

University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

---

Computer Science and Engineering: Theses,  
Dissertations, and Student Research

Computer Science and Engineering, Department of

---

12-2016

# Characterization of Molecular Communication Based on Cell Metabolism Through Mutual Information and Flux Balance Analysis

Zahmeeth Sayed Sakkaff

University of Nebraska - Lincoln, sszahmeeth@gmail.com

Follow this and additional works at: <http://digitalcommons.unl.edu/computerscidiss>

 Part of the [Biochemical and Biomolecular Engineering Commons](#), [Bioinformatics Commons](#), [Biological Engineering Commons](#), [Biology Commons](#), [Biomaterials Commons](#), [Biomechanics and Biotransport Commons](#), [Catalysis and Reaction Engineering Commons](#), [Computational Engineering Commons](#), [Computer Engineering Commons](#), [Molecular, Cellular, and Tissue Engineering Commons](#), [Nanoscience and Nanotechnology Commons](#), and the [Other Biomedical Engineering and Bioengineering Commons](#)

---

Sakkaff, Zahmeeth Sayed, "Characterization of Molecular Communication Based on Cell Metabolism Through Mutual Information and Flux Balance Analysis" (2016). *Computer Science and Engineering: Theses, Dissertations, and Student Research*. 114.  
<http://digitalcommons.unl.edu/computerscidiss/114>

This Article is brought to you for free and open access by the Computer Science and Engineering, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Computer Science and Engineering: Theses, Dissertations, and Student Research by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

CHARACTERIZATION OF MOLECULAR COMMUNICATION BASED ON  
CELL METABOLISM THROUGH MUTUAL INFORMATION AND FLUX  
BALANCE ANALYSIS

by

Zahmeeth Sayed Sakkaff

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Professor Massimiliano Pierobon

Lincoln, Nebraska

December, 2016

CHARACTERIZATION OF MOLECULAR COMMUNICATION BASED ON  
CELL METABOLISM THROUGH MUTUAL INFORMATION AND FLUX  
BALANCE ANALYSIS

Zahmeeth Sayed Sakkaff, M. S.

University of Nebraska, 2016

Adviser: Massimiliano Pierobon

Synthetic biology is providing novel tools to engineer cells and access the basis of their molecular information processing, including their communication channels based on chemical reactions and molecule exchange. Molecular communication is a discipline in communication engineering that studies these types of communications and ways to exploit them for novel purposes, such as the development of ubiquitous and heterogeneous communication networks to interconnect biological cells with nano and biotechnology-enabled devices, *i.e.*, the Internet of Bio-Nano Things. One major problem in realizing these goals stands in the development of reliable techniques to control the engineered cells and their behavior from the external environment. A possible solution may stem from exploiting the natural mechanisms that allow cells to regulate their metabolism, the complex network of chemical reactions that underlie their growth and reproduction, as a function of chemical compounds in the environment.

In this thesis, molecular communication concepts are applied to study the potential of cell metabolism, and its regulation, to channel information from the outside environment into the cell as function of chemical compounds in the environment, and quantify how much information of the internal state of the metabolic network can be perceived from the outside environment. For this, cell metabolism is characterized in

this work through two abstractions, namely, as a molecular communication encoder and a modulator, respectively. The former models the cell metabolism as a binary encoder of the mechanisms underlying the regulation of the cell metabolic network state in function of the chemical composition of the external environment. The latter models the metabolic network inside the cell as a digital modulator of metabolite exchange/growth according to the information contained in its state. Based on these abstractions, the aforementioned potential of cell metabolism is quantified with the information theoretic mutual information parameter obtained through the use of a well-known and computationally efficient metabolic simulation technique.

Numerical results are obtained through simulation of cell metabolism based on the standard processes of Genome Scale Modeling (GEM) and Flux Balance Analysis (FBA). These preliminary proof-of-concept results are based on the following three main cellular species: *Escherichia coli* (*E. coli*), the “standard” organism in microbiology, and two important human gut microbes studied in our collaborators’ lab, namely, the *Bacteroides thetaiotaomicron* (*B. theta*) and the *Methanobrevibacter smithii* (*M. smithii*), which provide a direct connection of this work to future practical applications.

## ACKNOWLEDGMENTS

At this place I would like to thank many people who helped me to complete this thesis successfully. First of all, Prof. Massimiliano Pierobon, who supervised me throughout this research and introduced me to the challenging Telecommunication Systems Modeling and Engineering of Cell Communication Pathways research area. I am very grateful to him for the immense amount of time that he dedicated for supervision, for his patience, for his continual support and motivation and knowledge leading me into the appropriate methods and techniques needed in this work. Your knowledge and enormous support has been the key to all my work, thanks! Next, I would like to thank my committee members Dr. Myra Cohen and Dr. Juan Cui, thank you for devoting your valuable time to improve the thesis. I would also like to thank Dr. Myra Cohen and Dr. Nicole Buan, Mikaela Cashman, Jennie Catlett and Dr. Christine Kelley for their constructive feedback, and for encouraging me to delve into the subject of cell metabolism, and flux balance analysis through the KBase software application suite.

I would like to thank all the Molecular and Biochemical Telecommunications (MBiTe) Lab colleagues, in particular, Aditya Immaneni who worked with me to visualizing the complex network of cell metabolisms. I would also like to thank the KBase developers team for their active support throughout the development of this research and US National Science Foundation for supporting this research through grant MCB-1449014. Apart from these, I would also like to acknowledge Allison Haindfield as the proof reader of this thesis, and I am gratefully indebted to her for her very valuable comments on this thesis. Finally, I must express my very profound gratitude to my family members, and friends for providing me with unfailing support and continuous encouragement throughout my years of study and through the pro-

cess of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

# Contents

<b>Contents</b>	<b>vi</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Background</b>	<b>5</b>
2.1 Motivation . . . . .	5
2.2 Biological Pathways . . . . .	6
2.3 Cell Metabolism . . . . .	9
2.4 The Mathematical Relation Between Transcription Factors and Enzymes	10
<b>3 Cell Metabolism as a Molecular Communication System</b>	<b>13</b>
3.1 Molecular Communication Abstraction of Cell Metabolism . . . . .	13
3.2 Steady-State Mutual Information . . . . .	18
3.2.1 Stage I . . . . .	18
3.2.2 Stage II . . . . .	19

<b>4</b>	<b><i>In silico</i> Characterization of the Mutual Information of Cell Metabolism through Flux Balance Analysis</b>	<b>21</b>
4.1	Estimation of Optimal Enzyme Expression Regulation through Flux Balance Analysis . . . . .	21
4.2	An Upper Bound Steady-State Mutual Information of Cell Metabolism	27
4.2.1	Upper Bound for Stage I . . . . .	27
4.2.2	Upper Bound for Stage II . . . . .	29
<b>5</b>	<b>Data Generation Workflow</b>	<b>31</b>
5.1	Data generation . . . . .	31
5.2	<i>In silico</i> experimentation for <i>E. coli</i> . . . . .	34
5.3	<i>In silico</i> experimentation for <i>B. theta</i> and <i>M. smithii</i> . . . . .	35
<b>6</b>	<b>Numerical Results</b>	<b>37</b>
6.1	Stage I . . . . .	38
6.1.1	Numerical Results for <i>E. coli</i> . . . . .	38
6.1.2	Numerical Results for <i>B. theta</i> and <i>M. smithii</i> . . . . .	42
6.2	Stage II . . . . .	45
6.2.1	Upper bound of steady-state mutual information over internal metabolic state changing reactions with respect to biomass only	47
6.2.2	Upper bound of steady-state mutual information over internal metabolic state changing reactions with respect to uptake and secretion of compounds and biomass . . . . .	48
6.2.3	Upper bound of steady-state mutual information over seven input compounds with respect to uptake and secretion of compounds and biomass . . . . .	49
6.3	Visualization . . . . .	51

<b>7 Conclusions and Future Work</b>	<b>55</b>
--------------------------------------	-----------

<b>Bibliography</b>	<b>57</b>
---------------------	-----------

# List of Figures

1.1	Proposed model structure for interpretation of biological system with molecular communication system . . . . .	3
2.1	Graphical representation of the interconnection of signal transduction, gene regulation and metabolic pathways. . . . .	7
3.1	Sketch of the proposed molecular communication abstraction of cell metabolism.	14
3.2	Sketch of the proposed molecular communication system based on cell metabolism. . . . .	15
3.3	Sketch of the proposed Metabolic Reaction Abstraction. . . . .	16
4.1	KEGG module of <i>Bacteroides thetaiotaomicron</i> metabolic pathways. <b>(a)</b> Summary of the biological processes shown in the pathway map of Glycolysis / Gluconeogenesis and Glyoxylate and dicarboxylate metabolism <b>(b)</b> Enlarged fine details of a section of a complete metabolic model, <b>(c)</b> Part of the complete KEGG database pathway maps of <i>Bacteroides thetaiotaomicron</i> . . . . .	23
4.2	From metabolic network to stoichiometric matrix. <b>(a)</b> Conceptual model of a metabolic network. <b>(b)</b> List of the reactions participating in the metabolic network. <b>(c)</b> Corresponding stoichiometric matrix. . . . .	24

4.3	Conceptual model of the FBA LP formulation for finding the optimal solution. . . . .	25
5.1	KBase Workflow for simulation data. . . . .	32
5.2	KBase workflow to generate silico data. . . . .	33
6.1	Optimal E.Coli K12 MG1655 growth as a function of the input flux of D-Glucose and Lactose in the environment. . . . .	39
6.2	FBA-estimated binary chemical reaction states $\{r_i^*\}_{i=1}^M$ for each combination of <i>D-Glucose</i> and <i>Lactose</i> input fluxes, where white = ON state; black = OFF state. . . . .	40
6.3	FBA-estimated binary chemical reaction states $\{r_i^*\}_{i=1}^M$ for each combination of <i>Glucose</i> , <i>Hematin</i> , <i>Formate</i> , $H_2$ , <i>VitaminB<sub>12</sub></i> , <i>Acetate</i> , and <i>Vitamin K</i> input fluxes, where yellow = ON state; violet = OFF state. . . . .	42
6.4	FBA-estimated binary chemical reaction states $\{r_i^*\}_{i=1}^M$ for each combination of <i>Glucose</i> , <i>Hematin</i> , <i>Formate</i> , $H_2$ , <i>VitaminB<sub>12</sub></i> , <i>Acetate</i> , and <i>Vitamin K</i> input fluxes, where yellow = ON state; violet = OFF state. . . . .	43
6.5	14 FBA groups of <i>B. theta</i> based on similar FBA-estimated chemical reaction states for each combination of <i>Glucose</i> , <i>Hematin</i> , <i>Formate</i> , $H_2$ , <i>VitaminB<sub>12</sub></i> , <i>Acetate</i> , and <i>Vitamin K</i> input fluxes. . . . .	45
6.6	31 FBA groups of <i>M. smithii</i> based on similar FBA-estimated chemical reaction states for each combination of <i>Glucose</i> , <i>Hematin</i> , <i>Formate</i> , $H_2$ , <i>VitaminB<sub>12</sub></i> , <i>Acetate</i> , and <i>Vitamin K</i> input fluxes. . . . .	46
6.7	A hive plot for the FBA group F is shown in the figure. The reactions are placed on the Z axis, the reactants on the X axis and the products on the Y axis. Further the External compounds are placed higher on the X and Y axes than the Internal compounds. . . . .	51

- 6.8 Shows differential hive plots of F vs G and F vs Z. The groups F and G in F vs G hive plot has the same biomass whereas, the groups F and Z in F vs Z hive plot have the least and highest biomass respectively. When a reaction is present in F and absent in G or Z the reaction is represented along with its links to the compounds. When a reaction is present in the other groups but absent in group F the reaction is shown as a node not connected to any other compounds. . . . . 53
- 6.9 Shows the state changing reactions represented as a network diagram. The size of a node is proportional to the number of links connecting to or from it. . . . . 54

## List of Tables

6.1	FBA group matrix of <i>B. theta</i> where the rows are the 14 groups and the columns represent the chemical compounds uptaken, secreted, and the generation biomass. . . . .	46
6.2	FBA group matrix of <i>M. smithii</i> where the rows are the 31 groups and the columns represent the chemical compounds uptaken, secreted, and the generated biomass. . . . .	47

# Chapter 1

## Introduction

Molecular communication is one of the latest frontiers in communication engineering [1], where tools from computer communications, information theory, signal processing, and wireless networking are applied to the domain of chemical reactions and molecule exchange. Recent results in molecular communication research range from theoretical studies of the communication channels and the expression of their communication capacity [36], [14], [43], [4], to the more practical design of suitable modulation and coding techniques [31], and networking protocols [16].

Synthetic biology is today providing novel tools for the design, realization, and control of biological processes through the programming of cells' genetic code [23]. These tools are allowing engineers to study and access the basis of molecular information processing in biological cells, which can be potentially utilized for the realization of practical molecular communication systems [35], [29]. The future pervasive deployment of genetically engineered cells and their interaction with other bio, micro and nano-technology enabled devices through molecular communication systems and networks has been recently envisioned as the novel paradigm of the Internet of Bio-Nano Things [2]. These ubiquitous and heterogeneous communications will

enable advanced applications in many fields, including medicine (*e.g.*, developing bio-compatible diagnosis and treatment systems), industry (*e.g.*, biologically-controlled food production), and agriculture (*e.g.*, monitoring and control of soil chemical and microbiological status).

One major problem in synthetic biology stands in the control from the external environment to the internal functionalities of genetically engineered cells. Various techniques to realize this control have been explored, such as the use of light, *i.e.*, optogenetics [45], magnetic fields, *i.e.*, magnetic nanoparticles [12], and dedicated signaling circuits [28]. A possible solution, may stem from exploiting the natural mechanisms involved in the regulation of cell metabolism. Cell metabolism is a complex network of chemical reactions that underlie the cell's growth and reproduction, which consumes and transforms chemical compounds present in its environment. Cells have mechanisms to regulate and optimize their metabolism (metabolic state) according to the chemical composition of the surrounding environment.

In order to achieve our goal of gaining more control of the internal cell functionalities from the external environment, we need a deeper understanding of how the information flows from the cell's environment to the metabolic state and how much information of the internal cell metabolic state can be perceived from the outside environment. This requires a deeper understanding of cell regulation mechanisms from beginning to end.

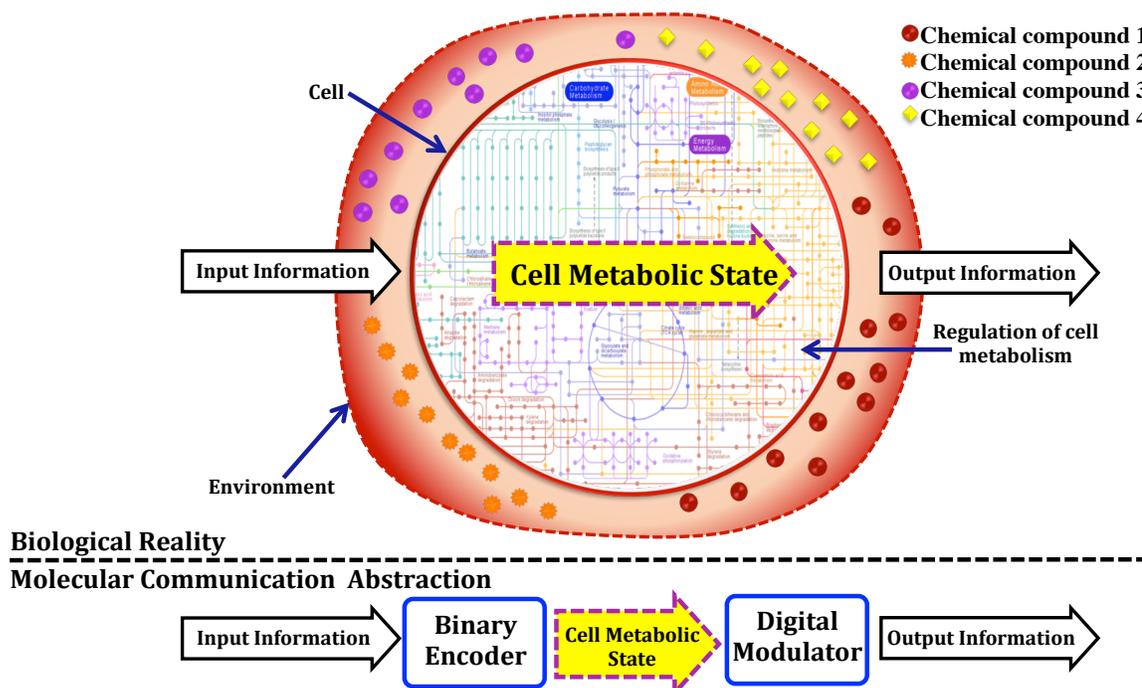


Figure 1.1: Proposed model structure for interpretation of biological system with molecular communication system

In this thesis, we apply molecular communication (MC) concepts to study the potential of cell metabolism, and its regulation, to channel information from the outside environment into the cell as shown in Figure 1.1. In the view of molecular communication systems, the potential of cell metabolism and its regulation can be understood by how much information (in bits) the cell can take from the external environment and encode into changes in its internal metabolic behavior. For this end, we abstract the cell metabolism as a cascade of two MC components, namely, i) as a binary encoder of mechanisms underlying the regulation of cell metabolic state as function of chemical compounds of the external environment, and ii) as a digital modulator of metabolite growth/exchange according to the information contained in the metabolic state of the cell.

