# A DECOMPOSITION APPROACH FOR COMPUTING ELEMENTARY FLUX MODES IN GENOME-SCALE METABOLIC NETWORKS

CHAN, SIU HUNG JOSHUA

M.Phil

The Hong Kong

Polytechnic University

2011

THE HONG KONG
POLYTECHNIC UNIVERSITY

DEPARTMENT OF INDUSTRIAL &
SYSTEMS ENGINEERING

# A DECOMPOSITION APPROACH FOR COMPUTING ELEMENTARY FLUX MODES IN GENOME-SCALE METABOLIC NETWORKS

**by**

**CHAN, Siu Hung Joshua**

**A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Philosophy**

**July 2011**

# Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_________________________(Signed)

CHAN, Siu Hung Joshua (Name of student)

# Abstract

The appearance of high-throughput experimental techniques to measure biological data in recent decades gives birth to Systems Biology which studies the emergent properties of biological systems by mathematical modelling.

The most ubiquitous structure in biological systems is the network structure. Among different biological networks, a particular important one is the metabolic network consisting of all the biochemical reactions and compounds in a cell. Reconstructed from the whole genome of a cell, the so-called genome-scale metabolic network successfully describes the cellular metabolism.

A fundamental computational framework applied to metabolic networks is the flux balance analysis (FBA) derived from the steady-state assumption. In FBA, the metabolic flux distribution, which is the vector containing all reaction rates in a metabolic network, can be obtained from solving a simple linear program given the stoichiometric information of reactions and a biological objective for optimization.

Metabolic pathway analysis (MPA) is a computational technique relevant to FBA to analyze metabolic pathways in metabolic networks. The first mathematically defined metabolic pathway, elementary flux mode (EFM), has theoretical as well as practical importance. One significant role of EFMs is that every flux distribution can be decomposed into a set of EFMs and a number of methods to study flux distributions originate from it. Yet finding such decompositions requires the complete set of EFMs, which is intractable in genome-scale metabolic networks due to combinatorial explosion.

In this research, we propose an algorithm to decompose flux distributions into EFMs in genome-scale networks. It is an iterative scheme of a mixed integer linear program. The algorithm is also able to approximate the EFM of largest contribution to an objective reaction in a flux distribution.

Complimentary to existing methods, our algorithm is capable of finding EFMs of flux distributions with complex structures, closer to the realistic case in which a cell is subject to various constraints. Our algorithm is first applied to study the growth of *Escherichia coli* (*E. coli*) under simple growth condition and we find that the employment of different EFMs is highly dynamic and sensitive to growth condition in order to achieve an optimal state of metabolism. This suggests a possible reason for the enormous redundancy of EFMs consuming the same set of uptake substrates and producing the same set of metabolites. A case of growth of *E. coli* in the Lysogeny broth (LB) medium in which the situation is complicated by the presence of various carbon sources is simulated and studied via our algorithm. Essential metabolites and their syntheses are located. Information on the contribution of each carbon source not obvious from the apparent flux distribution is also revealed. Finally, we apply our algorithm to analyze a real experimental flux distribution in mouse cardiomyocyte. Results consistent with literature are obtained. Interestingly, a mode of oxidative phosphorylation uncoupled from adenosine triphosphate (ATP) synthesis is discovered and this is not obvious from the flux distribution.

In conclusion, the algorithm can facilitate MPA in genome-scale metabolic networks. It provides an analytic method that prepares for the future breakthrough in experimental techniques to measure *in vivo* fluxes in a huge scale. One of the future directions is the improvement, refinement and further applications of the algorithm. Another possibility is the development of a more general algorithm to decompose a flux distribution into a set of EFMs with respect to a given optimization objective in a genome-scale metabolic network. Also, in the future, by further case studies and evaluations of different schemes for decomposition, a well-structured methodology may be established to analyze flux distributions in different situations as thorough as possible by their decompositions into EFMs.

# Acknowledgements

The completion of this research would be impossible without the help and support from many people whom I would like to thank. First of all, I must express my profound respect and gratitude towards my chief supervisor, Dr. P. Ji. He has brought me into a new exciting field of research which was unimaginable to me two years ago. During those unfruitful days, his patience, encouragement and suggestions have led me to the way out. I am in particular thankful for the trust and freedom that he has given to me so that I have become confident in doing research.

I thank my family and friends for their support and care. I thank my colleagues for the great working environment created. Thank Jack Wu for his much technical help and kind reminder. Special thanks are given to Wood Chan, with whom I experienced new positive impacts in research as well as an enjoyable university life. Sincere gratitude is given to Sandy Wong, who shares my joy and sorrow. The most importantly, I must thank God for His timely provision and His molding. He made a way in the waste land, and rivers in the dry country.

Finally, I would like to acknowledge The Hong Kong Polytechnic University for the financial support which enables my opportunity for doing this research.

# Table of Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| ADP | **:** Adenosine diphosphate |
| ATP | : Adenosine triphosphate |
| ATPM | : Adenosine triphosphate maintenance |
| BLP | : Bi-level program |
| DC | : Decomposability Check |
| EFM | : Elementary flux mode |
| EP | : Extreme pathway |
| *E. coli* | : *Escherichia coli* |
| FBA | : Flux balance analysis |
| GC | : Gas chromatography |
| GTP | : Guanosine triphosphate |
| LB medium | : Lysogeny broth medium |
| LP | : Linear program |
| MILP | : Mixed integer linear program |
| MFA | : Metabolic flux analysis |
| MPA | : Metabolic pathway analysis |
| mRNA | : Messenger ribonucleic acid |
| MS | : Mass spectrometry |
| NADPH | : Nicotinamide adenine dinucleotide phosphate |
| NMR | : Nuclear magnetic resonance |
| QP | : Quadratic program |
| TCA cycle | : tricarboxylic acid cycle |
| TF | : Transcriptional factor |
| TOF | : Time-of-flight |

# Chapter 1 Introduction

The appearance of high-throughput experimental techniques in the recent several decades gave birth to different ‘omics’ studies in biology, which analyze biological objects of interest with large-scale data. For example, genomics studies the genome of an organism by genome sequencing; transcriptomics investigates regulatory relations between transcriptomes by DNA microarray experiment; proteomics examines protein interactions by ChIP-chip experiment; metabolomics profiles metabolites by mass spectrometry (MS) and nuclear magnetic resonance (NMR), etc. [DeK06].

Since the last two decades, various large-scale data have been generated and become accessible to researchers through public online databases. Traditional methods are insufficient to extract useful information and build up biological knowledge from such a large amount of data. Instead, different computational approaches are adopted to cope with the scale of data. Systems biology is the paradigm to study the emergent properties of biological systems by integrating the huge amount of data, rather than to traditionally study the function of an individual entity [Vid04].

The most ubiquitous structure in biological systems is the network structure. For instance, the transcription of genes is governed by the transcriptional regulatory network behind or the gene regulatory network on the gene level; the metabolism in a cell is always described by a network structure called metabolic network; the signal transduction in a cell is carried out by its signaling network. With these networks accompanied by appropriate computational or mathematical models, quantitative analysis can be performed.

This research project focuses on the metabolic pathway analysis (MPA) in metabolic networks. Before stating the research objectives, these concepts and the relevant research challenges are briefly introduced.

## 1.1 Metabolic Networks

Proteins are produced from genes in organisms and they are essential for lives. All living organisms maintain lives by performing a large number of biochemical reactions. The most obvious is the respiration in cells generating energy in a form that cells can utilize to perform other important tasks. Almost all such biochemical reactions need enzymes, the most important type of proteins produced from organisms' genes. This stresses the essentiality of genes to an organism.

Metabolism refers to all biochemical reactions that occur in organisms. It is essential for the maintenance of life and performance of life activities, including growth, response to environment, reproduction, etc. All the compounds involved in these biochemical reactions are called metabolites. Together, the set of all metabolites and biochemical reactions form a metabolic network. In fact, it can be thought of the graphical representation of metabolism. The nodes are metabolites and the edges are reactions between metabolites. A more refined representation is a bipartite graph in which the first type of nodes represents metabolites, the second type represents reactions [Alm07]. An arc going to a reaction node from a metabolite node means the metabolite is a reactant in the reaction and an arc from a reaction to a metabolite means the metabolite is a product in the reaction. Reaction rates in metabolic network are called metabolic fluxes. Figure 1.1 shows a part of a metabolic network of *Escherichia coli* (*E. coli*) downloaded from Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/).

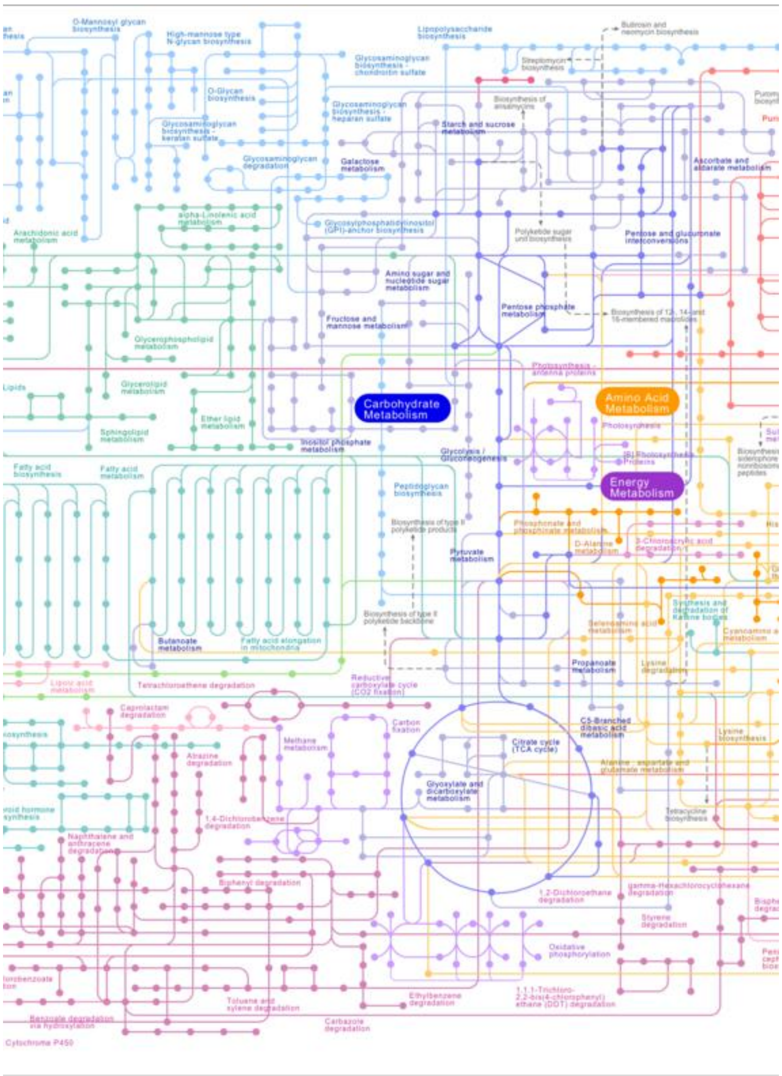

O-Mannosyl glycan biosynthesis
High-mannose type N-glycan biosynthesis
Lipopolysaccharide biosynthesis
Butirosin and neomycin biosynthesis
Glycosaminoglycan biosynthesis - chondroitin sulfate
Streptomycin biosynthesis
O-Glycan biosynthesis
Glycosaminoglycan biosynthesis - heparan sulfate
Biosynthesis of ansamycins
Glycosaminoglycan biosynthesis - keratan sulfate
Starch and sucrose metabolism
Glycosaminoglycan degradation
Galactose metabolism
Ascorbate and aldarate metabolism
alpha-Linolenic acid metabolism
Arachidonic acid metabolism
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis
Polyketide sugar unit biosynthesis
Pentose and glucuronate interconversions
Amino sugar and nucleotide sugar metabolism
Fructose and mannose metabolism
Pentose phosphate metabolism
Biosynthesis of 12-, 14- and 16-membered macrolides
Glycerophospholipid metabolism
Photosynthesis - antenna proteins
Lipids
Glycerolipid metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
Ether lipid metabolism
Photosynthesis
Sphingolipid metabolism
Inositol phosphate metabolism
Glycolysis / Gluconeogenesis
Fatty acid biosynthesis
Fatty acid metabolism
Energy Metabolism
Peptidoglycan biosynthesis
Biosynthesis of type II polyketide products
Phosphonate and phosphinate metabolism
D-Alanine metabolism
Pyruvate metabolism
Biosynthesis of type II polyketide backbone
Selenoamino acid metabolism
Butanoate metabolism
Fatty acid elongation in mitochondria
Propanoate metabolism
Lysine degradation
Lipoic acid metabolism
Tetrachloroethene degradation
Reductive carboxylate cycle (CO2 fixation)
Caprolactam degradation
Carbon fixation
C5-Branched dibasic acid metabolism
Methane metabolism
Citrate cycle (TCA cycle)
Lysine biosynthesis
Atrazine degradation
Alanine, aspartate and glutamate metabolism
Glyoxylate and dicarboxylate metabolism
1,4-Dichlorobenzene degradation
Tetracycline biosynthesis
Naphthalene and anthracene degradation
Biphenyl degradation
1,2-Dichloroethane degradation
gamma-Hexachlorocyclohexane degradation
Styrene degradation
Oxidative phosphorylation
Toluene and xylene degradation
Ethylbenzene degradation
1,1,1-Trichloro-2,2-bis(4-chlorophenyl)ethane (DDT) degradation
Benzoate degradation via hydroxylation
Carbazole degradation
Cytochrome P450

**Figure 1.1 Part of a Metabolic Network**

Part of a metabolic network of *E. coli* adopted from KEGG [Kan00].

The reconstruction of metabolic network lies in the recognition of biochemical reactions in the targeted cell. The present technology of reconstructing metabolic networks is fundamentally *in silico* [Edw00, Ree03, Fei07]. From the annotated genome of a certain organism, together with literature, metabolic database, all the gene products, mainly enzymes, possibly participating in metabolism are extracted. The reactions mediated by them and the metabolites involved are compiled as a large metabolic dataset. Metabolic networks reconstructed in this way are usually called genome-scale metabolic networks. Some of the genome-scale metabolic networks reconstructed by the Systems Biology Research Group in the University of California, San Diego are available online (http://bigg.ucsd.edu/). Different computational approaches had then been applied to study metabolism through genome-scale metabolic networks [Van06].

A widely accepted computational approach is flux balance analysis (FBA). It is a constraint-based optimization model relying on the stoichiometric information of chemical reactions and the mass balance principle in biochemistry. Since one metabolite is usually involved in more than one reaction, under the steady-state assumption that metabolite concentration remains unchanged, fluxes of different reactions are no longer independent, but satisfy a set of linear equations. Analysis of metabolic fluxes under this steady-state assumption alone is usually called metabolic flux analysis (MFA). However, such a system of linear equations is always underdetermined. Hence, with additional constraints and a biologically meaningful objective, a linear programming problem was formed and the optimal flux distribution was determined [Bon97]. This methodology is referred as FBA. The dynamic simulation dFBA derived from FBA showed satisfactory agreement with experimental data [Var94], validating the approach to a certain extent.

FBA has become the cornerstone in metabolic network analysis and various refinement and extension have been developed. For example, regulatory FBA incorporates effect of transcriptional regulation into FBA [Cov02]. Other examples include prediction of biological objective functions [Gia08] and prediction of flux bottlenecks in *E. coli* mutant strains [Her06] using bi-level optimization with FBA as the inner problem.

## 1.2 Metabolic Pathway Analysis

A series of biochemical reactions catalyzed by different enzymes form a metabolic pathway [Nel05], which is a part of a metabolic network. For example, in a cell, glycolysis (Figure 1.2), which is the process to release energy in the form of four **a**denosine **t**ri**p**hosphate (ATP) molecules generated from one glucose molecule and two ATP molecules, is an important example of metabolic pathways. There are ten reactions and eighteen metabolites involved.

Corresponding to the definition of metabolic pathways in the biochemistry context, two major mathematical definitions of metabolic pathways have also been proposed under the steady state assumption. The first one is the elementary flux mode (EFM) proposed in [Sch94]. EFMs by definition are flux distributions satisfying the steady-state assumption with minimal sets of reactions. However, the set of EFMs for a metabolic network is always very large and it is not systematically independent, i.e., a certain EFM can be represented as the positive sum of two others EFMs. Extreme pathways (EPs) had then been proposed as a smaller and systematically independent set of metabolic pathways [Sch00]. The set of EPs is always a subset of the set of EFMs. EFMs, however, seem to still be the most widely applied mathematical definition of metabolic pathways in a vast number of research (reviewed in [Dur09, Gag04, Lla10, Tri09b]).

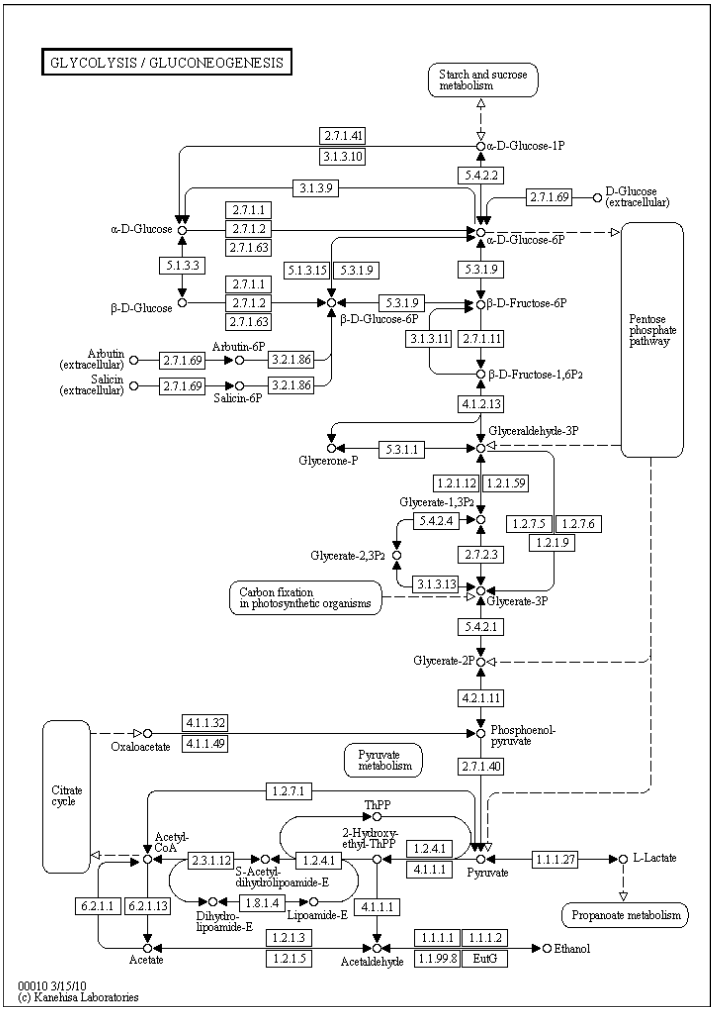


**Figure 1.2 Glycolysis Pathway**

The glycolysis pathway synthesizing pyruvate from glucose, adopted from KEGG [Kan00].

An important interpretation of EFMs is that every flux distribution can be decomposed as a positive sum of a subset of EFMs with the ‘no cancelation’ property that all EFMs have a particular component being zero wherever the flux distribution has that component being zero [Sch02]. The ‘no cancelation’ property is unique compared to other generating sets of the flux space like EPs and allows EFMs to answer questions regarding pathways producing or consuming specific metabolites [Lla10].

In many studies, flux distributions are analyzed by similar decompositions. Usually, such decompositions are chosen with respect to some optimization objectives. The first attempt is the α-spectrum to determine the range of attainable weights for each EP [Wib03]. This approach is also applicable to EFMs. Later, decomposing flux distributions with minimum sum of weights has been proposed [Schw05] and applied to study yeast glycolysis [Schw06]. Decompositions with respect to other objectives have also been adopted to investigate particular cases, for instances, maximum number of active EFMs [Noo07], minimum relative error [Wan07], maximum yield rate [Son09] and maximum entropy [Zha09b]. There are also other studies analyzing flux distributions by decompositions into EFMs e.g. [Kur07, Car09].

## 1.3 Challenge in the Computation of Elementary Flux Modes

These approaches based on decomposing flux distributions into sets of EFMs, despite the insights they have provided, have only limited capabilities because the calculation always requires the complete set of EFMs *in prior* given a metabolic network, whose computation is notoriously hard due to the combinatorial explosion when the network size grows. This drawback was first studied in [Kla02] and has been mentioned in many literatures, e.g. [Dur09, Tri09b, Yeu07]. Efforts have been put continuously to improve the computational speed and memory demand to compute EFMs, as discussed in Chapter 2. Despite the efforts, present algorithms still cannot cope with genome-scale metabolic networks reconstructed recently, usually consisting of at least a thousand of reactions and metabolites (e.g. [Fei07, Dua07]). This has motivated the present research.

## 1.4 Research Objectives

The aim of this research is to facilitate metabolic pathway analysis in genome-scale metabolic networks in which complete sets of EFMs cannot be found by existing algorithms. We in particular focus on EFMs that decompose a given flux distribution. There are two research objectives:

(1) To devise an algorithm to decompose a flux distribution into a set of EFMs in a genome-scale metabolic network;

(2) To analyze flux distributions in genome-scale metabolic networks by applying the algorithm to find decompositions into EFMs to gain new insight into cellular metabolism.

## 1.5 Importance of the Research

The two research objectives proposed are important in view of the following two aspects. Regarding the first research objective, it can facilitate the analysis of flux distributions by decompositions into EFMs, which has been found useful in the literature when applying to metabolic networks of small sizes but meanwhile has currently no corresponding method available in genome-scale metabolic networks. The first research objective will fill the research gap between the usefulness of the method and the lack of computational tool in genome-scale metabolic networks. Focus will be especially put on the computational bottleneck brought by the combinatorial explosion of the traditional methods, in which the full set of EFMs is required to be determined prior to the decomposition of a flux distribution into EFMs. This stresses the importance of the first research objective.

As for the second research objective, analyzing genome-scale flux distributions in genome-scale metabolic networks is expected to bring new insights in biology. It is exactly what the purpose of the systems biology approach is, namely to study biology at a system level. With further efforts, these insights can result in new knowledge in biology. Also, the applicability of the devised algorithm can be examined through studies on real cases in biology. Hence, the second objective is also important since it attempts to fulfill the noble purpose of the research area, systems biology, to obtain biological knowledge at a system level.

## 1.6 Scope of the Thesis

The rest of the thesis includes six more chapters.

In Chapter 2, a comprehensive literature review on metabolic networks, their reconstruction, metabolic pathway analysis and its development is presented.

In Chapter 3, a novel algorithm to decompose a flux distribution into a set of EFMs without finding the complete set of EFMs *in prior* is developed. It is an iterative scheme of a mixed integer linear program and guarantees certain nice properties of the decomposition.

In Chapter 4, the algorithm proposed is validated by computing the EFMs in a sample network [Cov01] in which existing methods are also feasible. To further test the capability of the algorithm, it is applied to decompose flux distributions into EFMs in the genome-scale metabolic network of *E. coli* strain K-12 MG1655 iAF1260 [Fei07]. The results are compared with the computation of the currently best methods up to the best of our understanding.

In Chapter 5, the algorithm is applied to investigate the growth of *E. coli* under two vastly different conditions, a glucose minimal medium and the Lysogeny broth (LB) medium, with a focus on the biological perspective.

In Chapter 6, an experimental flux distribution of mouse cardiomycyte [VoP06] is analyzed by the algorithm with an emphasis on the approximation of the largest contributing EFM relevant to the function of the mouse cardiomyocyte.

Finally, conclusions of the research and future work are given in Chapter 7.

# Chapter 2 Literature Review

## 2.1 From DNA to Metabolism

### 2.1.1 Gene and DNA

DNA, deoxyribonucleic acid, is the basic unit which stores genetic information in all known forms of lives. It consists of a sequence of nucleotides, each of them containing one base out of four possible choices, described by A, C, G and T. The well-known DNA sequence refers to the sequence of these four letters. One of the most important functions of such DNA sequence is its interaction with RNA, called transcription. During transcription, a messenger RNA (mRNA) copies information of a strand of DNA sequence. Units of three bases, or triplets, on the mRNA form codons and are translated in turn into amino acids in the molecular factory, ribosome. This process is called translation. After that, a sequence of amino acids is formed and then folded according to physical laws to become a protein, which is vital for every life on the Earth. A gene, conceptually referring to a unit of heredity, was discovered to be physically a strand of DNA sequence used to produce a protein. It is the functional unit of DNA and genome is just the whole collection of genes in an organism [Gri08].

### 2.1.2 Gene Expression and its Regulation

Gene expression refers to the whole sequential processes by which a gene product, usually protein or RNA, is produced using the information in a gene. The two main stages of gene expression, as stated in the central dogma of molecular biology [Cri70], are firstly the transcription and secondly the translation mentioned above. In addition, there are other modifications near the two stages. For example, after transcription and before translation, mRNA usually splices to remove introns on it, which are sequences containing no information for protein synthesis [Twy03].

