-limited</td>
<td>33</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>0.4h$^{-1}$</td>
<td>chemostat, C-limited</td>
<td>33</td>
<td>49</td>
</tr>
<tr>
<td>Yang et al. [66]</td>
<td>0.1h$^{-1}$</td>
<td>chemostat, C-limited</td>
<td>32</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>0.55h$^{-1}$</td>
<td>chemostat, C-limited</td>
<td>32</td>
<td>47</td>
</tr>
<tr>
<td>Ishii et al. [62]</td>
<td>0.1h$^{-1}$</td>
<td>chemostat, C-limited</td>
<td>47</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>0.4h$^{-1}$</td>
<td>chemostat, C-limited</td>
<td>47</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>0.7h$^{-1}$</td>
<td>chemostat, C-limited</td>
<td>47</td>
<td>67</td>
</tr>
<tr>
<td>Perrenoud et al. [65]</td>
<td>0.62h$^{-1}$</td>
<td>batch, aerobe</td>
<td>36</td>
<td>57</td>
</tr>
<tr>
<td>Holm et al. [64]</td>
<td>0.67h$^{-1}$</td>
<td>batch, aerobe</td>
<td>37</td>
<td>60</td>
</tr>
</tbody>
</table>

$a$m$_{core/gc}$: number of matched reactions for the *E. coli Core and Genome-scale model*.  
$b$m$_{Schuetz}$: number of matched reactions for the *Schuetz model*.

Usually the number of fluxes measured by these $^{13}$C-MFA is less than those of the model reconstructions. In order to evaluate the performance of FBA simulations, there needed a matching step from the predicted flux distribution to the reference dataset. The mapping of available measured fluxes to corresponding reactions in FBA model is given in Appendix A.

3.2 Cellular constraints

One problem of using FBA is that the simulation usually returns predictions which are diverse from reality. To solve this, biological or physiochemical restrictions are added to limit the solution space of the model. We systematically tested 6 single (or combined) constraints: P-to-O (P/O) ratio was set to 1 [67] by setting the energy-coupling NADH dehydrogenase I (Nuo) and cytochrome oxidase bo3 equal to zero [9]; the upper limit bound for the maximal oxygen consumption rate ($q_{O_{2, max}}$) was set to 15 mmol/gh as experimentally reported for chemostat cultures [62]; the bounds on cellular maintenance
energy (ATPM reaction) as the requirement for growth-independent was left at the default model value of 8.39 mmol/gh [6]; bound all fluxes (bounds) to maximal 200% of the glucose uptake rate as observed experimentally [68] and 35% of NADPH overproduction compared to the NADPH requirement for biomass production. Finally the combination of all above constraints is also included. They are empirical conditions which are synthesized gradually from literature and former researches as could be seen from Table 3.2

Empirical constraints related to the bounds of fluxes given in the original models were discarded and at least the relative substrate uptake rate (glucose) was constrained in each case.

Table 3.2: Cellular constraints used as additional information for FBA in this work. Most of them are adapted from [9].

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Meaning</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>P/O = 1</td>
<td>P-to-O ratio is set to unity due to known coupling efficiencies and expression levels of respiratory chain components [9]</td>
<td>[67]</td>
</tr>
<tr>
<td>qO₂ max ≤ 15</td>
<td>Maximum of oxygen uptake is set to 15 mmol/gh as another level of limitation about oxygen utilization. [17]</td>
<td></td>
</tr>
<tr>
<td>Maintenance constraint</td>
<td>ATPM reaction flux is set to 8.39 mmol/gh empirically as the requirement for growth-independent maintenance of the cell. [18, 63]</td>
<td></td>
</tr>
<tr>
<td>Bound</td>
<td>Bound all fluxes to maximal 200% of the glucose uptake rate as observed by experiments. [1, 68–70]</td>
<td></td>
</tr>
<tr>
<td>NADPH</td>
<td>35% of NADPH overproduced than needed for biomass production. [63]</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>Apply all above constraints.</td>
<td></td>
</tr>
</tbody>
</table>

3.3 Objective functions

This section presents the objective functions tested in the FBA models. There were not only the single objective function being examined, but the combinations of up to three of them were also involved at the same time in multi-objective optimization.

3.3.1 Single objective functions

For the experimental datasets selected we evaluated the predictive ability of eight different objective functions: maximization of biomass yield \(\max BM\) and ATP (maintenance reaction) \(\max ATP\), minimization of overall intracellular fluxes \(\ell_1\)-norm \(\min \sum |v|\), maximization of ATP and biomass yield per unit flux \(\max ATP / \sum v\) and \(\max BM / \sum v\), respectively, minimization of redox potential \(\min RD\), and minimization and maximization of ATP producing fluxes per unit substrate \(\min \sum ATP / S\) and \(\max \sum ATP / S\), respectively. The objective functions correspond to the most significant performed in [9, 50] and were mathematical defined as in Table 3.3.
Table 3.3: List of single objective functions tested and related references.

<table>
<thead>
<tr>
<th>Objective function</th>
<th>Description</th>
<th>Mathematical definition</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max BM</td>
<td>Maximization of biomass yield</td>
<td>( \text{max} \frac{\text{biomass}}{v_{\text{glucose}}} )</td>
<td>[19, 40, 53]</td>
</tr>
<tr>
<td>Max ATP</td>
<td>Maximization of ATP yield.</td>
<td>( \text{max} \frac{\text{ATP}}{v_{\text{glucose}}} )</td>
<td>[40, 54]</td>
</tr>
<tr>
<td>Min Flux</td>
<td>Minimization of ( l_1 )-norm of all fluxes.</td>
<td>( \text{min} \sum</td>
<td>v_i</td>
</tr>
<tr>
<td>Max ATP/flux</td>
<td>Maximization of ATP yield per flux unit.</td>
<td>( \text{max} \frac{\text{ATP}}{\sum v_i} )</td>
<td>[57]</td>
</tr>
<tr>
<td>Max BM/flux</td>
<td>Maximization of biomass yield per flux unit.</td>
<td>( \text{max} \frac{\text{biomass}}{\sum v_i} )</td>
<td>[9]</td>
</tr>
<tr>
<td>Min Rd</td>
<td>Minimization of redox potential.</td>
<td>( \text{min} \sum \frac{v_{\text{NADH}}}{v_{\text{glucose}}} )</td>
<td>[33]</td>
</tr>
<tr>
<td>Min ATPprod</td>
<td>Minimization of ATP producing fluxes.</td>
<td>( \text{min} \sum \frac{v_{\text{ATP}}}{v_{\text{glucose}}} )</td>
<td>[33]</td>
</tr>
<tr>
<td>Max ATPprod</td>
<td>Maximization of ATP producing fluxes.</td>
<td>( \text{max} \sum \frac{v_{\text{ATP}}}{v_{\text{glucose}}} )</td>
<td>[33, 58, 59]</td>
</tr>
</tbody>
</table>

For the two nonlinear objective functions \( \text{max} \frac{\text{ATP}}{\text{flux}} \) and \( \text{max} \frac{\text{BM}}{\text{flux}} \), the \( l_2 \)-norm (Euclidean norm) of the fluxes vector was used in accordance with [9] while \( \text{min} \text{Flux} \) refers to the \( l_1 \)-norm (Manhattan norm) of the same vector as in [50]. This is because using \( l_1 \)-norm reported better predictive results than that of \( l_2 \)-norm when it comes to minimization-of-all-fluxes objective function [50].