Inspired by [39], we apply information theory tools to express the performance

of this binary encoder (Stage I) and digital modulator (Stage II) in terms of steady-state mutual information. Subsequently, we define an upper bound to this mutual information that can easily be quantified *in silico* through the use of a well-known and computationally efficient metabolic simulation techniques, namely, Genome Scale Modeling (GEM) and Flux Balance Analysis (FBA), which relies only on the *a priori* knowledge of the cell's DNA code (genome) and epigenetic regulation.

Next, we present numerical results obtained by analyzing the binary encoder model of the *Escherichia coli* (abbreviated as *E. coli*) bacterium metabolism and its regulation with respect to two different input chemical compounds, namely, D-Glucose and Lactose. We also show the results of FBA in terms of growth rate computed for different combinations of values of input fluxes of the aforementioned two compounds. Further, we report the numerical results of Stage I and Stage II proposed for a case study with human gut microbes metabolism and its regulation named as *Bacteroides thetaiotaomicron* (abbreviated as *B. theta*) and *Methanobrevibacter smithii* (abbreviated as *M. smithii*) with respect to seven different input chemical compounds.

The rest of this thesis is organized as follows. Chapter 2 covers some necessary background in microbiology and mathematical relation between transcription factors and enzymes. Chapter 3 presents the cell metabolism as a molecular communication system. Here, we examine how cell metabolism is characterized by the two abstractions, namely, as an MC binary encoder and a digital modulator. Chapter 4 discusses the information theory to characterize the steady-state mutual information of the proposed abstractions. Chapter 5 discusses *in silico* experimentations to generate the data presented in this thesis. Chapter 6 presents the numerical results for three main cellular species, *E. coli*, *B. theta* and *M. smithii*, through several simulations and experiments. The final chapter covers the analysis of our findings, the conclusion, and details some future avenues.

# Chapter 2

## Background

### 2.1 Motivation

A living cell environment is composed of various biochemical compounds. Different compounds work together to encode a particular information by combining various biochemical signals [39], [33]. This environmental information is then sensed by the cell, which regulates its own internal metabolic network state consequently [39]. A cell is able to extract this information from the environment by means of biochemical processes present inside the cell. These biological processes are generally composed of signal transduction and gene regulation [18], [33], [7], which occur along a chain of chemical reactions known as biological pathways. This extracted information can be used to alter the gene expression, thereby modifying the cell metabolic network state [33], [39]. Hence, the final outcome of the modified cell metabolic network state is correlated to the information that the cell extracts from the environment. However, during this process there will be some limitation in the amount of information that the cell can extract from the environment. This is due to the generation of biochemical noise coming from feedback loops, cross talks, amplification, integration, and a delay

inherent within the aforementioned biological processes [18], [46]. As a result, the extracted information can be distorted, and the cell may not be successful in precisely modify its metabolic network accordingly.

In this thesis, we abstract the behavior of the aforementioned biological processes of a cell as a communication system and analyze them with communication engineering tools. The goal of our work is to characterize the aforementioned limitations of a cell in extracting input information from the environment. By stemming from this, we apply communication engineering concepts in order to characterize the potential of a cell to represent information from the external environment to the metabolic state and, subsequently, we quantify how much information of the internal state of the metabolic network can be perceived from the external environment as a form of biomass (growth), uptake, and secretion<sup>1</sup>. In order to quantify the limits of this information flow, inspired by the mathematical model proposed by [39], [46], [15], we apply Information theory tools. Based on the latter, we develop a mathematical framework to compute the amount of information flowing from the external environment through the cell's metabolic state to the cell growth, uptake, and secretion of chemical compounds. We envision that the results of this work will enable the future design of communication engineering techniques to operate a fine-tuned control of the cell behavior based on information transmission through its metabolism.

## 2.2 Biological Pathways

In this section, we briefly discuss what a biological pathway is and how the interaction between different biological pathways vary based on the variations in the chemical

---

<sup>1</sup>Biomass is a production of organic materials in the cell which maximize the growth. An uptake is an absorption of chemical compounds by the cell from the environment to help the chemical reactions during the cell metabolism process. Secretion is the release of chemical compounds, that were the products of some chemical reactions and not needed by the cell, into the environment.

compounds present in the environment. A better understanding of the workings of biological pathways will help us better understand the cell metabolism, which is the main focus of this study.

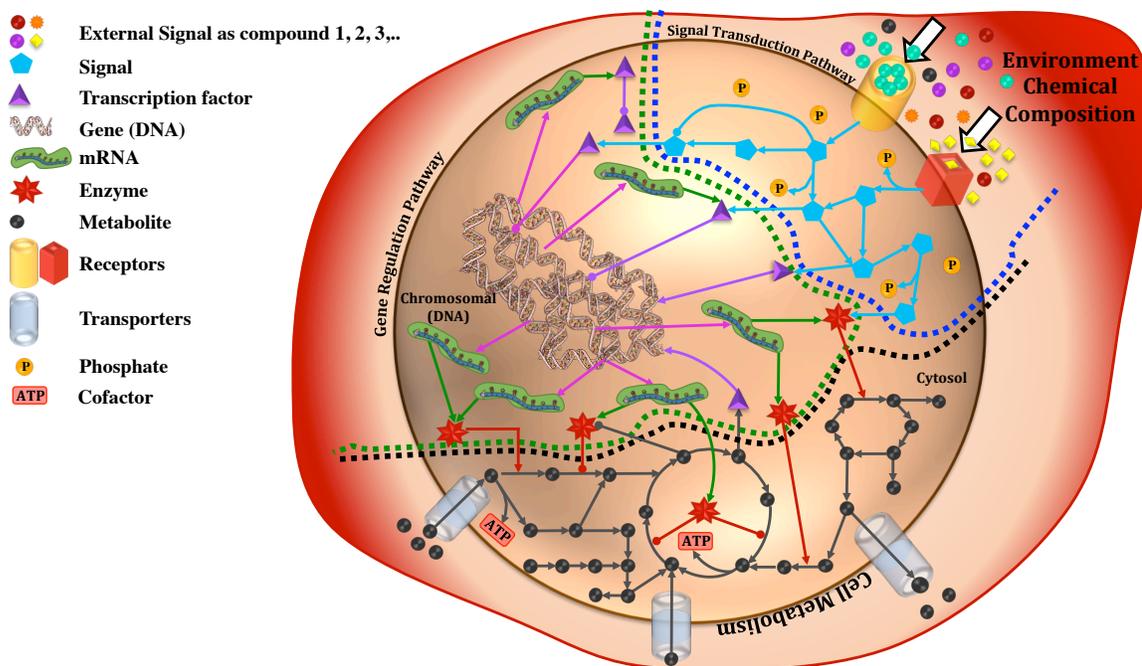


Figure 2.1: Graphical representation of the interconnection of signal transduction, gene regulation and metabolic pathways.

A biological pathway can be defined as a chain of reactions among chemical compounds in a cell which leads to the production of a certain product, such as protein or fat molecules. In some cases, a biological pathway can cause a change inside the cell such as turning a gene expression ON/OFF. In order to carry out their designated task, the compounds involved in biological pathways interact with each other and with chemical signals inside the cell. The biological pathways can essentially be categorized into three kinds: signal transduction pathways, gene regulation pathways, and cell metabolic pathways (which compose the aforementioned metabolic networks).

Signal transduction pathways transport information from the cell's environment to its interior. Cells have proteins on their surface, called receptors to which compounds from the environment bind. After this binding, the information about the compounds in the environment travels inside the cell where it is transported by specialized proteins that trigger specific reactions, these reactions are organized into cascades. This cascade of phosphorylations, which is the addition of a **phosphate** group to a molecule. This cascade of phosphorylation dictates the activation of transcription factors. Generally, the output of these reactions is relayed to one or more gene regulation pathways, detailed next. Figure 2.1 shows the chemical compounds (small structures outside of the cell) from the cell's environment and how they bind to the **receptors** (yellow cylinder and red cube) located on the surface of the cell membrane. This leads to the activation of the cell's signaling pathways (interconnected blue lines). Once these pathways are activated, special proteins (blue pentagons) transport the signal internally which regulate the gene expression by activating or inhibiting **transcription factors** (purple triangles) which are initially inactive proteins floats freely in the **cytosol**.

Gene regulation pathways consist of chemical reactions that govern the genes and their expression level into proteins. Some of these proteins are **enzymes** (red stars in Figure 2.1), that act as catalysts of specific biochemical reactions in metabolic pathways. They work by regulating the transcription and translation processes, where transcription factors interfere with the deoxyribonucleic acid (DNA) copy into messenger ribonucleic acid (mRNA) and the subsequent synthesis of the gene-encoded proteins [18]. The transcription factors can be either active or inactive based on the output of the aforementioned signal transduction pathways. Figure 2.1 also shows how the activated transcription factors interact with the **genes (DNA)**, which then produce molecules **mRNA**. These mRNA molecules are subsequently translated into

proteins. The transcription factors, genes, and mRNA form the gene regulatory pathways. The newly produced proteins are used for internal cellular mechanisms such as building cellular components, transporting the information in signal transduction pathways, and acting as catalysts for the metabolic reactions.

Metabolic pathways are chains of chemical reactions inside the cell that break down compounds present in the environment, such as sugar, minerals, or vitamins to obtain energy and materials to expand the cell (biomass). In Figure 2.1 the **cofactor** is an energy molecule (*e.g.* **ATP**) which drives the chemical reactions. Metabolic pathways are regulated by the aforementioned enzymes, that control rate of the reactions constituting them. As result of enzyme regulation, chemical compounds are exchanged between the cell and environment via special protein called **transporters**.

As we can see, the three main types of biological pathways such as signal transduction, gene regulation and metabolic pathways are tightly interconnected, and cells' behavior can only be properly understood by studying these three processes together. In this thesis we will apply on molecular communication concepts in order to study the potential of cell metabolism, and its regulation, to channel information from the outside environment into the cell. We also look at how the internal state changes of cell metabolism are perceived from the outside environment. In the next section, we detail the cell metabolism and its key aspects.

## 2.3 Cell Metabolism

Cell metabolism is a complex process comprising of pathways that break down the chemical compounds to produce energy, synthesize proteins, and produce building blocks required for the cell growth and reproduction [33]. Some of these compounds are consumed as inputs for subsequent chemical reactions in the pathways, while the

remaining metabolites are used to either generate energy or build cell components (biomass), while consuming energy or otherwise discarded through secretion. Additionally, these reactions also receive an uptake of chemical compounds from the cell's environment in the form of inputs to keep the chain of reactions going [42]. Most of the chemical reactions that compose the aforementioned metabolic pathways do not take place spontaneously, but are catalyzed by enzymes, defined above.

Cells have predefined mechanisms to control the activities of enzymes [30], mostly through the aforementioned signal transduction and gene expression pathways. For example, the cell can fine tune the rate at which the corresponding catalyzed reactions occur by controlling the expression of the corresponding enzyme-encoding gene, which in turn is controlled by the information about the compounds present in the external environment that signal transduction pathways propagate inside the cell. Among different adaptation mechanisms, we focus on the regulation of enzyme expression from their corresponding DNA genes as a function of the input chemical compounds.

As we stated, the metabolic process is regulated by enzymes. The enzyme expression mechanism should be controlled depending on the required outcome. This is achieved by regulating the transcription process, as discussed in the previous section.

## **2.4 The Mathematical Relation Between Transcription Factors and Enzymes**

The process underlying the expression of enzymes can be formalized mathematically. This formalization is important as it forms the basis of the molecular communication abstraction covered in the next section. Since the enzymes act as catalysts for most of the chemical reactions that happen during cell metabolism, the rate at which

each enzyme is expressed directly influences the rate of the corresponding catalyst reaction. According to a commonly accepted biological model, given a determinate concentration of active transcription factors  $[TF^*]$ , the rate  $R_e$  at which an enzyme is expressed is given by one of the following two sigmoidal expressions, called Hill's functions: [3], [44]:

$$R_e = \frac{\beta [TF^*]^n}{K_d^n + [TF^*]^n} \text{ if activation,} \quad (2.1)$$

$$R_e = \frac{\beta}{1 + \left(\frac{[TF^*]}{K_d}\right)^n} \text{ if repression;} \quad (2.2)$$

where  $\beta$  is the maximum expression level of the enzyme,  $n$  is the Hill's coefficient, having values between 1 and 4 depending on how many transcription factors cooperatively interact with the DNA gene and the equilibrium constant,  $K_d$  [44]. Equation (2.1) models the situation where a higher concentration of transcription factors increase the enzyme expression from zero to  $\beta$  (activation), while (2.2) models the opposite (repression). Due to the sigmoidal behavior of the above two expressions with respect to the active transcription factor concentration  $[TF^*]$ , they can be expressed with a logical approximation as follows [3]:

$$R_e \simeq \beta H([TF^*] - K_d) \text{ if activation,} \quad (2.3)$$

$$R_e \simeq \beta H(K_d - [TF^*]) \text{ if repression;} \quad (2.3)$$

where  $H(\cdot)$  is the Heaviside step function, equal to 1 when the argument is positive, and 0 *vice versa*. This logical approximation makes sense because the sigmoidal behavior expressed by (2.1) and (2.2) has the curve values close to two distinct states, namely, zero and maximum expression level  $\beta$ . According to the approximation in (2.3), the enzyme expression, and the rate of chemical reaction, can be separated into

distinct states: ON or OFF, depending on the concentration of active transcription factors, which in turns depends on the information about the compounds present in the environment propagated by the cell signal transduction pathways. The ON state represents the maximum enzyme expression rate and corresponding metabolic reaction rate, while OFF state represents no enzyme expression and absence of the corresponding chemical reactions in the cell metabolism. Hence, variations in the chemical compounds in the environment ultimately cause changes in the metabolic state by activating or deactivating the chemical reactions. These changes in the chemical reaction state (ON/OFF) are then reflected as variations in the uptake and secretion of chemical compounds and the biomass (cell growth), which can be experimentally measurable parameters when we consider the cell metabolic process from the perspective of the environment.

## Chapter 3

# Cell Metabolism as a Molecular Communication System

### 3.1 Molecular Communication Abstraction of Cell Metabolism

In order to understand how the information flows from the cell's environment to the metabolic state, and how much information of the internal cell metabolic state can be perceived from the outside environment, we need to understand the cell metabolic regulation mechanism from the beginning to end. For, this, we propose a molecular communication abstraction, sketched in Figure 3.1.

Part (a) in the below Figure 3.1 shows that the cell takes certain concentrations of chemical compounds into the cell metabolism, where the variations in these concentrations cause state changes in the cell's metabolic network, represented by dashed and solid lines connecting dots. The dots represent the chemical compounds involved in the metabolic reactions, or metabolites, the dashed lines represent the inactive

reactions, and the black and pink solid lines represent the active reactions, respectively. The pink solid lines represent the active state- changing reactions *i.e.*, the reactions that change their state as function of variations in the chemical compounds concentrations.