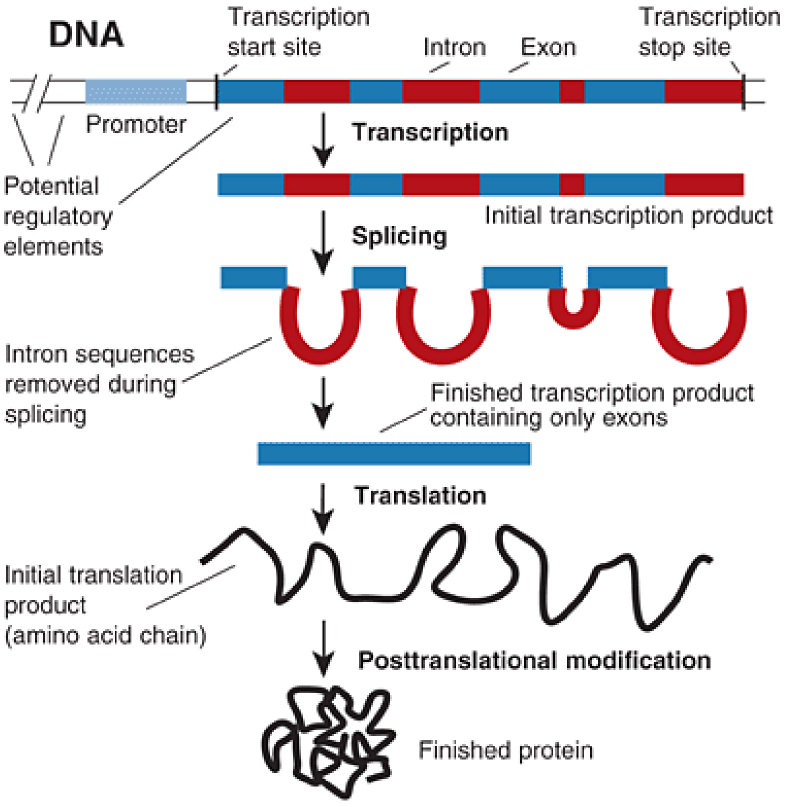

DNA
Transcription start site
Intron
Exon
Transcription stop site
Promoter
Potential regulatory elements
Transcription
Initial transcription product
Splicing
Intron sequences removed during splicing
Finished transcription product containing only exons
Translation
Initial translation product (amino acid chain)
Posttranslational modification
Finished protein

**Figure 2.1 Processes in Gene Expression**

Process in gene expression including transcription, alternative splicing, translation and posttranslational modification, adopted from NIAAA [NIA08].

Gene expression draws researchers' attention because not every gene is expressed equally probably and meanwhile not every organism expresses the same gene at the same level. The process of gene expression is different from species to species, strain to strain, and even cell to cell within a single multicellular organism [Twy03]. Jacob, Monod and Lwoff first noticed the existence of regulation mechanism in gene expression. They successfully manifested how gene expression was regulated and won the Nobel Prize in physiology or medicine in 1965 [Gri08].

Any step mentioned above during the gene expression process can be regulated. Among all the steps, the best studied one is the transcriptional regulation. Transcriptional factors (TFs), activated by phosphorylation, sometimes in response to extracellular stimuli, usually contain several DNA binding sites and bind to different genes to act as repressors or activators to repress or activate the transcription of certain genes (reviewed in [Kar94]). All regulation mechanisms can be formally described by a system with different states mathematically [Ros68] while the network indicating the interaction between different substances in regulation is the most widely accepted theoretical and graphical representation [deJ02]. For example, transcriptional regulatory network describes all the interactions between TFs, genes, RNA in transcription regulation. Basically, for any kind of regulations, all the substances, say TFs, are produced from genes at the very first step and all the species-specific characteristics, or heredity, are stored in genes [Bol02]. Hence, this starting from genes and ending at genes perspective has directed the development of gene regulatory network. Interactions in it may be explicit or implicit, showing the final influence between genes after considering all the regulation processes [Bra02].

### 2.1.3 Regulation of Gene Expression and Metabolism

The regulation mechanism is essential for life. It enables a cell to adapt to different external environment or stimulus [Kar94]. To perform different life activities, e.g. growth, respiration, reproduction, an organism depends on all the biochemical reactions taking place in it. The whole set of reactions is termed ‘metabolism’. It is thought to be an important aspect of the phenotype of an organism. Metabolism is always described by a network called metabolic network, with its nodes being metabolites, i.e., the compound involved in reactions [Che09]. Most of these reactions, including those most important, are enzymatic reactions relying on the presence and the activity of enzymes while enzymes are proteins produced from genes with their production rate being controlled by the gene expression regulation. Meanwhile, as an inverse relation, some metabolites produced during metabolism also contribute to the regulation of gene expression [Cov01]. In conclusion, genes, proteins, metabolite are interrelated objects in molecular networks and regulation of gene expression, metabolism are highly related process in organisms. Figure 2.2 gives an example of the network hierarchy [Bra02].

## 2.2 Systems Biology

The huge amount of objects and interactions being studied at the same time with an aim of a unifying framework give birth to systems biology, a new branch in biology in recent decades. In systems biology, biological systems, including the networks mentioned before, are modeled to be qualitative or quantitative, deterministic or stochastic, discrete or continuous and of other system characteristics, built upon biological knowledge and experimental data, in particular data from high–throughput technologies lately [Kli09]. It boosts our biological knowledge to a system level. It brings insight to biomedical research. In the remaining of the chapter, previous researches related to metabolic networks, which are the objects of interest in our research, are reviewed.

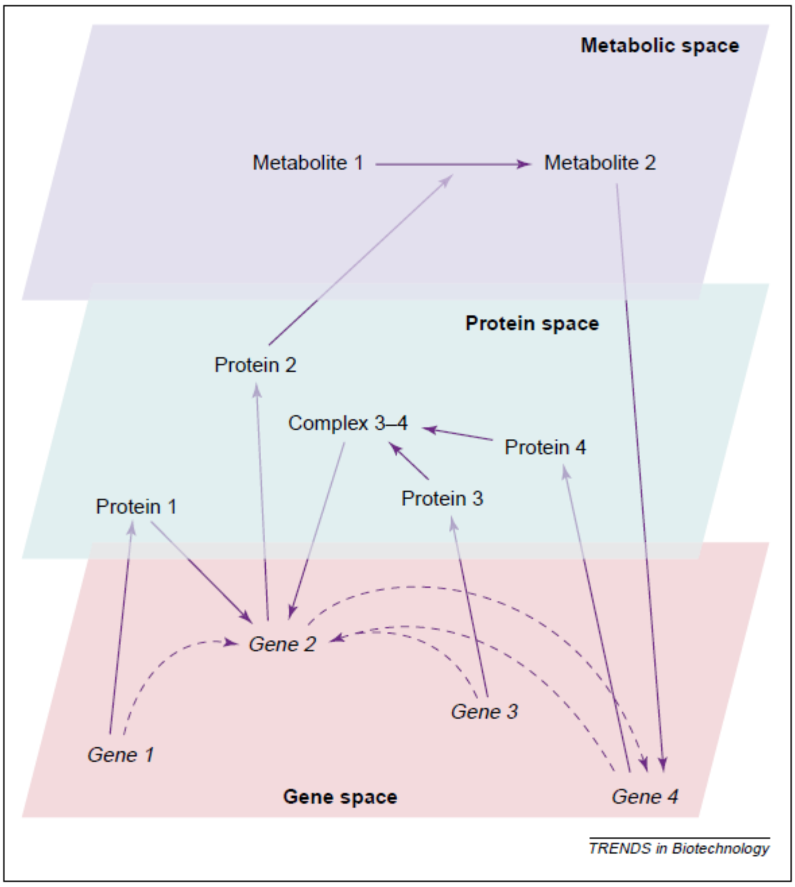


**Figure 2.2 Example of Network of Three Levels**

The interactions between genes, proteins and metabolites shown in the network of three levels, adopted from [Bra02].

## 2.3 Metabolic Networks

Metabolic network of a cell describes all biochemical reactions within the cell. Mathematically, it is represented by a bipartite graph with two types of nodes being metabolites and reactions [Alm07]. Let $R$ be the set of reaction nodes, $M$ be the set of metabolite nodes and $E$ be the set of arcs in the network. Then, the bipartite graph is given by $(R, M, E)$, where $(m, r) \in E$ for $m \in M, r \in R$ if and only if metabolite $m$ is a reactant of reaction $r$ and $(r, m) \in E$ if and only if metabolite $m$ is a product of reaction $r$. Different attributes can be assigned to the nodes and edges. The most common are the reaction stoichiometry assigned to edges and the reaction rates assigned to reaction nodes, which are called 'metabolic fluxes' or simply 'fluxes'. Reactions inside a cell are called internal or intracellular reactions while exchange reactions refer to the exchange of metabolites between a cell and its surrounding environment.

Between the whole network and individual reactions, the important concept of metabolic pathway is used to represent a subset of reactions in the network acting collaboratively to fulfill a certain function. The studies on metabolic network can reveal the metabolism of cells, identify metabolic pathways, predict the behavior of mutants, etc. All of these possess great values in the field of biomedical research, metabolic engineering, etc. In this section, the experimental technologies, the development of modeling and analyzing metabolic network are reviewed.

### 2.3.1 Experimental Technologies

The two types of elements in a metabolic network are metabolites and reactions. Hence two important measurements involved in experiments are values of metabolite concentration and metabolic fluxes. Metabolomics and fluxomics are the study of the whole profile of metabolite concentration [Fie02] and metabolic flux distribution [Sau04], respectively. The technologies applied are nuclear magnetic resonance (NMR) and mass spectrometry (MS) to label the pattern of carbon-13 (13C) [Sau06]. Later, the combination of gas chromatography (GC) and MS, abbreviated GC-MS, sometimes together with time-of-flight detectors (TOF), has allowed data generation of high throughput [Kel04]. By analyzing spectra obtained in experiments, relative ratios or absolute values can be found. For the detail of these experiments and the role of high throughput data in system biology, several reviews [Kel04, Sau04, Sau06, Fie02] are available.

### 2.3.2 *In silico* Metabolic Network Reconstruction

Metabolic networks used to be reconstructed only from biochemical data. After the completion of the genome sequencing for some microorganisms, for instances, the well-known *E. coli* and *S. cerevisiae*, genomic data has undertaken an important role in metabolic reconstruction [Edw00, Ree03, Fei07]. From the annotated genome of an organism, gene products, in particular enzymes, which are able to be produced from the genome, are identified. The possible reactions catalyzed by them are also found out from literature and enzyme databases. All reactions and metabolites involved together form a large dataset, which is the basis of the model. As the primary source of the data comes from genome sequence rather than *in vivo* experimental results, the model is *in silico*. Meanwhile, since it accounts for all possible reactions reflected from the genome, the model is a genome-scale one. A variety of analysis can be applied to conduct various studies based on this large dataset. One of the most popular methods is introduced in the following section.

## 2.4 Flux Balance Analysis

It is difficult to accurately measure fluxes, especially intracellular fluxes. It is more difficult to make simultaneous measurement on a number of fluxes. To overcome the situation, the metabolic flux balance method is adopted based on the steady state assumption [Hor72, Red88]. It assumes metabolism is in a steady state in general with constant flux values and constant metabolite concentrations [Hei77]. By this, the flux balance equation can be derived as below:

Let $\mathbf{S} = [S_{ij}] \in \mathbf{R}^{m \times n}$ be the stoichiometric matrix, with entries given:

1. If the number of metabolite i consumed in reaction j is $k$, then $S_{ij} = -k$.
2. If the number of metabolite i produced in reaction j is $k$, then $S_{ij} = k$.
3. If metabolite i does not participate in reaction j, then $S_{ij} = 0$.

Let $v_j$ be the flux of reaction j , $x_i$ be the concentration of metabolite i.

Then,

$$\frac{dx_i}{dt} = \sum_j S_{ij} v_j$$

By the steady state assumption, we have the flux balance equation:

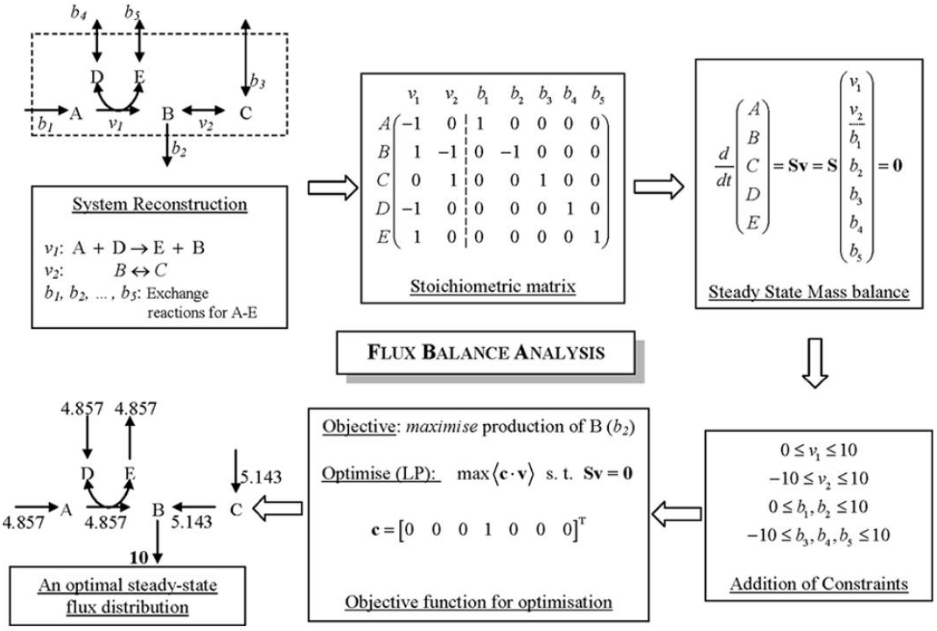
$$\frac{dx_i}{dt} = \sum_j S_{ij} v_j = 0$$

for all metabolites i. It can be written in matrix form:

$$\mathbf{Sv} = \mathbf{0}$$

where $\mathbf{v} = [v_1 \quad \cdots \quad v_n]^T$ is the flux vector.

In the stoichiometric matrix, columns for intracellular reactions must contain both positive and negative entries while each column for each exchange reaction by convention contains only one non-zero entry of '$-1$' for the particular metabolite being exchanged. The upper half of Figure 2.3 shows an example of transforming chemical equations into stoichiometric matrix [Ram09]. The first two columns stand for intracellular reactions and the last five for exchange reactions. The rows stand for the five metabolites A, B, C, D, and E.



$b_4$ $b_5$
D E
$b_3$
A B C
$b_1$ $v_1$ $v_2$
$b_2$
System Reconstruction
$v_1$: A + D → E + B
$v_2$: B ↔ C
$b_1, b_2, \ldots, b_5$: Exchange reactions for A-E
$v_1$ $v_2$ $b_1$ $b_2$ $b_3$ $b_4$ $b_5$
A −1 0 1 0 0 0 0
B 1 −1 0 −1 0 0 0
C 0 1 0 0 1 0 0
D −1 0 0 0 0 1 0
E 1 0 0 0 0 0 1
Stoichiometric matrix
$\frac{d}{dt}(A\ B\ C\ D\ E)^T = \mathbf{Sv} = \mathbf{S}(v_1\ v_2\ b_1\ b_2\ b_3\ b_4\ b_5)^T = \mathbf{0}$
Steady State Mass balance
FLUX BALANCE ANALYSIS
$0 \le v_1 \le 10$
$-10 \le v_2 \le 10$
$0 \le b_1, b_2 \le 10$
$-10 \le b_3, b_4, b_5 \le 10$
Addition of Constraints
Objective: maximise production of B ($b_2$)
Optimise (LP): $\max\langle \mathbf{c}\cdot\mathbf{v}\rangle$ s. t. $\mathbf{Sv} = \mathbf{0}$
$\mathbf{c} = [0\ 0\ 0\ 1\ 0\ 0\ 0]^T$
Objective function for optimisation
4.857 4.857
D E
5.143
4.857 A 4.857 B 5.143 C
10
An optimal steady-state flux distribution

**Figure 2.3 Steps in Flux Balance Analysis.**
The basic steps in FBA, adopted from [Ram09].

As seen in Figure 2.3, the system of linear equations is underdetermined and it is also the practical situation. To estimate the whole flux profile, upper and lower bounds of fluxes are determined first. Upper bounds are provided by enzyme capacity data while lower bounds, zero for irreversible reactions and negative for reversible ones, are obtained from the knowledge of thermodynamics [Pri04]. Biological objective functions are imposed to find optimal solutions from the feasible regions [Bon97]. The earliest used objective is the maximization of growth rate, or biomass synthesis and later, adenosine triphosphate (ATP) production, nicotinamide adenine dinucleotide phosphate (NADPH) production, the square sum of flux values and their combinations have been proposed [Sch07]. The lower half of Figure 2.3 shows an example of the procedure to estimate the fluxes. This whole linear programming (LP) approach is called ‘Flux Balance Analysis’ (FBA) today. It is the most common constraint-based analysis of metabolic network. LP is a well-studied problem and its structure is clean. Solutions are easy to obtain by well-known algorithms like simplex and barrier.

The simplicity and linearity of FBA make additional models and constraints easy to be incorporated. Also, different kinds of studies can be performed under this simple framework. Some of the important developments are introduced below.

### 2.4.1 Dynamic Simulations

An earlier extension of FBA was the dynamic FBA (dFBA) simulating the growth of *E. coli* [Var94]. dFBA divides a duration into a large number of small time steps. In each time step, a flux distribution is calculated by

FBA and is used together with ordinary differential equations to update the concentrations of metabolites which then influence the fluxes calculated in the next time step.

Noticing the above dFBA is static optimization-based, with the same steady-state assumption, a dynamic optimization-based problem has been formulated to simulate the diauxic growth of *E. coli* on glucose and acetate medium [Mah02]. With additional constraints on the rates of change of fluxes, both dynamic and static optimization-based dFBA are performed to compare their results.

### 2.4.2 Mutant Studies

For a strain of a species, especially microorganism which evolves quickly, computational method based on FBA can give valuable insights to the mutants of the strain. Usually in a mutant, some genes are mutated or can be virtually thought as deleted and this is sometimes called genetic perturbations. By constraining the flux of the reaction catalyzed by the respective enzyme to be zero, the flux distribution of a mutant can be estimated by FBA [Edw99]. From a biological intuition that the changes between an original strain and its mutant should be minimized, **MOMA**, a quadratic programming problem (QP) and **ROOM**, a mixed integer linear programming problem (MILP) were formulated to predict mutants' structures by minimizing the total flux change [Seg02] and the number of fluxes constrained to be zero [Shl05], respectively. As a reverse problem, **Optknock**, a bilevel programming problem (BLP) is proposed to find the best gene deletion strategy given the objective of minimizing the fluxes producing certain compounds [Bur03b].

### 2.4.3 Multiple Optimal Phenotypes

Multiple optimal solutions giving the same optimal objective function value often appear. Biologically, this situation can be comprehended as the existence of silent phenotypes [Raa01]. A first attempt to enumerate all the optimal solutions was made in 2000 [Lee00]. The enumeration is performed by a recursive MILP. Then an application on the *in silico* model of *E. coli* found certain properties across these multiple optima and some are matched with other experimental data [Ree04].

### 2.4.4 Additional Constraints

The basic equality constraints in FBA are actually the mass balance constraints, i.e. zero changes of mass for intracellular metabolites. Researchers keep exploring constraints from other aspects to refine the feasible region. It also helps to eliminate biologically unrealistic optima. Two first important additional constraints are the enzyme capacity which estimates the upper bounds of reactions and the reversibility of reactions determined by thermodynamic analysis [Pri04]. The incorporation of regulatory constraints [Cov01, Cov02, Cov03], based on the genome-scale regulatory network, also brings insights to metabolic network. Explicit thermodynamic constraints are added by considering the free energy in the loops of reactions with the cost of non-linearity of the feasible region [Bea02]. In the paper, an analogy to the Kirchoff’s law in Electricity is given in which the traditional mass balance is analogous to the junction law while the proposed energy balance is analogous to the loop law. Attempts to incorporate metabolite concentrations into FBA were also made [Hop07]. Recently, with the awareness of traditional FBA being flux or reaction-centered, flux sum analysis is proposed to consider the maximum and minimum absolute fluxes allowed for individual metabolites as complementary metabolite-centered constraints [Chu09]. Other modifications include establishing the stoichiometric matrix by carbon mole balance rather than mass balance [Jia07], locating a unique solution in FBA from a geometry analysis of the solution space [Sma09], etc.

### 2.4.5 Network Properties

Different analytic techniques related to FBA on the properties of metabolic networks have also been developed, including random sampling, flux variability, flux coupling and metabolic pathway analysis (listed in [Dur09]).

In the random sampling analysis, feasible flux distributions are randomly sampled and the distribution for a certain flux is fitted with a curve [Alm04]. Then by defining a specific measure on the means of fluxes to check their significance, a high-flux backbone of the metabolic network can be found.

The flux variability analysis reveals the ranges of flux values that are able to give the same optimal objective value in FBA [Mah03]. It first finds the optimal solution. Then the constraint of objective value equal to the optimal one is imposed. By minimizing and maximizing each flux as an objective respectively, the variability of each flux is found.

The flux coupling analysis finds the pairwise coupling relations between fluxes by examining the zeros and non-zeros in the solutions of an LP based on FBA [Bur04].

For metabolic pathway analysis, due to the large quantity of research conducted and an intended focus on it, it is separately reviewed in section 2.5.

### 2.4.6 Reverse Engineering from Experimental Data

The research based on genome-scale *in silico* models begins with different additional constraints and additional analysis on the model to gain insightful predictive information. Such information is meaningless without the support of experimental facts. To gain more insights from the available experimental data, computational methods directly accounting for the difference between *in silico* calculation and *in vivo* data, or reconciling the data into a model are developed. Below are some of these methods based on FBA.

One of the important while arguable basis of FBA is the assumed optimality of metabolic flux distribution with respect to a certain objective function, which is based on the theory of evolution. **ObjFind** is a bi-level programming problem (BLP) first developed to find the most consistent linear objective function given a set of experimental fluxes [Bur03a]. The inner problem is the usual FBA with the objective being a linear combination of fluxes, where the coefficients $c_j$ in the combination are determined by the outer problem to minimize the sum of squares of the difference between the experimental data and the optimal solution in the inner problem. The solution technique lies on transforming the problem into a single level non-linear problem by applying strong duality theorem and adding the dual constraints of the inner LP.

Recently, **BOSS** (biological objective solution search), another BLP to predict objectives from experimental flux data has been formulated [Gia08]. The predicted objective is not limited to a linear combination of the existing reactions, but can also be a new reaction with arbitrary stoichiometry on the existing metabolites. The effectiveness of the method is validated by predicting the objective of maximum growth rate in *S. cerevisiae* even when the biomass synthesis reaction is removed but the limitation comes from practical implementation due to the high non-linearity and non-convexity of the algorithm [Gia08].

In addition to predicting objectives, experimental flux data has also been used to identify the set of metabolic reactions *in vivo* from the large set *in silico* [Her06]. The algorithm used is called **OMNI** (optimal metabolic network identification). It is also a BLP with FBA as the inner problem and the outer problem determining which reactions are included in the network by minimizing the discrepancy between experimental flux data and the optimal flux distribution given by the inner problem. The algorithm finds the bottleneck reactions in the *in silico* model of *E. coli*. When the reactions are deleted, the model gives good prediction matching the experimental data well [Her06].

Experimental data of rates of change of metabolite concentration has also been reconciled into metabolic network model. In [Rag03], rates of change of metabolite concentration can be nonzero in contrast to the mass balance in usual FBA. By solving a BLP, intracellular fluxes can be determined. The inner problem resembles FBA with constraints $\sum_j S_{ij} v_j = 0$ relaxed to $\sum_j S_{ij} v_j = r_i$, where $r_i$ is the variable for the rate of change of the *i*-th metabolite concentration. The outer problem minimizes the sum of discrepancies between data and all $r_i$.

From above, it is seen that BLP is a fairly popular method effective to reconcile data into existing model and extract information from data with assumed models.

## 2.5 Metabolic Pathway Analysis

Metabolic pathway is an important concept in metabolic network. The major research conducted concerns the discovery of metabolic pathway. Before the genome-scale *in silico* model appeared, pathways can only be discovered by experimental means. Different computational approaches, however, have been developed today. They can be divided into two categories, path-finding and stoichiometric approaches [Pla08].