For linear optimization, there are usually non-unique sets of flux values \( v \) that give the same optimal value of objective function [71]. To avoid typical degeneracy of FBA solutions, the principle of parsimonious enzyme usage (pFBA) was used [7]. The detail formation of this constraint-based analysis has been given in Chapter 2. However, the purpose of pFBA application here is not necessarily to classify gene into five categories. Adapted to this work, the optimization problem is mathematically described as follows:

\[
\begin{align*}
\text{min } & \sum_{i=1}^{n} |v_i| \\
\text{s.t. } & Z = Z_{\text{optima}} \\
& S.v = 0 \\
& v^{ub} \geq v \geq v^{lb}
\end{align*}
\]

3.4 Multi-objective formulations

Usually in nature, optimization of one single objective function could not represent completely the behaviour of cell evolution. For example, cells tend to freely maximize growth rate in batch culture but within resource scarce environment, they also have to take the optimization of food usage into consideration. For that reason, several objective functions should be put together for studying at the same time. One possible method to deal with this problem is to use multi-objective or Pareto optimization instead of the single objective.
3.4 Multi-objective formulations

3.4.1 Pareto optimization

A mathematical formulation of Pareto optimization could be described as follows:

\[
\begin{align*}
\min / \max (Z_1, Z_2, ..., Z_n) \\
\text{s.t.} & \quad S v = 0 \\
& \quad v^{ub} \geq v \geq v^{lb}
\end{align*}
\] (3.2)

where \( Z_1, Z_2, ..., Z_n \) form the set of objective functions that must be targeted simultaneously in this problem. By doing this, it allows the trade-off between objective functions and since they usually competitive, there is no unique solution. The solutions of particular problem should be a set of varied \( Z_i \) and is known as noninferiority solutions set or simply Pareto front. It is defined as the solutions that improvement in one objective requires a degradation of another.

![Figure 3.1: Map from parameter space into objective function space. The red piece of curve is noninferior solutions in this case.](image)

Mathematical definition: \( v^* \in \Omega \) is a noninferior solution if there does not exist \( \Delta v \) such that \((v^* + \Delta v) \in \Omega \) and \( k, l \in \{1, ..., n\} \) such that

\[
\begin{align*}
Z_k(v^* + \Delta v) & \leq Z_k(v^*) \\
Z_l(v^* + \Delta v) & < Z_l(v^*)
\end{align*}
\] (3.3)

An example of Pareto optimization with only two objective functions of two variables is shown in Figure 3.1.

One method to find Pareto front of multiple objective functions optimization is the \( \epsilon \)-constraint algorithm. In this thesis, Pareto optimality is applied in FBA model with up to the three best single linear objective functions in each growth condition.

3.4.2 \( \epsilon \)-constraint algorithm

The method used in this thesis to calculate the Pareto front of multiple objective system is based on the idea of \( \epsilon \)-constraint algorithm [72] which is also the method used in [50]. \( \epsilon \)-constraint involves
Figure 3.2: A graphic illustration for $\epsilon$-constraint algorithm in case of bi-objective problem. Figure 3.2a presents a noninferior solution found by using $\epsilon$-constraint algorithm on a single point. Figure 3.2b shows the Pareto front of the whole problem represented by the red-highlighted region on the objective function space.

minimizing a primary objective while expressing the others in the from of inequality constraints:

$$\min_{v \in \Omega} Z_p(v)$$

s.t. $Z_i(v) \leq \epsilon_i$

$i = 1, \ldots, n, i \neq p$

where $Z_p$ is the primary objective to considered among the set of all functions $(Z_1, \ldots, Z_n)$. An example for the case of bi-objective Pareto optimization by using $\epsilon$-constraint is given in Figure 3.2a. In this illustration, $Z_1$ is the primary objective.

In order to adapt this method to solve the cases of interest, a numerical algorithm is employed. In which, the possible ranges of all non-primary objective functions are divided by $n_{\text{step}} = 1000$ then $\epsilon$-constraint will be applied to every $\epsilon \in \mathbb{R}^{n-1}$ retrieved by those partitions.

$$\epsilon = \{\epsilon_1, \ldots, \epsilon_{p-1}, \epsilon_{p+1}, \ldots, \epsilon_n\}$$

$$\epsilon_i = Z_{i,lb} + k(i) \frac{(Z_{i,ub} - Z_{i,lb})}{n_{\text{step}}}, \quad k(i) \in \{1, \ldots, n_{\text{step}}\}$$

$i = 1, \ldots, n, i \neq p$

where $Z_{i,lb}$ and $Z_{i,ub}$ is lower and upper bound of $Z_i$ respectively. A graphic illustration is shown in Figure 3.2b for the Pareto front of two objective functions. Due to the exponential iterations needed in $\epsilon$-constraint algorithm, the number of objective considered in Pareto optimization is limited to three. The only three single objective being chosen for this simulation are $\max BM$, $\max ATP$ and $\min Flux$.

### 3.5 Implementation

All calculations were implemented in Matlab 2012b (Mathworks Inc. Software) and simulations were performed using the Constraint-Based Reconstruction and Analysis (COBRA) toolbox (v. 2.0.5) [8]. In terms of optimization solvers, GLPK [73] was used for linear problems.
3.6 Calculating distance between predicted and experimental data

For the two non-linear non-convex objective functions, a numeric approximation method was used. For example, with the objective function \( \max \frac{\text{ATP}}{\text{flux}} \), firstly the range of ATP flux is calculated then 1000 value points are uniformly selected along the distance. After that, for each point, the ATP reaction rate is constrained to this value and the \( \ell_2 \)-norm of all fluxes in this system \( \sum v_i^2 \) is minimized simultaneously. The maximum of all 1000 ratios between ATP fluxes with their corresponding minimized flux norms is a good approximation for the objective \( \max \frac{\text{ATP}}{\text{flux}} \). Similar approach was used for \( \max \frac{\text{BM}}{\text{flux}} \).

Simulations were executed in parallel on a server machine with 8 AMD processors of 2.3GHz each. The libSBML [74] was the package used for reading SBML model files. Experimental datasets, metabolic models and ready-to-run matlab scripts are freely available at [https://github.com/hsnguyen/ObjComparison](https://github.com/hsnguyen/ObjComparison).

3.6 Calculating distance between predicted and experimental data

To evaluate the prediction ability, the definition of predictive fidelity [9] was used. This error is defined by:

\[
\begin{align*}
    d(v_{\text{comp}}, v_{\text{exp}}) &= \epsilon^T W \epsilon \\
    \epsilon &= \frac{v_{\text{comp}} - v_{\text{exp}}}{\sigma_{\text{exp}}} \\
    W_{i,i} &= \frac{1}{\sigma_i^2} \left( \sum_i \frac{1}{\sigma_i^2} \right)^{-1}
\end{align*}
\]

where \( d(v_{\text{comp}}, v_{\text{exp}}) \) is the standardized Euclidean distance between predicted fluxes \( v_{\text{comp}} \) and the experimental \textit{in vivo} \( v_{\text{exp}} \) fluxes weighted by their experimental variances \( \sigma_{\text{exp}} \). The set of compared reactions are given as Appendix A. Smaller predictive fidelity represent a better agreement between computational predicted and experimental fluxes.