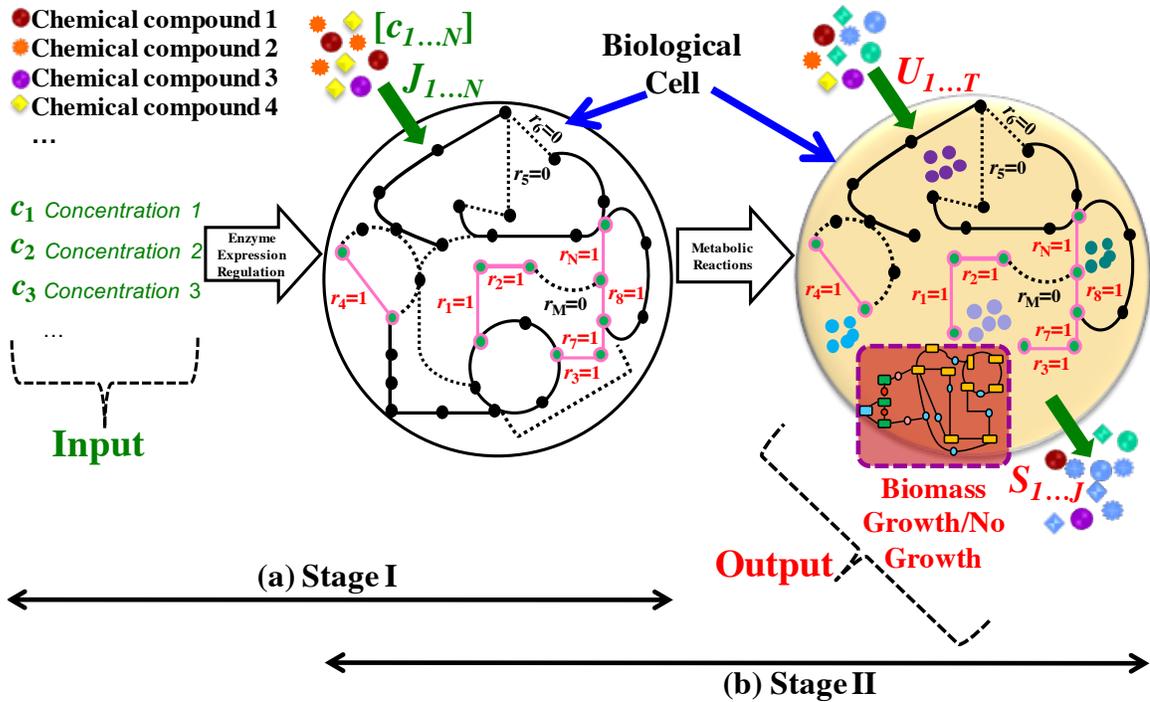


Figure 3.1: Sketch of the proposed molecular communication abstraction of cell metabolism.

We model Part (a), or Stage I, as the molecular communication binary encoder abstraction as in the form of expression of enzyme regulation and Part (b), or Stage II, as the molecular communication digital modulator abstraction. These state changes are reflected as variations in the uptake and secretion of chemical compounds and the biomass (cell growth), which can be measured when we consider the cell metabolic process from the perspective of the environment shown in Part (b) of the figure. The uptake of compounds are represented by  $U_w$ , while the secretion of compounds are represented by  $S_x$ . Hence, variations in the chemical compounds in the environment

can be modeled as input, whereas variations in the uptake, secretion, and biomass can be modeled as the output of the whole cell metabolic regulation process.

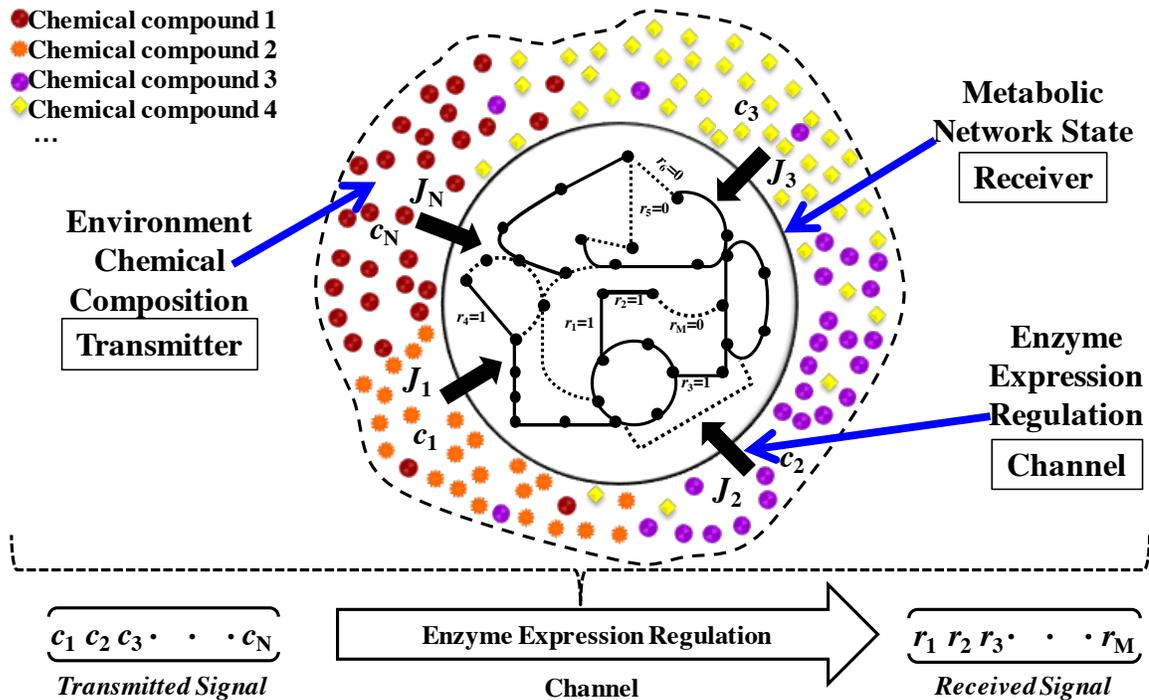


Figure 3.2: Sketch of the proposed molecular communication system based on cell metabolism.

Part (a), molecular communication binary encoder abstraction, also shown in Figure 3.2, abstracts the system into a Transmitter, a Receiver and a Channel. The goal of this abstraction is to model the enzyme regulation expression as a binary encoding of the information contained in the chemical composition of the cell’s environment “Some of this material appears in [38]”. The transmitter represents the environment surrounding the cell, where the transmitted signal is the set of chemical compounds present in the environment that are input of the pathways that compose the cell metabolic network. The channel represents the mechanisms that regulate the expression of determinate enzymes in function of the chemical compounds in input, and the receiver represents the cell metabolism, where the received signal is the result-

ing aforementioned activity (ON/OFF) of the chemical reactions catalyzed by these enzymes. This abstraction is more formally expressed as

$$\{c_1, c_2, \dots, c_N\} \xrightarrow[\text{Regulation}]{\text{Enzyme Expression}} \{r_1, r_2, \dots, r_M\}, \quad (3.1)$$

where  $c_i$  is the concentration (number of molecules per unit volume) of the chemical compound  $i$ ,  $N$  is the number of chemical compounds present in the environment surrounding the cell and input of the metabolic pathway network,  $r_i$  is a binary value equal to 1 if the enzyme-expression-regulated reaction  $i$  is ON, and equal to 0 if the same reaction is OFF,  $M$  is the number of enzyme-expression-regulated reactions that change their state upon variations in the concentrations of input chemical compounds  $c_i$ .

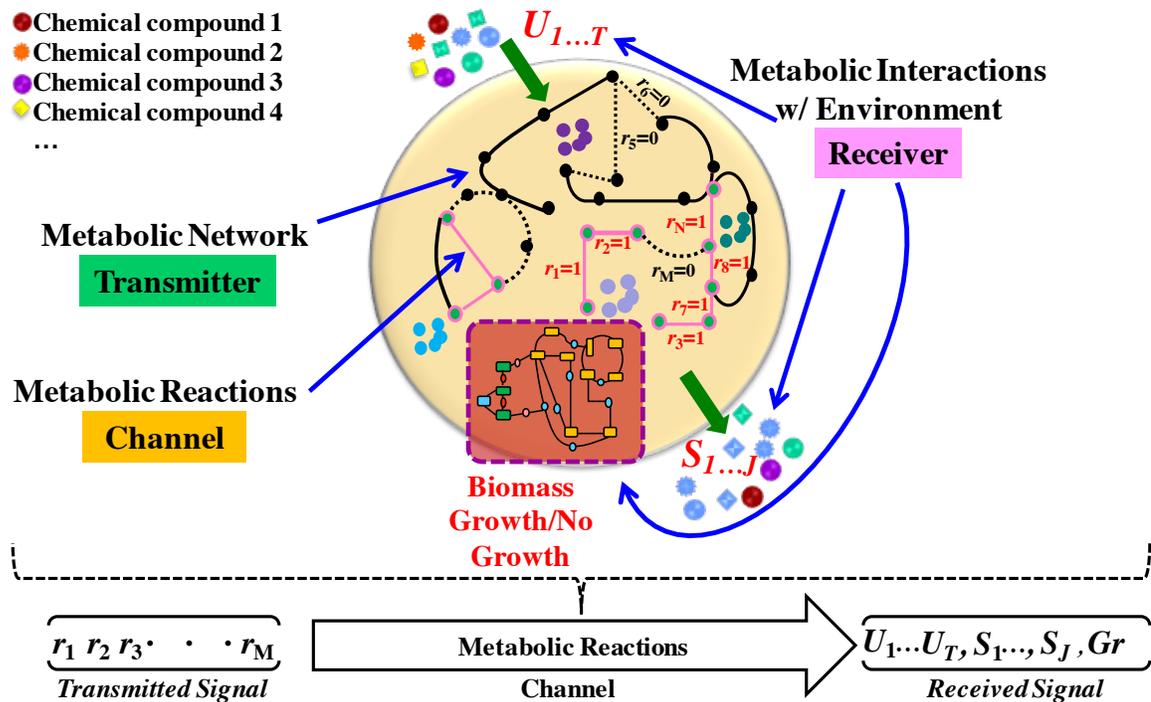


Figure 3.3: Sketch of the proposed Metabolic Reaction Abstraction.

In addition to the above abstraction, which lets us study the enzyme regulation to channel information from the cell's environment, we also need a model to under-

stand how much information of the internal cell metabolic state can be perceived from the outside environment. For this, we propose another abstraction model called digital modulator abstraction (Part b, Figure 3.1), also shown in Figure 3.3 which models the metabolic network inside the cell as a digital modulator of metabolite exchange/growth according to the information contained in the metabolic network state. In this abstraction, the transmitter represents the cell itself, where the transmitted signal is the metabolic state of the cell, which means the ON/OFF activity of the state-changing enzyme-regulated reactions. The channel represents mechanisms involved in the metabolic reactions as a function of the chemical compounds within the cell's input. The receiver represents i) the cell's environment, and ii) the biomass (as a form of cell growth) where the received signal is the variation in the uptake and biomass, respectively. This is more formally expressed as

$$\{r_1, r_2, \dots, r_M\} \xrightarrow[\text{Reactions}]{\text{Metabolic}} \{U_1, \dots, U_k, S_1, \dots, S_j, Gr\}, \quad (3.2)$$

The channel model of the above mentioned abstraction is shown in Equation (3.2) where  $r_i$  is a binary value equal to 1 if the enzyme-expression-regulated reaction  $i$  is ON, and equal to 0 if the same reaction is OFF,  $M$  is the number of enzyme-expression-regulated reactions.  $U_w$  is the flux, the velocity of molecule concentration propagating in space (*e.g.*, from environment to cell), of metabolites uptaken from the environment,  $S_x$  is the flux of metabolites excreted by the cell into its environment and *Growth* ( $Gr$ ) represents the flux of added components to the cell in the form of biomass.

We can conclude that the system shown in Figure 3.1 illustrates that the cell's behavior can be modeled as a transceiver. This is because the cell i) acts as a receiver when its internal metabolic state changes as it takes as input a signal in the form of

variations of the compounds in its environment, and ii) the cell acts as a transmitter when the state changes of its metabolic network reactions act as a signal, which is transmitted into the output changes in chemical compounds uptake and secretion from/into environment and in the biomass.

## 3.2 Steady-State Mutual Information

### 3.2.1 Stage I

We define the steady-state mutual information  $I$  of the molecular communication system abstraction (explained in Chapter 3) for the aforementioned Stage I as the amount of information about the chemical composition of the cell's environment measured in bits that a cell is able to represent in the binary state of its enzyme-expression-regulated metabolic reactions at steady state, after any evolution of the enzyme-expression regulation channel "Some of this material appears in [38]". According to information theory [9], the mutual information for Stage I can be defined as follows:

$$I(\{c_i\}_{i=1}^N; \{r_i\}_{i=1}^M) = H(\{c_i\}_{i=1}^N) - H(\{c_i\}_{i=1}^N | \{r_i\}_{i=1}^M), \quad (3.3)$$

where the input entropy  $H(\{c_i\}_{i=1}^N)$  can be defined as

$$H(\{c_i\}_{i=1}^N) = - \int P(\{c_i\}_{i=1}^N) \log_2 P(\{c_i\}_{i=1}^N) d\{c_i\}_{i=1}^N, \quad (3.4)$$

where the integration  $\int$  is performed throughout the possible values that the set of chemical compound concentrations  $\{c_i\}_{i=1}^N$  can assume. The conditional entropy of

the input given the output  $H(\{c_i\}_{i=1}^N | \{r_i\}_{i=1}^M)$  is then defined as follows:

$$H(\{c_i\}_{i=1}^N | \{r_i\}_{i=1}^M) = - \sum_{k=1}^K P \left( \left[ \{r_i\}_{i=1}^M \right]_k \right) \int P \left( \{c_i\}_{i=1}^N \mid \left[ \{r_i\}_{i=1}^M \right]_k \right) \log_2 P \left( \{c_i\}_{i=1}^N \mid \left[ \{r_i\}_{i=1}^M \right]_k \right) d\{c_i\}_{i=1}^N ; \quad (3.5)$$

respectively, where  $K$  is equal to the total number of different sets of binary values at the output of the system  $\left[ \{r_i\}_{i=1}^M \right]_k$  resulting from the all the possible values that the input chemical compound concentrations  $\{c_i\}_{i=1}^N$  can assume, and  $P(\cdot)$  is the probability distribution of the argument random variable/s. In the aforementioned definition of mutual information, we are ignoring possible memory in the system, *i.e.*, the values in  $\{r_i\}_{i=1}^M$  could depend on the past trajectory of the values of the input concentrations  $\{c_i\}_{i=1}^N$ . This might be the effect of hysteretic behaviors in the gene regulatory functions, which we currently ignore all the assumptions are in Section 2, with the justification that many of these mechanisms are even poorly understood in biology [10].

### 3.2.2 Stage II

We define the steady-state mutual information  $I$  for the aforementioned Stage II as the amount of information of the internal binary cell metabolic state that can be perceived from the outside environment through the metabolic-state-modulated values of the fluxes of uptaken and secreted metabolites, and the biomass (growth).

According to information theory [9], this can be defined as

$$I(\{r_i\}_{i=1}^M; \{U_t\}_{t=1}^T, \{S_j\}_{j=1}^J, Gr) = H(\{r_i\}_{i=1}^M) - H(\{r_i\}_{i=1}^M | \{U_t\}_{t=1}^T, \{S_j\}_{j=1}^J, Gr), \quad (3.6)$$

where the input entropy  $H(\{r_i\}_{i=1}^M)$  can be defined as

$$H(\{r_i\}_{i=1}^M) = - \sum_{k=1}^K P(\{r_{i_k}\}_{i=1}^M) \log_2 P(\{r_{i_k}\}_{i=1}^M), \quad (3.7)$$

where the summation  $\sum$  is performed throughout input of the FBA Groups based on similar binary enzyme-expression-regulated state changing  $\{r_i\}_{i=1}^M$  within cell metabolism. The conditional entropy of the input given the output  $H(\{r_i\}_{i=1}^M | \{U_t\}_{t=1}^T, \{S_j\}_{j=1}^J, Gr)$  is then defined as follows:

$$\begin{aligned} H(\{r_i\}_{i=1}^M | \{U_t\}_{t=1}^T, \{S_j\}_{j=1}^J, Gr) &= - \sum_{q=1}^Q P\left(\left[\{U_t\}_{t=1}^T, \{S_j\}_{j=1}^J, Gr\right]_q\right) \quad (3.8) \\ &\sum_{k=1}^K P\left(\{r_{i_k}\}_{i=1}^M \left| \left[\{U_t\}_{t=1}^T, \{S_j\}_{j=1}^J, Gr\right]_q\right.\right) \\ &\log_2 P\left(\{r_{i_k}\}_{i=1}^M \left| \left[\{U_t\}_{t=1}^T, \{S_j\}_{j=1}^J, Gr\right]_q\right.\right); \end{aligned}$$

respectively, where  $Q$  is equal to the total number of different sets of flux values at the output of the system  $\{U_i\}_{i=1}^k, \{S_i\}_{i=1}^j, Growth (Gr)$  resulting from the enzyme-expression-regulated reactions  $\{c_i\}_{i=1}^N$  within cell metabolism, and  $P(\cdot)$  is the probability distribution of the argument random variable/s.

## Chapter 4

# *In silico* Characterization of the Mutual Information of Cell Metabolism through Flux Balance Analysis

### 4.1 Estimation of Optimal Enzyme Expression Regulation through Flux Balance Analysis

Flux Balance Analysis (FBA) is a widely used and computationally efficient mathematical method to estimate the chemical reactions that might be active in the metabolic network, a cell metabolic network, given determinate environmental conditions [34], [5]. FBA uses a set of linear equations and applies optimization techniques to estimate the metabolic network state that results in the maximum growth rate under given determinate conditions. As a consequence, FBA can be utilized for es-

timating differences in optimal metabolic network states of a cell corresponding to different environmental conditions. As detailed in the following, these estimates are the basis for our *in silico* mutual information “Some of this material appears in [38]”.