### 2.5.1 Path-Finding Approaches

Path-finding approaches typically enumerate all paths from a huge database of annotated genomes, enzymes, reactions and metabolites, for example, **PathFinder** [Goe02] and **PathMiner** [McS03]. These enumeration methods encounter the problem that the number of possible paths is too large for a systematic analysis due to the combinatorial nature of the composition of paths [Kuf00]. For more biologically meaningful paths, connectivity of metabolite nodes, a network property is considered [Cro05]. Connectivity in graph theory is the number of edges connected to a node, so it describes the number of reactions that a metabolite participates in. Methods have been developed to find paths minimizing the connectivity of the metabolites involved [Cro06, Pla09a]. These methods yield much better results.

### 2.5.2 Stoichiometric Approaches: Elementary Flux Modes

As noted in the name, stoichiometric approaches make use of the stoichiometric information in chemical equations. An earlier stoichiometry method is the elementary flux mode (EFM), or simply elementary mode [Sch94]. Mathematically, the feasible solution space in FBA is a convex polyhedral cone. By convex analysis, if we treat a reversible reaction as two reactions with non-negative fluxes, then every feasible solution can be represented as the non-negative sum of the extreme rays of the convex polyhedral cone. These rays define EFMs and they are pathways in metabolic networks. EFMs satisfy the property of genetic independence, or called non-decomposability or elementarity. This means an EFM cannot be decomposed as the sum of two other flux modes whose reactions with non-zero fluxes are subsets of the reactions with non-zero fluxes in the EFM. Practically, EFMs have been applied to a variety of studies: investigating network structures and robustness [Ste02], dynamic properties of pathways [Ste07], pathway efficiency [Car07] and modularity [Yoo07], exploring new pathways [Sch99, deF09b] and hence suggesting rational strain design [Kla04, Tri08 Tri09a], predicting mutants' behavior [Zha09a], identifying interactions with other networks [Kla06], etc. (reviewed in [Dur09, Gag04, Lla10, Tri09b]).

### 2.5.3 Stoichiometric Approaches: Extreme Pathway

In the set of EFMs, a certain EFM can virtually be the combination of other EFMs when the decomposed forward and backward reactions of the same reversible reaction can cancel out each other [Pap04], i.e., when they are transferred back to the original flux space. As a complementary definition, extreme pathway, abbreviated as EP, is defined to be a basis of the convex flux cone that cannot be decomposed into the non-negative sum of other convex bases [Sch00]. Hence the number of extreme pathways is always no larger than that of elementary modes. Figure 2.4 [Pap04] shows a clear distinction between two definitions. EIMo 4 in Figure 2.4b is the resultant of ExPa 1 and ExPa 2 in Figure 2.4c [Pap04].

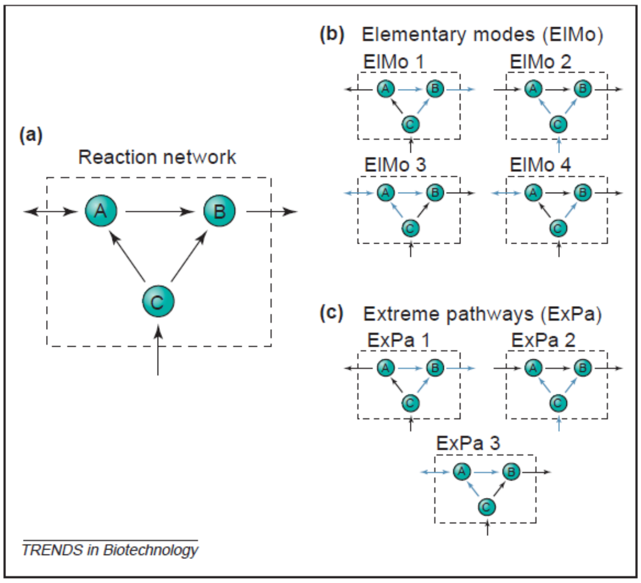


**Figure 2.4 Elementary Modes and Extreme Pathways**

(a) Sample Network (b) Elementary Modes (c) Extreme Pathways [Pap04]

### 2.5.4 Flux Distributions Composed of Metabolic Pathways

An important interpretation of EFMs is that every flux distribution $\mathbf{v} = [v_1 \cdots v_n]^T$ can be decomposed as a positive sum of a subset of EFMs $\{\mathbf{e}_i = [e_{i1} \ \cdots \ e_{in}]^T\}$ with the 'no cancelation' property that all $\mathbf{e}_j$ have a particular component being zero wherever **v** has that component being zero [Sch02]. That is,

$$\mathbf{v} = \sum_j w_j \mathbf{e}_j, \ w_j > 0 \ \text{and} \ Z(\mathbf{v}) \subset Z(\mathbf{e}_j) \ \text{for all}\, j,$$

where $w_j$ is the weight for $\mathbf{e}_j$ and $Z(\mathbf{v}) = \{i \mid v_i = 0\}$ is the index set for the zero components of **v**. The 'no cancelation' property is unique compared to other generating sets of the flux space like EPs and allows EFMs to answer questions regarding pathways producing or consuming specific metabolites [Lla10].

In many studies, flux distributions are analyzed by similar decompositions. Because such decomposition is in general not unique, a particular decomposition is usually chosen with respect to an optimization objective. The optimization problem can be stated as follows:

$$\begin{aligned} &\min \ f(\mathbf{w}) \\ &\text{subject to} \qquad \mathbf{Ew} = \mathbf{v} \\ &\qquad\qquad\qquad\ \ \mathbf{w} \geq \mathbf{0} \end{aligned}$$

where $\mathbf{E} = [\mathbf{e}_1 \ \cdots \ \mathbf{e}_K]$ is the matrix with all EFMs as its columns; $\mathbf{w} = [w_1 \ \cdots \ w_K]^T$ is the weight vector for EFMs; $f(\mathbf{w})$ is the objective function dependent on **w**. It is remarked that each $\mathbf{e}_j$ is usually scaled to the largest possible fluxes given the upper bound for each reaction. The first attempt is the α-spectrum to determine the range of attainable weights for each EP [Wib03]. This approach is also applicable to EFMs. Later, decomposing flux distributions with minimum sum of weights has been proposed [Schw05] and applied to study yeast glycolysis [Schw06]. Decompositions with respect to other objectives have also been adopted to investigate particular cases, for instances, maximum number of active EFMs [Noo07], minimum relative error [Wan07], maximum yield rate [Son09] and maximum entropy [Zha09b]. There are also other studies analyzing flux distributions by decompositions into EFMs e.g. [Kur07, Car09]. Table 2.1 summarizes all different objectives.

| Objective | Objective function $f(\mathbf{w})$ | Ref. |
|---|---|---|
| α-spectrum | $\pm w_i$ for each $i = 1, \ldots, K$ | [Wib03] |
| Min. no. of EFMs | $\sum_{i=1}^{K} a_i$ where $a_i = \begin{cases} 1 \text{ if } & w_i > 0 \\ 0 \text{ if } & w_i = 0 \end{cases}$ | [Wib03] |
| Min. sum of squares of weights | $\mathbf{w}^T\mathbf{w} \ \left( = \sum_{i=1}^{K} w_i^{\ 2} \right)$ | [Schw05]<br>[Schw06] |
| Max. no. of EFMs | $-\sum_{i=1}^{K} a_i$ where $a_i = \begin{cases} 1 \text{ if } & w_i > 0 \\ 0 \text{ if } & w_i \le 0 \end{cases}$ | [Noo07] |
| Max. yield rate | $\mathbf{w}^T\mathbf{w} \ \left( = \sum_{i=1}^{K} w_i^{\ 2} \right)$,<br>each $\mathbf{e}_j$ scaled to a particular uptake reaction. | [Son09] |
| Max. entropy | $\sum_{i=1}^{K} \rho_i \ln \rho_i$<br>where $\rho_i = \frac{e_{iu_s}}{v_{u_s}} w_i$ and $u_s$ is the uptake reaction of the extracellular metabolite *s*. | [Zha09b] |

**Table 2.1 Different Objectives to Decompose Flux Distributions in Literature**

### 2.5.5 Computation of Elementary Flux Modes

The approaches mentioned in section 2.5.4 based on decomposing flux distributions into sets of EFMs, despite the insights they have provided, have limited capabilities because the calculation always requires the complete set of EFMs *in prior* given a metabolic network, whose computation is notoriously hard due to the combinatorial explosion when the network size grows. This drawback was first studied by Klamt and Stelling [Kla02] and has been mentioned in many literatures, e.g. [Dur09, Kla02, Lla10, Ter09, Tri09b, Yeu07].

To overcome the difficulty, efforts have been put to improve the computational speed and memory demand for computing EFMs. The first algorithm was presented by Schuster *et al.* [Sch96]. Improvements have then been made continuously by introducing the nullspace approach to reduce computational cost [Wag04,Urb05]; the binary approach to save a great deal of memory and allow bit operations [Gag04]; the elementary testing by matrix rank to accelerate computations [Kla05]; the bit pattern trees to hasten elementary testing [Ter06] and the latest recursive enumeration approach based on bit pattern trees [Ter08]. Software implementing these approaches has also been engineered, including METATOOL [Pfe99, von06], FluzAnalyzer [Kla03], CellNetAnalyzer [Kla07], *EFMtool* [Ter08]. Despite these improvements, these algorithms still cannot cope with genome-scale metabolic networks reconstructed recently, usually consisting of at least a thousand of reactions and metabolites (e.g. [Fei07, Dua07]).

One characteristic in common of the mentioned approaches to calculate

EFMs is that they all compute the complete set of EFMs given the network stoichiometry and this is the major reason of the computational infeasibility in genome-scale networks because the size of the full set grows extremely fast due to combinatorial explosion [Kla02]. As an alternative, recently there have been attempts to find specific metabolic pathways from stoichiometric information by optimization modeling. They are able to cope with the scale and furthermore to identify pathways with specific requirements. Pathways matched with literature are found [Bea07] by balancing 'low presence' compounds and minimizing the number of reactions involved and ATP consumed which is supported by evolution viewpoint. Later by further considering connectivity of compounds, linear pathways were located [Pla09b]. It is able to reconstruct most of the metabolic paths in *E. coli*. Also, the K-shortest EFMs in genome-scale networks have been calculated successfully by an optimization model [deF09a]. More recently, the K-shortest generating flux modes, a subset of EFMs, have also been investigated by a similar model [Rez11].

## 2.6 Summary

The basic definition and representation of metabolic network have been introduced in this chapter. Among the existing methods to investigate it, FBA is one of the most fundamental while simple methods. Major studies based on FBA have been reviewed.

A relevant and very important area studying the properties of metabolic networks is metabolic pathway analysis. Definitions, interpretations, applications as well as challenges in metabolic pathway analysis in particular EFMs have been presented. It is remarked that while the computation of the full sets of EFMs has come to a bottleneck, recent attempts to find metabolic pathways by optimization models has been successful. This inspires the proposed research. Detailed reviews in metabolic pathway analysis are available [Lla10, Pla08, Pap04, Sch96, Sch00, Tri09b].

# Chapter 3 A Decomposition Approach

## 3.1 Algorithm to Decompose Flux Distributions

In this chapter, we present an algorithm to find a set of EFMs that decomposes a given flux distribution in a genome-scale metabolic network without finding the complete set of EFMs *in prior*. The algorithm solves optimization models recursively with data stored in a stack structure. Its implementation is highly dynamic. By editing the model being solved, a set of EFMs with certain properties can be found. In addition to the basic decomposition, the algorithm is applied to approximate EFMs with largest contributions to a particular objective reaction in a given flux distribution.

## 3.2 Decomposability

Suppose the stoichiometric matrix is given by $\mathbf{S}_0 = [\mathbf{S}_{irr} \mid \mathbf{S}_{rev}] \in \mathbf{R}^{m\times(n_1+n_2)}$ with $\mathbf{S}_{irr} \in \mathbf{R}^{m\times n_1}$ being the columns for irreversible reactions and $\mathbf{S}_{rev} \in \mathbf{R}^{m\times n_2}$ for reversible reactions. A flux distribution, or equivalently a flux mode, is a vector **v** satisfying the steady state assumption and the thermodynamic constraint regarding the reversibility of reactions, i.e.

$$\mathbf{v} \in \mathbf{F}_0 = \left\{ \mathbf{v} = \begin{bmatrix} \mathbf{v}_{rev} \\ \mathbf{v}_{irr} \end{bmatrix} \mid \mathbf{S}_0\mathbf{v} = \mathbf{0}, \mathbf{v}^{irr} \geq \mathbf{0}, \mathbf{v}_{irr} \in \mathbf{R}^{n_1}, \mathbf{v}_{rev} \in \mathbf{R}^{n_2} \right\} \subseteq \mathbf{R}^{n_1+n_2}$$

where $\mathbf{v}^{rev}$, $\mathbf{v}^{irr}$ are the vectors for reversible and irreversible reactions respectively. $\mathbf{F}_0$ is called the flux space. Note that $\mathbf{F}_0$ is inside the null space of $\mathbf{S}_0$.

An EFM by definition is a flux mode unable to be decomposed as the sum of two other flux modes, whose active reactions are proper subsets of the active reactions of that flux mode [Sch94]. By 'active' reactions, we mean reactions with non-zero fluxes. The definition is also called non-decomposability, elementarity or genetic independence in literature. Mathematically, a flux mode **e** is an EFM if there are no flux modes $\mathbf{v}_1$, $\mathbf{v}_2$ such that

$$\mathbf{e} = \mathbf{v}_1 + \mathbf{v}_2 \text{ with } Z(\mathbf{e}) \underset{\neq}{\subset} Z(\mathbf{v}_1) \text{ and } Z(\mathbf{e}) \underset{\neq}{\subset} Z(\mathbf{v}_2)$$

where $Z(\mathbf{v}) = \{i \mid v_i = 0\}$ is the index set for the zero components of **v**. If for a flux mode **e**, such $\mathbf{v}_1$, $\mathbf{v}_2$ exist, they are said to be bounded by **e**. The algorithm proposed exploits this decomposability. In what follows, first, an optimization model to check the decomposability of a given flux mode is formulated, followed by the algorithm integrating the model to decompose flux distributions. Then, modifications of the algorithm to approximate EFMs of largest contributions are presented.

## 3.3 Decomposability Check

In the optimization model formulated in this section, all reactions are assumed to be irreversible. This can be done easily in reality by replacing a reversible reaction by two irreversible reactions with stoichiometry negative to each other. Explicitly, we define a new stoichiometric matrix **S** by $\mathbf{S} = [\mathbf{S}_{irr} \mid \mathbf{S}_{rev} \mid -\mathbf{S}_{rev}] = [S_{ij}] \in \mathbf{R}^{m \times n}$ where $n = n_1 + 2n_2$. The corresponding new flux space $\mathbf{F}$ is given by

$$\mathbf{F} = \left\{ \mathbf{v} = \begin{bmatrix} \mathbf{v}_{irr} \\ \mathbf{v}_{rev}^{+} \\ \mathbf{v}_{rev}^{-} \end{bmatrix} \middle| \, \mathbf{Sv} = \mathbf{0}, \mathbf{v} \geq \mathbf{0}, {\mathbf{v}_{rev}^{+}}^{T} \mathbf{v}_{rev}^{-} = 0, \mathbf{v}_{irr} \in \mathbf{R}^{n_1}, \mathbf{v}_{rev}^{+}, \mathbf{v}_{rev}^{-} \in \mathbf{R}^{n_2} \right\} \subseteq \mathbf{R}^{n}$$

Note that $\mathbf{v}_{rev} = \mathbf{v}_{rev}^{+} - \mathbf{v}_{rev}^{-}$. The conditions $\mathbf{v} \geq \mathbf{0}$ and ${\mathbf{v}_{rev}^{+}}^{T} \mathbf{v}_{rev}^{-} = 0$ ensures that for each reversible reaction, only one of the entries in $\mathbf{v}_{rev}^{+}, \mathbf{v}_{rev}^{-}$ is non-zero.

In the model, the decomposability of a given flux mode $\mathbf{v} = [v_1 \ \cdots \ v_n]^T$ is determined by examining whether a flux mode $\mathbf{p} = [p_1 \ \cdots \ p_n]^T$ bounded by **v** can be found. First, **p** must satisfy the usual steady state condition:

$$\sum_{j=1}^{n} S_{ij} p_j = 0 \text{ for } i = 1, \ldots, m \tag{3.1}$$

Each reaction is assigned with a binary integer variable $a_j$ specifying the on/off condition by the following constraints:

$$a_j \leq p_j \text{ for } j = 1, \ldots, n \tag{3.2}$$

$$p_j \leq M \delta_j a_j \text{ for } j = 1, \ldots, n \tag{3.3}$$

where *M* is a large positive number and $\delta_j = \begin{cases} 1 & \text{if } v_j > 0 \\ 0 & \text{if } v_j = 0 \end{cases}$. If $a_j = 0$, from (3.3), we have $p_j = 0$ and thus reaction *j* is not involved, else if $a_j = 1$, from (3.2), we have $p_j \geq 1$ and reaction *j* must have a positive flux. Constraints similar to (3.2 – 3.3) have been employed in optimization models to find pathways before [Bea07, deF09am Rez10]. The difference of our approach lies in the introduction of $\delta_j$. If $\delta_j = 0$, $a_j$ and $p_j$ are forced to be zero by (3.2 – 3.3). In other words, if $v_j = 0$, then $p_j = 0$. This implies $Z(\mathbf{v}) \subset Z(\mathbf{p})$ which is a necessary condition for **p** to be bounded by

**v**. For the large number *M*, it can be interpreted as the greatest stoichiometric ratio permissible within the flux mode **p**, so it should be sufficiently large to allow all flux modes bounded by **v** to satisfy (3.2 – 3.3).

To ensure **p** is non-trivial, i.e. has at least one non-zero flux, we have:

$$\sum_{j=1}^{n} a_j \geq 1 \tag{3.4}$$

For **p** to be properly bounded by **v**, the number of non-zero fluxes in **p** must be strictly less than the number of non-zero fluxes in **v**, so we have:

$$\sum_{j=1}^{n} a_j \leq \sum_{j=1}^{n} \delta_j - 1 \tag{3.5}$$

Finally, since all reactions are irreversible, all fluxes must be positive:

$$p_j \geq 0 \text{ for } j = 1, \ldots, n \tag{3.6}$$

Constraints (3.1 – 3.6) suffice to define **p** as a flux mode bounded by **v**. Hence any objective can work. We simply choose maximizing zero:

$$\max 0 \tag{3.7}$$

(3.1 – 3.7) form a mixed integer linear program that we call Decomposability Check (DC). Given a flux mode **v**, if any feasible solution can be found in solving DC, then **v** is decomposable. Otherwise, we conclude that **v** is an EFM.

Different from an early study [deF09a], **p** is not required to be integers though this may lengthen computational time. In some of our investigations, fractional stoichiometric coefficients are involved, like the biomass composition, and turning these coefficients into integral will make them too large in magnitude and inconvenient for computation.

## 3.4 Decomposition of Flux Distributions

The algorithm to decompose flux distributions is an iterative scheme of DC. Flux modes involved are stored in a stack structure. The decomposability of the given flux distribution is first examined by DC. If it is decomposable, the flux mode bounded by it returned by DC will be stacked up and become the current flux mode. Else if it is non-decomposable, an EFM is reached and it leaves the stack. The EFM found is then used to update each intermediate flux mode by subtracting a scalar multiple of the EFM. After updating, intermediate flux modes unable to contribute to the first flux mode are removed. This procedure is repeated until all flux modes leave the stack. A set of EFMs decomposing the flux distribution is then obtained.

### 3.4.1 Algorithmic Steps

Let *N* be the number of flux modes in the stack, *K* be the number of EFMs found. The *s*-th flux mode in the stack is denoted as $\mathbf{fm}_s$ with the flux of the *j*-th reaction being $fm_{sj}$. $\mathbf{efm}_k$ and $efm_{kj}$ are similarly defined for the *k*-th EFM. The steps of the algorithm can be summarized as follows.

Step 0. Initialize with $N = 1$, $K = 0$ and $\mathbf{fm}_1 = \mathbf{v}$.

Step 1. Solve DC with $\mathbf{fm}_N$ as input.

If there is a feasible solution **p**, go to step 2. Otherwise, go to step 3.

Step 2. Update $N$ by $N+1$. Set $\mathbf{fm}_N = \mathbf{p}$. Go to step 1.

Step 3. Update $K$ by $K+1$. Set $\mathbf{efm}_K = \mathbf{fm}_N$.

If $N = 1$, terminate the algorithm, else go to step 4.

Step 4. For $s=1,\ldots,N-1$, update $\mathbf{fm}_s$ by $\mathbf{fm}_s - r_s^K \times \mathbf{efm}_K$

where $r_s^K = \min_j \left\{ fm_{sj} / efm_{Kj} \middle| efm_{Kj} > 0 \right\}$. Go to step 5.

Step 5. Remove $\mathbf{fm}_N$ from the stack.

If $N > 2$, for $s = 2,\ldots,N-1$, check if $Z(\mathbf{fm}_1) \underset{\neq}{\subset} Z(\mathbf{fm}_s)$.

If not, remove $\mathbf{fm}_s$ from the stack.

Set $N$ to be the current size of the stack. Go to step 1.

An example of the algorithm can be found in A.1 in the appendices. The algorithm can be interpreted as decomposing $\mathbf{fm}_1$ repeatedly until $\mathbf{fm}_1$ becomes an EFM. Step 1 checks the decomposability of the current flux mode $\mathbf{fm}_N$. If a feasible solution exists, the flux mode is not an EFM and it is replaced by a new flux mode bounded by it which is found by DC as indicated in step 2. The procedure is repeated until an EFM is reached.

Once an EFM is found, in step 3, first we check whether it is the only flux mode in the stack. If it is, this means it is the last EFM and the algorithm is terminated. Otherwise, it will be used to update all flux modes in the stack as in step 4. Note that after each updating step, $\mathbf{fm}_1$ is the flux mode remaining to be further decomposed. This updating process serves to eliminate the largest possible fluxes able to be contributed by the EFM found in each preceding flux mode. The number of non-zero entries in each flux mode is diminished. This step can accelerate computations because there are less non-zero entries needed to be dealt with in each flux mode. In practice, step 4 can be efficiently performed by matrix multiplications.

After updating flux modes, some intermediate flux modes may have positive entries that have become zero in $\mathbf{fm}_1$. The ‘no cancellation’ property states that these flux modes cannot contribute to $\mathbf{fm}_1$ anymore, so they are removed from the stack in step 5, which can be carried out by bit operations in practice. In fact, they can still contribute to **v**, but it turns out that step 4 and step 5, besides saving memory and computational cost, guarantee that the finally resulting set of EFMs has the four nice properties described in the following subsection.

### 3.4.2 Properties of the Solution Found by the Algorithm

The first two properties are the ‘denseness’ and ‘uniqueness’ of the solution. No EFM found is redundant and each EFM has a unique positive weight in the decomposition. The other two properties of greater theoretical interest are the linear independence and systemic independence of the solution set, i.e. the set of solution EFMs forms a linear basis as well as a convex basis for the flux distribution. These four properties, which follow from the fact that our algorithm decomposes a flux distribution by stepwise finding an EFM and reducing the flux mode, also hold in the original flux space $\mathbf{F}_0$. They together give an exact role for each EFM in the solution. The cooperation between different pathways can be clearly revealed. This brings more exact biological interpretations. These properties are proved in the following theorem.

**Theorem.** *Given a flux distribution* $\mathbf{v} = [v_1 \ \cdots \ v_n]^T \in \mathbf{F}$, *the set of EFMs found by the proposed algorithm is a dense and unique decomposition of the flux distribution, i.e. each EFM has a unique positive weight in the decomposition. Furthermore, the set of EFMs is linearly and systematically independent. All of these properties also hold in the original flux space* $\mathbf{F}_0$.

Proof:

The *s*-th flux mode in the stack is denoted as $\mathbf{fm}_s$ with the flux of the *j*-th reaction being $fm_{sj}$. Similarly, the *k*-th EFM found is denoted as $\mathbf{efm}_k$ with the flux of the *j*-th reaction being $efm_{kj}$.