For the original definition of predictive fidelity in [9], there are different Euclidean distances to the \textit{in vivo} fluxes due to alternative optima. Consequently the fidelity “range” is defined by the minimum and maximum Euclidean distances from predicted fluxes and corresponding \textit{in vivo} fluxes.

\[
\begin{align*}
    \min / \max d(v, v_{\text{exp}}) \\
    \text{s.t.} \quad Z = Z_{\text{optima}} \\
    S v = 0 \\
    v_{ub} \geq v \geq v_{lb}
\end{align*}
\]

where \( d(v, v_{\text{exp}}) \) is the Euclidean distance between fluxes \( v \) and the available \textit{in vivo} \( v_{\text{exp}} \). In this thesis, by applying pFBA in every simulations, the alternative optima phenomenon is efficiently neutralized and the fidelity “range” could be considered as a single erroneous value of the prediction.

A summary of the method applied in this thesis is shown in Figure 3.3. In total, there are three model reconstructions, each of them would be tested with 8 single objective functions together with 7 single cellular constraints plus 10 pairwise constraints. Thus for the first round there would be
Figure 3.3: Workflow presents the summary performed in the thesis.

$3 \times 8 \times 17 = 408$ simulations needed to be executed. The simulations for the multi-objective models is computationally expensive due to the application of Pareto optimization problem.
Results and Discussion

Contents

4.1 The impact of objective functions by FBA in different conditions .......................... 30
4.2 The impact of objective functions by FBA in different metabolic models ............... 33
4.3 Evaluation of pairwise constraints ............................................................................. 36
4.4 Evaluation of multi-objective functions ................................................................. 36
Chapter 4. Results and Discussion

FBA uses optimization of cellular criteria to find a subset of optimal states in the possibly large solution space shaped by mass balance and capacity constraints [19]. The cellular objective functions for an organism are strongly dependent on the input data, comparison-data and size/type of the metabolic models. Thus, across three different metabolic models and using available intracellular flux measurements for *E. coli*, we examine the optimal criteria derived in a previous study [9]. Furthermore, the effect of single (or pairs) standard constraints and multiple objective functions on FBA are also evaluated. The metabolic models used are three stoichiometric models of *E. coli* metabolism [6, 9, 20] and the flux distributions for eight different objective functions. The experimental conditions included batch cultures as well as chemostat cultures under aerobic glucose and ammonium limitation.

The results obtained in the FBA simulations for each case were ordered according to the fidelity error (Equation 3.4) between predicted and experimental fluxes. Thus the estimation of predictive result is based on two perspectives: (i) the predictive errors represented by plots for different simulations and (ii) scatter plots of separated flux-by-flux comparisons to see how each experimental flux distribution matched to the corresponding predictions.

4.1 The impact of objective functions by FBA in different conditions

The following section describes the predictive fidelities on three *E. coli* metabolic models for various objective and constraint combinations in different experimental conditions. It can be observed that for simulations using different growth conditions, the corresponding error patterns can differ considerably. The 8 objective functions showed a great difference in the accuracy.

4.1.1 Carbon (C)-limited chemostat cultures

Under nutrient scarcity (chemostat cultures) in glucose-limited conditions, the linear maximization of the biomass yield (\(\text{max} \, \text{BM}\)) achieved the best predictive accuracy comparing to the other objective functions and for the most of constraints (Figure 4.1 and the remaining datasets and models are given in Appendix 5). For the "none" constraint case, the predicted flux values are plotted against the experimental flux values (Figure 4.1b, 4.1d, 4.1f and 4.1h). The predictive errors of \(\text{max} \, \text{BM}\) on average always reported the lowest values when compared to other objective functions. In fact, by comparing the predicted fluxes against the experimental flux values (Figure 4.1b, 4.1d, 4.1f and 4.1h), it is clear that for the \(\text{max} \, \text{BM}\) objective function, there is a correlation which reflects the highly agreement between *in silico* to *in vivo* fluxes. This result agrees with previous works that supported the use of maximization of biomass as the most important objective function for FBA in continuous cultures [1, 9, 40].

Another interesting criteria is the parsimony criteria (\(\text{min} \sum |v|\)). The finding that minimization of the overall intracellular flux may play a key role, as shown for the genome-scale model (e.g. Figure 4.1c), is in line with a previous work for hybridoma cells in continuous culture [55]. The two nonlinear
4.1. The impact of objective functions by FBA in different conditions

Figure 4.1: Predictive Fidelities for objective functions using different constraints (a,c,e,g) and corresponding individual flux predictions (b,d,f,h) between predicted and experimental fluxes for all the objective functions evaluated without additional constraint of four E. coli (C)-limited chemostat cultures on the Genome-scale model. Missing lines represent infeasible solutions.
objectives \( \text{max } BM/\text{flux} \) and \( \text{max } ATP/\text{flux} \) were also proved to be useful for this growth condition while \( \text{max } ATP_{\text{prod}} \) in all cases revealed large discrepancies between \textit{in silico} and \textit{in vivo} fluxes.

Moreover, the effects of each of the single cellular constraints to every cellular objective is also shown in the Figure 4.1. With the exception of \( \text{max } BM, \text{max } ATP \) and \( \text{min } \text{Flux} \), there are no significant changes in the predictive fidelities for all the constraints tested. Our analysis indicated that the single standard constraints namely none, \( q_{O_2} \), \textit{maintenance energy} or \( \text{NADPH} \) reported better predictive fidelities than the other constraints. Although common properties with the previous work [9], some relative alterations in this work can be observed for the \textit{Schuetz} model and Nanchen dataset. For instance, the behaviour of maximization of ATP per unit flux was consistent and able to work well instead of being useless without additional constraints (see Appendix B Figure B.2). On the other hand, the maximization of ATP was not a good choice as expected in continuous cultures. The probable reason could be the use of the relative flux values instead of the split flux ratios used by \textit{Schuetz} et al [9] when comparing computational and experimental results.

### 4.1.2 Ammonium (N)-limited chemostat cultures

The predictive fidelities for the N-limited chemostat cultures are shown in Figure 4.2 (remaining datasets are given in Appendix B). For the N-limited chemostat culture, a similar pattern was observed

![Figure 4.2: Predictive fidelities for objective functions with constraints (a,c) and corresponding individual flux predictions (b,d) between predicted and experimental fluxes for all the objective functions evaluated without additional constraint of \textit{E. coli} N-limited chemostat culture conditions on the \textit{Core} and \textit{Genome-scale} model. Missing lines represent infeasible solutions in the simulation.](image-url)
4.2 The impact of objective functions by FBA in different metabolic models

at the simulation results. The $\text{max BM}$ was located at the group of top best objective functions but less accurate than the previous aerobic chemostat cultures. The $\text{max ATP}_{\text{prod}}$ objective function provided the worse prediction of flux distribution for the Core and Genome-scale model (Figure 4.2). Instead, the performance of two non-linear objective and minimization of the net of fluxes were improved to the same level of $\text{max BM}$. However, when considering scatter plots on the right then $\text{max BM}$ objective function, without any additional constraints, gives a higher prediction and will be most useful.

Overall, as shown in Figure 4.2a,4.2c the best objective for the ammonium-limited continuous culture was obtained by minimization of flux distribution or by maximization of the biomass. However when considering the individual flux prediction as in scatter plots (Figure 4.2b,4.2d), minimization of flux distribution did not perform significantly as expected.

Unlike the C-limited conditions, some constraints improved the prediction power of FBA compared to “none” scenario. However, the pattern was not consistent through all objectives or model reconstructions that were tested. For example, in Figure 4.2a the “maintenance” constraint indeed improved the predictive fidelity on $\text{max BM}$, but is useless for other objectives (e.g. $\text{max ATP}$).