Through FBA we are able to obtain an estimate of the state  $\{r_i^*\}_{i=1}^M$  of the aforementioned enzyme-expression-regulated chemical reactions that results into an overall maximum biomass production. The FBA-estimated chemical reaction states  $\{r_i^*\}_{i=1}^M$  are those that maximize the growth of the cell given a chemical composition of the surrounding environment  $\{c_i\}_{i=1}^N$ , and represent the best regulation of these chemical reactions that the cell might ever achieve. The aforementioned mechanisms of activation or repression that might be in place for the regulation of enzyme expression have been most probably acquired through evolution. Although they tend to reach this optimal solution, they might just realize a subset of the needed reaction state adaptations [11]. The estimation of the optimal enzyme expression regulation through FBA can be formalized as follows:

$$\{c_1, c_2, \dots, c_N\} \xrightarrow[\text{Analysis}]{\text{Flux Balance}} \{r_1^*, r_2^*, \dots, r_M^*\}, \quad (4.1)$$

The FBA computation stems from the construction of a GEnome-scale Model (GEM) for the organism under analysis. The construction of a GEM greatly benefits from the availability of cell genome information and the latest advances in bioinformatics techniques [5]. In particular, this is realized by searching for known genes that encode metabolic enzymes, defined in Section 2.2, which are possibly activating metabolic chemical reactions. These reactions are described in extensively curated online catalogs. Subsequently, further chemical reactions are included in the GEM through comparisons with the genomes and the corresponding known metabolic pathways of other similar organisms that have been already extensively studied and an-

notated. Subsequently, further chemical reactions are included in the GEM through comparisons with the genomes and the corresponding known metabolic pathways of other similar organisms that have been already extensively studied and annotated.

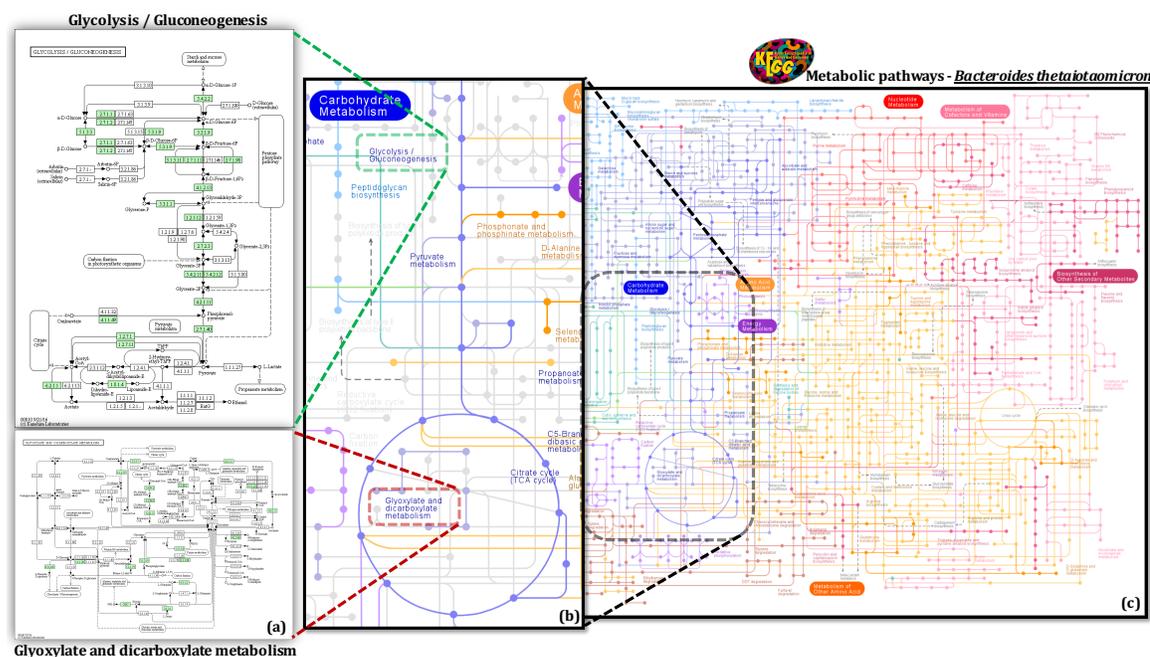


Figure 4.1: KEGG module of *Bacteroides thetaiotaomicron* metabolic pathways. (a) Summary of the biological processes shown in the pathway map of Glycolysis / Gluconeogenesis and Glyoxylate and dicarboxylate metabolism (b) Enlarged fine details of a section of a complete metabolic model, (c) Part of the complete KEGG database pathway maps of *Bacteroides thetaiotaomicron*.

Figure 4.1 (c) shows a graphical representation of parts of a GEM (on the right) for the organisms *Bacteroides thetaiotaomicron*, also referred to as *B. theta*, used in our study, which is obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [26]. The nodes represent compounds that are inputs/outputs to the reactions, and edges represent the chemical reactions. Inputs from the environment are taken by the organism are involved in and are involved in the reactions of metabolic pathways, resulting in the exchange of fluxes with the environment (uptake and secretion), or in the production of biomass (growth) to in the exchange of fluxes

with the environment (uptake and secretion), or in the production of biomass.

The set of possible metabolic reactions inside a GEM can be expressed through a stoichiometric matrix  $S$ , where each row represents a chemical compound and each column represents a possibly active metabolic reaction in cell metabolism. Figure 4.2 shows an example of how a stoichiometric matrix  $S$  can be obtained from a metabolic network. Figure 4.2. (a) shows a conceptual model of a metabolic network where the internal chemical reactions are represented by  $R_i$ , and the exchange fluxes with the environment are represented by  $E_j$ , as shown in Figure 4.2 (b). The stoichiometric matrix  $S$  in Figure 4.2. (c) is organized in a way that each row corresponds to a chemical compound in the metabolic network, while each column corresponds the reaction or flux exchanged with the environment. Each entry of the stoichiometric matrix  $S$  is the stoichiometric coefficient that indicates how many molecules of a chemical compound, represented by row entry, are consumed (coefficient  $< 0$ ) or produced (coefficient  $> 0$ ) in one of the possible reactions,

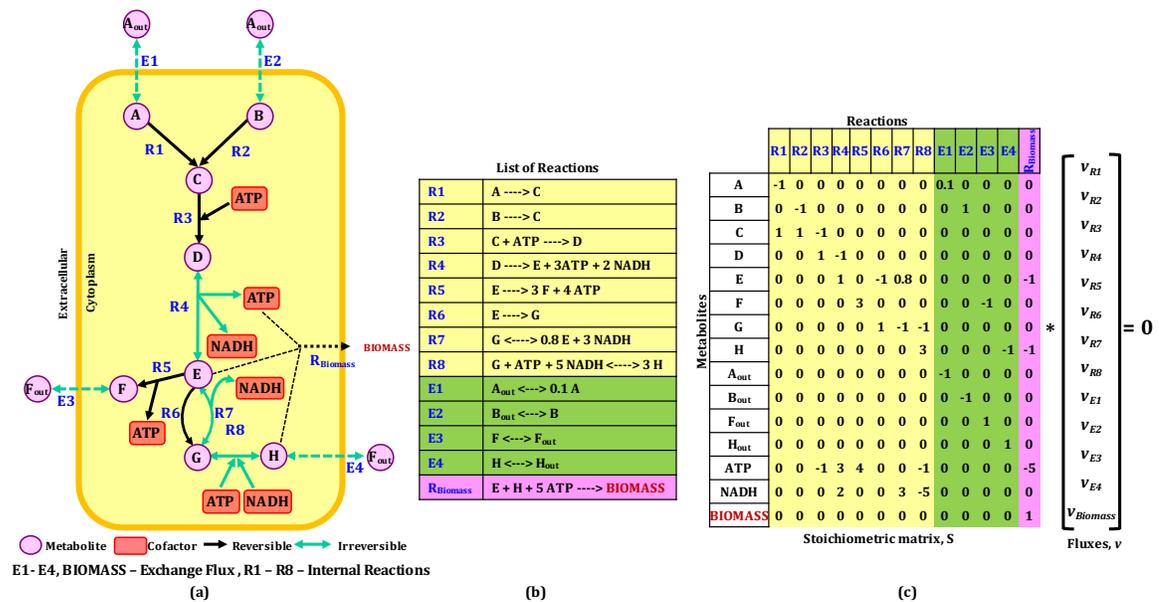


Figure 4.2: From metabolic network to stoichiometric matrix. (a) Conceptual model of a metabolic network. (b) List of the reactions participating in the metabolic network. (c) Corresponding stoichiometric matrix.

The FBA solution is expressed in terms of  $\mathbf{v}^*$  which is a column vector that contains the optimal flux of each reaction, defined as the number of molecules per unit volume and unit time that are consumed/produced by that reaction. It is obtained through a Linear Program (LP), formalized as follows [34]:

$$\begin{aligned} & \text{maximize} && a' \mathbf{v} \\ & \text{subject to} && S \mathbf{v} = 0 \\ & && \mathbf{v}_{\min} \leq \mathbf{v} \leq \mathbf{v}_{\max}, \end{aligned}$$

where  $a$  is a column vector that contains the weight coefficients of the fluxes that the FBA optimizes. In our case, the entries of  $a$  are equal to 1 only at the indexes corresponding to the chemical compounds that are considered part of the aforementioned biomass, produced by the cell and responsible for cell growth, while the other entries are equal to 0. The column vectors  $\mathbf{v}_{\min}$  and  $\mathbf{v}_{\max}$  constrain the minimum and maximum flux, respectively, of each corresponding reaction considered in the FBA, and define the space where the LP searches for the optimal solution. This constraint based modeling can be explained by the Figure 4.3 [34]:

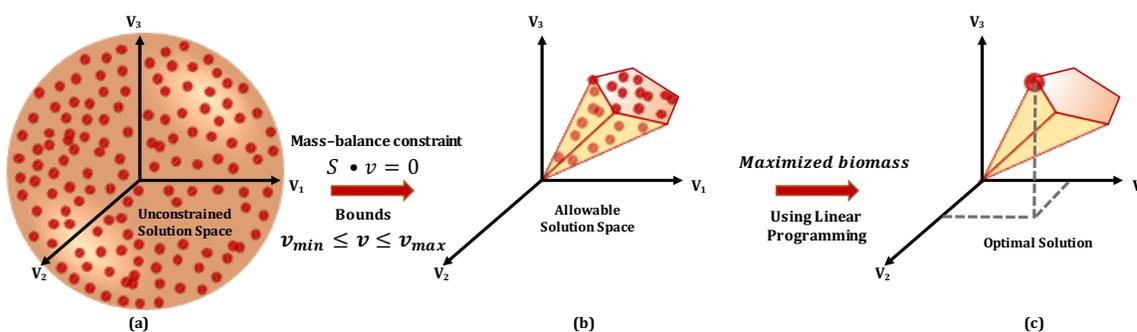


Figure 4.3: Conceptual model of the FBA LP formulation for finding the optimal solution.

When growth/flux constraints are applied by the lower and upper bounds  $\mathbf{v}_{\min}$  and  $\mathbf{v}_{\max}$ , in addition to the constraints from the aforementioned stoichiometric matrix  $S$ , the entire solution space (shown in (a)) of the metabolic network can be reduced to a smaller allowable solution space, shown in (b). Then, optimization is applied on the objective function  $a'\mathbf{v}$  to identify a single optimal solution, which is a flux distribution lying on the edge of the allowable solution space as shown in (c).

The values of  $\mathbf{v}_{\min}$  and  $\mathbf{v}_{\max}$  are set to reasonable biological limiting values [34], with the exception of the reaction corresponding to the uptake of the input chemical compounds present in the surrounding environment  $\{c_i\}_{i=1}^N$  for which we are estimating the chemical reaction states  $\{r_i^*\}_{i=1}^M$ . This is expressed as follows:

$$v_{min,i} = v_{max,i} = J_i \left( \{c_i\}_{i=1}^N \right), \quad i = 1, \dots, N \quad (4.2)$$

where  $J_i$  is in general a function of all the input concentrations  $\{c_i\}_{i=1}^N$  that returns the flux of input chemical compound  $i$ , and depends on the particular method employed by cells to uptake this chemical compound (e.g. facilitated diffusion or active transport through the cell membrane as shown in Figure 2.1). The expression of  $J_i(\cdot)$  is in general known from biochemistry literature. As an example, the expressions for the input glucose  $J_{gl}$  and lactose  $J_{lac}$  fluxes considered in part of the numerical examples of this thesis are as follows [40]:

$$\begin{aligned} J_{gl}(c_{gl}) &= \frac{J_{gl}^{max} c_{gl}}{\Phi_{gl} + c_{gl}}, \\ J_{lac}(c_{gl}, c_{lac}) &= \frac{J_{lac}^{max} c_{lac}}{k_{lac} + c_{lac}} \left( 1 - \phi_{gl} \frac{J_{gl}^{max} c_{gl}}{k_{gl} + c_{gl}} \right), \end{aligned} \quad (4.3)$$

where the parameter values can be found in Table 3 of [40].

The estimates of the chemical reaction states  $\{r_i^*\}_{i=1}^M$  are finally computed from the

optimal flux vector  $\mathbf{v}^*$ , which is obtained by the FBA given the chemical composition of the surrounding environment  $\{c_i\}_{i=1}^N$ , as follows:

$$r_i^* = \begin{cases} 0, & \text{if } v_i^* = 0 \\ 1, & \text{otherwise} \end{cases}, \quad (4.4)$$

## 4.2 An Upper Bound Steady-State Mutual Information of Cell Metabolism

### 4.2.1 Upper Bound for Stage I

Given the optimal estimates of the chemical reaction states  $\{r_i^*\}_{i=1}^M$  obtained through the FBA from the knowledge of the cell's genome for all the values that our input set of chemical compound concentrations  $\{c_i\}_{i=1}^N$  can assume, we can compute the following steady-state mutual information “Some of this material appears in [38]”:

$$I(\{c_i\}_{i=1}^N; \{r_i^*\}_{i=1}^M) = H(\{c_i\}_{i=1}^N) - H(\{c_i\}_{i=1}^N | \{r_i^*\}_{i=1}^M), \quad (4.5)$$

where  $H(\{c_i\}_{i=1}^N | \{r_i^*\}_{i=1}^M)$  is computed through Equation (3.5) by substituting the chemical reaction states  $\{r_i\}_{i=1}^M$  resulting from the real regulation of the enzyme expression with the FBA-estimated chemical reaction states  $\{r_i^*\}_{i=1}^M$ .

In this thesis, we consider the mutual information in Equation (4.5) computed with the results of the FBA as an upper bound to the real steady-state mutual information in Equation (3.3) that we would obtain in reality as a result of the enzyme expression regulation. This is formalized as follows:

$$I(\{c_i\}_{i=1}^N; \{r_i^*\}_{i=1}^M) \geq I(\{c_i\}_{i=1}^N; \{r_i\}_{i=1}^M). \quad (4.6)$$

The expression in Equation (4.6) can be proven through the Data Processing Inequality from information theory [9], which states that the aforementioned inequality holds true if the steady-state chemical reaction states  $\{\hat{r}_i\}_{i=1}^M$  given a set  $\{\hat{c}_i\}_{i=1}^N$  of values for the input concentrations can be probabilistically determined from the sole knowledge of the chemical reaction states  $\{\hat{r}_i^*\}_{i=1}^M$ , without the need of having knowledge of the input concentrations. This is expressed as follows [9]:

$$P\left(\{r_i\}_{i=1}^M \mid \{c_i\}_{i=1}^N, \{r_i^*\}_{i=1}^M\right) = P\left(\{r_i\}_{i=1}^M \mid \{r_i^*\}_{i=1}^M\right). \quad (4.7)$$

Equation (4.7) can be explained by considering that the chemical reaction states  $\{\hat{r}_i^*\}_{i=1}^M$  are those that underlie the optimally regulated cell metabolism that maximizes the cell growth rate (or biomass production) given a set of values for the input concentrations  $\{\hat{c}_i\}_{i=1}^N$ . In reality, when subject to the same input concentrations  $\{\hat{c}_i\}_{i=1}^N$ , a cell reaches the steady-state chemical reaction states  $\{\hat{r}_i\}_{i=1}^M$ , which might be in general different from the aforementioned optimal states. If these states are indeed not optimal, the cell will not grow (produce biomass) and reproduce at the maximum rate possible given the input concentrations  $\{\hat{c}_i\}_{i=1}^N$ . When considering multiple cells in a population subject to the same input concentrations, if different cells show different steady-state chemical reaction states (because of cell-cell variability), those that have states closer to the optimal states will grow faster, and ultimately outnumber other cells. As a consequence, cells have evolved gene expression regulation mechanisms, such as those described in Section. 2.3, through which they adapt their steady-state chemical reaction states as close as possible to optimality given a set of input concentrations [10]. Given the optimal chemical reaction states  $\{\hat{r}_i^*\}_{i=1}^M$  and the knowledge of the gene expression regulation mechanisms in place in a particular cell species, we are theoretically able to estimate the probability distribution of the

steady-state chemical reaction states  $\{\hat{r}_i\}_{i=1}^M$ , which are those that best approximate the optimal states. As a consequence, under the aforementioned assumptions, the conditional probability of the steady-state chemical reaction states  $\{\hat{r}_i\}_{i=1}^M$  given the input concentrations  $\{\hat{c}_i\}_{i=1}^N$  and chemical reaction states  $\{\hat{r}_i^*\}_{i=1}^M$  is equal to the same probability but only conditioned to the chemical reaction states  $\{\hat{r}_i^*\}_{i=1}^M$ , as expressed in Equation (4.7).