During the algorithm, whenever an EFM is found, in step 4, it is used to update other flux modes $\mathbf{fm}_s$ in the stack by $\mathbf{fm}_s - r_s^K \times \mathbf{efm}_K$, where $r_s^K = \min_j \left\{ fm_{sj} / efm_{Kj} \middle| efm_{Kj} > 0 \right\}$, in particular the first flux mode $\mathbf{fm}_1$ which in the very beginning is the original flux distribution. Hence, when the *K*-th EFM is found, we have:

$$\mathbf{v} = \mathbf{fm}_1 + \sum_{k=1}^{K} r_1^k \mathbf{efm}_k \ . \tag{3.8}$$

For $s = 1, \ldots, N_K - 1$, where $N_K$ is the size of the stack when the *K*-th EFM is found, let

$$j_s^K = \arg\min_j \left\{ fm_{sj} / efm_{Kj} \middle| efm_{Kj} > 0 \right\}. \tag{3.9}$$

Note that $j_s^K$ is the reaction whose flux in $\mathbf{fm}_s$ can be totally contributed by the EFM found. In step 5, intermediate flux modes unable to contribute to $\mathbf{fm}_1$ anymore are deleted, i.e. those flux modes $\mathbf{fm}_s$ not satisfying $Z(\mathbf{fm}_1) \underset{\neq}{\subset} Z(\mathbf{fm}_s)$. Therefore, whenever $efm_{Kj} > 0$, $fm_{1j} > 0$ because:

$$Z(\mathbf{fm}_1) \underset{\neq}{\subset} Z(\mathbf{fm}_2) \underset{\neq}{\subset} Z(\mathbf{fm}_3) \underset{\neq}{\subset} \cdots \underset{\neq}{\subset} Z(\mathbf{fm}_{N_K}) = Z(\mathbf{efm}_K).$$

This ensures $r_1^K > 0$ and $fm_{1 j_1^K} > 0$. Assume there are $K_0$ EFMs in the final solution. From (3.8), finally we have

$$\mathbf{v} = \sum_{k=1}^{K_0} r_1^k \mathbf{efm}_k \tag{3.10}$$

Note that $\mathbf{efm}_{K_0}$ is just $\mathbf{fm}_1$ at the last step of the algorithm and $r_1^{K_0} = 1$. This proves the denseness of the decomposition because each $r_1^k > 0$.

For the uniqueness of the weight of each EFM, consider again when the *K*-th EFM is found. After updating the flux mode in step 4, by definition (3.9), the flux of reaction $j_1^K$ in the new $\mathbf{fm}_1$ becomes

$$fm_{1j_1^K} - r_1^K efm_{Kj_1^K} = fm_{1j_1^K} - \left(fm_{1j_1^K} / efm_{Kj_1^K}\right) efm_{Kj_1^K} = 0 . \tag{3.11}$$

From (3.11), it can be seen that step 4 will make at least one positive entry in $\mathbf{fm}_1$ ( $j_1^K$ ) become zero. The optimization model used in our algorithm then ensures that the EFMs found later during the algorithm must have that entry ( $j_1^K$ ) equal to zero due to $Z(\mathbf{fm}_1) \underset{\neq}{\subset} Z(\mathbf{efm}_k)$ for all $k > K$ while the present EFM has that entry being positive, i.e.

$$efm_{Kj_1^K} > 0 \text{ and } efm_{kj_1^K} = 0 \text{ for all } k > K . \tag{3.12}$$

To prove the uniqueness, assume there is another set of weights $(w_1, \ldots, w_{K_0})$ for the same set of EFMs such that

$$\sum_{k=1}^{K_0} w_k \mathbf{efm}_k = \sum_{k=1}^{K_0} r_1^k \mathbf{efm}_k .$$

Then we have

$$\sum_{k=1}^{K_0} (w_k - r_1^k) \mathbf{efm}_k = \mathbf{0} .$$

For each $K = 1, \ldots, K_0$, look at reaction $j_1^K$:

$$\sum_{k=1}^{K_0} (w_k - r_1^k) efm_{kj_1^K} = \sum_{k=1}^{K} (w_k - r_1^k) efm_{kj_1^K} = 0 \tag{3.13}$$

The first equality results from (3.12) since $efm_{kj_1^K} = 0$ for all $k > K$.

For $K=1$, since from (3.12), we have $efm_{1j_1^1} > 0$, together with (3.13), we have

$$(w_1 - r_1^1)efm_{1j_1^1} = 0 \Rightarrow w_1 = r_1^1$$

When $K=2$, we have

$$\sum_{k=1}^{2}(w_k - r_1^k)efm_{kj_1^2} = (w_1 - r_1^1)efm_{1j_1^2} + (w_2 - r_1^2)efm_{2j_1^2} = (w_2 - r_1^2)efm_{2j_1^2} = 0\,.$$

This implies $w_2 = r_1^2$. Similarly, for $K = 1,\ldots,K_0$, by looking at reaction $j_1^1,\ldots,j_1^{K_0}$ respectively, we have $w_K = r_1^K$ and hence the set of weights is unique.

From above, it is obvious that the set of EFMs is linearly independent. Assume

$$\sum_{k=1}^{K_0}\alpha_k \mathbf{efm}_k = \mathbf{0}\,.$$

By looking at reaction $j_1^1, j_1^2,\ldots,j_1^{K_0}$ respectively as above, we have $a_k = 0$ for all *k* and hence the set of EFMs is linearly independent. The set is also systematically independent, following from the linear independence because systematic independence requires that for any $k_1 \neq k_2 \neq k_3 \neq k_1$, there does not exist $\alpha,\beta > 0$ such that

$$\alpha\mathbf{efm}_{k_1} + \beta\mathbf{efm}_{k_2} = \mathbf{efm}_{k_3}\,.$$

The linear independence naturally guarantees this property.

In the original flux space $\mathbf{F}_0$, for each pair of reactions representing for a reversible reaction in $\mathbf{F}_0$, at most one of the reactions has non-zero flux. Define the function mapping the indices of reactions in $\mathbf{F}_0$ and $\mathbf{F}$:

$$f:\{1,\ldots,n\}\to\{1,\ldots,n_1+n_2\} \text{ such that } f(j)=\begin{cases} j & \text{if } 0\le j\le n_1+n_2 \\ j-n_2 & \text{if } n_1+n_2 < j \le n \end{cases}.$$

Note that $f\left(j_1^K\right)$ are distinct for all $1 \le K \le K_0$ as $\mathbf{efm}_{K,rev}^{+}{}^{T}\mathbf{efm}_{K,rev}^{-} = 0$ holds for all EFMs. Hence, the condition similar to (3.12) holds in $\mathbf{F}_0$ :

$$efm_{Kf\left(j_1^K\right)} \begin{cases} > 0 \text{ if } & 0 \le j_1^K \le n_1 + n_2 \\ < 0 \text{ if } & n_1 + n_2 < j_1^K \le n \end{cases} \text{ and } efm_{kf\left(j_1^K\right)} = 0 \text{ for all } k > K . \tag{3.14}$$

From (3.14), exactly the same argument for the uniqueness, linear independence and the consequent systematic independence applies. For the 'denseness', it is obvious that (3.10) still holds in $\mathbf{F}_0$ with each $r_1^k > 0$ . □

There are two major differences of the proposed algorithm compared with previous optimization models to find pathways. First, it is an iterative scheme to solve optimization models instead of solving a single optimization model. Second, any feasible solution can finally become an EFM and optimality is not necessarily needed. Any optimization objective works. In what follows, we propose one for a specific application.

## 3.5 Approximation of EFMs of Largest Contributions

In flux balance analysis (FBA), cellular metabolism is assumed to achieve an optimal state with respect to an objective like the maximum growth rate and ATP production (reviewed in [Sch07]).

Conversely, for a flux distribution (maybe experimentally measured, or simulated by methods other than FBA), if a biologically reasonable objective of the cell can be assumed, it will be insightful to decompose the flux distribution into EFMs that have considerable contributions to that objective. By 'contribution', we mean the flux of the objective reaction provided by an EFM in the flux distribution. Finding largest contributing EFMs can reveal principal operational modes in cells. We applied our algorithm to approximate such decompositions with only two modifications required. The first is the replacement of objective (3.7) by:

$$\max\ p_{j_0} \tag{3.7'}$$

where $j_0$ is the objective reaction that we are interested. The resulting solution flux mode then has the greatest flux at reaction $j_0$. Nonetheless, the flux mode can possibly contribute little to $\mathbf{fm}_1$. This comes to the second modification which replaces constraint (3.3) by:

$$p_j \le M'\, fm_{1j} a_j \text{ for } j = 1, \dots, n \tag{3.3'}$$

Here $\delta_j$ is replaced by $fm_{1j}$. This takes the flux values of $\mathbf{fm}_1$ into account. Intuitively, the solution flux mode should have a better contribution because upper bounds for fluxes are not the same but proportional to the fluxes of $\mathbf{fm}_1$. In fact, the stepwise solution has maximum contribution to the objective reaction flux in $\mathbf{fm}_1$ among all flux modes bounded by $\mathbf{fm}_1$. This also forms the rationale of applying the algorithm to approximate EFMs with largest contributions: in each step, the best flux mode bounded by the current flux mode is found until an EFM is reached. Hence, it is a greedy approach. For the large number $M'$, it should be chosen large enough to properly scale $\mathbf{fm}_1$ to allow all feasible flux modes.

## 3.6 Implementations

Both versions of the optimization model DC [version 1: objective (3.7) subject to constraints (3.1 – 3.6); version 2: objective (3.7’) subject to constraints (3.1 – 3.2), (3.3’), (3.4 –3.6)] are mixed integer linear programs

(MILPs). It can be solved efficiently by various software packages. ILOG CPLEX® was used in all the investigations reported in this thesis. For the whole algorithm, all the data and operations other than solving optimization models were processed in MATLAB®. For hardware, all computations were performed in a computer with a 2.67 GHz CPU and 24 GB of RAM.

## 3.7 Summary

In this chapter, by exploiting the definition of non-decomposability, an optimization to check the decomposability of a flux distribution and an algorithm to decompose a flux distribution into a set of EFMs without first finding the complete set of EFMs have been devised. The set of solution EFMs found by the algorithm is dense, unique, linearly and systematically independent. A characteristic of the algorithm is that the elementarity of the flux modes found is independent of the optimality of the optimization model. This enables the algorithm to be applied to approximate EFMs of the largest contributions to a particular reaction of interest.

# Chapter 4 Method Validation

## 4.1 Introduction

In this chapter, the algorithm proposed in Chapter 3 is applied to three metabolic networks of different sizes to test its capability. The first is the sample network originally used to demonstrate regulatory FBA [Cov01] and later applied to study $\alpha$-spectrum [Wib03]. The full set of EFMs of this network of small size can be computed easily by existing methods. The second is the core metabolic network of *E. coli* strain K-12 MG1655 iAF1260 [Fei07]. It is more realistic with a size considerably larger than the first one but the subset of all EFMs consuming certain substrate can still be computed. These two networks are used as benchmarks to examine the validity of our algorithm.

Then, to highlight the usefulness of our algorithm complementary to existing methods, a computational experiment is conducted. Flux distributions of optimal growth rate subject to various substrate availability in the complete *E. coli* MG1655 iAF1260 metabolic network [Fei07] are generated and are used to compare the performance of our algorithm and *EFMtool* [Ter08], which is presently the most efficient algorithm to find full sets of EFMs up to the best of our understanding.

## 4.2 Benchmarks

For each of the two networks for benchmarks, we first randomly sample 2000 flux distributions [Alm04] and decompose them by the two versions of our algorithm. We then check whether the solutions are true EFMs that are able to be located by existing methods. For the approximation of EFMs of largest contributions, we choose the biomass production, or equivalently the growth rate, as the objective reaction. We check the rankings of contributions of the EFMs found by our algorithm among the complete set of EFMs. The sets of EFMs for the two networks will be calculated by *EFMtool* [Ter08] and the random sampling is implemented with the COBRA Toolbox [Bec07].

### 4.2.1 Test Procedure for Benchmarks

The test follows the procedure below:

1. Randomly sample 2000 flux distributions;
2. Run our algorithm (version 1 and version 2);
3. Check whether all solutions are true EFMs able to be found by *EFMtool*;
4. For solutions from version 2, in each of the sample flux distributions, we check the ranking of the best contributing EFM in the solution as follows. Suppose the full set of EFMs is $\mathbf{EFM} = \{\mathbf{e}_k \mid k = 1,\ldots,K\}$ and the sample flux distribution is $\mathbf{v} = [v_1 \ \cdots \ v_n]^T$. For each $\mathbf{e} \in \mathbf{EFM}$, compute $c_{\mathbf{e}} = r \times e_{growth}$ where $r = \min_j \{v_j / e_j | e_j > 0\}$. Note that $c_{\mathbf{e}}$ is the maximum possible

contribution to $v_{growth}$ by **e**. Denote **efm** the best contributing EFM in the solution with a weight $w$. The contribution by **efm** is $w \times efm_{growth}$. Then the ranking of **efm** is given by the number of elements in the following set:

$$\{c_{\mathbf{e}} \mid \mathbf{e} \in \mathbf{EFM} \text{ and } c_{\mathbf{e}} \geq w \times efm_{growth}\}$$

In other words, the number of different contributions larger than or equal to the contribution by **efm**. Also, the percentage difference of the contribution by **efm** to the contribution by the overall best contributing $\mathbf{e} \in \mathbf{EFM}$ is calculated:

$$\frac{\left(w \times efm_{growth} - \max(c_{\mathbf{e}})\right)}{\max(c_{\mathbf{e}})} \times 100\%$$

### 4.2.2 Sample Metabolic Network

There are 20 reactions and 19 metabolites in the sample metabolic network [Cov01]. The information of the metabolic network is given in section A.1 in the appendices. The number of EFMs found by *EFMtool* was 82, the same as in Wiback *et al.* [Wib03]. In that paper, an algorithm for extreme pathways was used, but since the authors had treated each reversible reaction as two irreversible reactions, the resulting set of pathways is in fact the set of EFMs [Lla10].

In average, version 1 of the algorithm took 5s to decompose a flux distribution and version 2 took 6s. It was verified that all solutions belong to the set of EFMs computed by *EFMtool*.

For version 2, among the 2000 samples, EFMs of maximum contributions to growth rate in the flux distributions are the first EFMs calculated in 98% of the solutions. The corresponding percentage for version 1 which does not optimize for any objective is only about 15%. All other solutions by version 2 except one include EFMs of the second or third largest contributions. For each sample, the percentage differences between the best contributions to the objective in the solutions by both versions of the algorithm and the best contribution in the full set of EFM are calculated. The comparison between both versions shows that version 2 of our algorithm has a distinguishable power to find EFMs of best contributions (Table 4.1).

| Network | Version | Mean | Median | S.D. |
|---|---|---|---|---|
| Sample Network | 1 | -27.2% | -28.5% | 18.7% |
| | 2 | -0.3% | 0% | 2.3% |
| Core *E. coli* Network | 1 | -19.1% | -8.9% | 23.6 |
| | 2 | -0.7% | 0% | 4.9% |

**Table 4.1 Percentage difference of the contribution of the first EFM by the algorithm to the contribution by the overall best EFM in each sample.**

### 4.2.3 Core *E. coli* K-12 MG1655 iAF1260 Metabolic Network

The core *E. coli* metabolic network contains 95 reactions and 72 metabolites. The complete set of EFMs for the network is unable to be computed by *EFMtool* due to insufficient memory in the test computer. The available uptake substrates are then restricted to glucose, phosphate, $CO_2$, $H^+$, $H_2O$, $NH_4$ and $O_2$ only. Over 100,000 EFMs were found by *EFMtool*. The number is comparable to previous studies with similar configurations [Kla02]. One issue complicating the computation is the fractional stoichiometry in the biomass reaction estimated from molecular content

[Fei07]. Magnitudes of coefficients in the reaction highly vary. The ratio of the greatest to the smallest is 840.

In average, version 1 of the algorithm took 12s to finish and version 2 took 18s. All solutions are verified to be true EFMs.

For version 2, in the 2000 trials, the greatest contributing EFMs are first found in 97% of solutions. 80% of the remaining solutions contain EFMs whose contributions rank top 10% among all EFMs. This further confirms the utility of our algorithm, especially in view of the number of different possible contributions by all EFMs in each sample, which is over 7000 in average (Table 4.2).

| Network | No. of different contributions | Percentile of the first EFM | |
|---|---|---|---|
| | | Version 1 | Version 2 |
| Sample Network | 32 | 0.7% | 0.1% |
| Core *E. coli* Network | 7010 | 2.9% | 0.2% |

**Table 4.2 Average number of different contributions of all EFMs and average percentile of the first EFM by the algorithm in each sample.**

### 4.2.4 Conclusion on Benchmarks

Benchmark results for the two networks with quite different sizes show that our algorithm is able to decompose flux distributions into sets of EFMs. Furthermore, it is capable of approximating the best contributing EFM with respect to a certain reaction and the success rate of finding the largest contributing EFM is significant.

## 4.3 Computational Experiment

In cell culture experiments, minimal media containing one carbon source and necessary inorganic compounds only are often preferred due to the ease in analysis. Correspondingly, flux distributions of optimal growth rate simulating these cases consist of only few EFMs regardless of the network complexity. In this case, existing tools are able to find the set of EFMs that contributes to a flux mode by considering the subnetwork formed by the active reactions in the flux distribution (called subnetwork of a flux distribution for simplicity). Realistic metabolic fluxes are, however, shaped by factors like thermodynamics, gene regulation and in particular heterogeneous nutrients [Zam10]. Structures of flux distributions are thus more complicated and existing tools may not cope with them. To test the capability of our algorithm to analyze these cases, we perform a computational experiment in the complete *E. coli* K-12 MG1655 iAF1260 genome-scale metabolic network.

### 4.3.1 Complete *E. coli* K-12 MG1655 iAF1260 Metabolic Network

The complete *E. coli* K-12 MG1655 iAF1260 genome-scale metabolic network reconstructed by Feist *et al.* [Fei07] is a huge network much more complex than the core one studied in section 4.2.3. There are 1039 metabolites and 2382 reactions, of which 852 are reversible. After compartmentalization and transformation of each reversible reaction into two irreversible reactions, there are 3234 irreversible reactions and 1668 metabolites, of which 299 are extracellular metabolites for uptake. The detailed information of the metabolic network is available [Fei07].

### 4.3.2 Procedure of the Computational Experiment

In each trial of the experiment, in addition to inorganic compounds necessary for growth, a random set of extracellular carbon sources is selected to be available for uptake and a flux distribution of optimal growth rate is determined by FBA. It is then decomposed by version 2 of our algorithm with the growth rate as the objective. Meanwhile, we try to find the complete set of EFMs of the subnetwork of the flux distribution by *EFMtool*.

Let $\mathbf{S} = [\mathbf{s}_1 \ \cdots \ \mathbf{s}_n]$ be the stoichiometric matrix. Let $\mathbf{UT}_{carbon}$be the index set for the uptake reactions for extracellular carbon sources ($|\mathbf{UT}_{carbon}| = 278$). The steps for each trial of the test are as follows:

1. Randomly select a number of $k$ carbon sources from $\mathbf{UT}_{carbon}$ to form the set $\mathbf{UT}_{carbon}^{select}$.
2. Obtain a flux distribution $\mathbf{v} = [v_1 \ \cdots \ v_n]^T$ of optimal growth by the following optimization model:

$$\begin{array}{lll} \max v_{growth} & & \\ \text{subject to} & \mathbf{Sv} = \mathbf{0} & \\ & 0 \le v_j \le 1000 & \text{for } j \notin \mathbf{UT}_{carbon} \\ & 0 \le v_j \le 1 & \text{for } j \in \mathbf{UT}_{carbon}^{select} \\ & v_j = 0 & \text{for} \end{array}$$

$j \in \mathbf{UT}_{carbon} \setminus \mathbf{UT}_{carbon}^{select}$

3. Decompose $\mathbf{v}$ by the version 2 of our algorithm with the growth rate as the objective.
4. Assume without loss of generality that for $1 \le j \le N_r \le n$, $v_j > 0$

where $N_r$ is the number of active reactions in $\mathbf{v}$. The stoichiometric matrix of the subnetwork of $\mathbf{v}$ is given by $\mathbf{S}_\mathbf{v} = [\mathbf{s}_1 \ \cdots \ \mathbf{s}_{N_r}]$. Examine whether *EFMtool* is able to find the full set of EFMs of $\mathbf{S}_\mathbf{v}$ with all reactions indicated as irreversible reactions.

The trials of step 1 to step 4 are repeated with for $k = 10, 20, \ldots, 270$ (step 1). For each fixed *k*, the trials are repeated 10 times, i.e. ten different sets of carbon sources are randomly selected.

### 4.3.3 Results of the Computational Experiment

In the computational experiment, the number of active reactions $N_r$, partially reflecting the complexity of a flux distribution, increases with the number of carbon sources consumed $N_c$ linearly (Figure 4.1). In general, when $N_c \geq 30$, *EFMtool* is unable to find the full set of EFMs of the subnetwork of the flux distribution. In contrast, our algorithm succeeded to decompose every test flux distribution.

In certain cases with large $N_r$ in the flux distributions but a small number of contributing EFMs $N_{efm}$ (<50) in the solution found by our algorithm, *EFMtool* is unable to find the sets of EFMs of the subnetworks while our algorithm can find EFMs decomposing the flux distributions in relatively short time (Figure 4.2, ‘+’s with small $N_{efm}$). This corresponds to the situation in which a flux distribution having an enormous number of contributing EFMs can actually be represented as a convex sum of only a few of those EFMs. Our algorithm is very useful and efficient in this case.

Moreover, the memory demand is not expensive as the most complicated cases in the computational experiment can also be solved by our algorithm in a computer with 2 GB memory only.

The complexity of our algorithm is also examined. The computational time increases with $N_{efm}$ as well as $N_r$ (Figure 4.2). Simple linear regression indicates a satisfactory linear relationship between the computational time and the product $N_{efm} \times N_r$ ($R^2 = 0.97$, Figure 4.3).

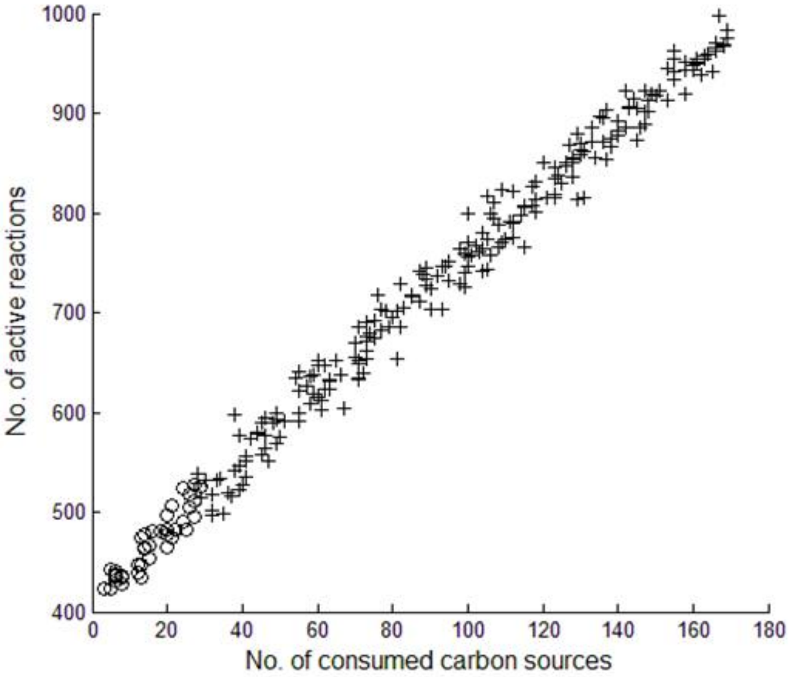


**Figure 4.1 The number of active reactions against the number of consumed carbon sources in the test flux distribution.**

'o' represents the case in which the set of EFMs of the subnetwork of active reactions can be calculated by *EFMtool*. '+' represents the corresponding case in which the set cannot be calculated by *EFMtool*.

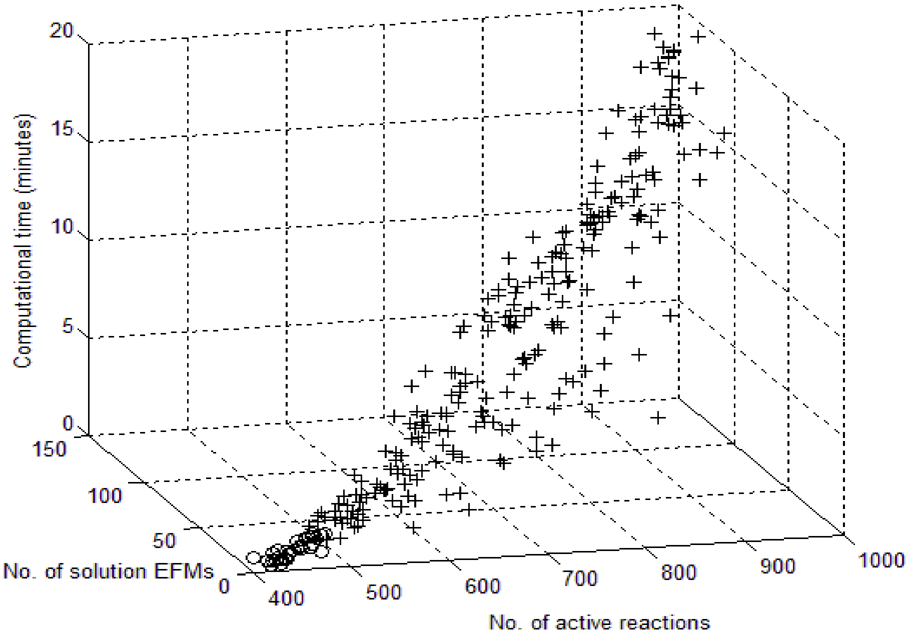


**Figure 4.2 The computational time of the proposed algorithm against the number of active reactions and the number of EFMs in the solution found.**

'o' represents the case in which the set of EFMs of the subnetwork of active reactions can be calculated by *EFMtool*. '+' represents the corresponding case in which the set cannot be calculated by *EFMtool*.