4.1.3 Batch cultures

The predictive fidelities for the batch cultures in the genome-scale model are shown in Figure 4.3. For batch cultures the results suggest that the cells “prefer” the maximization of biomass yield per flux unit. In contrast to a previous work [33], it can be observed that the minimization of redox potential did not reflect a good objective function for E. coli batch cultures. Furthermore, the minimization and maximization of ATP producing fluxes were the worst choices. The $\text{max BM}$ and $\text{max BM}/\text{flux}$ objective were the three best functions for the simulation of E. coli batch cultures.

Additionally, consistent and promising fidelities on the two batch cultures (Holm and Perrenoud dataset) were observed also for $\text{max ATP}/\text{flux}$. A similar result was reported by Schuetz et al. [9]. However, some differences to their conclusions can be observed. First, the $\text{max BM}/\text{flux}$ was slightly better than $\text{max ATP}/\text{flux}$ making this objective function the best choice, rather than the reported $\text{max ATP}/\text{flux}$ for the batch culture. Second, in contrast to previous work [9], $\text{max ATP}_{\text{prod}}$ was considered a bad objective function in batch cultures (Figure 4.3).

Overall, the two nonlinear objectives seem to be reliable objectives in the cases tested, while maximization of biomass and minimization of net flux also presents consistent performances. Additionally, in accordance with results from Schuetz et al [9], the effect of the constraints appeared to be mostly insignificant to the choice of the objective function for all the conditions.

4.2 The impact of objective functions by FBA in different metabolic models

We next examine whether any of these findings are consistent across different metabolic models of the same system. Here, predictive fidelities were presented to focus the comparison of three different metabolic models (Core model, Schuetz model and iAF1260 Genome-scale model).
Figure 4.3: Predictive fidelities for objective functions with constraints (a,c) and corresponding individual flux predictions (b,d) between predicted and experimental fluxes for all the objective functions evaluated without additional constraint of *E. coli* batch culture conditions on the Genome-scale model.

Figure 4.4 shows the effect of the metabolic models type and size on the qualitative and quantitative predictions for the same dataset. It is clear that the predictive fidelity patterns varies depending on the metabolic network and the dataset used. In general, for the chemostat cultures (Figure 4.4a 4.4b 4.4c 4.4d), the *Genome-scale* model reported the best fidelities since their prediction errors seems to be smallest. A possible reason is that for more detailed model reconstruction a better agreement between predictive and experimental fluxes is obtained. When comparing the predictive performance of FBA simulations on the *Core* and *Schuetz* model, there were similarities beside several discrepancies. These two reconstructions are on the same level of scale and share the common metabolic structure, but there exist some differences on their specific pathway reactions. In fact, Schuetz et al. [9] had constructed a highly interconnected stoichiometric network model based on known reactions of *E. coli* central carbon metabolism. This could be the reason for the superior of prediction results on *Schuetz* model when compare to other with the same scale. However, in term of extension, *Genome-scale* model *iAF1260* is on a higher level than that of *Schuetz* model since the number of reactions is about tenfold than the original *Core* reconstruction. The results on Figure 4.4 especially 4.4a and 4.4b supported this comparison since the improvement made by changing from *Core* model to *Genome-scale* model is even better than that when only replacing the former by *Schuetz* model. The complete results for all datasets is presented in Section C of Appendix B.
4.2. The impact of objective functions by FBA in different metabolic models

![Graph](image)

**Figure 4.4:** Predictive fidelities of FBA simulation for each objective function in several datasets with three different metabolic models. Values above 1 are out of range.

To conclude, the difference reflects the fact that beside agreements on the most effective functions for certain situations, using different metabolic models could significantly affect the conclusions for certain cases. The factors related to the model reconstructions such as the biomass composition can...
be one possible reason for this difference.

4.3 Evaluation of pairwise constraints

In this section, the effect of adding cellular constraints to the genome-scale system are evaluated. Here FBA was run for pairwise constraints to predict the fluxes. In order to understand how the phenotype predictions vary across the different constraints, a particular case is selected, namely the chemostat fermentation under the highest dilution rate of $0.7\,h^{-1}$. This is a typical case where FBA simulations are less accurate, since the cells were sub-optimally grown due to overflow metabolism \cite{39}. For this case the flux predictions behaved qualitatively well for the same optimal criteria and is well-described by maximizing biomass yield (Figure 4.5a). It can be observed in Figure 4.5a that all single or even pairwise constraints to this model did not significantly improve flux prediction in a consistent manner for all tested cases. The same behaviour could be observed in the other objective functions. In fact, the uselessness of pairwise cellular criteria also can be observed on other datasets or model reconstructions (see Section B of Appendix B).

![Figure 4.5](image)

**Figure 4.5:** Comparison between pFBA predictive fidelities using single \(4.5a\) and pairwise \(4.5b\) constraints for the dataset from Ishii under the dilution rate of $0.7\,h^{-1}$.

In general, no improvement in predictive fidelity for any of the single objective functions when the combination of two constraints are used instead. On the other hand, when the complete set of constraints “all constraints scenario” are used simultaneously, its contribution was not significant compared to the “none” constrain scenario as reported previously in \cite{9}. Identical conclusions were obtained when repeating all these simulations with the other datasets (see Appendix B).

4.4 Evaluation of multi-objective functions

In this section, every possible combinations of the objective functions $\max BM$, $\max ATP$ and $\min Flux$ was examined instead of the use of only single objective in the previous simulations. These three objective functions had been reported as the best in a previous multi-objective research \cite{50}. However, in this previous work, metabolic flux analysis simulations were limited. Therefore, obtained studies are needed to determine the optimality principles in different metabolic models. There are also good
4.4. Evaluation of multi-objective functions

biological meanings to evaluate the Pareto fronts of these three objective. In fact, growth rate, energy usage and enzymatic parsimony are among the most important but usually conflict aspects of metabolism. It means that the improvement of one of them usually lead to the degradation of the other two. Thus, using the Pareto optimization to understand the trade-off between these three meaningful objective was required.

Table 4.1, 4.2 and 4.3 show the best predictive fidelities from experimental fluxes to the optimize solutions of using single objective and to the Pareto fronts of combinations of two or three objective functions for the Core model, Schuetz model and Genome-scale model, respectively.

Table 4.1: Predictive fidelities of pFBA simulation using single and multi-objective functions in the Core model.