## 4.2.2 Upper Bound for Stage II

In order to compute the amount of information of the internal optimal binary metabolic state that can be perceived from the external environment (uptake/secretion fluxes, biomass) of the cell, the calculation of the steady-state mutual information for Stage II, defined in Section 3.2.2, is performed by considering the optimal chemical reaction states  $\{r_i^*\}_{i=1}^M$  as input, and the optimal values of the fluxes of uptaken  $\{U_t^*\}_{t=1}^T$  and secreted  $\{S_j^*\}_{j=1}^J$  metabolites, and biomass  $Gr^*$ . The latter values are obtained from the FBA solution as the values in the aforementioned optimal flux vector  $\mathbf{v}^*$  corresponding to uptake, secretion, or biomass reactions, respectively. This mutual information is expressed as

$$I(\{r_i^*\}_{i=1}^M; \{U_t^*\}_{t=1}^T, \{S_j^*\}_{j=1}^J, Gr^*) = \tag{4.8}$$

$$H(\{r_i^*\}_{i=1}^M) - H(\{r_i^*\}_{i=1}^M | \{U_t^*\}_{t=1}^T, \{S_j^*\}_{j=1}^J, Gr^*),$$

where  $H(\{r_i^*\}_{i=1}^M)$  and  $H(\{r_i^*\}_{i=1}^M | \{U_t^*\}_{t=1}^T, \{S_j^*\}_{j=1}^J, Gr^*)$  are computed through Equations (3.7) and (3.8) by substituting the chemical reaction states  $\{r_i^*\}_{i=1}^M$  and the fluxes  $\{U_t^*\}_{t=1}^T$ ,  $\{S_j^*\}_{j=1}^J$ , and biomass  $Gr^*$  in place of those resulting from the real regulation of cell metabolism. In general, the values computed through the expression in Equation (4.5) might differ from those from Equation 3.6. Nevertheless, since Stage

II is in series to Stage I, and the optimal flux vector  $\mathbf{v}^*$  can be deterministically computed from the optimal chemical reaction states  $\{r_i^*\}_{i=1}^M$  by LP optimization, we only need to consider Equation (4.8) to obtain an upper bound to the mutual information of the overall system.

A detailed description of the computation of Equations (4.5) and (4.8) for the organisms considered in this work are provided in Chapter 6.

# Chapter 5

## Data Generation Workflow

### 5.1 Data generation

For *in silico*, we generated simulation data using KBase [25] which is an open source software and data platform that allows users to upload their data, collaborate by sharing with other users, perform automated analysis, and publish their results and conclusions. The simulation being an execution of a static model, allows us to collect the chemical reactions that occur in the metabolic model, under selected configurations.

We follow a similar data generation process for the three organisms *E. coli*, *B. theta* and *M. smithii*. As shown in Figure 5.1, first we build an initial metabolic model for these organisms. To achieve this, we run the Build Metabolic Model app provided by KBase, which uses a protein phylogeny database to translate both the organism's genome into protein sequences. This initial metabolic model is a genome-scale metabolic network of biochemical reactions reconstructed from functional protein annotations derived from biochemistry literature and similarity with other known protein sequences. These protein functions are then mapped onto biochemical reactions

in the KEGG model [26]. This initial model is a draft model and only a starting point because this metabolic model might be missing some crucial reactions and might have incorrect protein functional annotations in its network that are required to generate biomass. A total of 121 initial draft models were created for *E. coli*, one for each of the 121 configurations. For *B. theta* and *M. smithii*, a total of 128 initial draft models were created, one for each of the 128 configurations. The details on how these configurations are created will be discussed in detail in the Sections 5.2 and 5.3.

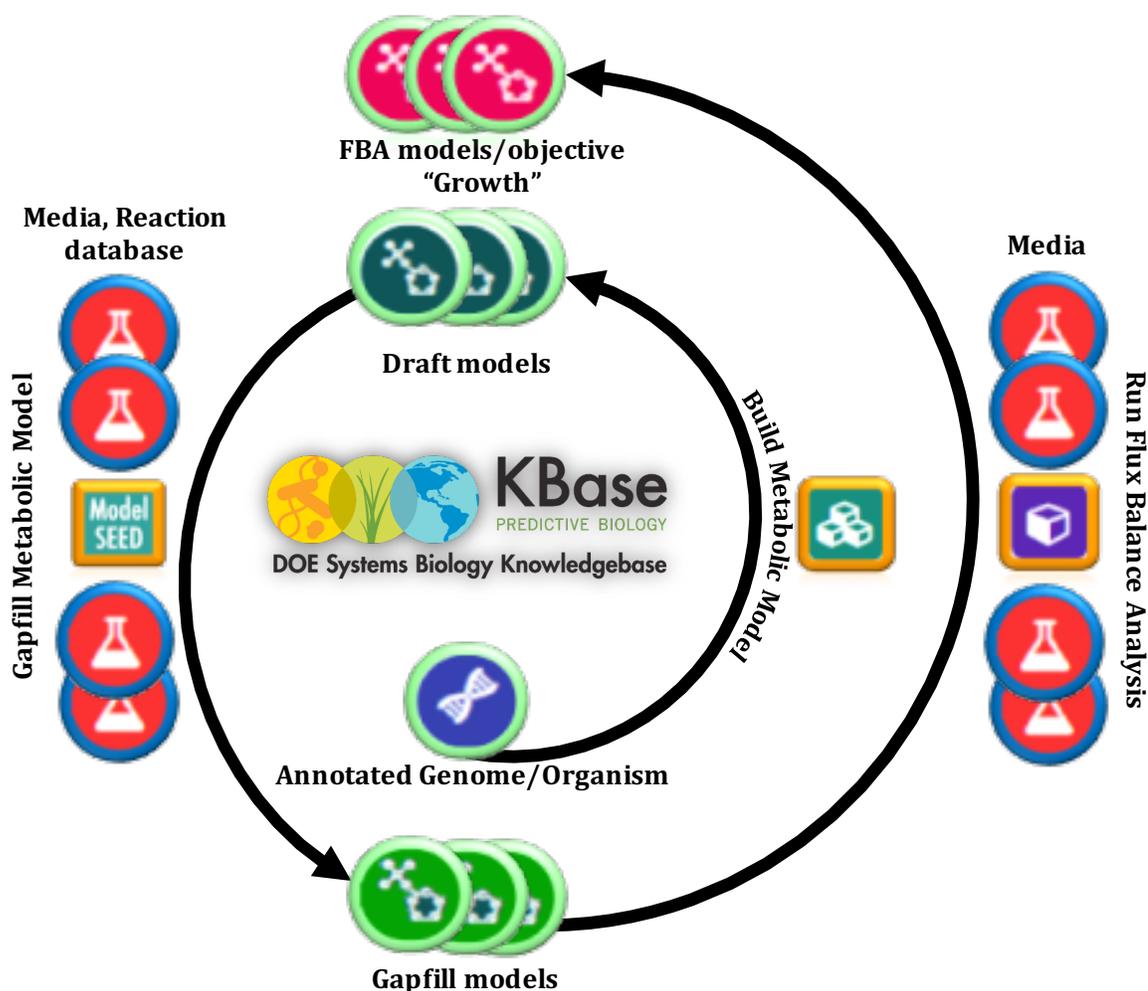


Figure 5.1: KBBase Workflow for simulation data.

The next step is the gap filling process [20] where a minimal number of bio-

chemical reactions and compounds are added to the initial metabolic model with the goal of making the metabolic network able to synthesize the biomass in a specified medium. For each of the configurations, gap filled models were produced from their corresponding initial draft models.

Finally the growth rate of an organism (biomass production rate) in a specified medium can be obtained by running the FBA. FBA simulates how metabolites flow through the metabolic network of an organism in a specified medium and uses constraint-based approach discussed in Section 4.1 a to estimate the steady-state biomass production rate. The biomass information resulting from the FBA is used to quantify the mutual information parameter which will be discussed in the following section.

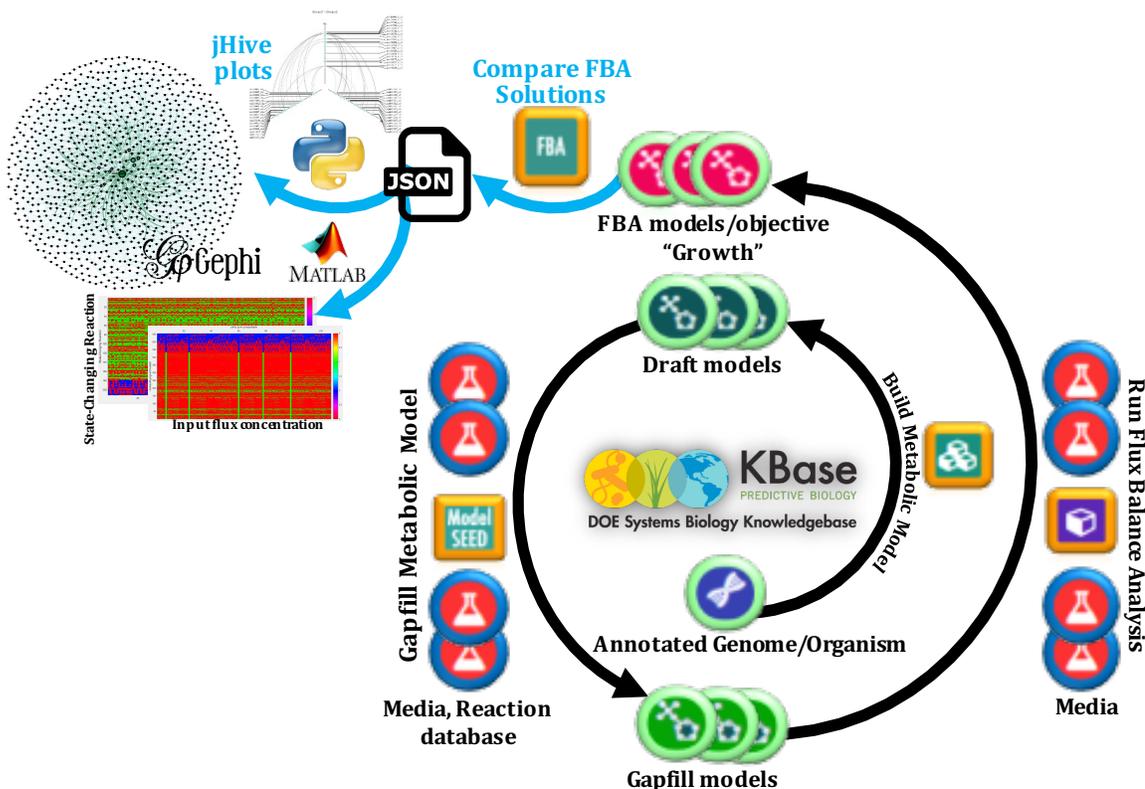


Figure 5.2: KBBase workflow to generate silico data.

The flow for the process involved in generating the Mutual Information for Stage I

and II can be summarized by the figure above. The flow is very similar to the process explained in Figure 5.1, except that it involves three additional steps at the end. We now describe these last three steps.

The output of the FBA will be 128 files, for each of the organisms, *B. theta* and *M. smithii*. For *E. coli*, the output of the FBA will be 121 files. These are FBA objects that display the growth of the model, reaction fluxes, biomass compounds and coefficient values, gene IDs, compound fluxes and gene knockout data in a table-based format. We take these FBA objects and send them to the Compare FBA Solutions app in KBase to compare flux profiles predicted by the FBA. This app compares the objective value, flux through each reaction in FBA, uptake, and secretion of metabolites in each of the FBA outputs. The reaction fluxes inside the FBA outputs are categorized into “not in model,” “no flux,” “forward flux,” and “reverse flux” whereas metabolite fluxes are categorized into “not in model,” “no flux,” “uptake,” and “secretion.”

The results pertaining to reaction fluxes and metabolite fluxes were analyzed using a script written in MATLAB and python respectively, to generate what we call a binary state changing matrix. This matrix shows whether the state changing reactions are ON or OFF for each of the FBAs. Additionally, we also use our binary state changing matrices to visualize the relation between compounds and reactions, the details and implications of which will be discussed in the following section.

## 5.2 *In silico* experimentation for *E. coli*

For our study, we initially focused on the *E. coli* bacterium strain K-12 MG1655, which is considered one of the golden standard model systems in synthetic biology labs, and whose genome is completely known [19] “Some of this material appears

in [38]”. By stemming from this genome, we built the corresponding GEM, and subsequently performed the FBA by using the KBase (Department of Energy Systems Biology Knowledgebase) software application suite [25]. To obtain our numerical results, we developed a model of the external environment containing the known minimal set of chemical compounds necessary for this *E. coli* strain to grow, *i.e.*, produce biomass, and at the same time allowing the variation of the concentrations of key compounds that result in changes in the optimal FBA-computed states of metabolic reactions. In particular, we obtained our numerical results by performing FBA on every combination of input fluxes of two input compounds D-Glucose and Lactose ranging from 0 to 100 [mmol/g CDW/hr] with increments of 10 [mmol/g CDW/hr], for a total of 121 different combinations of input fluxes.

### 5.3 *In silico* experimentation for *B. theta* and *M. smithii*

We also studied two other organisms, namely, the *B. theta* and the *M. smithii*, both found in the human gut, with the goal of understanding their individual metabolic networks and their interactions with each other and the human body [8], [37].

A set of nutritional compounds such as food, nutrients, toxins which are required as inputs and products of each organism’s metabolism were identified as Independent Variables. Seven such compounds were selected to obtain ground truth by running the entire configuration space in the laboratory. These seven compounds which are hypothesized to impact whether or not the organisms will grow are Glucose, Hematin, Formale, H<sub>2</sub>, Vitamin *B*<sub>12</sub>, Acetate and Vitamin K. These compounds can either be present in the solution (ON) or not (OFF). These seven compounds along with the

base compounds (which should be present always / minimal media) form the input media for the experimentation. Each of these seven compounds may or may not be present in the media, giving us a total of  $2^7 = 128$  media [37].

The growth of the organisms was studied as the Dependent Variable of the experimentation, which is measured as the biomass dry weight in grams of cells *in vitro* and as the sum of flux through the biomass reactions using an FBA analysis [20].

# Chapter 6

## Numerical Results

In this section, we present a proof-of-concept numerical example of both abstractions (Stage I and Stage II) and their analysis method proposed in this thesis. In particular, we report the results of different combinations of values of the input fluxes of D-Glucose and Lactose in terms of growth rate of the *E. coli* bacterium. For this, we based our environment on the K-12 MG1655 minimal media [13], enriched with metal tracers common to other two standard media, namely, the Lysogeny Broth (LB) and the Carbon-D-Glucose media [6]. All compound fluxes contained in  $\mathbf{v}_{\min}$  and  $\mathbf{v}_{\max}$  were set -100 and 100 [mmol per gram cell dry weight per hour] ([mmol/g CDW/hr]), respectively. On top of the defined media environment, we introduced two other input compounds, namely, D-Glucose and Lactose, for which we simulated a variation in their concentration, and consequent corresponding values in  $\mathbf{v}_{\min}$  and  $\mathbf{v}_{\max}$  according to Equation (4.2) and Equation (4.3). Also, we compute the upper bound steady-state mutual information for Stage I and Stage II for the same organism.