### 4.3.4 Remarks on the Computational Experiment

The computational experiment has shown that our algorithm is more advantageous than existing methods in analyzing flux distributions with complex structures, which are closer to realistic cases in which a cell is subject to complicated factors.

## 4.4 Conclusion

In this chapter, we have demonstrated the ability of the proposed algorithm to decompose flux distributions and approximate EFMs with largest contributions to flux distributions in genome-scale metabolic networks. In particular, it is more advantageous than existing methods in respect of finding EFMs of flux distributions with complex structures. We conclude that the proposed algorithm is able to be applied to analyze flux distributions in the genome-scale metabolic network in which the full set of EFMs is always intractable.

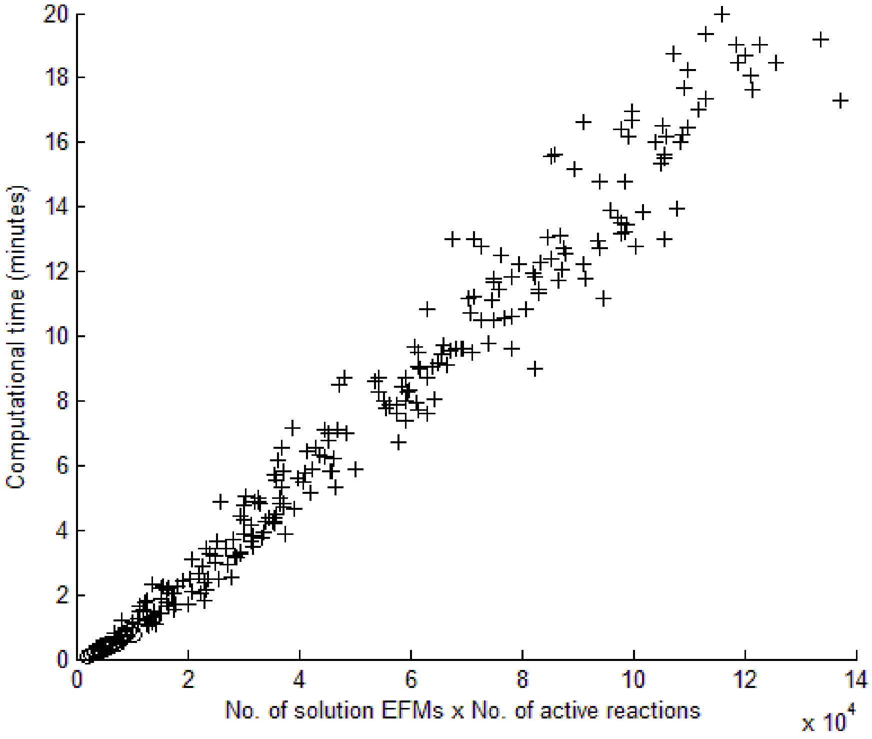


**Figure 4.3 Computational time against the number of solution EFMs multiplied by the number of active reactions.**

'o' represents the case in which the set of EFMs of the subnetwork of active reactions can be calculated by *EFMtool*. '+' represents the corresponding case in which the set cannot be calculated by *EFMtool*.

# Chapter 5 The Growth of *E. Coli*

## 5.1 Introduction

In this chapter, we apply the algorithm to the complete *E. coli* MG1655 iAF1260 metabolic network [Fei07] to investigate the growth of *E. coli*. Different to section 4.3, we aim to look into the flux distributions from a biological perspective. First, flux distributions in a minimal medium with glucose as the only carbon source are simulated and analyzed by our algorithm. Second, a flux distribution simulating a much more complicated case of growth of *E. coli* on the Lysogeny broth (LB) medium [Bae06a, Bae06b] is studied in detail with our algorithm.

## 5.2 Growth of *E. coli* on the Glucose Minimal Medium

To simulate the flux distribution of growth, adenosine triphosphate maintenance (ATPM) cost, biomass composition and maximum uptake rates for glucose and oxygen, all experimentally determined in [Fei07], are adopted. ATPM is represented by a reaction in the model which hydrolyzes the high energy phosphate bond in ATP into adenosine diphosphate (ADP) and phosphate. It is the usual way to generate energy in most organisms. The flux value of ATPM can be interpreted as the production rate of extra ATP used for hydrolysis for purposes not shown in the model. Uptake conditions for the two cases of growth, aerobic and anaerobic growth, are also provided [Fei07]. Based on the information, we first optimize for the growth rate to obtain in silico flux distributions for the two cases of growth. Then our algorithm is applied to decompose the flux distributions. Finally, a sensitivity analysis on how the uptake rate limits and the ATPM cost influence the cooperation of EFMs to maximize biomass production is performed by our algorithm.

### 5.2.1 Aerobic Growth

In the aerobic growth condition, glucose is the only carbon source with its uptake rate capped by 8 mmol gDW$^{-1}$ h$^{-1}$. The maximum oxygen uptake rate is 18.5 mmol gDW$^{-1}$ h$^{-1}$. Other ions and small molecules are also available for uptake. The ATPM cost is 8.39 mmol gDW$^{-1}$ h$^{-1}$. The cost is imposed by setting the value as the lower bound for the reaction.

With the biomass production as the objective reaction, after applying our algorithm to decompose the flux distribution for optimal aerobic growth, two EFMs were found. The first EFM is responsible for 84% of the biomass production and the second EFM is responsible for the rest of the growth and all ATPM. The two EFMs use nearly the same set of reactions. Each EFM has only one reaction different from each other.

Since there is only glucose present in the medium as the single carbon source, the structures of the flux distributions generated are expected to be relatively simple. The small number of EFMs involved is thus not surprising in the light of the optimal nature of the flux distribution calculated. Since the only constraint on the flux distribution is the ATPM cost, the solution can be interpreted as first fulfilling the ATP requirement by an EFM which is balanced in producing ATP and biomass, and then producing biomass by the most efficient EFM using the rest of the resources.

### 5.2.2 Anaerobic Growth

In the anaerobic case, all conditions remain the same except the oxygen uptake is prohibited. In the optimized flux distribution, four extracellular organic compounds, acetate, ethanol, formate and succinate, are produced as a result of fermentation. Again, application of our algorithm yielded two EFMs. This time the roles of the EFMs are more obvious. The first mode is responsible for biomass production only and the second mode, which is a mixed acid fermentation mode producing acetate, ethanol and formate, is responsible for the ATP maintenance only.

### 5.2.3 Sensitivity Analysis

The factors determining the optimal flux distributions in the two cases and the resulting decomposition into EFMs in the model include the ATPM cost, oxygen and glucose uptake rate limits. For example, if there is no ATPM cost, the first EFM in the previous aerobic case producing the majority of biomass will be adopted entirely as the flux distribution. In contrast, if the ATPM cost is too high, we can anticipate that a pathway with high efficiency for ATPM will be chosen and little biomass will be produced using the rest of the glucose and oxygen. By varying these coefficients and applying our algorithm, different modes of pathway cooperation can be revealed.

Figure 5.1 shows the result of sensitivity analysis by fixing the glucose uptake rate limit and varying the oxygen uptake rate limit and the ATPM cost. From the boundary of the scatter diagram, it can be seen that the maximum ATPM cost sustainable increases quite linearly with the maximum oxygen uptake rate. Although there are only five types of EFMs involved regarding what they consume and produce, the modes of their cooperation vary largely (Table 5.1). An interesting point during the analysis is that an optimal flux distribution is always composed of different suboptimal EFMs regarding their efficiencies in growth or maintenance. This highlights the importance of suboptimal pathways and suggests a possible reason for the high redundancy of pathways in organisms. Their cooperation can help the metabolism to best adapt to different environmental conditions.

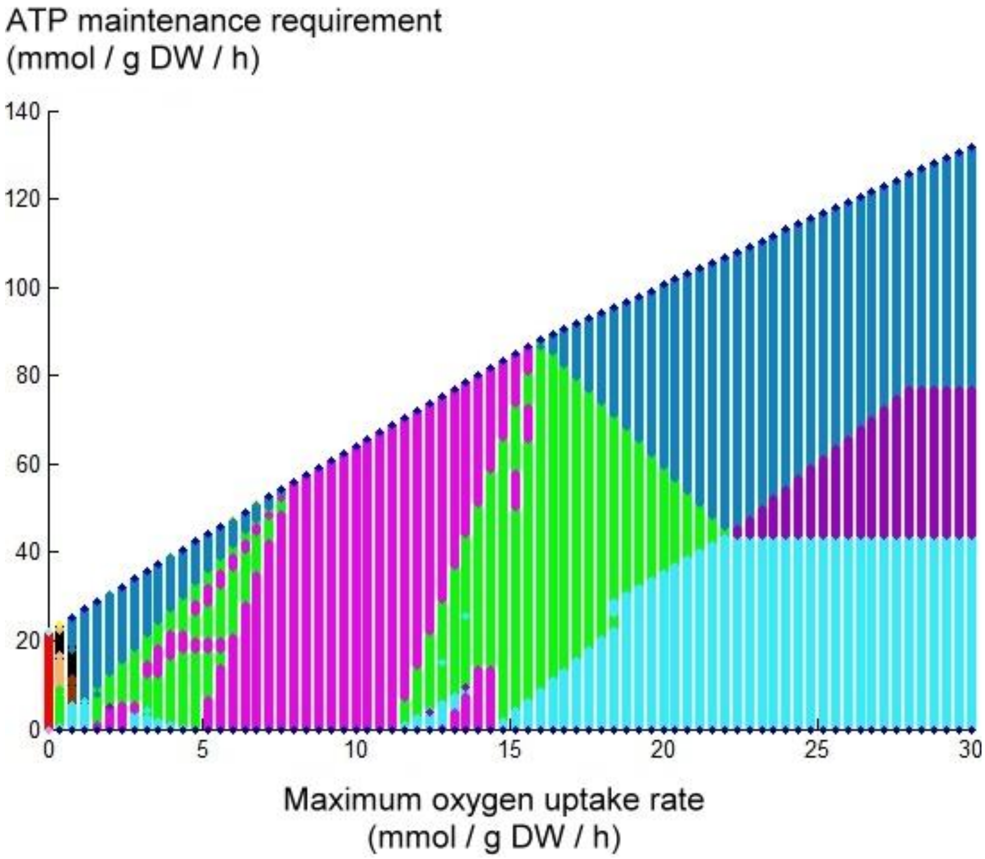


**Figure 5.1 Scatter diagram for different modes of cooperation between EFMs**

Different modes are plotted with different icons. The legend is given in Table 5.1.

| EFM | 🟦 | 🟦 | 🟪 | 🟪 | 🟪 | 🟦 | 🟦 | 🟦 | 🟩 | 🟨 | 🟧 | 🟥 | 🟫 | ⬛ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ARG | √ |  |  | √ |  | √ |  |  | √ |  | √ |  | √ |  |
| ARM |  | √ |  | √ |  |  |  | √ | √ | √ | √ |  |  | √ |
| ARGM |  |  | √ |  |  | √ |  | √ | √ |  |  |  | √ | √ |
| ANG |  |  |  |  | √ |  |  |  |  |  |  | √ |  |  |
| ANM |  |  |  |  |  |  | √ |  |  | √ | √ | √ | √ | √ |

**Table 5.1 Legend for the different modes of cooperation in Figure 5.1**

‘AR’: aerobic; ‘AN’: anaerobic; ‘G’: growth; ‘M’: ATP maintenance.
For example, a tick in the box for ‘ARGM’ means an aerobic EFM contributing to both growth and ATP maintenance is involved in that case.

## 5.3 Growth of *E. coli* on the Lysogeny Broth Medium

The Lysogeny broth (LB) medium, or usually called Luria-Bertani medium, is one of the most common complex media used in bacterial growth. It is composed of tryptone, yeast extract and NaCl. In their studies, Baev *et al.* monitored the utilization of sugars, alcohols, organic acids [Bae06a] and amino acids, peptides, nucleotides [Bae06b] indirectly by transcriptional microarrays during the growth of *E. coli* MG1655 in the LB medium. Interestingly, during about 3 hour to 5 hour of fermentation, simultaneous assimilation of a set of carbon sources was observed [Bae06a]. Combining the result in Baev *et al.* [Bae06b], a large number of sugars, amino acid and other carbon sources were absorbed during the time period. We simulate a flux distribution describing the metabolism of *E. coli* under such conditions by optimizing the growth rate in FBA and allowing the list of carbon sources in Table 5.2 and some inorganic compounds for uptake. Maximum oxygen uptake rate (18.5 mmol gDW$^{-1}$) and ATP maintenance cost (8.39 mmol gDW$^{-1}$ h$^{-1}$) under growth condition determined experimentally in Feist *et al.* [Fei07] are adopted. All maximum uptake rates of carbon sources are set to be 0.5 mmol gDW$^{-1}$ h$^{-1}$ because information on enzyme capacities is not given and under this constraint, the resulting flux distribution consumed all 25 carbon sources while a test on a range of values of maximum uptake rates (≤ 20 mmol gDW$^{-1}$ h$^{-1}$, randomly assigned to each source repeated for 100 times) does not show significant differences in the structure of the flux distribution. We denote the simulated flux distribution by $f_0$ to avoid confusion.

| **Sugar** | N-Acetyl-glucosamine (acgam) | |
|---|---|---|
| L-Arabinose (arab-L) | L-Fucose (fuc-L) | D-Galactose (gal) |
| D-Glucosamine (gam) | D-Mannose (man) | Melibiose (melib) |
| L-Rhamnose (rmn) | Trehalose (tre) | |
| **Sugar alcohol** | D-Mannitol (mnl) | Glycerol (glyc) |
| **Organic acid** | D-Lactate (lac-D) | L-Lactate (lac-L) |
| **Amino acid** | L-Arginine (arg-L) | L-Aspartate (asp-L) |
| L-Glutamine (gln-L) | Glycine (gly) | L-Histidine (his-L) |
| L-Methionine (met-L) | L-Serine (ser-L) | |
| **Nucleotide** | | Adenosine monophosphate (amp) |
| Cytidine monophosphate (cmp) | | Guanosine monophosphate (gmp) |
| Inosine monophosphate (imp) | | Uridine monophosphate (ump) |

| **Inorganic compound** | | Cob(I)alamin | Molybdate | Phosphate | Tungstate |
|---|---|---|---|---|---|
| $Ca^{2+}$ | $Cl^{-}$ | CO2 | $Co^{2+}$ | $Cu^{2+}$ | $Fe^{2+}$ |
| $Fe^{3+}$ | $H_2O$ | $H^{+}$ | $K^{+}$ | $Mg^{2+}$ | $Mn^{2+}$ |
| $Na^{+}$ | $NH_4^{+}$ | O2 | $SO_4^{2-}$ | $Zn^{2+}$ | |

**Table 5.2 List of extracellular metabolites available in the LB medium**
Abbreviations adopted from [Fei07] are written in brackets.

### 5.3.1 Test of Growth by Individual Carbon Sources

There are 493 active reactions in $f_0$. To understand the complexity and the difference of $f_0$ due to multiple sources, we examine whether there is any flux mode bound by $f_0$ that is contributed solely by a single carbon source. Let $\mathbf{C} = \{C_1, \ldots, C_{25}\}$ be the index set of the uptake reactions of the 25 extracellular carbon sources in Table 5.2. The flux mode with maximum sum of fluxes bounded by $f_0$ contributed by each individual carbon source is found by solving the following optimization model for $k = 1, \ldots, 25$:

$$\max \sum_{j=1}^{n} v_j$$

$$\begin{aligned} \text{subject to} \quad & \mathbf{Sv} = \mathbf{0} \\ & 0 \le \mathbf{v} \le \mathrm{f}_0 \\ & v_j = 0 \qquad \text{for } j \in \mathbf{C} \setminus \{C_k\} \end{aligned}$$

Surprisingly, only three non-zero flux modes (atpm1~3 in Table 5.3) consuming L-fucose, D-lactate and L-lactate respectively are resulted, each verified to be EFMs by *EFMtool*. They together provide the whole ATPM flux in $\mathrm{f}_0$ and no single carbon source is independently consumed for growth.

Meanwhile, we test the ability of each carbon source to give non-zero growth rate independently by solving the following optimization model for $k = 1, \ldots 25$:

$$\max v_{growth}$$

$$\begin{aligned} \text{subject to} \quad & \mathbf{Sv} = \mathbf{0} \\ & 0 \le v_j \le 1000 \quad \text{for } j = 1, \ldots, n \\ & v_j = 0 \qquad \text{for } j \in \mathbf{UT}_{carbon} \setminus \{C_k\} \\ & 0 \le v_{C_k} \le 1 \end{aligned}$$

It turns out that all carbon sources except L-histidine and L-methionine can generate growth independently.

From these two results, we can conclude that first, more efficient growth can be achieved by the simultaneous assimilation of the carbon sources in Table 5.2 if they are the only sources available, rather than the assimilation of only a single source at a time. This provides a rationale for the switch of the mode of assimilation from a sequential one to a simultaneous one as observed in Baev *et al.* [Bae06a]. Second, $f_0$ cannot be simply decomposed into flux modes of individual carbon sources. This necessitates the use of our algorithm.

| FM | $N_r$ | $N_c$ | ATPM | GR | FM | $N_r$ | $N_c$ | ATPM | GR |
|---|---|---|---|---|---|---|---|---|---|
| $f_0$ | 493 | 25 | 8.39 | 0.9975 | gr11 | 422 | 12 | 0 | 0.0050 |
| atpm1 | 35 | 1 | 5.375 | 0 | gr12 | 412 | 12 | 0 | 0.0047 |
| atpm2 | 23 | 1 | 2.25 | 0 | gr13 | 417 | 14 | 0 | 0.0025 |
| atpm3 | 23 | 1 | 0.765 | 0 | gr14 | 419 | 12 | 0 | 0.0017 |
| $f_1$ | 482 | 23 | 0 | 0.9975 | gr15 | 422 | 12 | 0 | 0.0009 |
| gr1 | 431 | 16 | 0 | 0.1918 | gr16 | 417 | 14 | 0 | 0.0006 |
| gr2 | 422 | 13 | 0 | 0.1726 | gr17 | 420 | 14 | 0 | 0.0004 |
| gr3 | 428 | 15 | 0 | 0.1652 | gr18 | 415 | 13 | 0 | 0.0003 |
| gr4 | 416 | 14 | 0 | 0.1522 | gr19 | 417 | 12 | 0 | 0.0003 |
| gr5 | 421 | 14 | 0 | 0.1134 | gr20 | 417 | 13 | 0 | 0.0002 |
| gr6 | 411 | 12 | 0 | 0.0825 | gr21 | 417 | 13 | 0 | 0.0002 |
| gr7 | 422 | 14 | 0 | 0.0565 | gr22 | 414 | 13 | 0 | 0.0001 |
| gr8 | 417 | 12 | 0 | 0.0262 | | | | | |
| gr9 | 420 | 12 | 0 | 0.0147 | | | | | |
| gr10 | 417 | 12 | 0 | 0.0055 | | | | | |

**Table 5.3 Flux modes responsible for ATPM and growth**

FM: flux mode; $N_r$: number. of active reactions; $N_c$: number of consumed carbon sources; ATPM: ATPM flux (mmol $gDW^{-1}$); GR: growth rate (mmol $gDW^{-1}$). atpm1~3 and gr1~5 account for all ATPM flux and 80% of the growth rate respectively. All flux modes except $f_0$, $f_1$ are EFMs.

### 5.3.2 Results of the Decomposition by the Algorithm

$f_0$ is then simplified into $f_1$ by subtracting atpm1~3 in Table 5.3. We try to find the set of EFMs of the subnetwork of $f_1$ by *EFMtool* but it is unsuccessful due to insufficient memory. We then decompose $f_1$ by version

2 of our algorithm. 22 EFMs are found. All are verified to be true EFMs by *EFMtool*. The first EFM, also the one with the largest growth rate in the solution, accounts for 19% of the growth rate.

Consistent with the test of individual carbon sources in section 5.3.1, all EFMs of growth consume multiple sources simultaneously, at least twelve (Table 5.4).

The 22 EFMs consist of quite similar sets of reactions. In fact, among the 482 active reactions in $f_1$, 376 reactions are found to be shared by all EFMs, called the 'backbone' reactions (Figure 5.2). They can be interpreted as the necessary reactions for optimal growth on the medium. The structures are complicated, as expected from the detailed biomass composition containing 63 metabolites. Various biochemical pathways are involved: glycolysis, pentose phosphate pathway, non-mevalonate pathway, glycerophospholipid metabolism, lipopolysaccharide biosynthesis, cell envelope biosynthesis, biosynthesis of different amino acids, cofactor and prosthetic group, etc. They are mainly connected by branch point metabolites like pyruvate (pyr) and chorismate (chor). Interestingly, in the backbone, besides the uptake of 8 extracellular carbon sources, some cytosolic metabolites are always synthesized in each EFM and then enter into the backbone acting as source nodes, including L-aspartate, L-serine, D-glucosamine 6-phosphate, glyceraldehyde 3-phosphate, D-glucose 6-phosphate, pyruvate, alpha-D-ribose 1-phosphate, uridine, D-xylulose 5-phosphate. Meanwhile, we find that outside the backbone, there is not any single extracellular carbon source used by all EFMs. This means that these cytosolic sources of the backbone must thus be first synthesized regardless of the carbon sources. Hence, the ability to synthesize them represents the

| gr | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| acgam | √ | √ | √ |  | √ |  | √ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| amp | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ |
| arab_L | √ |  |  |  |  |  |  |  |  |  | √ |  |  | √ | √ | √ |  |  |  |  |  | √ |
| arg_L | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ |
| asp_L | √ |  | √ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| cmp | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ |
| fuc_L |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| gal |  |  |  | √ |  |  |  | √ |  | √ |  |  | √ |  |  |  | √ | √ | √ | √ | √ |  |
| gam |  |  |  | √ |  | √ |  | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ |
| gln_L | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ |
| gly | √ | √ | √ | √ | √ |  | √ |  |  |  |  |  | √ |  |  |  | √ |  |  |  |  |  |
| glyc | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ |
| gmp | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ |
| his_L | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ |
| imp | √ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| lac_D |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| lac_L |  |  |  |  |  |  |  | √ |  |  |  |  |  |  |  |  |  |  | √ |  |  |  |
| man |  |  |  | √ |  | √ |  |  | √ |  |  | √ |  |  |  |  |  |  |  |  |  |  |
| melib | √ |  | √ |  |  |  | √ |  | √ |  | √ |  |  | √ | √ |  |  |  |  |  |  |  |
| met_L | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ |
| mnl |  |  | √ |  |  |  |  |  |  | √ |  |  | √ |  |  | √ | √ | √ |  | √ | √ |  |
| rmn |  | √ |  |  | √ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ser_L | √ | √ | √ | √ | √ | √ | √ |  |  |  |  | √ | √ |  |  | √ | √ | √ |  | √ | √ | √ |
| tre |  | √ |  |  | √ |  | √ |  |  |  |  |  |  |  |  | √ |  |  |  |  |  | √ |
| ump | √ |  | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ |
| N$_c$ | 16 | 13 | 15 | 14 | 14 | 12 | 14 | 12 | 12 | 12 | 12 | 12 | 14 | 12 | 12 | 14 | 14 | 13 | 12 | 13 | 13 | 13 |

**Table 5.4 Consumed carbon sources of the 22 EFMs contributing to growth**

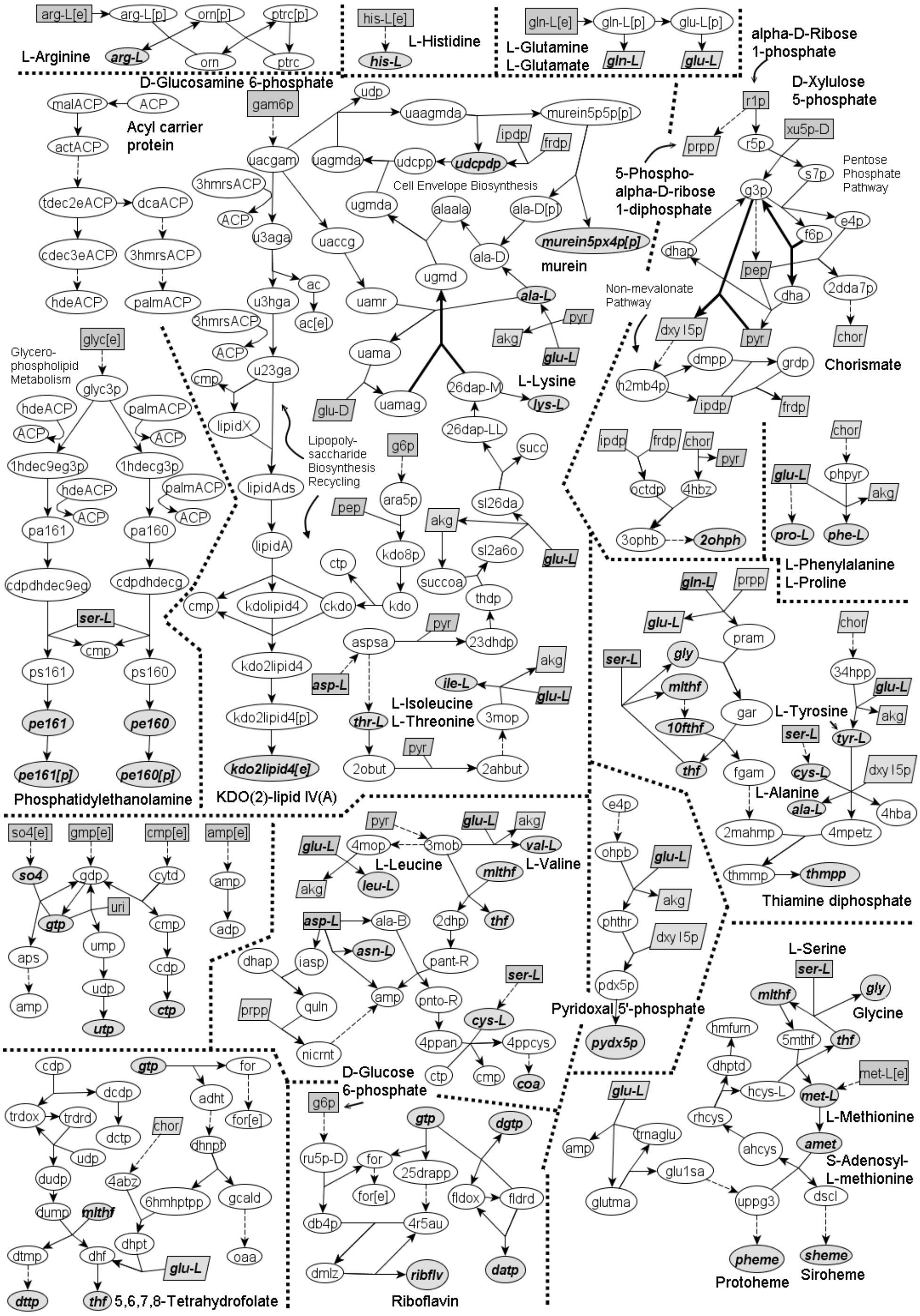


**Figure 5.2 'Backbone' reactions for $f_1$**

Sources entering the backbone are in squares. Metabolites involved in many pathways are in parallelograms. Metabolites comprising biomass are in *bold and italic text*. Independent intersecting edges are bolded for clarity. Small molecules, cofactors and their reactions are not shown for simplicity. Abbreviations are the same as in [Fei07].

ability for the cell to grow efficiently. The reactions outside the backbone of the five EFMs with largest growth, accounting for 80% of the growth rate, are shown in Figure 5.3, from which different pathways to synthesize those necessary cytosolic metabolites entering the backbone can be seen. All EFMs are in general different arrangements of these pathways to strike delicate balance to produce biomass.