<table>
<thead>
<tr>
<th></th>
<th>Emmerling h⁻¹, N-limited</th>
<th>Ishii h⁻¹, C-limited</th>
<th>Ishii h⁻¹, C-limited</th>
<th>Ishii h⁻¹, C-limited</th>
<th>Perrenoud h⁻¹, batch</th>
<th>Holm h⁻¹, batch</th>
</tr>
</thead>
<tbody>
<tr>
<td>max BM</td>
<td>0.4085</td>
<td>0.1373</td>
<td>0.0327</td>
<td>0.0688</td>
<td>0.2307</td>
<td>0.3046</td>
</tr>
<tr>
<td>max ATP</td>
<td>0.3458</td>
<td>0.3725</td>
<td>0.1891</td>
<td>0.2452</td>
<td>0.6508</td>
<td>0.7334</td>
</tr>
<tr>
<td>min Flux</td>
<td>0.1491</td>
<td>0.6321</td>
<td>0.5951</td>
<td>0.6612</td>
<td>0.2408</td>
<td>0.3212</td>
</tr>
<tr>
<td>max BM + max ATP</td>
<td>0.1169</td>
<td>0.1373</td>
<td>0.0212</td>
<td>0.0677</td>
<td>0.2284</td>
<td>0.3046</td>
</tr>
<tr>
<td>max BM + min Flux</td>
<td>0.1490</td>
<td>0.1367</td>
<td>0.0327</td>
<td>0.0677</td>
<td>0.1820</td>
<td>0.0707</td>
</tr>
<tr>
<td>max ATP + min Flux</td>
<td>0.1023</td>
<td>0.3559</td>
<td>0.1843</td>
<td>0.1807</td>
<td>0.1967</td>
<td>0.1399</td>
</tr>
<tr>
<td>max BM + max ATP + min Flux</td>
<td>0.0856</td>
<td>0.1367</td>
<td>0.0212</td>
<td>0.0649</td>
<td>0.1490</td>
<td>0.0321</td>
</tr>
</tbody>
</table>

As observed in a previous work [50] no dual combination could describe all measured fluxes adequately, and only the three efficiency objectives (max BM, max ATP and min Flux) achieved the highest optimality for the datasets and models analyzed (Table 4.1, 4.2 and 4.3). Our results show that the multi-objective approach improved the best prediction obtained with traditional FBA using a single objective function. However, the improvement is not the same for different datasets. Overall, our study showed that the multi-objective approach using two objective is better than single objective but usually inferior than the combination of three objectives simultaneously.

Table 4.2: Predictive fidelities of pFBA simulation using single and multi-objective functions in the Schuetz model.

<table>
<thead>
<tr>
<th></th>
<th>Emmerling h⁻¹, N-limited</th>
<th>Ishii h⁻¹, C-limited</th>
<th>Ishii h⁻¹, C-limited</th>
<th>Ishii h⁻¹, C-limited</th>
<th>Perrenoud h⁻¹, batch</th>
<th>Holm h⁻¹, batch</th>
</tr>
</thead>
<tbody>
<tr>
<td>max BM</td>
<td>0.3412</td>
<td>0.2551</td>
<td>0.1595</td>
<td>0.1491</td>
<td>0.3099</td>
<td>0.1841</td>
</tr>
<tr>
<td>max ATP</td>
<td>0.3369</td>
<td>0.6662</td>
<td>0.3166</td>
<td>0.3072</td>
<td>0.5433</td>
<td>0.6457</td>
</tr>
<tr>
<td>min Flux</td>
<td>0.2298</td>
<td>0.4019</td>
<td>0.3237</td>
<td>0.2761</td>
<td>0.3530</td>
<td>0.2131</td>
</tr>
<tr>
<td>max BM + max ATP</td>
<td>0.1544</td>
<td>0.2551</td>
<td>0.1517</td>
<td>0.1491</td>
<td>0.2952</td>
<td>0.1501</td>
</tr>
<tr>
<td>max BM + min Flux</td>
<td>0.2282</td>
<td>0.2551</td>
<td>0.1595</td>
<td>0.1491</td>
<td>0.3099</td>
<td>0.1681</td>
</tr>
<tr>
<td>max ATP + min Flux</td>
<td>0.2291</td>
<td>0.3146</td>
<td>0.2514</td>
<td>0.2499</td>
<td>0.3481</td>
<td>0.2052</td>
</tr>
<tr>
<td>max BM + max ATP + min Flux</td>
<td>0.1545</td>
<td>0.2551</td>
<td>0.1521</td>
<td>0.1491</td>
<td>0.2953</td>
<td>0.1505</td>
</tr>
</tbody>
</table>

Similar results are shown on Table 4.2 when Schuetz model reconstruction is applied. Moreover, for the Genome-scale model, the multi-objective approach significantly improve the flux prediction for
all datasets in a more consistent manner (Table 4.3).

**Table 4.3:** Predictive fidelities of pFBA simulation using single and multi-objective functions in the *Genome-scale* model.

<table>
<thead>
<tr>
<th></th>
<th>Emmerling 0.09 h⁻¹, N-limited</th>
<th>Ishii 0.1 h⁻¹, C-limited</th>
<th>Ishii 0.4 h⁻¹, C-limited</th>
<th>Ishii 0.7 h⁻¹, C-limited</th>
<th>Perrenoud 0.65 h⁻¹, batch</th>
<th>Holm 0.67 h⁻¹, batch</th>
</tr>
</thead>
<tbody>
<tr>
<td>max BM</td>
<td>0.3658</td>
<td>0.1401</td>
<td>0.0375</td>
<td>0.0697</td>
<td>0.2314</td>
<td>0.3302</td>
</tr>
<tr>
<td>max ATP</td>
<td>0.3563</td>
<td>0.3741</td>
<td>0.1882</td>
<td>0.2472</td>
<td>0.6782</td>
<td>0.7327</td>
</tr>
<tr>
<td>min Flux</td>
<td>0.2332</td>
<td>0.1295</td>
<td>0.1213</td>
<td>0.1403</td>
<td>0.4984</td>
<td>0.3519</td>
</tr>
<tr>
<td>max BM + max ATP</td>
<td>0.2314</td>
<td>0.1401</td>
<td>0.0304</td>
<td>0.0693</td>
<td>0.1247</td>
<td>0.3302</td>
</tr>
<tr>
<td>max BM + min Flux</td>
<td>0.2241</td>
<td>0.1007</td>
<td>0.0348</td>
<td>0.0679</td>
<td>0.2304</td>
<td>0.2962</td>
</tr>
<tr>
<td>max ATP + min Flux</td>
<td>0.2904</td>
<td>0.1288</td>
<td>0.1207</td>
<td>0.1267</td>
<td>0.1334</td>
<td>0.1545</td>
</tr>
<tr>
<td>max BM + max ATP + min Flux</td>
<td>0.2238</td>
<td>0.0982</td>
<td>0.0261</td>
<td>0.0661</td>
<td>0.1127</td>
<td>0.1047</td>
</tr>
</tbody>
</table>

Overall, the multi-objective function is very helpful in improving the prediction of pFBA model. However, the pattern is not the same with different model reconstructions and datasets. The common behaviour that could be observed is that the efficiency of objective combination depends on the performance of every single objective. It means, for example, if one objective among them is outperform the remaining then the result of multi-objective pFBA is close to that of the superior function alone and the improvement seems not so significant. In some circumstances, all the objective functions giving similar prediction errors and the combination of them would make quite strong promotion of the fidelity. This phenomenon happens frequently and gives a reasonable meaning for the trade-off between biological behaviours of living organisms. Sometimes, the cell needs to focus on one dominating objective but in certain conditions, several objectives are considered simultaneously in an equal manner.
Conclusions and Future Work

Contents

5.1 Main conclusions and work summary ........................................... 40
5.2 Ongoing and future work ......................................................... 40
Chapter 5. Conclusions and Future Work

5.1 Main conclusions and work summary

The work developed during this thesis explored the effect of different optimal principles in FBA for three metabolic models using different conditions and comparison dataset than previous evaluations. Although the fidelity patterns of FBA can differ considerably under different conditions, the classical “maximization of biomass” was shown in general as one of the best objectives. Moreover, our results show that the model size has a high impact on the predictive fidelities.