Moreover, we extend our analysis on the two members of human gut microbiota, the *B. theta* bacterium, and *M. smithii*, which is a methanogenic archaeon. The genome of both these organisms are completely known [41], [47]. We first built the

GEM model, and performed the FBA analysis using the KBase [25]. For modeling the external environment, we have to consider the known minimal set of compounds necessary for both organisms to grow (i.e. produce biomass) while also allowing the variation of the concentrations of key compounds that result in changes in the optimal FBA-computed states of metabolic reactions. For this, we based our environment on the minimal media [24], enriched with seven compounds, namely Carbon-D-Glucose (G), Hematin (He), Formate (F),  $H_2$ , *Vitamin B<sub>12</sub>* ( $B_{12}$ ), Acetate (A), and Vitamin K ( $V_k$ ), to obtain the ground truth for our study. Each of these seven compounds may or may not be present in the media, giving us a total of 128 media. Along with these important compounds, there are other compounds that are common in all media, namely, Calcium, Chloride ion, Carbon dioxide, Cobalt, Copper, Ferrous ion, Water, H+, Potassium, L-Cysteine, Magnesium, Manganese, Sodium, Ammonium, Nickel, Phosphate, Sodium bicarbonate, Sulfate, Zinc, and Ferric ion.

All compound fluxes contained in  $\mathbf{v}_{\min}$  were set to -100 [mmol per gram cell dry weight per hour] ([mmol/g CDW/hr]), and  $\mathbf{v}_{\max}$  values varied between a low of 0.0000037 and max of 46166.8875 [mmol per gram cell dry weight per hour] ([mmol/g CDW/hr]). We generated the numerical results for both abstractions by performing FBA on every combination of input fluxes for all seven compounds mentioned earlier for a total of 128 different combinations.

## 6.1 Stage I

### 6.1.1 Numerical Results for *E. coli*

In Figure 6.1 we show the results of the FBA in terms of growth rate, or equivalently, output flux of produced biomass, as defined in Section 4.1, computed for different

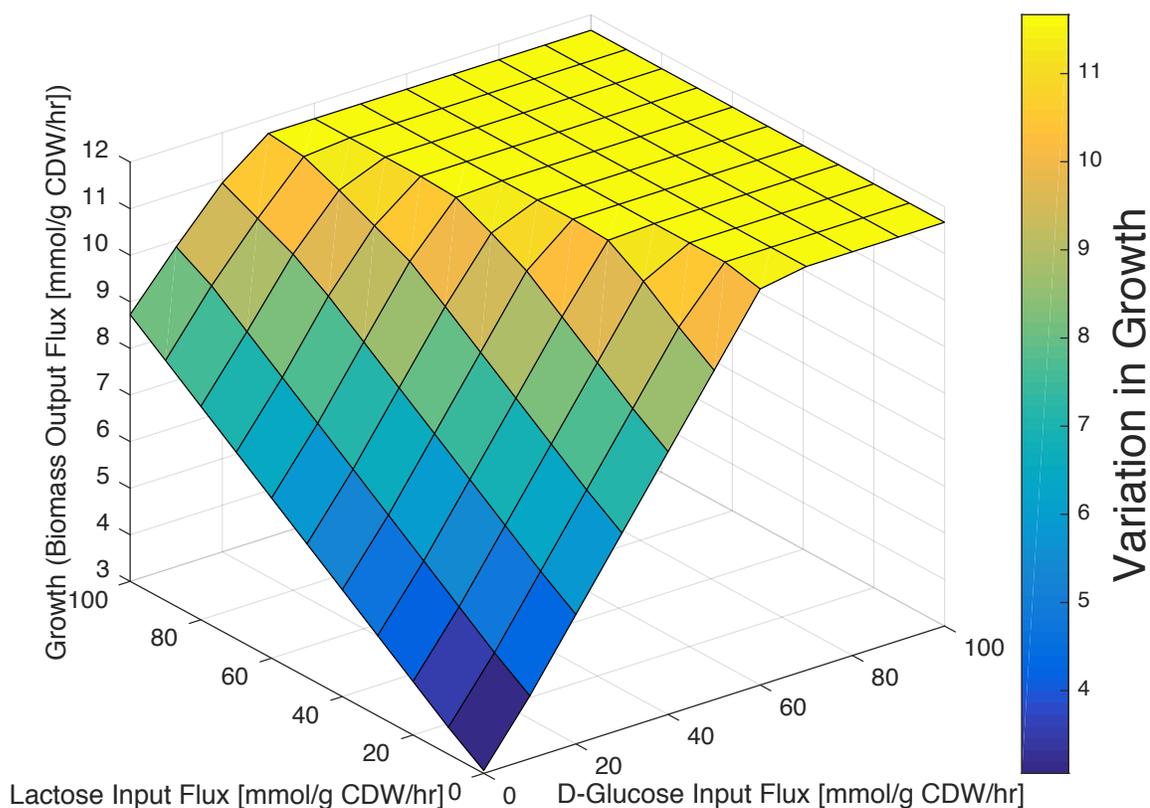


Figure 6.1: Optimal E.Coli K12 MG1655 growth as a function of the input flux of D-Glucose and Lactose in the environment.

combinations of values of the input fluxes of D-Glucose and Lactose. In these results, the variation in optimal growth rate, which is dependent on the optimal metabolic reaction states, as discussed below, varies from a minimum value of 3.061 [mmol/g CDW/hr] when the fluxes of D-Glucose and Lactose are absent from the environment, to a maximum value of 11.67 [mmol/g CDW/hr]. These curves show also a saturation in the optimal growth rate for D-Glucose fluxes on the higher end of the range, and the minimal value of D-Glucose flux to obtain this saturation varies as function of the Lactose flux, from a minimum of 30 [mmol/g CDW/hr] to a maximum of 70 [mmol/g CDW/hr].

In Figure 6.2, for each of the 121 tested combinations of the input fluxes of D-Glucose and Lactose, one for each column of the matrix, we show the binary values

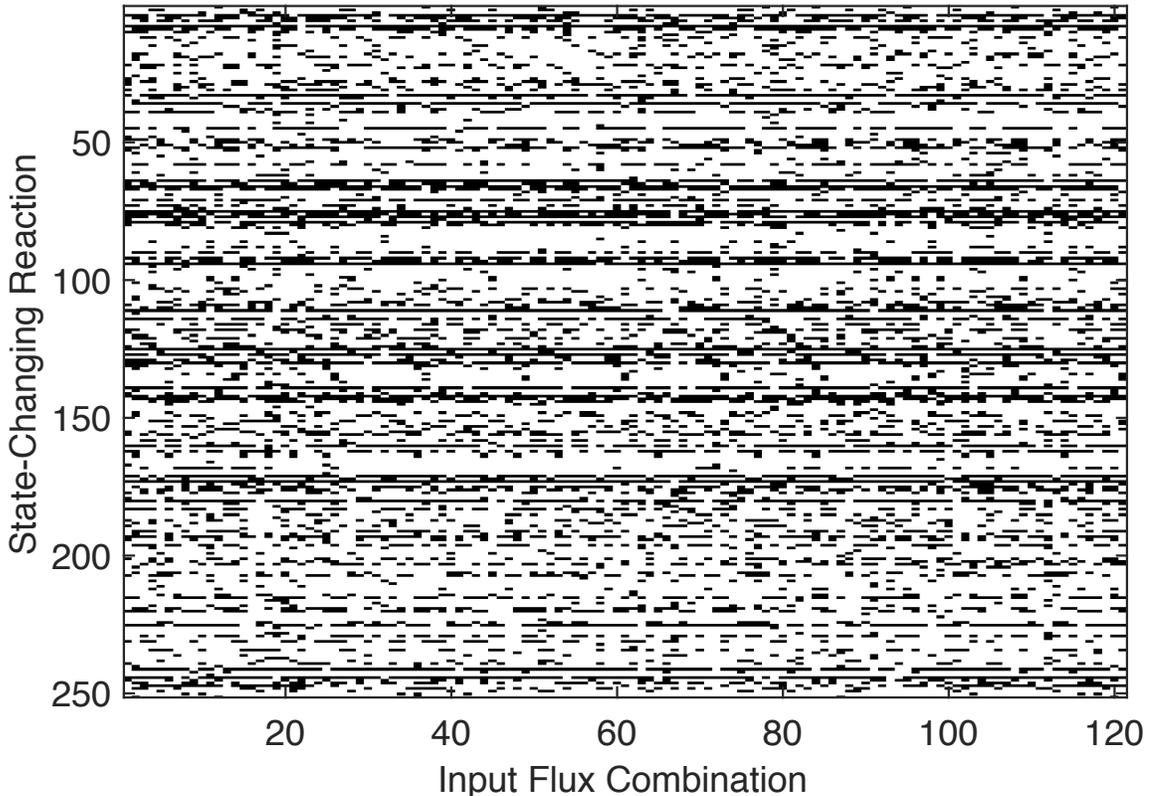


Figure 6.2: FBA-estimated binary chemical reaction states  $\{r_i^*\}_{i=1}^M$  for each combination of *D-Glucose* and *Lactose* input fluxes, where white = ON state; black = OFF state.

of FBA-estimated chemical reaction states  $\{r_i^*\}_{i=1}^M$  as defined in Section 4.1, one for each column, where the number of metabolic reactions  $M$  that show a state change within the considered combination of input fluxes of D-Glucose and Lactose is equal to 251.

The computation of the upper bound of the steady-state mutual information is finally realized by applying the expressions in Equations (4.5), (3.4), and (3.5), taking into account that the possible combinations of input fluxes of D-Glucose and Lactose are drawn from a discrete set. For these preliminary results, we make the assumption that these combinations are equiprobable. As a consequence, the corresponding combinations  $\{c_i\}_{i=1}^N = \{c_{gl}, c_{lac}\}$  computed through Equation (4.3) can be

as well considered equiprobable with probability density  $P(\{c_{gl}, c_{lac}\}) = 1/(\text{\#of input combinations}) = 1/121$ . The resulting input entropy  $H(\{c_i\}_{i=1}^N)$ , where  $N = 2$ , is then computed through Equation (3.4) by substituting the integral with a summation over the number of input combinations, which results into  $\log_2(121) = 6.92$  bits. To compute the conditional entropy of the input given the output  $H(\{c_i\}_{i=1}^N | \{r_i\}_{i=1}^M)$ , where  $N = 2$  and  $M = 251$ , we translated Equation (3.5) into the following formula:

$$H(\{c_{gl}, c_{lac}\} | \{r_i^*\}_{i=1}^M) = - \sum_{y=1}^Y P_Y \sum_{x=1}^{X_y} P(x|y) \log_2 P(x|y), \quad (6.1)$$

where  $Y$  and  $P_Y$  correspond to the number of times and the probability, respectively, that a reaction state combination is found more than once in the data shown in Figure 6.2,  $X_y$  is the number of times the reaction state combination  $y$  is found in the data, and  $P(x|y)$  is the probability of having a combination of input concentrations  $x = \{c_{gl,x}, c_{lac,x}\}$  given the reaction state  $y$ , which we consider as a uniform distribution in the number of different combinations of input concentrations that have been found resulting into the same the reaction state  $y$ . In our data we found a total of 4 different reaction combinations that are repeated twice, which results in a conditional entropy of the input given the output  $H(\{c_{gl}, c_{lac}\} | \{r_i^*\}_{i=1}^M) = 0.53$  bits. Finally, the following value is found for the upper bound of the steady-state mutual information:

$$I(\{c_{gl}, c_{lac}\}; \{r_i^*\}_{i=1}^{115}) = 6.39 \text{ bits} \quad (6.2)$$

### 6.1.2 Numerical Results for *B. theta* and *M. smithii*

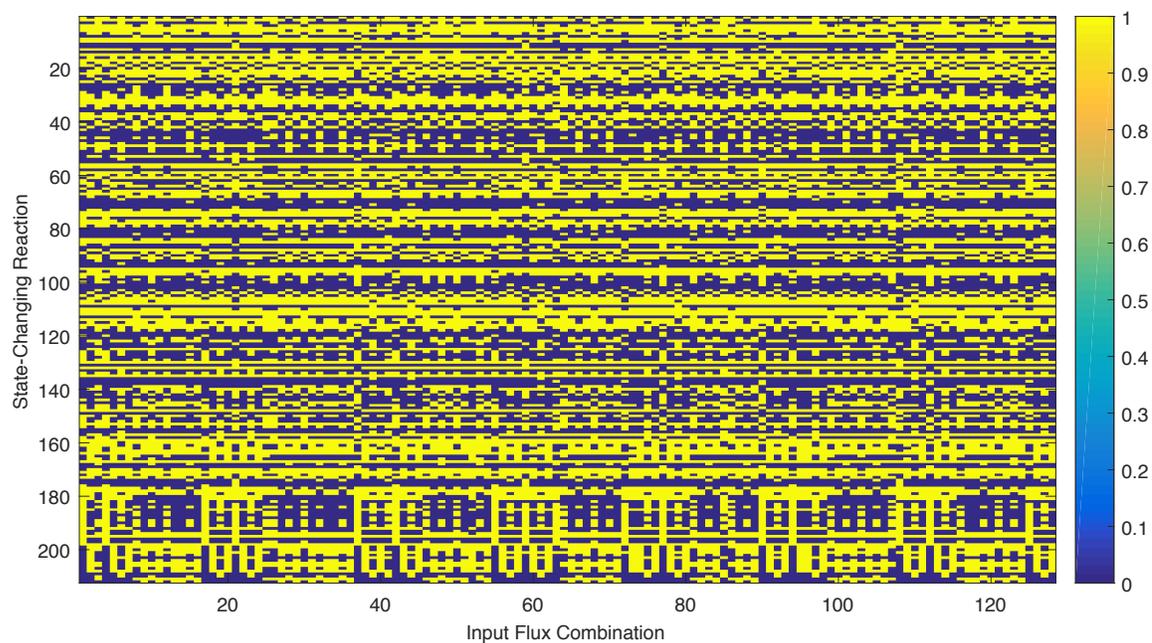


Figure 6.3: FBA-estimated binary chemical reaction states  $\{r_i^*\}_{i=1}^M$  for each combination of *Glucose*, *Hematin*, *Formate*,  $H_2$ , *VitaminB<sub>12</sub>*, *Acetate*, and *Vitamin K* input fluxes, where yellow = ON state; violet = OFF state.

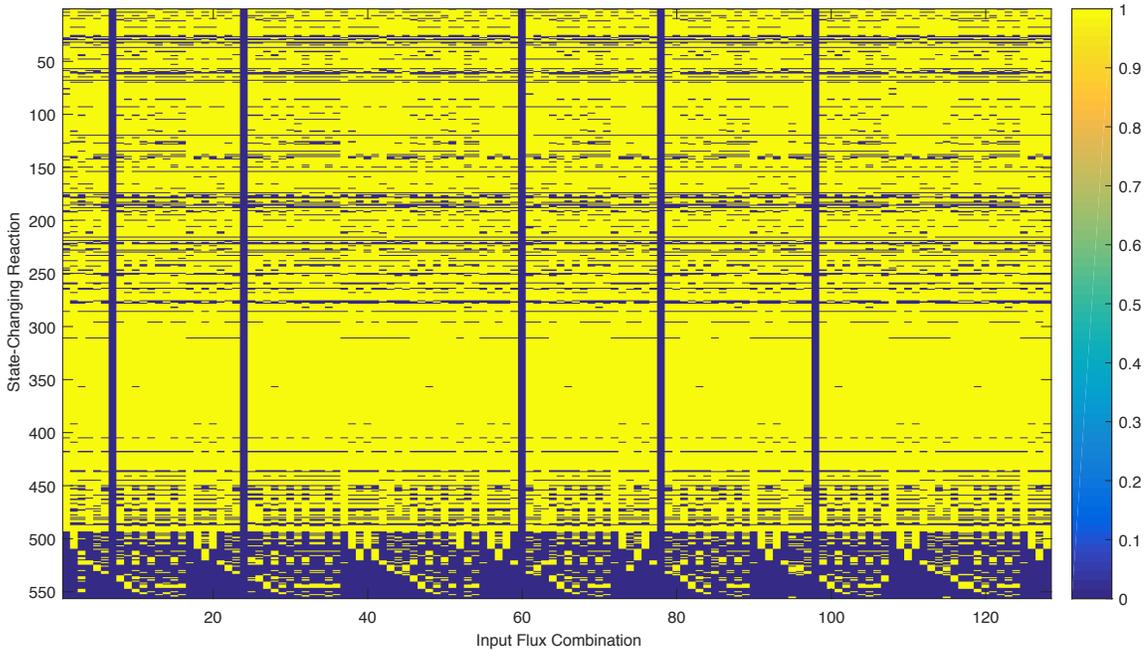


Figure 6.4: FBA-estimated binary chemical reaction states  $\{r_i^*\}_{i=1}^M$  for each combination of *Glucose*, *Hematin*, *Formate*,  $H_2$ , *VitaminB<sub>12</sub>*, *Acetate*, and *Vitamin K* input fluxes, where yellow = ON state; violet = OFF state.