We further look into the contribution of each carbon source. Recall that $\mathbf{C} = \{C_1, \ldots, C_{25}\}$ is the index set of the uptake reactions of the 25 extracellular carbon sources. Given the flux distribution $f_1$, the apparent relative contribution by each carbon source can be simply defined as:

$$RC_k^{\mathbf{f}_1} = v_{C_k} \Big/ \sum_{k=1}^{25} v_{C_k} \ .$$

Under the resolution of the decomposition of $f_1$ into the 25 EFMs, the 'marginal relative contribution' of each carbon source is calculated by summing up the relative contribution by that source in each EFM:

$$RC_k^{\mathbf{EFM}} = \sum_{i=1}^{25} \left( \frac{efm_{i,C_k}}{\sum_{k=1}^{25} efm_{i,C_k}} \times \frac{\alpha_i efm_{i,growth}}{v_{growth}} \right)$$

where $v_{growth} = \sum_{i=1}^{25} \alpha_i efm_{i,growth}$ . The values of the relative contributions are shown in Table 5.5. The carbon source with largest relative contribution to biomass is N-Acetyl-glucosamine, over 5.4% and L-histidine has the lowest contribution equal to 0.9% only. The information cannot be obtained from $f_0$ or $f_1$ alone in which nearly all uptake fluxes of carbon sources are equal. The value of marginal relative contribution, however, seems to change with the non-unique decomposition by EFMs. Further analysis will be needed to formally address this issue.

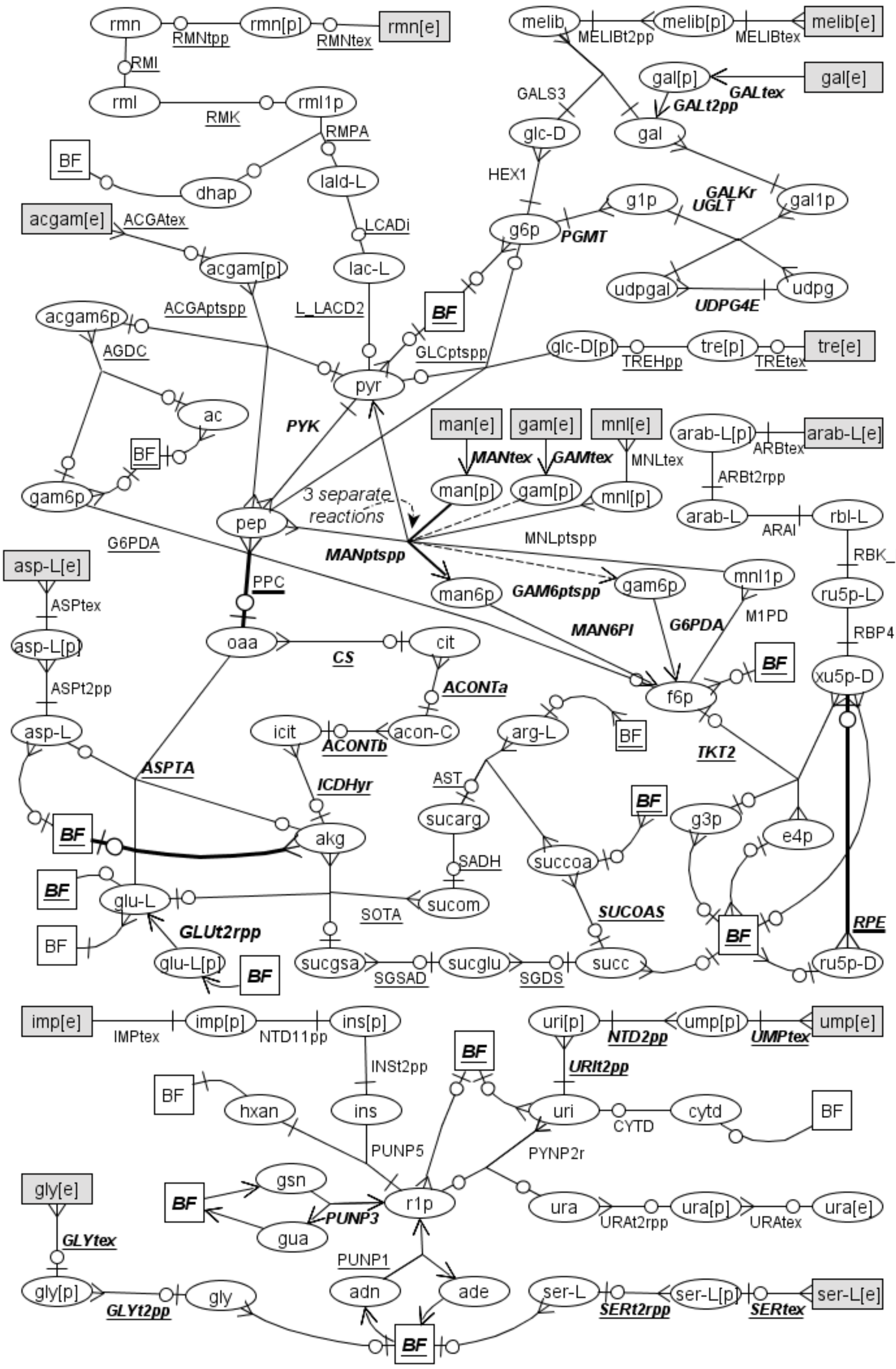


**Figure 5.3 Reactions outside the backbone of the top five EFMs contributing to growth**

Extracellular sources are in squares. ‘BF’ stands for ‘backbone fluxes’. Legend for EFMs (source to sink is from left to right): 1st: —+; 2nd: —o; 3rd: ≻—; 4th: reaction name ***bold and italic***; 5th: reaction name <u>underscored</u>. Independent intersecting edges are bolded for clarity. Small molecules, cofactors and their reactions are not shown for simplicity. Abbreviations are the same as in [Fei07].

It is remarked that the above analysis focuses on $f_0$ and $f_1$. In-depth studies are required to prove the claims from it. For example, alternative optimal flux distributions and different maximum uptake rates chosen may alter the results on the dependence of carbon sources and the necessity of certain metabolites. Still, the resolution brought by the present decomposition has revealed some information not observed from the apparent flux distribution.

## 5.4 Conclusion

In this chapter, we have applied the proposed algorithm to study the growth of *E. coli* under two vastly different conditions, the glucose minimal medium and the LB medium. For the growth on the glucose minimal medium, the sensitivity analysis has succeeded to reveal various modes of cooperation of EFMs that are very sensitive to changes in uptake conditions in the *E. coli* MG1655 iAF1260 network. Also, from the analysis, a rationale for the tremendous redundancy of metabolic pathways has been suggested. For the case of growth of *E. coli* on the LB medium, it gives an exemplary flux distribution with a complex structure in which a simultaneous mode of assimilation of carbon sources is seen. By our algorithm, essential reactions and metabolites, contribution of carbon sources to the flux distribution have been studied under the resolution of the decomposition by EFMs. We conclude that the algorithm can facilitate metabolic pathway analysis in genome-scale metabolic networks. It provides an analytic method that prepares for the future breakthrough in experimental techniques to measure *in vivo* fluxes in a huge scale.

| Carbon Source | $RC_k^{\mathbf{f}_1}$ (%) | $RC_k^{\mathbf{EFM}}$ (%) |
|---|---|---|
| acgam | 4.2560 | 5.3863 |
| amp | 4.2560 | 5.1745 |
| arab_L | 4.2560 | 4.5966 |
| arg_L | 4.2560 | 4.9183 |
| asp_L | 4.2560 | 4.8024 |
| cmp | 4.2560 | 5.2073 |
| fuc_L | 4.2560 | 0 |
| gal | 4.2560 | 4.4227 |
| gam | 4.2560 | 4.3578 |
| gln_L | 4.2560 | 4.7229 |
| gly | 4.2560 | 4.9817 |
| glyc | 4.2560 | 3.7351 |
| gmp | 4.2560 | 4.6646 |
| his_L | 0.8044 | 0.9148 |
| imp | 4.2560 | 4.6727 |
| lac_D | 4.2560 | 0 |
| lac_L | 4.2560 | 1.4088 |
| man | 4.2560 | 4.4594 |
| melib | 4.2560 | 4.8377 |
| met_L | 1.3069 | 1.4863 |
| mnl | 4.2560 | 4.9286 |
| rmn | 4.2560 | 5.3487 |
| ser_L | 4.2560 | 4.8582 |
| tre | 4.2560 | 5.3817 |
| ump | 4.2560 | 4.7329 |

**Table 5.5 Apparent and marginal relative contributions of each carbon source**

# Chapter 6 Mouse Cardiomyocyte

## 6.1 Introduction

In this chapter, the algorithm is applied to analyze a realistic experimental flux distribution of the mouse cardiac muscle, or called mouse cardiomyocyte, published in an early study [VoP06]. The approximation of the EFM of largest contribution to a particular reaction relevant to the function of the mouse cardiomyocyte is particular focused.

## 6.2 Mouse Cardiomyocyte Metabolic Network

The mouse cardiomyocyte metabolic network was reconstructed to interpret mass isotopomer data and a flux distribution in a mouse heart perfused with labeled substrates was experimentally estimated [VoP06]. There are 257 reactions and 240 metabolites. It is compartmentalized into three parts: cytoplasm, mitochondrion and extracellular. The stoichiometric matrix, experimental flux distribution and other information of the network is given as supplementary information in [VoP06].

Obviously, the primary role of cardiomyocyte is to generate energy for heart contraction to maintain life. In the reconstructed network, ATP hydrolysis, which is the primary biochemical reaction in organisms to release energy from the high-energy phosphoanhydridic bonds, is represented by the reaction DMatp. It is therefore reasonable to assume that

DMatp is the cellular objective. We then apply our algorithm to approximate the largest ATP-producing EFM in the given experimental flux distribution.

## 6.3 Experimental Flux Distribution

The experimental flux distribution in [VoP06] follows the steady state assumption quite well though some unsteadiness exists. For computational convenience, we adjusted the flux distribution slightly to satisfy the steady state assumption by minimizing the sum of squared differences. This yields the following quadratic programming problem:

$$\min \left(\mathbf{v}-\mathbf{v}_{\exp}\right)^T \left(\mathbf{v}-\mathbf{v}_{\exp}\right)$$

$$\begin{aligned} \text{subject to} \quad & \mathbf{Sv} = 0 \\ & \mathbf{v} \le M \operatorname{sgn}\left(\mathbf{v}_{\exp}\right) \\ & \mathbf{v} \ge 0 \end{aligned}$$

where $\mathbf{v}_{\exp}$ is the experimental flux distribution; M is a large number and $\operatorname{sgn}\left(\mathbf{v}_{\exp}\right)$ is the sign of $\mathbf{v}_{\exp}$. The second constraint aims to keep the zeros in the flux distribution zero.

## 6.4 Results of the Decomposition by the Algorithm

In our solution, we find three EFMs account for the ATP hydrolysis DMatp. The EFM with the largest DMatp flux in the solution, also the first EFM detected by our algorithm, is the combination of fatty acid oxidation, the tricarboxylic acid cycle (TCA cycle) and the oxidative phosphorylation in the mitochondria. The EFM with 42 reactions consumes the long-chain fatty acid oleate and contributes 85% to the DMatp flux. This is actually expectable because long-chain fatty acid has been found to be the major source in cardiac muscle and the inability of its catabolism can cause ischemia [van92]. Figure 6.1 describes the EFM. Uptake oleate is activated by coenzyme A and becomes oleoyl-CoA (FACOAL181i). After that, oleoyl-CoA enters the matrix through the carnitine shuttle (C181CPT1, C181CRNt, C181CPT2) and then yields an acetyl-CoA, which initiates the TCA cycle, as well as the intermediate coenzyme A ester, palmitate, which is in turn oxidated into 8 molecules of acetyl-CoA (FAOXC181, FAOXC160). The oxidation of palmitate at the same time generates 7 molecules of FADH2 which are also generated in the TCA cycle by succinate-Q reductase (SUCD1m) and serve as electron donors in the oxidative phosphorylation. This explains the standard cooperation between the three subsystems and the ability of the EFM to synthesize ATP at a high rate.

The second EFM contributing to DMatp is the standard anaerobic respiration. The EFM consisting of 17 reactions contributes 11% to the ATP demand. Glucose is catabolized into pyruvate during glycolysis which simply turns into lactate at last. Each mole of glucose leads to two moles of ATP in this EFM.

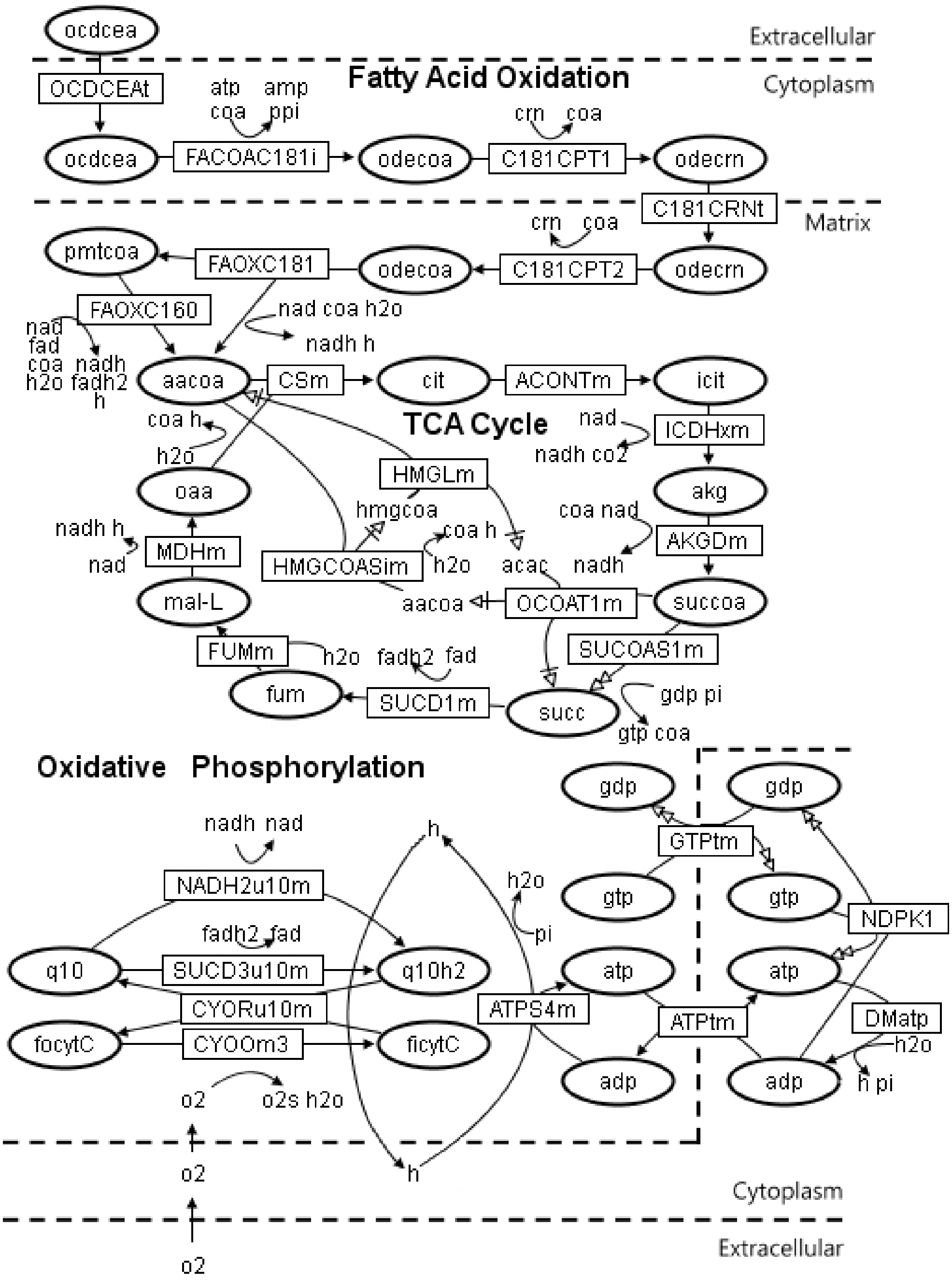


**Figure 6.1 1st and 3rd EFMs found in the flux distribution of mouse cardiomyocyte.**

Normal arrows indicate reactions shared by both EFMs. Arrows with double hollow triangles represent reactions used by the first EFM only and those with a hollow triangle and a hyphen are used by the third EFM only. Circled nodes are main metabolites and rectangles stand for reactions. All abbreviations are the same as in the original network [VoP06]. Some metabolites and their reactions are not shown, like $CO_2$, $H_2O$ and phosphate.

The remaining 3% of the DMatp flux is provided by the third EFM very similar to the first one (Figure 6.1). It also consumes oleate and involves the TCA cycle and oxidative phosphorylation. The major difference is the replacement of the reaction SUCOAS1m catalyzed by succinate thiokinase which produces succinate and guanosine triphosphate (GTP). The reaction OCOAT1m by β-ketoacyl-CoA transferase is used instead. Although both reactions transform succinyl-CoA into succinate, the former generates GTP which can directly convert into ATP while the latter consumes another important energy source derived from fatty acid, the ketone body acetoacetate, and produces acetoacetyl-CoA. This reaction is actually a part of the inter-conversion between acetoacetate and acetoacetyl-CoA (OCOAT1m, HMGCOASim, HMGLm). This was discovered as pseudoketogenesis in literature [Fin88, VoP06]. It maintains the TCA cycle and indicates the limit of the activity of the standard TCA cycle due to the capacity of succinate thiokinase. This agrees with the fact that the bottlenecks in the first EFM are GTPm and SUCOAS1m.

These three EFMs suffice to produce all ATP used for hydrolysis. The sum of fluxes by the three EFMs contributes to 54% of the total fluxes. This suggests that the mouse cardiac muscle performs metabolic activities other than ATP hydrolysis for contractile function and ion pumps. From the remaining flux distribution, productions of two ketone bodies, acetoacetate and D-3-hydroxybutyrate, two intermediate metabolites of the TCA cycle, citrate and succinate, as well as pyruvate and lactate are seen. To contextualize the remaining flux distribution, instead of sticking to the objective of maximizing DMatp, we maximized the production of each substrate accordingly. When a set of EFMs producing one substrate is found, we move to the next substrate. In this way, EFMs found can be categorized by their production (Table 6.1).

We first optimize for the production of the two ketone bodies. Five EFMs of ketogenesis are found (EFM4–8) and the majority of extracellular succinate and all pyruvate are also produced by these EFMs. Productions of the rest of the succinate (EFM9), L-lactate (EFM10–11) and finally H+ (EFM12–13) are optimized in turn and together five other EFMs are found. The remaining one (EFM14) is an isolated internal cycle describing the exchange of L-citrulline and ornithine between mitochondria and cytoplasm. One characteristic of all EFMs found with zero DMatp fluxes, except the cycle, i.e. EFM4–13, is that all productions involve both glycolysis from glucose and the fatty acid oxidation of oleate.

An interesting point is that we discover that one of these EFMs performs oxidative phosphorylation without synthesizing ATP. This implies the possibility of the existence of the uncoupling of oxidative phosphorylation and ATP synthesis. Actually, it is consistent with the finding that long-chain fatty acids lead to this uncoupling for other purposes like heat generation [Sku91]. Such mode of operation cannot be easily revealed from the whole flux distribution in which the P/O ratio is as high as 4.2.

| Extracellular metabolites/ Dmatp | EFM1 | EFM2 | EFM3 | EFM4 | EFM5 | EFM6 | EFM7 |
|---|---|---|---|---|---|---|---|
| bhb | 0 | 0 | 0 | 0.381 | 0.076 | 0 | 0 |
| acac | 0 | 0 | 0 | 0 | 0 | 0.152 | 0.103 |
| cit | 0 | 0 | 0 | 0.017 | 0 | 0 | 0 |
| pyr | 0 | 0 | 0 | 0 | 0 | 0.621 | 0 |
| succ | 0 | 0 | 0 | 0 | 0.0031 | 0 | 0 |
| lac-L | 0 | 1.8894 | 0 | 0 | 0 | 0 | 0 |
| h | 0.0637 | 1.8894 | 0.0027 | 0.3813 | 0.0757 | 0.7738 | 0.1027 |
| co2 | 2.162 | 0 | 0.092 | 0.062 | 0.047 | 0 | 0.104 |
| h2o | 1.95 | 0 | 0.083 | 0 | 0.034 | 0.723 | 0.189 |
| o2 | -3.02 | 0 | -0.13 | -0.49 | -0.1 | -0.56 | -0.19 |
| glc | 0 | -0.9447 | 0 | -0.0402 | -0.0233 | -0.3106 | -0.0518 |
| ocdcea | -0.1201 | 0 | -0.0051 | -0.0804 | -0.0123 | -0.0339 | -0.0113 |
| DMatp | 14.19 | 1.889 | 0.557 | 0 | 0 | 0 | 0 |

| Extracellular metabolites/ Dmatp | EFM8 | EFM9 | EFM10 | EFM11 | EFM12 | EFM13 | EFM14 | Flux distribution |
|---|---|---|---|---|---|---|---|---|
| bhb | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.457 |
| acac | 0.011 | 0 | 0 | 0 | 0 | 0 | 0 | 0.266 |
| cit | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.017 |
| pyr | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.621 |
| succ | 0.0005 | 0.0009 | 0 | 0 | 0 | 0 | 0 | 0.0044 |
| lac-L | 0 | 0 | 0.2109 | 0.0004 | 0 | 0 | 0 | 2.1007 |
| h | 0.0107 | 0.0021 | 0.2174 | 0.0005 | 0.0169 | 0.002 | 0 | 3.5389 |
| co2 | 0.004 | 0.015 | 0.22 | 0.004 | 0.456 | 0.054 | 0 | 3.218 |
| h2o | 0.012 | 0.014 | 0.199 | 0.003 | 0.419 | 0.05 | 0 | 3.676 |
| o2 | -0.02 | -0.02 | -0.31 | -0.01 | -0.59 | -0.07 | 0 | -5.49 |
| glc | -0.0018 | -0.0004 | -0.1054 | -0.0002 | -0.019 | -0.0022 | 0 | -1.4998 |
| ocdcea | -0.0021 | -0.0009 | -0.0122 | -0.0002 | -0.019 | -0.0022 | 0 | -0.2997 |
| DMatp | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16.63 |

**Table 6.1 DMatp fluxes and exchange rates of extracellular metabolites in the flux distribution and the decomposed EFMs.**

acac: acetoacetate; bhb: (R)-3-Hydroxybutanoate; cit: citrate; co2: carbon dioxide; glc: glucose; h: H+; h2o: H2O; lacL: L-lactate; o2: O2; ocdcea: octadecenoate (n-C18:1) (oleate); pyr: pyruvate; succ: succinate; DMatp: ATP demand. A positive flux represents excretion and a negative flux represents uptake.