Despite the observed variations, several generalities emerged. In agreement with previous studies, the single objective of maximization of biomass yield achieves the best predictive accuracy. For the batch growth condition the most consistent optimality criteria appears to be described by the maximization of the biomass yield per flux or by the objective of maximization of ATP yield per flux (see above definitions). Moreover, under N-limited continuous cultures, the criteria minimization of the flux distribution or maximization of biomass yield was determined as the most significant. On the other hand, the predictions obtained by flux balance analysis using additional combined standard constraints are not better than those obtained using the single constraint or even none of them. For the multi-objective optimization by Pareto-optimal flux distributions improve the best predictions obtained by each single objective function alone. The improvement, however, is not the same for all models with different reconstructions or datasets. The combination in some cases provide significant advances but there exist situations when it returns equal performance with the best single objective function.

Although some optimal criteria gives reasonable predictions under certain conditions, there is no universal criteria that performs well under all conditions. Therefore, systems biologists should perform a careful evaluation and analysis of the objective functions case-by-case for each particular condition and application.

Finally, Matlab code for investigation the effect of cellular objective function and constrains on metabolic models has been also developed. This implementation of all objective functions and constraints can be easily adapted to test new metabolic systems and can be evaluated by comparing its results with those reported here.

5.2 Ongoing and future work

There are some limitations in this project that should be improved in future work. Firstly, nonlinear objective functions were not included in Pareto optimization due to the computational cost. By taking these options into consideration, one could study more about the trade-off mechanisms of cell under different conditions. Ongoing efforts are currently directed for investigating the challenge of multi-objective optimization formulations to reduce the computational cost of the simulations performed. It also would be a big help if the trade-off behaviours of single objective in multi-objective modelling is thoroughly elaborated. Also, new big data (fluxomics), should be added to make a comprehensive knowledge about the metabolic flux distribution of E. coli more accurate. Another possible improvement is the application of more genome-scale reconstruction models of other strains/species, together
with different experimental \textit{in vivo} datasets. The aim is to have more information to investigate the predictive power of the models and also the hidden factors involved in the optimizing mechanism of other cells. Additionally, cells with genetic perturbation should be included in the simulations to study the essential relationship between genotype and phenotype in metabolism. Finally, a consensus study is required to have a systematic framework for this case-by-case analysis. It should contain a common standard formulation for all flux prediction methods, as well as reference dataset and the way to evaluate the performance of each algorithm. It is hoped that the work developed in this thesis contributes with methods to apply Systems Biology in biotechnology industries, and it becomes an open door for new applications using the presented study.
Bibliography


Bibliography


Bibliography


Experimental Datasets and Matches between experimental and model reactions
<table>
<thead>
<tr>
<th>Reaction</th>
<th>0.0% C-limited</th>
<th>0.0% C-limited</th>
<th>0.0% C-limited</th>
<th>0.0% C-limited</th>
<th>0.0% C-limited</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Table A.1:</strong> Experimental flux data for growth of E. coli in different conditions from different sources (Isbell and Emmerling)</td>
<td>0.0% C-limited</td>
<td>0.0% C-limited</td>
<td>0.0% C-limited</td>
<td>0.0% C-limited</td>
<td>0.0% C-limited</td>
</tr>
<tr>
<td>№</td>
<td>Reaction</td>
<td>Nanchen</td>
<td>Yang</td>
<td>Perrenoud</td>
<td>Holm</td>
</tr>
<tr>
<td>----</td>
<td>---------------------------</td>
<td>---------------</td>
<td>--------------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$0.09h^{-1}$</td>
<td>$0.40h^{-1}$</td>
<td>$0.1h^{-1}$</td>
<td>$0.55h^{-1}$</td>
</tr>
<tr>
<td>1</td>
<td>GLC + ATP → G6P</td>
<td>100.0 ± 10.0</td>
<td>100.0 ± 5.0</td>
<td>100.0 ± 10.0</td>
<td>100.0 ± 10.0</td>
</tr>
<tr>
<td>2</td>
<td>G6P → 6PG + NADPH</td>
<td>30.0 ± 5.0</td>
<td>36.0 ± 3.0</td>
<td>34.0 ± 3.4</td>
<td>26.0 ± 2.6</td>
</tr>
<tr>
<td>3</td>
<td>6PG → P5P + CO$_2$ + NADPH</td>
<td>23.0 ± 5.0</td>
<td>29.0 ± 3.0</td>
<td>34.0 ± 3.4</td>
<td>26.0 ± 2.6</td>
</tr>
<tr>
<td>4</td>
<td>G6P → F6P</td>
<td>69.0 ± 9.0</td>
<td>64.0 ± 4.0</td>
<td>65.0 ± 6.5</td>
<td>73.0 ± 7.3</td>
</tr>
<tr>
<td>5</td>
<td>6PG → T3P + PYR</td>
<td>7.0 ± 6.0</td>
<td>7.0 ± 4.0</td>
<td>0.0 ± 0.0</td>
<td>0.0 ± 0.0</td>
</tr>
<tr>
<td>6</td>
<td>F6P + ATP → 2*T3P</td>
<td>82.0 ± 11.0</td>
<td>78.0 ± 5.0</td>
<td>83.0 ± 8.3</td>
<td>85.0 ± 8.5</td>
</tr>
<tr>
<td>7</td>
<td>2P5P → S7P + T3P</td>
<td>8.0 ± 2.0</td>
<td>9.0 ± 1.0</td>
<td>11.0 ± 1.1</td>
<td>8.0 ± 0.8</td>
</tr>
<tr>
<td>8</td>
<td>P5P + E4P → F6P + T3P</td>
<td>5.0 ± 2.0</td>
<td>6.0 ± 1.0</td>
<td>8.0 ± 0.8</td>
<td>4.0 ± 0.4</td>
</tr>
<tr>
<td>9</td>
<td>S7P + T3P → E4P + F6P</td>
<td>8.0 ± 2.0</td>
<td>9.0 ± 1.0</td>
<td>11.0 ± 1.1</td>
<td>8.0 ± 0.8</td>
</tr>
<tr>
<td>10</td>
<td>T3P → PGA + ATP + NADH</td>
<td>176.0 ± 21.0</td>
<td>168.0 ± 9.0</td>
<td>173.0 ± 17.3</td>
<td>172.0 ± 17.2</td>
</tr>
<tr>
<td>11</td>
<td>PGA → PEP</td>
<td>168.0 ± 21.0</td>
<td>156.0 ± 9.0</td>
<td>163.0 ± 16.3</td>
<td>161.0 ± 16.1</td>
</tr>
<tr>
<td>12</td>
<td>PEP → PYR + ATP</td>
<td>201.0 ± 26.0</td>
<td>120.0 ± 9.0</td>
<td>130.0 ± 13.0</td>
<td>124.0 ± 12.4</td>
</tr>
<tr>
<td>13</td>
<td>PYR → AcCoA + CO$_2$ + NADH</td>
<td>190.0 ± 28.0</td>
<td>113.0 ± 9.0</td>
<td>107.0 ± 10.7</td>
<td>93.0 ± 9.3</td>
</tr>
<tr>
<td>14</td>
<td>OAA + AcCoA → ICT</td>
<td>121.0 ± 26.0</td>
<td>96.0 ± 9.0</td>
<td>87.0 ± 8.7</td>
<td>70.0 ± 7.0</td>
</tr>
<tr>
<td>15</td>
<td>ICT → OGA + CO$_2$ + NADPH</td>
<td>65.0 ± 27.0</td>
<td>96.0 ± 9.0</td>
<td>87.0 ± 8.7</td>
<td>70.0 ± 7.0</td>
</tr>
<tr>
<td>16</td>
<td>OGA → FUM + CO$_2$ + 1.5<em>ATP + 2</em>NADH</td>
<td>56.0 ± 29.0</td>
<td>86.0 ± 9.0</td>
<td>78.0 ± 7.8</td>
<td>58.0 ± 5.8</td>
</tr>
<tr>
<td>17</td>
<td>FUM → MAL</td>
<td>113.0 ± 27.0</td>
<td>86.0 ± 9.0</td>
<td>78.0 ± 7.8</td>
<td>58.0 ± 5.8</td>
</tr>
<tr>
<td>18</td>
<td>MAL → OAA + NADH</td>
<td>169.0 ± 28.0</td>
<td>78.0 ± 8.0</td>
<td>75.0 ± 7.5</td>
<td>58.0 ± 5.8</td>
</tr>
<tr>
<td>19</td>
<td>MAL → PYR + CO$_2$ + NADH</td>
<td>0.0 ± 0.4</td>
<td>8.0 ± 2.0</td>
<td>3.0 ± 0.3</td>
<td>0.0 ± 0.0</td>
</tr>
<tr>
<td>20</td>
<td>OAA + ATP → PEP + CO$_2$</td>
<td>70.0 ± 10.0</td>
<td>15.0 ± 0.2</td>
<td>67.0 ± 6.7</td>
<td>23.0 ± 2.3</td>
</tr>
<tr>
<td>21</td>
<td>PEP + CO$_2$ → OAA</td>
<td>31.0 ± 6.0</td>
<td>45.0 ± 4.0</td>
<td>94.0 ± 9.4</td>
<td>52.0 ± 5.2</td>
</tr>
<tr>
<td>22</td>
<td>AcCoA → Acetate + ATP</td>
<td>0.0 ± 1.0</td>
<td>0.0 ± 0.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>NADPH → NADH</td>
<td>29.0 ± 55.0</td>
<td>45.0 ± 18.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>O$_2$ + 2NADH → 2Pi/2*ATP</td>
<td>372.0 ± 80.0</td>
<td>301.0 ± 29.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Biomass production</td>
<td>5.0 ± 1.0</td>
<td>7.0 ± 1.0</td>
<td>7.1 ± 0.7</td>
<td>8.6 ± 0.9</td>
</tr>
<tr>
<td>26</td>
<td>ICT + AcCoA → MAL + FUM + NADH</td>
<td>56.0 ± 10.0</td>
<td>0.0 ± 1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>DHAP → PYR</td>
<td>0.0 ± 10.0</td>
<td>0.0 ± 10.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>ETH → ETHxt</td>
<td>0.0 ± 1.0</td>
<td>0.0 ± 1.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.2: (continue) from Nanchen, Yang, Perrenoud and Holm.
### Table A.3: Mapping from experimental reactions to the corresponding ones in model reconstructions. The "-" symbol indicates opposite direction of the reaction. Note that besides this common mapping, some datasets may provide information about additional specific reactions in the model reconstruction also.