In Figure 6.3 and 6.4, for each of the 128 tested combinations of the input fluxes of seven compounds, one for each column of the matrix, we show the binary values of FBA-estimated chemical reaction states  $\{r_i^*\}_{i=1}^M$  as defined in Section 4.1. In addition one for each column, where the number of metabolic reactions  $M$  that show a state change within the considered combination of input fluxes of seven compounds is equal to 212 and 556 for *B. theta* and *M. smithii*, respectively.

The computation of the upper bound of the steady-state mutual information is finally realized by applying the expressions in Equations (4.5), (3.4), and (3.5), taking into account that the possible combinations of input fluxes of the seven compounds are drawn from a discrete set. For these preliminary results, we make the assumption that these combinations are equiprobable. As a consequence, the corresponding combinations  $\{c_i\}_{i=1}^N = \left\{c_G, c_{He}, c_F, c_{H_2}, c_{B_{12}}, c_A, c_{V_k}\right\}$  computed through Equation (4.3) can be as well considered equiprobable with probability density

$P\left(\left\{c_G, c_{He}, c_F, c_{H_2}, c_{B_{12}}, c_A, c_{V_k}\right\}\right) = 1/(\text{\#of input combinations}) = 1/128$ . The resulting input entropy  $H(\{c_i\}_{i=1}^N)$ , where  $N = 7$ , is then computed through Equation (3.4) by substituting the integral with a summation over the number of input combinations, which results into  $\log_2(128) = 7$  bits.

To compute the conditional entropy of the input for both organisms, given the output  $H(\{c_i\}_{i=1}^N | \{r_i\}_{i=1}^M)$ , where for *B. theta*  $N = 7$  and  $M = 212$ , and for *M. smithii*  $N = 7$  and  $M = 556$ , we translated Equation (3.5) into the following formula:

$$H(\{c_G, c_{He}, c_F, c_{H_2}, c_{B_{12}}, c_A, c_{V_k}\} | \{r_i^*\}_{i=1}^M) = - \sum_{y=1}^Y P_Y \sum_{x=1}^{X_y} P(x|y) \log_2 P(x|y), \quad (6.3)$$

where  $Y$  and  $P_Y$  correspond to the number of times and the probability, respectively, that a reaction state combination is found more than once in the data shown in Figure 6.3 and 6.4.  $X_y$  is the number of times the reaction state combination  $y$  is found in the data, and  $P(x|y)$  is the probability of having a combination of input concentrations  $x = \{c_{G,x}, c_{He,x}, c_{F,x}, c_{H_2,x}, c_{B_{12},x}, c_{A,x}, c_{V_k,x}\}$  given the reaction state  $y$ , which we consider as a uniform distribution in the number of different combinations of input concentrations that have been found resulting into the same reaction state  $y$ .

The experimental data for *B. theta* shows that there are a total of 114 different reaction combinations that are repeated more than once, which results in a conditional entropy of the input given the output  $H(\{c_G, c_{He}, c_F, c_{H_2}, c_{B_{12}}, c_A, c_{V_k}\} | \{r_i^*\}_{i=1}^M) = 3.6933$  bits. Finally, the following value is found for the upper bound of the steady-state mutual information:

$$I(\{c_G, c_{He}, c_F, c_{H_2}, c_{B_{12}}, c_A, c_{V_k}\}; \{r_i^*\}_{i=1}^{113}) = 3.3068 \text{ bits} \quad (6.4)$$

On the other hand, for *M. smithii* we found a total of 97 different reaction combinations that are repeated more than once, which results in a conditional entropy of the input given the output  $H(\{c_G, c_{He}, c_F, c_{H_2}, c_{B_{12}}, c_A, c_{V_k}\} | \{r_i^*\}_{i=1}^M) = 2.4778$  bits. Finally, the following value is found for the upper bound of the steady-state mutual information:

$$I(\{c_G, c_{He}, c_F, c_{H_2}, c_{B_{12}}, c_A, c_{V_k}\}; \{r_i^*\}_{i=1}^{136}) = 4.5222 \text{ bits} \quad (6.5)$$

## 6.2 Stage II

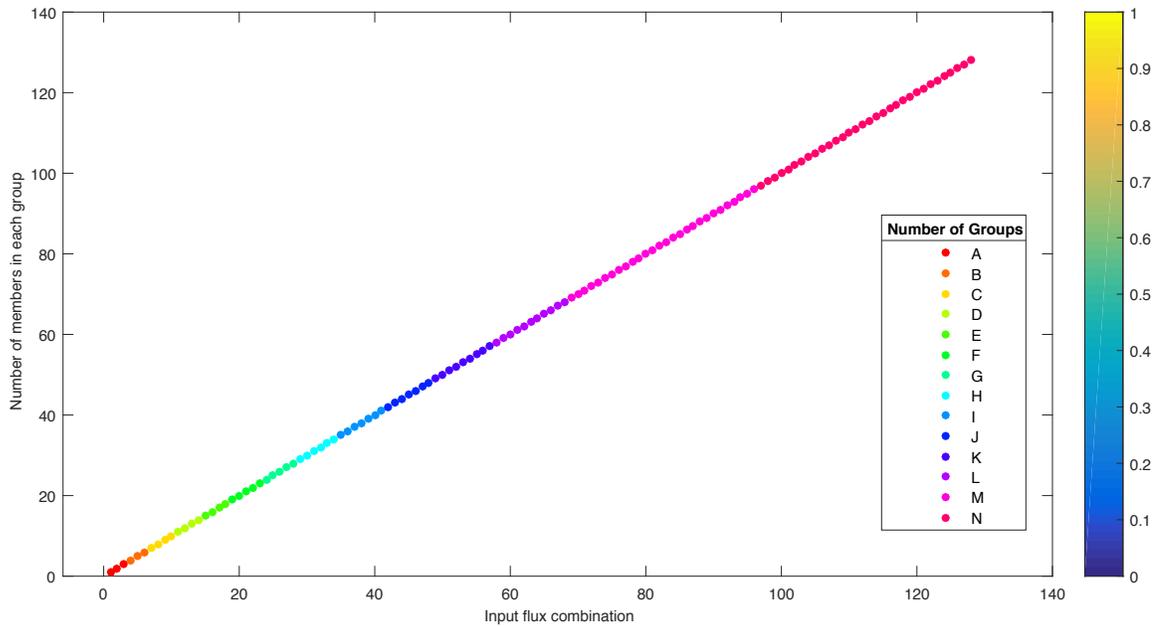


Figure 6.5: 14 FBA groups of *B. theta* based on similar FBA-estimated chemical reaction states for each combination of *Glucose*, *Hematin*, *Formate*,  $H_2$ , *VitaminB<sub>12</sub>*, *Acetate*, and *Vitamin K* input fluxes.

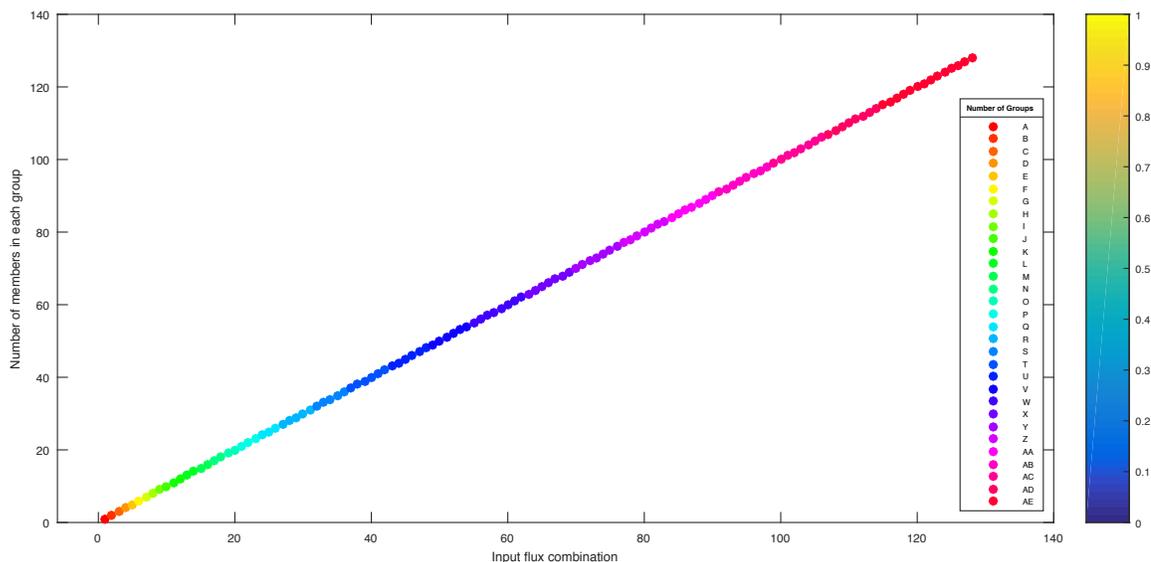


Figure 6.6: 31 FBA groups of *M. smithii* based on similar FBA-estimated chemical reaction states for each combination of *Glucose*, *Hematin*, *Formate*,  $H_2$ , *VitaminB<sub>12</sub>*, *Acetate*, and *Vitamin K* input fluxes.

From the matrix shown in Figure 6.5 and 6.6 we grouped the FBAs based on the similar FBA-estimated chemical reaction states, giving us 14 groups for *B. theta* and 31 groups for *M. smithii*. We generate a new matrix, shown in Table 6.1 and 6.2, where the rows are the groups derived in the previous step and the columns represent the chemical compounds uptaken, secreted, and the biomass generated during the metabolism.

FBA	Env	glc	hem	form	h2	vitb12	acet	vitk	biom						
g1	0.0021953	-1	0.0021953	0.0021953	0.713144	7.33267	0	0.0021953	0.0021953	0	0.0021953	0	0	0	0.87827
g2	0.0021953	-1	0.0021953	0.0021953	0.713144	7.33267	0	0.0021953	0.0021953	0	0.0021953	0.0021953	0	0	0.87827
g3	0.0021953	-1	0.0021953	0.0021953	0.713144	7.33267	0	0.0021953	0.0021953	0	0.0021953	0.0021953	0	0	0.87827
g4	0.0021953	-1	0.0021953	0.0021953	0.713144	7.33267	0	0.0021953	0.0021953	0	0.0021953	0.0021953	0	0	0.87827
g5	0.0021953	-1	0.0021953	0.0021953	0.713144	7.33267	0	0.0021953	0.0021953	0	0.0021953	0.0021953	0	0	0.87827
g6	0.0021953	-1	0.0021953	0.0021953	0.713144	7.33267	0	0.0021953	0.0021953	0	0.0021953	0.0021953	0	0	0.87827
g7	0.0021953	-1	0.0021953	0.0021953	0.713144	7.33267	0	0.0021953	0.0021953	0	0.0021953	0.0021953	0	0	0.87827
g8	0.0021953	-1	0.0021953	0.0021953	0.713144	7.33267	0	0.0021953	0.0021953	0	0.0021953	0.0021953	0	0	0.87827
g9	0.0021953	-1	0.0021953	0.0021953	0.713144	7.33267	0	0.0021953	0.0021953	0	0.0021953	0.0021953	0	0	0.87827
g10	0.0021953	-1	0.0021953	0.0021953	0.713144	7.33267	0	0.0021953	0.0021953	0	0.0021953	0.0021953	0	0	0.87827
g11	0.0021953	-1	0.0021953	0.0021953	0.713144	7.33267	0	0.0021953	0.0021953	0	0.0021953	0.0021953	0	0	0.87827
g12	0.0021953	-1	0.0021953	0.0021953	0.713144	7.33267	0	0.0021953	0.0021953	0	0.0021953	0.0021953	0	0	0.87827
g13	0.0021953	-1	0.0021953	0.0021953	0.713144	7.33267	0	0.0021953	0.0021953	0	0.0021953	0.0021953	0	0	0.87827
g14	0.0021953	-1	0.0021953	0.0021953	0.713144	7.33267	0	0.0021953	0.0021953	0	0.0021953	0.0021953	0	0	0.87827

Table 6.1: FBA group matrix of *B. theta* where the rows are the 14 groups and the columns represent the chemical compounds uptaken, secreted, and the generation biomass.



where for *B. theta*  $K = 14$  and  $M = 8$  and for *M. smithii*  $K = 31$  and  $M = 15$ , we use the same formula shown in Equation 3.8 over only biomass, where  $M$  is number of groups based on similar biomass flux values shown in the tables above. The conditional entropy for *B. theta* was found to be 1 bit and *M. smithii* was found to be 1.1455 bits.

The information loss in function of the input state changing reactions within internal cell metabolism and output biomass is then be calculated as a difference of input entropy and conditional entropy of the input given the output, which was found to be 2.8074 bits for *B. theta* and 3.8087 bits for *M. smithii*.

## 6.2.2 Upper bound of steady-state mutual information over internal metabolic state changing reactions with respect to uptake and secretion of compounds and biomass

We further group the FBA groups based on the similar biomass values. This gave us 8 groups for *B. theta* and 15 groups for *M. smithii*. As we described in Sections 3.2.2 and 4.2.2 we use the same equations to compute the upper bound of the steady-state mutual information with respect to chemical compounds uptaken, secreted and biomass flux values.

The corresponding groups are  $\{g_i\}_{i=1}^K = \{r_{1_k}^*, \dots, r_{M_k}^*\}_{i=1}^K$ . Here  $\{g_i\}_{i=1}^K$  represents the list of groups, which are 14 for *B. theta* and 31 *M. smithii*. The resulting input entropy  $H(\{g_i\}_{i=1}^K)$ , where  $K = 14$  and 31, is then computed through Equation (3.4) by substituting the integral with a summation over the number of input combinations, which results into  $\log_2(14) = 3.8074$  bits for *B. theta* and  $\log_2(31) = 4.9542$  bits for

*M. smithii*. To compute the conditional entropy of the input for both organisms *B. theta* and *M. smithii* given the output  $H(\{g_i\}_{i=1}^K | \{U_k\}_{k=1}^K, \{S_j\}_{j=1}^K, Gr)$ , where for *B. theta*  $K = 14$  and  $M = 13$ , and for *M. smithii*  $K = 31$  and  $M = 16$ , we use the same formula shown in Equation 3.8 over uptaken compounds, secreted compounds, and biomass with the only change that  $M$  is number of groups based on similar compounds secreted, uptaken and biomass flux values shown in the tables above. The conditional entropy for *B. theta* was found to be 0.1429 bits and *M. smithii* was found to be 1.0810 bits.

The information loss in the function of input chemical compounds and output as uptaken compounds, secreted compounds, and biomass could be calculated as a difference of input entropy and the conditional entropy of the input given the output, which was found to be 3.6645 bits for *B. theta* and 3.8732 bits for *M. smithii*.

### 6.2.3 Upper bound of steady-state mutual information over seven input compounds with respect to uptake and secretion of compounds and biomass

We further group the FBA groups based on the similar biomass values. This gave us 8 groups for *B. theta* and 15 groups for *M. smithii*. As we described in Sections 3.2.2 and 4.2.2 we use the same equations to compute the upper bound of the steady-state mutual information over seven input chemical compounds with respect to uptaken, secreted and biomass flux values.

The corresponding groups are  $\{g_i\}_{i=1}^K = \{r_{1_k}^*, \dots, r_{M_k}^*\}_{i=1}^K$ . Here  $\{g_i\}_{i=1}^K$  represents the list of groups, which are 14 for *B. theta* and 31 *M. smithii*. The resulting input entropy  $H(\{g_i\}_{i=1}^K)$ , where  $K = 14$  and 31, is then computed through Equation (3.4) by substituting the integral with a summation over the number of input

combinations, which results into  $\log_2(128) = 7$  bits. To compute the conditional entropy of the input for both organisms *B. theta* and *M. smithii* given the output  $H(\{g_i\}_{i=1}^K | \{U_k\}_{k=1}^K, \{S_j\}_{j=1}^K, Gr)$ , where for *B. theta*  $K = 14$  and  $M = 8$ , and for *M. smithii*  $K = 31$  and  $M = 15$ , we use the same formula shown in Equation 3.8 over uptaken compounds, secreted compounds, and biomass with the only change that  $M$  is number of groups based on similar compounds secreted, uptaken and biomass flux values shown in the tables above. The conditional entropy for *B. theta* was found to be 4.2644 bits and *M. smithii* was found to be 4.3802 bits.

The information loss in the function of input chemical compounds and output as uptaken compounds, secreted compounds, and biomass could be calculated as a difference of input entropy and the conditional entropy of the input given the output, which was found to be 2.7356 bits for *B. theta* and 2.6198 bits for *M. smithii*.