## 6.5 Conclusion

The above analysis on the experimental flux distribution of mouse cardiomyocyte shows that although there are only a small number of carbon sources available, a realistic flux distribution can have a certain complexity, as reflected by the number of EFMs, compared to a simulated flux distribution optimized for a particular objective only. In this sense, the decomposition of a flux distribution into a set of EFMs is very useful to study the structure of the flux distribution, break down it into different components and reveal information that is not obvious from the apparent flux distribution.

# Chapter 7 Conclusion and Future Work

## 7.1 Conclusion of the Research

### 7.1.1 Algorithm to Decompose a Flux Distribution into EFMs

In this research, we developed a novel algorithm to decompose flux distributions into EFMs in genome-scale networks. It is an iterative scheme of a mixed integer linear program with the processed flux modes stored in a stack structure. Each time the mixed integer linear program returns a feasible flux mode bounded by the current flux mode and the process is repeated until no ‘smaller’ flux mode exists. Then an EFM is reached. Each EFM found is used to update the flux modes in the stack so that the computational cost is reduced and properties like denseness, uniqueness, linear and systemic independence are guaranteed. The algorithm is also able to approximate the EFM of largest contribution to an objective reaction in a flux distribution.

The algorithm brings a breakthrough of the bottleneck of metabolic pathway analysis in genome-scale metabolic networks caused by the combinatorial explosion of the number of EFMs as the network size grows. While traditional methods to find EFMs rely on the double description method to find the full set of EFMs, the proposed algorithm can find individual EFMs with specific properties directly. The benchmarks reported in section 4.2 have confirmed the correctness of our algorithm to find EFMs and its usefulness to approximate the EFM with the largest contribution to a particular objective reaction. Moreover, the computational experiment reported in section 4.3 has shown the capability of our algorithm compared to the best of all the existing methods. Our algorithm has succeeded to find EFMs contributing to flux distributions in a genome-scale metabolic network in which existing methods are unable to compute the complete set of EFMs.

From these studies, we conclude that the algorithm is valid and makes an advance in the computation of EFMs. Hence, the first research objective which is to devise an algorithm to decompose a flux distribution into a set of EFMs has been achieved.

### 7.1.2 Case Studies

With the algorithm, we have performed two case studies, the growth of *E. coli* and the metabolism of mouse cardiomyocyte. For the growth of *E. coli*, two vastly different conditions have been investigated, the glucose minimal medium and the LB medium. The sensitivity analysis on the glucose minimal medium has revealed various modes of cooperation of EFMs that are very sensitive to changes in uptake conditions and it has brought insight into the observation that the redundancy of metabolic pathways is always immense.

For the growth of *E. coli* on the LB medium, the structure of the simulated flux distribution is very complex. Still, by our algorithm, essential reactions and metabolites, contribution of carbon sources to the flux distribution have been located under the resolution of the decomposition by EFMs.

For the mouse cardiomyocyte, by assuming the ATP hydrolysis, which is responsible to release energy in many organisms, to be the cellular objective, we have looked into the EFMs that contribute to the ATP hydrolysis. Metabolic pathways consistent with literature and knowledge in biochemistry have been successfully located. Also, a possibly hidden phenomenon supported by literature, the uncoupling of oxidative phosphorylation and ATP synthesis, has been detected.

In view of the results of the case studies, we conclude that the second research objective which is to analyze flux distributions in genome-scale metabolic networks by the algorithm has also been achieved.

### 7.1.3 Contributions of the Research

There are three particular contributions of this research. First, regarding the first research objective to devise an algorithm, this provides an alternative way complementary to traditional methods to compute EFMs that decompose a flux distribution. The devised algorithm and the ideas behind have their own theoretical value, including formulating the elementarity of EFMs as the feasibility of the optimization model DC and finding EFMs by iteratively restricting the values of fluxes to zeros.

Second, the devised algorithm facilitates the analysis of flux distributions in genome-scale metabolic networks by decompositions into EFMs, which has been found useful in the literature when applying to metabolic networks of small sizes but meanwhile has currently no corresponding method available in genome-scale metabolic networks. It is a novel approach to answer the open question as to the functional relevance of

EFMs. In comparison with previous methods, the main advantage is that it does not require the complete determination of the full set of EFMs. This constitutes a progress over the state of the art.

Third, regarding the second research objective, analyzing genome-scale flux distributions by applying the devised algorithm can generate new insights and hypotheses in biology, fulfilling the purpose of the systems biology approach which aims to study biology at a system level. These insights and hypotheses can possibly be developed into new biological knowledge with further experimental verifications. This kind of information obtained at a system level is unique with respect to the traditional sheer experimental studies in biology. At the same time, the devised algorithm can also be validated by examining its applicability to real cases in biology. Hence, this demonstrates and concretizes the usefulness of the devised algorithm by trying to attain the goal of systems biology which is to study biology at a system level.

In conclusion, the algorithm proposed in this research can facilitate metabolic pathway analysis in genome-scale metabolic networks and this can shed light on cellular metabolism. It provides an analytic method that prepares for the future breakthrough in experimental techniques to measure *in vivo* fluxes in a huge scale.

## 7.2 Future Work

Metabolic pathway analysis in genome-scale metabolic networks is challenging. There are several research possibilities that can further facilitate metabolic pathway analysis by overcoming the computational barrier and make it a more useful, unified analytic technique.

First, the algorithm proposed in this research still has room for improvement, for example, the computation cost and memory demand. The iterative scheme of the current algorithm is simple and straightforward, analogous to a depth first search in the branch-and-bound method for optimization. It can be modified to improve its efficiency. For instance, the updating process can be performed earlier before an EFM is found. In this way, the data structure becomes a binary tree instead of a simple stack and different strategies used in branch-and-bound can then be applied. Also, multi-core computation may be implemented. Besides, other use of the objective function applicable to particular cases in biology may exist and this may add new value to the algorithm.

Second, efforts should be put on solving a more general and meaningful problem which is to decompose a flux distribution into a set of EFMs with respect to an optimization objective, like the α-spectrum [Wib03] and other problems listed in Table 2.1. Techniques in linear or even non-linear optimization will be required. The algorithm proposed in this research may serve as a reference.

Third, comparison between different methods or objectives to decompose a flux distribution into EFMs is desirable. Since it is known that the decomposition of a flux distribution into EFMs is generally not unique, different objectives of decomposition are expected to yield different results. This non-uniqueness of the representation by EFMs undermines its applicability and the reliability of the conclusion drawn from the analysis since alternative interpretation from another set of EFMs can exist. By investigating solutions from different objectives in different cases from a more biologically relevant perspective, the applicability of different objectives in different cases can be evaluated. If some well-defined invariants are located during the studies, this will be of even greater importance and may lead to a well-structured methodology to analyze flux distributions by EFMs. The usefulness and significance of metabolic pathway analysis will then be largely increased.

# References

[Alm04] Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabasi AL. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* **427**:839–43. 2004.

[Alm07] Almaas E, Holder A, Livingstone K. Introduction to systems biology for mathematical programmers. In: Lim GJ, Lee EK (eds.) *Optimization in Medicine and Biology*, Taylor & Francis Group, LLC. 2007.

[Bae06a] Baev MV, Baev D, Radek AJ, Campbell JW. Growth of *Escherichia coli* MG1655 on LB medium: monitoring utilization of sugars, alcohols, and organic acids with transcriptional microarrays. *Applied Microbiology and Biotechnology* **71**:310–316. 2006a.

[Bae06b] Baev MV, Baev D, Radek AJ, Campbell JW. Growth of *Escherichia coli* MG1655 on LB medium: monitoring utilization of amino acids, peptides, and nucleotides with transcriptional microarrays. *Applied Microbiology and Biotechnology* **71**:317–322. 2006b.

[Bea02] Beard DA, Liang SD, Qian H. Energy balance for analysis of complex metabolic networks. *Biophysical Journal* **83**:79–86. 2002.

[Bea07] Beasley JE, Planes FJ. Recovering metabolic pathways via optimization. *Bioinformatics* **23**:92–8. 2007.

[Bec07] Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nature Protocols* **2**:727–38. 2007.

[Bol02] Bolouri H, Davidson EH. Modeling transcriptional regulatory networks. *BioEssays* **24**:1118–29. 2002.

[Bon97] Bonarius HPJ, Schmid G, Tramper J. Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends* in *Biotechnology* **15**:308–14. 1997.

[Bra02] Brazhnik P, Fuente A, Mendes P. Gene networks: How to put the function in genomics. *Trends in Biotechnology* **20**:467–72. 2002.

[Bur03a] Burgard AP, Maranas CD. Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnology and Bioengineering* **82**:670–7. 2003a.

[Bur03b] Burgard AP, Pharkya P, Maranas CD. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering* **84**:647–57. 2003b.

[Bur04] Burgard AP, Nikolaev EV, Schilling CH, Maranas CD. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Research* **14**:301–312. 2004.

[Car07] Carlson RP. Metabolic systems cost-benefit analysis for interpreting network structure and regulation. *Bioinformatics* **23**:1258–64. 2007.

[Car09] Carlson RP. Decomposition of complex microbial behaviors into resource-based stress responses. *Bioinformatics* **25**:90–7. 2009.

[Che09] Chen L, Wang RS, Zhang XS. *Biomolecular Networks: Methods and Applications in Systems Biology*. John Wiley and Sons, Inc. 2009.

[Chu09] Chung BKS, Lee DY. Flux-sum analysis: a metabolite-centric approach for understanding the metabolic network. *BMC Systems Biology* **3**:117–26. 2009.

[Cov01] Covert MW, Schilling CH, Palsson BØ. Regulation of gene expression in flux balance models of metabolism. *Journal of Theoretical Biology* **213**:73–88. 2001.

[Cov02] Covert MW, Palsson BØ. Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *Journal of Biological Chemistry* **277**:28058–64. 2002.

[Cov03] Covert MW, Palsson BØ. Constraints-based models: Regulation of gene expression reduces the steady-state solution space. *Journal of Theoretical Biology* **221**:309-25. 2003.

[Cri70] Crick F. Central dogma of molecular biology. *Nature* **227**:561-3. 1970.

[Cro05] Croes D, F, Wodak SJ, van Helden Jacques. Metabolic PathFinding: inferring relevant pathways in biochemical networks. *Nucleic Acids Research* **33**:W326-30. 2005.

[Cro06] Croes D, Couche F, Wodak SJ, van Helden Jacques. Inferring meaningful pathways in weighted biochemical networks. *Journal of Molecular Biology* **356**:222–36. 2006.

[deF09a] de Figueiredo LF, Podhorski A, Rubio A, Kaleta C, Beasley JE, Schuster S, Planes FJ. Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics* **25**:3158–65. 2009a.

[deF09b] de Figueiredo LF, Schuster S, Kaleta C, Fell DA. Can sugars be produced from fatty acids? A test case for pathway analysis tools. *Bioinformatics* **25**:152–8. 2009b

[deJ02] de Jong H. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology* **9**:67. 2002.

[DeK06] De Keersmaecker SCJ, Thijs IMV, Vanderleyden J, Marchal K. Integration of omics data: how well does it work for bacteria? *Molecular Microbiology* **62**:1239-50. 2006.

[Dua07] Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America* **104**:1777–82. 2007.

[Dur09] Durot M, Bourguignon PY, Schachter V. Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiology Review* **33**:164–90. 2009.

[Edw99] Edwards JS, Palsson BØ. Systems properties of the Haemophilus influenzae Rd metabolic genotype. *Journal of Biological Chemistry* **274**:17410–6. 1999.

[Edw00] Edwards JS, Palsson BØ. The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences of the United States of America* **97**:5528–33. 2000.

[Fei07] Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology* **3**:121. 2007.

[Fie02] Fiehn O. Metabolomics – the link between genotypes and phenotypes. *Plant Molecular Biology* **48**:155–71. 2002.

[Fin88] Fink G, Desrochers S, Des Rosiers C, Garneau M, David F, Daloze T, Landau BR, Brunengraber H. Pseudoketogenesis in the perfused rat heart. *Journal of Biological Chemistry* **263**:18036–42. 1988.

[Gag04] Gagneur J, Klamt S. Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics* **5**:175. 2004.

[Gia08] Gianchandani EP, Oberhardt MA, Burgard AP, Maranas CD, Papin JA. Predicting biological system objectives *de novo* from internal state measurements. *BMC Bioinformatics* **9**:43. 2008.

[Goe02] Goesmann A, Haubrock M, Meyer F, Kalinowski J, Giegerich R. PathFinder: reconstruction and dynamic visualization of metabolic pathways. *Bioinformatics* **18**:124–9. 2002.

[Gri08] Griffiths AJF, Wessler SR, Lewontin RC, Carroll SB. *Introduction to Genetic Analysis (9th ed.)*. W. H. Freeman and Company, New York. 2008.

[Hei77] Heinrich R, Rapoport SM, Rapoport TA. Metabolic regulation and mathematical models. *Progress in Biophysics and Molecular Biology* **32**:1–82. 1977.

[Her06] Herrgard MJ, Fong SS, Palsson BØ . Identification of genome-scale metabolic network models using experimentally measured flux profiles. *PLoS Computational Biology* **2**:e72. 2006.

[Hop07] Hoppe A, Hoffmann S, Holzhutter HG. Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. *BMC Systems Biology* **1**:23. 2007.

[Hor72] Horn F, Jackson R. General mass action kinetics. *Archive for Rational Mechanics and Analysis* **47**:81–116. 1972.

[Jia07] Jiang D, Zhou S, Guan J. A novel method for flux distribution computation in metabolic networks. In: Hochreiter S, Wagener R (eds.) *Bioinformatics Research and Development*. Springer. 2007.

[Kan00] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**:27–30. 2000.

[Kar94] Karin M. Signal transduction from the cell surface to the nucleus through the phosphorylation of transcription factors. *Current Opinion in Cell Biology*. **6**:415–24. 1994.

[Kau03] Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. *Current Opinion in Biotechnology* **14**:491–6. 2003.

[Kel04] Kell DB. Metabolomics and systems biology: making sense of the soup. *Current Opinion in Microbiology* **7**:296–307. 2004.

[Kim07] Kim H, Lee JK, Park T. Boolean networks using the chi-square test for inferring large-scale gene regulatory networks. *BMC Bioinformatics* **8**:37. 2007.

[Kla02] Klamt S, Stelling J. Combinatorial complexity of pathway analysis in metabolic networks. *Molecular Biology Reports* **29:**233–6. 2002.

[Kla03] Klamt S, Stelling J, Ginkel M, Gilles ED. FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps. *Bioinformatics* **19**:261–9. 2003.

[Kla04] Klamt S, Gilles ED. Minimal cut sets in biochemical reaction networks. *Bioinformatics* **20**:226–34. 2004.

[Kla05] Klamt S, Gagneur J, von Kamp A. Algorithmic approaches for computing elementary modes in large biochemical reaction networks. *IEE Proceedings - Systems Biology* **152**:249–55. 2005.

[Kla06] Klamt S, Saez-Rodriguez J, Lindquist JA, Simeoni L, Gilles ED. A methodology for the structural and functional analysis of signaling and regulatory networks. *BMC Bioinformatics* 7:56. 2006.

[Kla07] Klamt S, Saez-Rodriguez J, Gilles ED. Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Systems Biology* **1**:2. 2007.

[Kli09] Klipp E, Liebermeister W, Wierling C, Kowald A, Lehrach H, Herwig R. *Systems Biology*. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim. 2009.

[Kuf00] Kuffner R, Zimmer R, Lengauer T. Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics*, **16**:825–36. 2000.

[Kur07] Kurata H, Zhao Q, Okuda R, Shimizu K. Integration of enzyme activities into metabolic flux distributions by elementary mode analysis. *BMC Systems Biology* **1**:31. 2007.

[Lee00] Lee S, Phalakornkule C, Domach MM, Grossmann IE. Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. *Computers and Chemical Engineering* **24**:711–6. 2000.

[Lee06] Lee JM, Gianchandani EP, Papin JA. Flux balance analysis in the era of metabolomics. *Briefing in Bioinformatics* **7**:140-50. 2006.

[Lla10] Llaneras F, Picó J. Which metabolic pathways generate and characterize the flux space? A comparison among elementary modes, extreme pathways and minimal generators. *Journal of Biomedicine and Biotechnology* **2010**:753904. 2010.

[Mah02] Mahadevan R, Edwards JS, Doyle FJ. Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophysical Journal* **83**:1331-40. 2002.

[Mah03] Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering* **5**:264–76. 2003.

[McS03] McShan DC, Rao S, Shah I. PathMiner: predicting metabolic pathways by heuristic search. *Bioinformatics* **19**:1692–8. 2003.

[NIA08] National Institute of Alcohol Abuse and Alcoholism. “Gene structure and gene expression in higher organisms”, National Institute on Alcohol Abuse and Alcoholism [Internet]. [Published 2008 Dec, cited 2010 June 10]. Available from: http://www.niaaa.nih.gov/.

[Nel05] Nelson DL, Cox MM. *Lehninger Principles of Biochemistry*, 4th ed. Worth Publishers, NY. 2005.

[Noo07] Nookaew I, Meechai A, Thammarongtham C, Laoteng K, Ruanglek V, Cheevadhanarak S, Nielsen J, Bhumiratana S. Identification of flux regulation coefficients from elementary flux modes: a systems biology tool for analysis of metabolic networks. *Biotechnology and Bioengineering* **97**:1535–49. 2007.

[Ort10] Orth JD, Fleming RMT, Palsson BØ. 18 Feb 2010, posting date. Chapter 10.2.1, Reconstruction and use of microbial metabolic networks: the core *Escherichia coli* metabolic model as an educational guide. In Böck A, Curtiss III R,.Kaper JB, Karp PD, Neidhardt FC, Nyström T, Slauch JM, Squires CL, Ussery D (ed.), *EcoSal—Escherichia coli and Salmonella: Cellular and Molecular Biology*. http://www.ecosal.org. ASM Press, Washington, DC.

[Pap04] Papin JA, Stelling J, Price ND, Klamt S, Schuster S, Palsson BØ. Comparison of network-based pathway analysis methods. *Trends in Biotechnology* **22**:400–5. 2004.

[Pfe99] Pfeiffer T, Sánchez-Valdenebro I, Nuño JC, Montero F, Schuster S. METATOOL: For Studying Metabolic Networks. *Bioinformatics* **15**:251–7. 1999.

[Pla08] Planes FJ, Beasley JE. A critical examination of stoichiometric and path-finding approaches to metabolic pathways. *Briefing in Bioinformatics* **9**:422–36. 2008.

[Pla09a] Planes FJ, Beasley JE. Path finding approaches and metabolic pathways. *Discrete Applied Mathematics* **157**:2244-56. 2009a.

[Pla09b] Planes FJ, Beasley JE. An optimization model for metabolic pathways. *Bioinformatics* **25**:2723–9. 2009b.

[Pri04] Price ND, Reed JL, Palsson BØ. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology* **2**:886-97. 2004.

[Raa01] Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff HV, van Dam K, Oliver SG.. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology* **19**:45–50. 2001.

[Rag03] Raghunathan AU, Perez-Correa JR, Biegler LT. Data reconciliation and parameter estimation in flux-balance analysis. *Biotechnology and Bioengineering* **84**:700–9. 2003.

[Ram09] Raman K, Chandra N. Flux balance analysis of biological systems: applications and challenges. *Briefing in Bioinformatics* **10**:435-49. 2009.

[Red88] Reder C. Metabolic control theory: a structural approach. *Journal of Theoretical Biology* **135**:175-201. 1988.

[Ree03] Reed JL, Vo TD, Schilling CH, Palsson BØ. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biology* 2003, **4**:R54. 2003.

[Ree04] Reed JL, Palsson, BØ. Genome-scale *in silico* models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome Research*. **14**:1797–805. 2004.

[Rez11] Rezola A, de Figueiredo LF, Brock M, Pey J, Podhorski A, Wittmann C, Schuster S, Bockmayr A, Planes FJ. Exploring metabolic pathways in genome-scale networks via generating flux modes. *Bioinformatics* **27**:534–40. 2011.

[Ros68] Rosen R. Recent developments in the theory of control and regulation of cellular processes. In: Bourne GH (ed.) *International Review of Cytology*, New York: Academic Press. 1968.

[Sau04] Sauer U. High-throughput phenomics: experimental methods for mapping fluxomes. *Current Opinion in Biotechnology* **15**:58–63. 2004.

[Sau06] Sauer U. Metabolic networks in motion: 13C-based flux analysis. *Molecular Systems Biology* **2**:62. 2006.

[Sch94] Schuster S, Hilgetag C. On the elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems* **2**:165-82. 1994.

[Sch96] Schuster S, Hilgetag C, Woods JH, Fell DA. Elementary modes of functioning in biochemical networks. In Cuthbertson R, Holcombe M, Paton R. (eds.) *Computation in Cellular and Molecular Biological Systems*, World Scientific, Singapore, pp.151–65. 1996.

[Sch99] Schuster S, Dandekar T, Fell DA. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends in Biotechnology* **17**:53–60. 1999.

[Sch00] Schilling CH, Letscher D, Palsson BØ . Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology* **203**:229-48. 2000.

[Sch02] Schuster S, Hilgetag C, Woods JH, Fell DA. Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism. *Journal of Mathematical Biology* **45**:153–81. 2002.

[Sch07] Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Molecular Systems Biology* **3**:119. 2007.

[Schw05] Schwartz JM, Kanehisa M. A quadratic programming approach for decomposing steady-state metabolic flux distributions onto elementary modes. *Bioinformatics* **21**(Suppl2):ii204–ii205. 2005.

[Schw06] Schwartz JM, Kanehisa M. Quantitative elementary mode analysis of metabolic pathways: the example of yeast glycolysis. *BMC Bioinformatics* **7**:186. 2006.

[Seg02] Segre D, Vitkup D, Church GM. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America* **99**:15112–7. 2002.

[Shl05] Shlomi T, Berkman O, Ruppin E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proceedings of the National Academy of Sciences of the United States of America* **102**:7695–700. 2005.

[Sku91] Skulachev VP. Fatty acid circuit as a physiological mechanism of uncoupling of oxidative phosphorylation. *FEBS Letters* **294**:158–62. 1991.

[Sma09] Smallbone K, Simeonidis E. Flux balance analysis: A geometric perspective. *Journal of Theoretical Biology* **258**:311-5. 2009.

[Son09] Song HS, Ramkrishna D. Reduction of a set of elementary modes using yield analysis. *Biotechnology and Bioengineering* **102**:554–68. 2009.

[Ste02] Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED. Metabolic network structure determines key aspects of functionality and regulation. *Nature* **420**:190-3. 2002.

[Ste07] Steuer R, Nesi AN, Fernie AR, Gross T, Blasius B, Selbig J. From structure to dynamics of metabolic pathways: application to the plant mitochondrial TCA cycle. *Bioinformatics* **23**: 1378–85. 2007.

[Ter06] Terzer M, Stelling J. Accelerating the computation of elementary modes using pattern trees. In Bucher,P. and Moret,B.M.E. (eds.) WABI, Vol. 4175 of *Lecture Notes in Computer Science*. Springer, Heidelberg, pp. 333–43. 2006.

[Ter08] Terzer M, Stelling J. Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics* **24**:2229–35. 2008.

[Ter09] Terzer M, Maynard ND, Covert MW, Stelling J. Genome-scale metabolic network. *Wiley Interdisciplinary Reviews Systems Biology and Medicine* **1**:285–97. 2009.

[Tri08] Trinh CT, Unrean P, Srienc F. Minimal Escherichia coli cell for the most efficient production of ethanol from hexoses and pentoses. *Applied and Environmental Microbiology* **74**:3634–43. 2008.