<table>
<thead>
<tr>
<th>No</th>
<th>Experimental reaction</th>
<th>Model reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GLC + ATP → G6P</td>
<td>R glk, ptsG/H/I</td>
</tr>
<tr>
<td>2</td>
<td>G6P → 6PG + NADPH</td>
<td>zwf, pgl</td>
</tr>
<tr>
<td>3</td>
<td>6PG → P5P + CO + 2NADPH</td>
<td>gnd</td>
</tr>
<tr>
<td>4</td>
<td>G6P → F6P</td>
<td>pgi</td>
</tr>
<tr>
<td>5</td>
<td>6PG → T3P + PYR</td>
<td>edd, eda</td>
</tr>
<tr>
<td>6</td>
<td>F6P + ATP → 2*T3P</td>
<td>pfkA, pfkB, fbaA, fbaB, tpiA</td>
</tr>
<tr>
<td>7</td>
<td>2P5P → S7P + T3P</td>
<td>tktA, tktB</td>
</tr>
<tr>
<td>8</td>
<td>P5P + E4P → F6P + T3P</td>
<td>r2, tktB</td>
</tr>
<tr>
<td>9</td>
<td>S7P + T3P → E4P + F6P</td>
<td>talA, talB</td>
</tr>
<tr>
<td>10</td>
<td>T3P → PGA + ATP + NADH</td>
<td>gapA, pgk</td>
</tr>
<tr>
<td>11</td>
<td>PGA → PEP</td>
<td>eno</td>
</tr>
<tr>
<td>12</td>
<td>PEP → PYR + ATP</td>
<td>pykF , pykA</td>
</tr>
<tr>
<td>13</td>
<td>PYR → AcCoA + CO + 2NADH</td>
<td>aceEF</td>
</tr>
<tr>
<td>14</td>
<td>OAA + AcCoA → ICT</td>
<td>gltA, prpC, acnA, acnA</td>
</tr>
<tr>
<td>15</td>
<td>ICT → OGA + CO + 2NADH</td>
<td>icd</td>
</tr>
<tr>
<td>16</td>
<td>OGA → FUM + CO + 1.5*ATP + 2NADH</td>
<td>sucAB, sucCD</td>
</tr>
<tr>
<td>17</td>
<td>FUM → MAL</td>
<td>fumA, fumB, fumC</td>
</tr>
<tr>
<td>18</td>
<td>MAL → OAA + NADH</td>
<td>mdh, mgo</td>
</tr>
<tr>
<td>19</td>
<td>MAL → PYR + CO + 2NADH</td>
<td>maeA, maeB</td>
</tr>
<tr>
<td>20</td>
<td>OAA + ATP → PEP + CO 2</td>
<td>pck</td>
</tr>
<tr>
<td>21</td>
<td>PEP → OAA</td>
<td>ppc</td>
</tr>
<tr>
<td>22</td>
<td>AcCoA → Acetate + ATP</td>
<td>pta, ackA, ackB, tdcD, purT, acs, ac</td>
</tr>
<tr>
<td>23</td>
<td>NADPH → NADH</td>
<td>pntAB, udhA</td>
</tr>
<tr>
<td>24</td>
<td>O2 + 2NADH → 2P/O*ATP</td>
<td>o2</td>
</tr>
<tr>
<td>25</td>
<td>Biomass production</td>
<td>Biomass Ecoli core w/GAM 59p81M</td>
</tr>
</tbody>
</table>
Table A.4: Specific reaction fluxes used in each objective function. Reactions with ‘.f’ and ‘.b’ suffixes mean forward and backward irreversible reactions respectively which are originated from a reversible reaction.