[Tri09a] Trinh CT, Srienc F. *Metabolic Engineering* of Escherichia coli for efficient conversion of glycerol to ethanol. *Applied and Environmental Microbiology* **75**:6696–705. 2009a.

[Tri09b] Trinh CT, Wlaschin A, Srienc F. Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. *Applied Microbiology and Biotechnology* **81**:813–26. 2009b.

[Twy03] Twyman R. Gene Expression. Wellcome Trust Human Genome [Internet]. London: Welcome Trust; [published 2003 Aug 1; cited 2010 June 9]. Available from: http://genome.wellcome.ac.uk/.

[Urb05] Urbanczik R, Wagner C. An improved algorithm for stoichiometric network analysis: theory and applications. *Bioinformatics* **21**:1203–10. 2005.

[van92] van der Vusse GJ, Glatz JF, Stam HC, Reneman RS. Fatty acid homeostasis in the normoxic and ischemic heart. *Physiological Reviews* **72**:881–940. 1992.

[Van06] Van Dien S, Schilling CH. Bringing metabolomics data into the forefront of systems biology. *News and Views, Molecular Systems Biology* (2006) doi:10.1038/msb4100078. 2006.

[Var94] Varma A, Palsson BØ. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Applied and Environmental Microbiology* **60**:3724–31. 1994.

[Vid04] Vidal M, Furlong EEM. From Omics to Systems Biology. *Nature Review Genetics* **5** (poster). 2004.

[von06] von Kamp A, Schuster S. Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics* **22**:1930–1. 2006.

[VoP06] Vo TD, Palsson BØ. Isotopomer analysis of myocardial substrate metabolism: a systems biology approach. *Biotechnology and Bioengineering* **95**:972–83. 2006.

[Wag04] Wagner C. Nullspace approach to determine the elementary modes of chemical reaction systems. *Journal of Physical Chemistry B* **108**:2425-31. 2004.

[Wan07] Wang Q. Yang Y, Ma H, Zhao X. Metabolic network properties help assign weights to elementary modes to understand physiological flux distributions. *Bioinformatics* **23**:1049–52. 2007.

[Wib03] Wiback S.J, Mahadevan R, Palsson BØ. Reconstructing metabolic flux vectors from extreme pathways: defining the α-spectrum. *Journal of Theoretical Biology* **224**:313–24. 2003.

[Yeu07] Yeung M, Thiele I, Palsson BØ. Estimation of the number of extreme pathways for metabolic networks. *BMC Bioinformatics* **8**:363. 2007.

[Yoo07] Yoon J, Si Y, Nolan R, Lee K. Modular decomposition of metabolic reaction networks based on flux analysis and pathway projection. *Bioinformatics* **23**: 2433–40. 2007.

[Zam10] Zamboni N. 13C metabolic flux analysis in complex systems. *Current Opinion in Biotechnology* **22**:103–8. 2010.

[Zha09a] Zhao Q, Kurata H. Genetic modification of flux for flux prediction of mutants. *Bioinformatics* **25**:1702–8. 2009a.

[Zha09b] Zhao Q, Kurata H. Maximum entropy decomposition of flux distribution at steady state to elementary modes. *Journal of Bioscience and Bioengineering* **107**:84–9. 2009b.

# Appendices

## A.1 Example of the Algorithm

The network used to illustrate the algorithm is the sample network in [Cov01], which has also been used as benchmark in chapter 4.

### A.1.1 Information of the Sample Metabolic Network

| Reaction | Lower Bound | Upper Bound | Equation |
|---|---|---|---|
| R1 | 0 | 200 | A + ATP -> B |
| R2a | 0 | 200 | B -> C + 2 ATP + 2 NADH |
| R2b | 0 | 200 | C + 2 ATP + 2 NADH -> B |
| R3 | 0 | 200 | B -> F |
| R4 | 0 | 200 | C -> G |
| R5a | 0 | 200 | G -> 0.8 C + 2 NADH |
| R5b | 0 | 200 | 0.8 C + 2 NADH -> G |
| R6 | 0 | 200 | C -> 3 D + 2 ATP |
| R7 | 0 | 200 | C + 4 NADH -> 3 E |
| R8a | 0 | 200 | G + ATP + 2 NADH -> H |
| R8b | 0 | 200 | H -> G + ATP + 2 NADH |
| Rres | 0 | 200 | NADH + O2 -> ATP |
| Tc1 | 0 | 10.5 | Carbon1 -> A |
| Tc2 | 0 | 10.5 | Carbon2 -> A |
| Tf | 0 | 5 | Fext -> F |
| Td | 0 | 12 | D -> Dext |
| Te | 0 | 12 | E -> Eext |
| Th | 0 | 5 | Hext -> H |
| To2 | 0 | 15 | Oxygen -> O2 |
| Growth | 0 | 200 | C + F + H + 10 ATP -> Biomass |

**Table A1 Information of the Sample Metabolic Network.**
Reactions, lower and upper bounds and the equations adopted from [Cov01].

For the visual representation of the sample metabolic network, interested readers are referred to [Cov01].

### A.1.2 Example

The following example demonstrates the first few steps to decompose a flux distribution and all operations in the algorithm are involved.

| | Stack |
|---|---|
| Reaction | fm1 |
| R1 | 6.2636 |
| R2a | 5.0242 |
| R2b | 0 |
| R3 | 1.2394 |
| R4 | 3.1421 |
| R5a | 5.7775 |
| R5b | 0 |
| R6 | 0.6455 |
| R7 | 4 |
| R8a | 0 |
| R8b | 2.6354 |
| Rres | 10.874 |
| Tc1 | 3.3753 |
| Tc2 | 2.8882 |
| Tf | 0.6191 |
| Td | 1.9365 |
| Te | 12 |
| Th | 4.4939 |
| To2 | 10.874 |
| Growth | 1.8585 |

→

| Stack | | |
|---|---|---|
| fm1 | fm2 | fm3 |
| 6.2636 | 3.3753 | 1.1066 |
| 5.0242 | 3.3753 | 1.1066 |
| 0 | 0 | 0 |
| 1.2394 | 0 | 0 |
| 3.1421 | 1.0151 | 0 |
| 5.7775 | 1.0151 | 0 |
| 0 | 0 | 0 |
| 0.6455 | 0.5964 | 0.6455 |
| 4 | 2.0683 | 0 |
| 0 | 0 | 0 |
| 2.6354 | 0 | 0 |
| 10.874 | 0.5076 | 2.2131 |
| 3.3753 | 3.3753 | 1.1066 |
| 2.8882 | 0 | 0 |
| 0.6191 | 0.5076 | 0.4611 |
| 1.9365 | 1.7892 | 1.9365 |
| 12 | 6.2049 | 0 |
| 4.4939 | 0.5076 | 0.4611 |
| 10.874 | 0.5076 | 2.2131 |
| 1.8585 | 0.5076 | 0.4611 |

Initially, the stack contains only the original flux distribution (left). Decompose until an EFM (fm3) is found (Step 1 & 2, right). The shaded boxes are reactions that are shut down.

| | Stack | | EFM |
|---|---|---|---|
| Reaction | fm1 | fm2 | efm1 |
| R1 | 5.157 | 3.1215 | 1.1066 |
| R2a | 3.9176 | 3.1215 | 1.1066 |
| R2b | 0 | 0 | 0 |
| R3 | 1.2394 | 0 | 0 |
| R4 | 3.1421 | 1.0151 | 0 |
| R5a | 5.7775 | 1.0151 | 0 |
| R5b | 0 | 0 | 0 |
| R6 | 0 | 0.4483 | 0.6455 |
| R7 | 4 | 2.0683 | 0 |
| R8a | 0 | 0 | 0 |
| R8b | 2.6354 | 0 | 0 |
| Rres | 8.6609 | 0 | 2.2131 |
| Tc1 | 2.2687 | 3.1215 | 1.1066 |
| Tc2 | 2.8882 | 0 | 0 |
| Tf | 0.158 | 0.4018 | 0.4611 |
| Td | 0 | 1.345 | 1.9365 |
| Te | 12 | 6.2049 | 0 |
| Th | 4.0328 | 0.4018 | 0.4611 |
| To2 | 8.6609 | 0 | 2.2131 |
| Growth | 1.3974 | 0.4018 | 0.4611 |

→

| Stack | EFM |
|---|---|
| fm1 | efm1 |
| 5.157 | 1.1066 |
| 3.9176 | 1.1066 |
| 0 | 0 |
| 1.2394 | 0 |
| 3.1421 | 0 |
| 5.7775 | 0 |
| 0 | 0 |
| 0 | 0.6455 |
| 4 | 0 |
| 0 | 0 |
| 2.6354 | 0 |
| 8.6609 | 2.2131 |
| 2.2687 | 1.1066 |
| 2.8882 | 0 |
| 0.158 | 0.4611 |
| 0 | 1.9365 |
| 12 | 0 |
| 4.0328 | 0.4611 |
| 8.6609 | 2.2131 |
| 1.3974 | 0.4611 |

fm3 leaves the stack (Step 3, left of the table). Update all flux modes (Step 4, 'Stack' column in the left of the table). It is checked that fm2 cannot contribute to fm1 since it does not contain all the zeros of fm1. Remove it from the stack (Step 5, right of the table). The shaded boxes are reactions that fm1 has zero fluxes but fm2 does not.

# A.2 Core *E. coli* K-12 MG1655 iAF1260 Model

## A.2.1 Network Information

| Abbreviation | Official Name | Equation |
|---|---|---|
| ACALD | acetaldehyde dehydrogenase (acetylating) | [c] : acald + coa + nad <==> accoa + h + nadh |
| ACALDt | acetaldehyde reversible transport | acald[e] <==> acald[c] |
| ACKr | acetate kinase | [c] : ac + atp <==> actp + adp |
| ACONTa | aconitase (half-reaction A, Citrate hydro-lyase) | [c] : cit <==> acon-C + h2o |
| ACONTb | aconitase (half-reaction B, Isocitrate hydro-lyase) | [c] : acon-C + h2o <==> icit |
| ACt2r | acetate reversible transport via proton symport | ac[e] + h[e] <==> ac[c] + h[c] |
| ADK1 | adenylate kinase | [c] : amp + atp <==> (2) adp |
| AKGDH | 2-Oxogluterate dehydrogenase | [c] : akg + coa + nad --> co2 + nadh + succoa |
| AKGt2r | 2-oxoglutarate reversible transport via symport | akg[e] + h[e] <==> akg[c] + h[c] |
| ALCD2x | alcohol dehydrogenase (ethanol) | [c] : etoh + nad <==> acald + h + nadh |
| ATPM | ATP maintenance requirement | [c] : atp + h2o --> adp + h + pi |
| ATPS4r | ATP synthase (four protons for one ATP) | adp[c] + (4) h[e] + pi[c] <==> atp[c] + (3) h[c] + h2o[c] |
| BIOMASS | Biomass Objective Function with GAM | [c] : (1.496) 3pg + (3.7478) accoa + (59.8100) atp + (0.3610) e4p + (0.0709) f6p + (0.1290) g3p + (0.2050) g6p + (0.2557) gln-L + (4.9414) glu-L + (59.8100) h2o + (3.5470) nad + (13.0279) nadph + (1.7867) oaa + (0.5191) pep + (2.8328) pyr + (0.8977) r5p --> (59.8100) adp + (4.1182) akg + (3.7478) coa + (59.8100) h + (3.5470) nadh + (13.0279) nadp + (59.8100) pi |
| CO2t | CO2 transporter via diffusion | co2[e] <==> co2[c] |
| CS | citrate synthase | [c] : accoa + h2o + oaa --> cit + coa + h |
| CYTBD | cytochrome oxidase bd (ubiquinol-8: 2 protons) | (2) h[c] + (0.5) o2[c] + q8h2[c] --> (2) h[e] + h2o[c] + q8[c] |
| D_LACt2 | D-lactate transport via proton symport | h[e] + lac-D[e] <==> h[c] + lac-D[c] |
| ENO | enolase | [c] : 2pg <==> h2o + pep |
| ETOHt2r | ethanol reversible transport via proton symport | etoh[e] + h[e] <==> etoh[c] + h[c] |
| EX_ac(e) | Acetate exchange | [e] : ac <==> |
| EX_acald(e) | Acetaldehyde exchange | [e] : acald <==> |
| EX_akg(e) | 2-Oxoglutarate exchange | [e] : akg <==> |
| EX_co2(e) | CO2 exchange | [e] : co2 <==> |
| EX_etoh(e) | Ethanol exchange | [e] : etoh <==> |
| EX_for(e) | Formate exchange | [e] : for <==> |
| EX_fru(e) | D-Fructose exchange | [e] : fru <==> |
| EX_fum(e) | Fumarate exchange | [e] : fum <==> |
| EX_glc(e) | D-Glucose exchange | [e] : glc-D <==> |

| | | |
|---|---|---|
| EX_gln_L(e) | L-Glutamine exchange | [e] : gln-L <==> |
| EX_glu_L(e) | L-Glutamate exchange | [e] : glu-L <==> |
| EX_h(e) | H+ exchange | [e] : h <==> |
| EX_h2o(e) | H2O exchange | [e] : h2o <==> |
| EX_lac_D(e) | D-Lactate exchange | [e] : lac-D <==> |
| EX_mal_L(e) | L-Malate exchange | [e] : mal-L <==> |
| EX_nh4(e) | Ammonium exchange | [e] : nh4 <==> |
| EX_o2(e) | O2 exchange | [e] : o2 <==> |
| EX_pi(e) | Phosphate exchange | [e] : pi <==> |
| EX_pyr(e) | Pyruvate exchange | [e] : pyr <==> |
| EX_succ(e) | Succinate exchange | [e] : succ <==> |
| FBA | fructose-bisphosphate aldolase | [c] : fdp <==> dhap + g3p |
| FBP | fructose-bisphosphatase | [c] : fdp + h2o --> f6p + pi |
| FORt2 | formate transport via proton symport (uptake only) | for[e] + h[e] --> for[c] + h[c] |
| FORti | formate transport via diffusion | for[c] --> for[e] |
| FRD7 | fumarate reductase | [c] : fum + q8h2 --> q8 + succ |
| FRUpts2 | Fructose transport via PEP:Pyr PTS<br>(f6p generating) | fru[e] + pep[c] --> f6p[c] + pyr[c] |
| FUM | fumarase | [c] : fum + h2o <==> mal-L |
| FUMt2_2 | Fumarate transport via proton symport (2 H) | fum[e] + (2) h[e] --> fum[c] + (2) h[c] |
| G6PDH2r | glucose 6-phosphate dehydrogenase | [c] : g6p + nadp <==> 6pgl + h + nadph |
| GAPD | glyceraldehyde-3-phosphate dehydrogenase | [c] : g3p + nad + pi <==> 13dpg + h + nadh |
| GLCpts | D-glucose transport via PEP:Pyr PTS | glc-D[e] + pep[c] --> g6p[c] + pyr[c] |
| GLNS | glutamine synthetase | [c] : atp + glu-L + nh4 --> adp + gln-L + h + pi |
| GLNabc | L-glutamine transport via ABC system | atp[c] + gln-L[e] + h2o[c] --> adp[c] + gln-L[c] + h[c] + pi[c] |
| GLUDy | glutamate dehydrogenase (NADP) | [c] : glu-L + h2o + nadp <==> akg + h + nadph + nh4 |
| GLUN | glutaminase | [c] : gln-L + h2o --> glu-L + nh4 |
| GLUSy | glutamate synthase (NADPH) | [c] : akg + gln-L + h + nadph --> (2) glu-L + nadp |
| GLUt2r | L-glutamate transport via proton symport, reversible (periplasm) | glu-L[e] + h[e] <==> glu-L[c] + h[c] |
| GND | phosphogluconate dehydrogenase | [c] : 6pgc + nadp --> co2 + nadph + ru5p-D |
| H2Ot | H2O transport via diffusion | h2o[e] <==> h2o[c] |
| ICDHyr | isocitrate dehydrogenase (NADP) | [c] : icit + nadp <==> akg + co2 + nadph |
| ICL | Isocitrate lyase | [c] : icit --> glx + succ |
| LDH_D | D-lactate dehydrogenase | [c] : lac-D + nad <==> h + nadh + pyr |
| MALS | malate synthase | [c] : accoa + glx + h2o --> coa + h + mal-L |
| MALt2_2 | Malate transport via proton symport (2 H) | (2) h[e] + mal-L[e] --> (2) h[c] + mal-L[c] |
| MDH | malate dehydrogenase | [c] : mal-L + nad <==> h + nadh + oaa |
| ME1 | malic enzyme (NAD) | [c] : mal-L + nad --> co2 + nadh + pyr |
| ME2 | malic enzyme (NADP) | [c] : mal-L + nadp --> co2 + nadph + pyr |

| | | |
|---|---|---|
| NADH16 | NADH dehydrogenase (ubiquinone-8 & 3 protons) | (4) h[c] + nadh[c] + q8[c] --> (3) h[e] + nad[c] + q8h2[c] |
| NADTRHD | NAD transhydrogenase | [c] : nad + nadph --> nadh + nadp |
| NH4t | ammonia reversible transport | nh4[e] <==> nh4[c] |
| O2t | o2 transport via diffusion | o2[e] <==> o2[c] |
| PDH | pyruvate dehydrogenase | [c] : coa + nad + pyr --> accoa + co2 + nadh |
| PFK | phosphofructokinase | [c] : atp + f6p --> adp + fdp + h |
| PFL | pyruvate formate lyase | [c] : coa + pyr --> accoa + for |
| PGI | glucose-6-phosphate isomerase | [c] : g6p <==> f6p |
| PGK | phosphoglycerate kinase | [c] : 3pg + atp <==> 13dpg + adp |
| PGL | 6-phosphogluconolactonase | [c] : 6pgl + h2o --> 6pgc + h |
| PGM | phosphoglycerate mutase | [c] : 2pg <==> 3pg |
| PIt2r | phosphate reversible transport via proton symport | h[e] + pi[e] <==> h[c] + pi[c] |
| PPC | phosphoenolpyruvate carboxylase | [c] : co2 + h2o + pep --> h + oaa + pi |
| PPCK | phosphoenolpyruvate carboxykinase | [c] : atp + oaa --> adp + co2 + pep |
| PPS | phosphoenolpyruvate synthase | [c] : atp + h2o + pyr --> amp + (2) h + pep + pi |
| PTAr | phosphotransacetylase | [c] : accoa + pi <==> actp + coa |
| PYK | pyruvate kinase | [c] : adp + h + pep --> atp + pyr |
| PYRt2r | pyruvate reversible transport via proton symport | h[e] + pyr[e] <==> h[c] + pyr[c] |
| RPE | ribulose 5-phosphate 3-epimerase | [c] : ru5p-D <==> xu5p-D |
| RPI | ribose-5-phosphate isomerase | [c] : r5p <==> ru5p-D |
| SUCCt2_2 | succinate transport via proton symport (2 H) | (2) h[e] + succ[e] --> (2) h[c] + succ[c] |
| SUCCt3 | succinate transport out via proton antiport | h[e] + succ[c] --> h[c] + succ[e] |
| SUCDi | succinate dehydrogenase (irreversible) | [c] : q8 + succ --> fum + q8h2 |
| SUCOAS | succinyl-CoA synthetase (ADP-forming) | [c] : atp + coa + succ <==> adp + pi + succoa |
| TALA | transaldolase | [c] : g3p + s7p <==> e4p + f6p |
| THD2 | NAD(P) transhydrogenase | (2) h[e] + nadh[c] + nadp[c] --> (2) h[c] + nad[c] + nadph[c] |
| TKT1 | transketolase | [c] : r5p + xu5p-D <==> g3p + s7p |
| TKT2 | transketolase | [c] : e4p + xu5p-D <==> f6p + g3p |
| TPI | triose-phosphate isomerase | [c] : dhap <==> g3p |

**Table A2 Reactions of the Core *E. coli* K-12 MG1655 iAF1260 Model.**

The abbreviation, official name and equation for each reaction are adopted from [Fei07].

| Abbreviation | Official Name | Formula |
|---|---|---|
| 13dpg | 3-Phospho-D-glyceroyl phosphate | C3H4O10P2 |
| 2pg | D-Glycerate 2-phosphate | C3H4O7P |
| 3pg | 3-Phospho-D-glycerate | C3H4O7P |
| 6pgc | 6-Phospho-D-gluconate | C6H10O10P |
| 6pgl | 6-phospho-D-glucono-1,5-lactone | C6H9O9P |
| ac | Acetate | C2H3O2 |
| ac[e] | Acetate (extracellular) | C2H3O2 |
| acald | Acetaldehyde | C2H4O |
| acald[e] | Acetaldehyde (extracellular) | C2H4O |
| accoa | Acetyl-CoA | C23H34N7O17P3S |
| acon-C | cis-Aconitate | C6H3O6 |
| actp | Acetyl phosphate | C2H3O5P |
| adp | ADP | C10H12N5O10P2 |
| akg | 2-Oxoglutarate | C5H4O5 |
| akg[e] | 2-Oxoglutarate (extracellular) | C5H4O5 |
| amp | AMP | C10H12N5O7P |
| atp | ATP | C10H12N5O13P3 |
| cit | Citrate | C6H5O7 |
| co2 | CO2 | CO2 |
| co2[e] | CO2 (extracellular) | CO2 |
| coa | Coenzyme A | C21H32N7O16P3S |
| dhap | Dihydroxyacetone phosphate | C3H5O6P |
| e4p | D-Erythrose 4-phosphate | C4H7O7P |
| etoh | Ethanol | C2H6O |
| etoh[e] | Ethanol (extracellular) | C2H6O |
| f6p | D-Fructose 6-phosphate | C6H11O9P |
| fdp | D-Fructose 1,6-bisphosphate | C6H10O12P2 |
| for | Formate | CH1O2 |
| for[e] | Formate (extracellular) | CH1O2 |
| fru[e] | D-Fructose (extracellular) | C6H12O6 |
| fum | Fumarate | C4H2O4 |
| fum[e] | Fumarate (extracellular) | C4H2O4 |
| g3p | Glyceraldehyde 3-phosphate | C3H5O6P |
| g6p | D-Glucose 6-phosphate | C6H11O9P |
| glc-D[e] | D-Glucose (extracellular) | C6H12O6 |
| gln-L | L-Glutamine | C5H10N2O3 |
| gln-L[e] | L-Glutamine (extracellular) | C5H10N2O3 |
| glu-L | L-Glutamate | C5H8NO4 |
| glu-L[e] | L-Glutamate (extracellular) | C5H8NO4 |
| glx | Glyoxylate | C2H1O3 |
| h2o | H2O | H2O |
| h2o[e] | H2O (extracellular) | H2O |
| h | H+ | H |
| h[e] | H+ (extracellular) | H |
| icit | Isocitrate | C6H5O7 |

| | | |
|---|---|---|
| lac-D | D-Lactate | C3H5O3 |
| lac-D[e] | D-Lactate (extracellular) | C3H5O3 |
| mal-L | L-Malate | C4H4O5 |
| mal-L[e] | L-Malate (extracellular) | C4H4O5 |
| nad | Nicotinamide adenine dinucleotide | C21H26N7O14P2 |
| nadh | Nicotinamide adenine dinucleotide - reduced | C21H27N7O14P2 |
| nadp | Nicotinamide adenine dinucleotide phosphate | C21H25N7O17P3 |
| nadph | Nicotinamide adenine dinucleotide phosphate - reduced | C21H26N7O17P3 |
| nh4 | Ammonium | H4N |
| nh4[e] | Ammonium (extracellular) | H4N |
| o2 | O2 | O2 |
| o2[e] | O2 (extracellular) | O2 |
| oaa | Oxaloacetate | C4H2O5 |
| pep | Phosphoenolpyruvate | C3H2O6P |
| pi | Phosphate | HO4P |
| pi[e] | Phosphate (extracellular) | HO4P |
| pyr | Pyruvate | C3H3O3 |
| pyr[e] | Pyruvate (extracellular) | C3H3O3 |
| q8 | Ubiquinone-8 | C49H74O4 |
| q8h2 | Ubiquinol-8 | C49H76O4 |
| r5p | alpha-D-Ribose 5-phosphate | C5H9O8P |
| ru5p-D | D-Ribulose 5-phosphate | C5H9O8P |
| s7p | Sedoheptulose 7-phosphate | C7H13O10P |
| succ | Succinate | C4H4O4 |
| succ[e] | Succinate (extracellular) | C4H4O4 |
| succoa | Succinyl-CoA | C25H35N7O19P3S |
| xu5p-D | D-Xylulose 5-phosphate | C5H9O8P |

**Table A3 Metabolites of the Core *E. coli* K-12 MG1655 iAF1260 Model.** The abbreviation, official name and formula for each reaction are adopted from [Fei07].

For the visual representation of the core *E. coli* K-12 MG1655 iAF1260 metabolic network, interested readers are referred to [Ort10].