<table>
<thead>
<tr>
<th>Objective function</th>
<th>Involved reactions</th>
<th>Model reaction</th>
<th>Core model</th>
<th>Schuetz model</th>
<th>Genome-scale model</th>
</tr>
</thead>
<tbody>
<tr>
<td>max BM</td>
<td>Biomass objective function</td>
<td>Biomass. Ecoli.core.w.GAM</td>
<td>biomass</td>
<td>Ec_biomass_jAF1260_core_59p81M</td>
<td></td>
</tr>
<tr>
<td>max BM/flux</td>
<td>ATP maintenance requirement</td>
<td>ATPM</td>
<td>maint</td>
<td>ATPM</td>
<td></td>
</tr>
<tr>
<td>min Flux</td>
<td>Intracellular reactions</td>
<td>All</td>
<td>All</td>
<td>All</td>
<td></td>
</tr>
<tr>
<td>min Redox</td>
<td>All reactions that produce NADH or NADPH</td>
<td>GAPD.f, PHD, ME2, AKGDH, MDH.f, NADTRHD, LDH.D.f, ACALD.f, ALCD2x.f, G6PDH2r_r, GND ICDHyr.f, THD2, FRD7, SUCDi</td>
<td>gapA.f, aceEF, maeA, sucAB, mdh.f, udhA, fdhF, fdoGHI, fdnGHI_r2, ldhA.f, adhE.b, mhpF.b, adhP.b, adhC.b, maeB, awf.f, gcd, icd.f, pntAB, frdABCD.f, sdhAB_r, dld, sdhABCD_b</td>
<td>GAPD_f, PDH, ME2, AKGDH, MDH_f, NADTRHD, LDH.D.f, ACALD.f, ALCD2x.f, ME2, G6PDH2r_r, GND, ICDHyr.f, THD2pp, FRD2, FRD3, SUCDi</td>
<td></td>
</tr>
<tr>
<td>max ATPprod</td>
<td>ATP producing reactions</td>
<td>PGK_f, PYK, SUCOAS.f, ATPS4r_f, ACKr_b</td>
<td>pgk_f, pykA, pykF, sucCD_f, atp, ackA_f, ackB_f, tdcD_f, purT_f</td>
<td>PGK_f, PYK, SUCOAS_f, ATPS4rpp_f, ACKr_b</td>
<td></td>
</tr>
</tbody>
</table>
A. Experimental Datasets and Matches between experimental and model reactions
FBA simulation results
B. FBA simulation results

Simulation with single constraint

Figure B.1: Predictive fidelities for objective functions with single constraints for *E. coli* Core model.

(a) Ishii 0.1h⁻¹, C−limited, Core model
(b) Ishii 0.4h⁻¹, C−limited, Core model
(c) Ishii 0.7h⁻¹, C−limited, Core model
(d) Emmerling 0.09h⁻¹, N−limited, Core model
(e) Emmerling 0.09h⁻¹, C−limited, Core model
(f) Emmerling 0.4h⁻¹, C−limited, Core model
Predictive Fidelity

Nanchen 0.09h\(^{-1}\), C-limited, Core model

Yang 0.1h\(^{-1}\), C-limited, Core model

Perrenoud 0.65h\(^{-1}\), batch, Core model

Holm 0.67h\(^{-1}\), batch, Core model

P/O=1

qO\(_2\)max=15

Maintenance

Bounds

NADPH

All constraints
Figure B.2: Predictive fidelities for objective functions with single constraints for E. coli Schuetz model.

(a) Ishii 0.1h\(^{-1}\), C—limited, Schuetz’s model

(b) Ishii 0.4h\(^{-1}\), C—limited, Schuetz’s model

(c) Ishii 0.7h\(^{-1}\), C—limited, Schuetz’s model

(d) Emmerling 0.09h\(^{-1}\), N—limited, Schuetz’s model

(e) Emmerling 0.09h\(^{-1}\), C—limited, Schuetz’s model

(f) Emmerling 0.4h\(^{-1}\), C—limited, Schuetz’s model
Nanchen 0.09h$^{-1}$, C−limited, Schuetz’s model

Yang 0.1h$^{-1}$, C−limited, Schuetz’s model

Perrenoud 0.65h$^{-1}$, batch, Schuetz’s model

Holm 0.67h$^{-1}$, batch, Schuetz’s model
B. FBA simulation results

Figure B.3: Predictive fidelities for objective functions with single constraints for *E. coli* iAF1260 Genome-scale model.
B. FBA simulation results

Simulation with pairwise constraint

Figure B.4: Predictive fidelities for objective functions with pairwise constraints for *E. coli* Core model.

(a) Ishii 0.1h\(^{-1}\), C\(-\)limited, Core model

(b) Ishii 0.4h\(^{-1}\), C\(-\)limited, Core model

(c) Ishii 0.7h\(^{-1}\), C\(-\)limited, Core model

(d) Emmerling 0.09h\(^{-1}\), N\(-\)limited, Core model

(e) Emmerling 0.09h\(^{-1}\), C\(-\)limited, Core model

(f) Emmerling 0.4h\(^{-1}\), C\(-\)limited, Core model
Predictive Fidelity

- BM
- ATP
- Flux
- BM/flux
- ATP/flux
- Rd
- ATPprod

Conditions:
- P/O=1 & qO2max=15
- P/O=1 & Maintenance
- P/O=1 & Bounds
- P/O=1 & NADPH
- qO2max=15 & Maintenance
- qO2max=15 & Bounds
- qO2max=15 & NADPH
- Maintenance & Bounds
- Maintenance & NADPH
- Bounds & NADPH

Models:
- Nanchen 0.09h⁻¹, C-limited, Core model
- Nanchen 0.4h⁻¹, C-limited, Core model
- Yang 0.1h⁻¹, C-limited, Core model
- Yang 0.55h⁻¹, C-limited, Core model
- Perrenoud 0.65h⁻¹, batch, Core model
- Holm 0.67h⁻¹, batch, Core model
B. FBA simulation results

Figure B.5: Predictive fidelities for objective functions with pairwise constraints for *E. coli* Schuetz model.

(a) Ishii 0.1h$^{-1}$, C- limited, Schuetz’s model

(b) Ishii 0.4h$^{-1}$, C- limited, Schuetz’s model

(c) Ishii 0.7h$^{-1}$, C- limited, Schuetz’s model

(d) Emmerling 0.09h$^{-1}$, N- limited, Schuetz’s model

(e) Emmerling 0.09h$^{-1}$, C- limited, Schuetz’s model

(f) Emmerling 0.4h$^{-1}$, C- limited, Schuetz’s model
B. FBA simulation results

Figure B.6: Predictive fidelities for objective functions with pairwise constraints for *E. coli* iAF1260 Genome-scale model.
B. FBA simulation results

Scatter plots

Figure B.7: Individual flux predictions between predicted and experimental fluxes for all the objective functions evaluated without additional constraint of the Core model.
B. FBA simulation results

Figure B.8: Individual flux predictions between predicted and experimental fluxes for all the objective functions evaluated without additional constraint of the Schuetz model.
(g) Nanchen 0.09 h\(^{-1}\), C-limited, Schuetz’s model

(h) Nanchen 0.4 h\(^{-1}\), C-limited, Schuetz’s model

(i) Yang 0.1 h\(^{-1}\), C-limited, Schuetz’s model

(j) Yang 0.50 h\(^{-1}\), C-limited, Schuetz’s model

(k) Perrenoud 0.65 h\(^{-1}\), batch, Schuetz’s model

(l) Helm 0.67 h\(^{-1}\), batch, Schuetz’s model
B. FBA simulation results

Figure B.9: Individual flux predictions between predicted and experimental fluxes for all the objective functions evaluated without additional constraint of the Genome-scale model.
B. FBA simulation results

Comparing three model reconstructions

Figure B.10: Predictive fidelities of pFBA simulation for each objective function in six remaining datasets with three different metabolic models. Values above 1 are out of range.

(a) Ishii 0.4h⁻¹, C−limited

(b) Nanchen 0.4h⁻¹, C−limited

(c) Nanchen 0.09h⁻¹, C−limited

(d) Emmerling 0.4h⁻¹, C−limited

(e) Yang 0.55h⁻¹, C−limited

(f) Yang 0.55h⁻¹, C−limited
Published abstract and article
C. Published abstract and article