## 6.3 Visualization

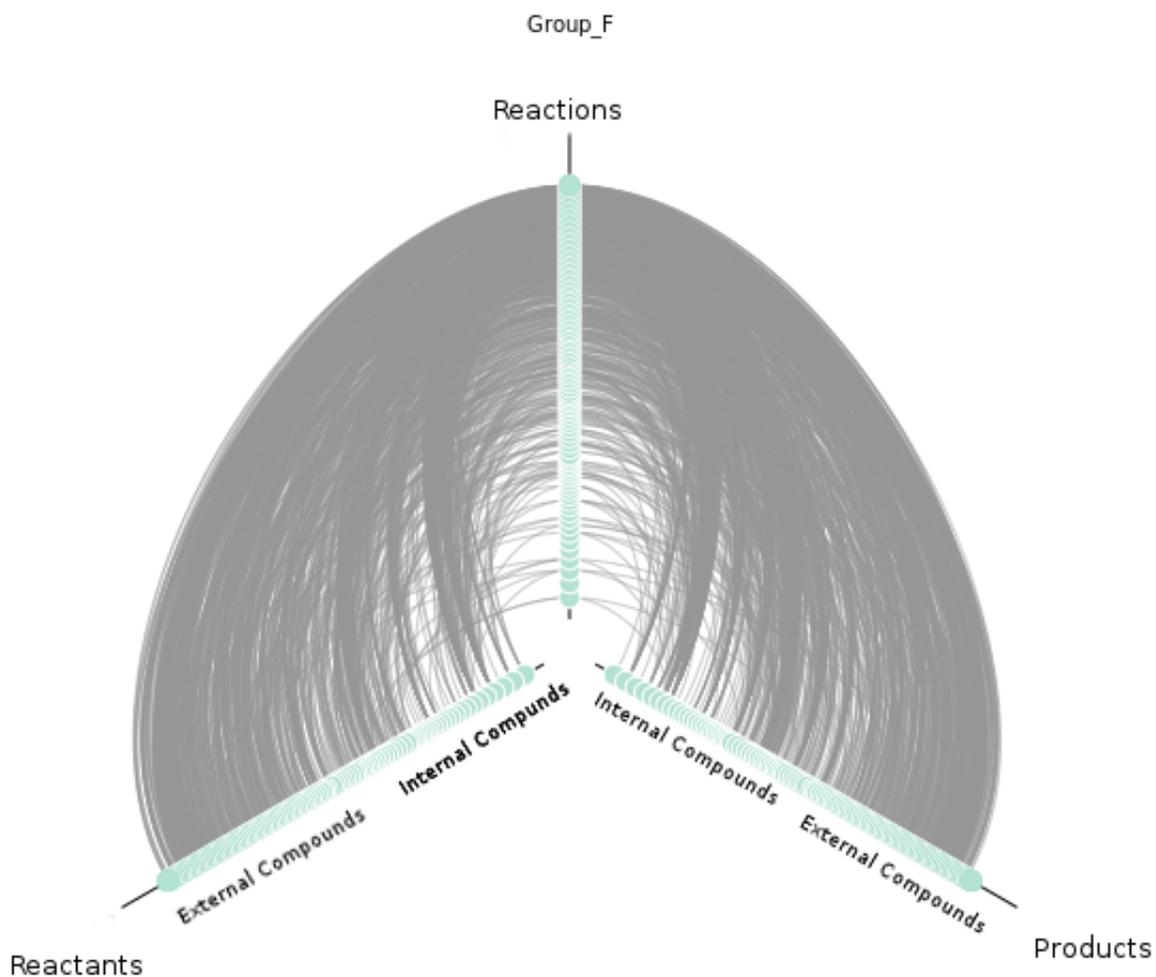


Figure 6.7: A hive plot for the FBA group F is shown in the figure. The reactions are placed on the Z axis, the reactants on the X axis and the products on the Y axis. Further the External compounds are placed higher on the X and Y axes than the Internal compounds.

Hive plots [27] were utilized to visualize the complex network of reactions taking place within the cell. In a hive plot the nodes of a network are arranged along a predetermined number of axes. The hive plot shown is a representation of the FBA group for *M. smithii*. We obtained 31 FBA-estimated chemical reaction states groups and labeled them A-AE as shown in Table 6.2. It was determined that three axes

were sufficient to represent the entire set of data available. The first axis consists of all the state changing reactions present across all the FBA groups. The second and third axes represent the same set of compounds arranged in the same order in both the axes. However, the second axes represents the compounds which are products of a reaction while the third axes represents the compounds which are reactants in a reaction. This differentiation helps to better see the difference between multiple FBA groups. Also, the compounds that are external (uptaken or secreted) to the cell are arranged at higher positions on the axes when compared to the internal compounds.

The Software Jhive 3.0 [22] utilized to generate the hive plots. Jhive is implemented in Java and therefore can be used on any operating system that supports Java. Jhive can also generate differential hive plots which allows the comparison of two different hive plots with ease. Figures show the comparison of FBA groups F, G and Z. While group F and G have the same value of production of biomass, group Z has the least production of biomass.

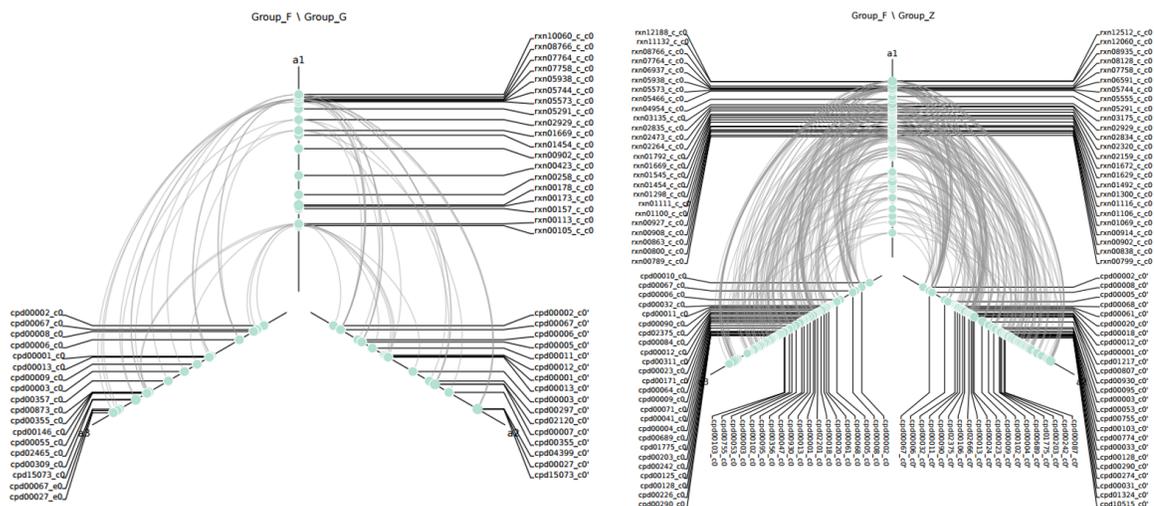


Figure 6.8: Shows differential hive plots of F vs G and F vs Z. The groups F and G in F vs G hive plot has the same biomass whereas, the groups F and Z in F vs Z hive plot have the least and highest biomass respectively. When a reaction is present in F and absent in G or Z the reaction is represented along with its links to the compounds. When a reaction is present in the other groups but absent in group F the reaction is shown as a node not connected to any other compounds.

The comparison shows that groups F and G differ by a few reactions while F and Z vary by more reactions. The figure 6.9 represents the same FBA group as the first figure but as a network diagram instead. The hive plots show clearly that there is still internal change in the ones with same biomass and much more (but expected) change when the biomass differs. The software Gephi [17] was used to plot the network diagram. The size of a node is proportional to the number of links to that node. The network nodes represent reactions as well as compounds. It is difficult to represent the large set of data as a network diagram. Hive plots provide a way to arrange the nodes based on predetermined classification.

Visualization of cell metabolic network helps in observing the changes due to different combinations of FBAs. By stemming from the aforementioned abstraction, we

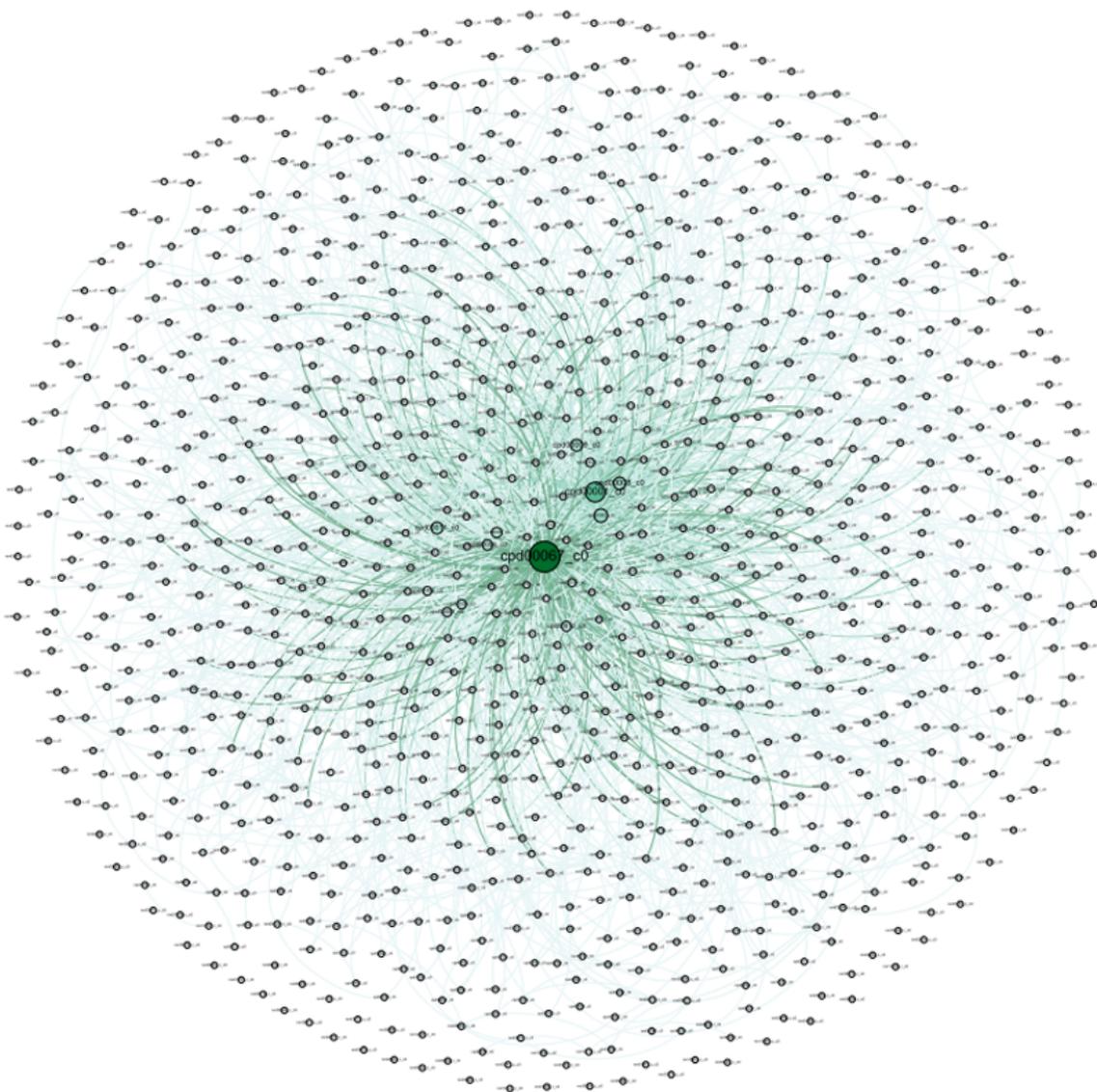


Figure 6.9: Shows the state changing reactions represented as a network diagram. The size of a node is proportional to the number of links connecting to or from it.

presented an *in silico* computation of the aforementioned upper bound to the mutual information by using the FBA metabolic simulation technique based on the knowledge of the genome of three different sample organisms. This allowed to estimate the information capacity for Stage I and Stage II, and ultimately, the flow of information measured in bits through cell metabolism.

## Chapter 7

# Conclusions and Future Work

In this thesis, we have introduced a method to obtain an upper bound to the steady-state mutual information of a communication system based on cell metabolism and its regulation. This upper bound stands as theoretical limit of the ability of a cell to internally represent the information contained in the chemical composition of the external environment, and to convey this information to experimentally observable parameters underlying its interactions with the environment. For this, we introduced an abstraction of cell metabolism based on the theory of molecular communication systems. In this abstraction, the cell metabolism is modeled by two subsequent stages. Stage I models the regulation of the chemical reaction activity in cell metabolism as a binary encoder of the external concentration of chemical compounds. Stage II models the cell metabolic interaction with the external environment as a digital modulator of the state of the binary metabolic reactions into the metabolite uptake/secretion and the growth of the cell itself.

The abstraction and analysis method developed in this thesis will contribute to the characterization of organisms in biology by introducing a novel point of view, based on communication engineering abstractions, to reason about and quantify how organisms

process information. In particular, we envision that our *in silico* computation can be utilized to understand how organisms can be fine tuned and controlled by properly varying the chemical composition of their living environment. In the long run, this approach will potentially help the design of techniques to control functionalities in cells engineered through genetic circuits [32]. Additionally, we provided a sample visualization of the metabolic network parameters used to help in computation of the mutual information based on Hive plots.

Future work will be focused on a thorough modeling and evaluation of this molecular communication system, including models of the noise source and the dynamic behavior of metabolic regulation using Dynamic Flux balance Analysis [21], including the investigation of its information theoretical capacity. We will also focus on developing a model to study how to extract information from lab data and compare it to the optimal (upper bound) solution, and set the basis for the design of the aforementioned techniques to operate a fine-tuned control on the cell behavior based on information transmission through its metabolism.

# Bibliography

- [1] I. F. Akyildiz, J. M. Jornet, and M. Pierobon. Nanonetworks: A new frontier in communications. *Communications of the ACMs*, 54(11):84–89, November 2011.
- [2] I. F. Akyildiz, M. Pierobon, S. Balasubramaniam, and Y. Koucheryavy. The internet of bio-nano things. *IEEE Communications Magazine*, 53(3):32–40, March 2015.
- [3] Uri Alon. *An introduction to systems biology: design principles of biological circuits*. CRC press, 2006.
- [4] G. Aminian, H. Arjmandi, A. Gohari, M.N. Kenari, and U. Mitra. Capacity of lti-poisson channel for diffusion based molecular communication. In *In Proc. of 2015 IEEE International Conference on Communications (ICC)*, June 2015.
- [5] Gino JE Baart and Dirk E Martens. Genome-scale metabolic models: reconstruction and analysis. *Neisseria meningitidis: Advanced Methods and Protocols*, pages 107–126, 2012.
- [6] G. Bertani. Lysogeny at mid-twentieth century: P1, P2, and other experimental systems. *Journal of Bacteriology*, 186(3):595–6008, 2004.
- [7] N Campbell and J Reece. *Biology* 7th edition, ap, 2005.

- [8] Mikaela Cashman, Jennie L. Catlett, Myra B. Cohen, Nicole Buan, Zahmeeth Sakkaff, Massimiliano Pierobon, and Christine Kelley. Sampling and Inference in Configurable Biological Systems: A Software Testing Perspective. Technical Report TR-UNL-CSE-2016-0007, University of Nebraska-Lincoln, 2016.
- [9] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory, 2nd Edition*. Wiley, 2006.
- [10] E. H. Davidson. *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. Academic Press, Elsevier, 2006.
- [11] Eric H Davidson. *The regulatory genome: gene regulatory networks in development and evolution*. Academic press, 2010.
- [12] Jon Dobson. Remote control of cellular behaviour with magnetic nanoparticles. *Nature Nanotechnology*, 3:139–143, 2008.
- [13] EcoCyc. Escherichia coli K-12 substr. MG1655 Growth Medium: M9 medium with 2% glycerol. <http://biocyc.org/ECOLI/NEW-IMAGE?type=Growth-Media&object=MIX0-59>. [Online; accessed 13-March-2016].
- [14] A. Einolghozati, M. Sardari, and F. Fekri. Capacity of diffusion-based molecular communication with ligand receptors. In *IEEE Information Theory Workshop (ITW)*, October 2011.
- [15] Francesco Fabris. Shannon information theory and molecular biology. *Journal of Interdisciplinary Mathematics*, 12(1):41–87, 2009.
- [16] L. Felicetti, M. Femminella, G. Reali, T. Nakano, and A.V. Vasilakos. Tcp-like molecular communications. *IEEE Journal on Selected Areas in Communications*, 32(12):2354–2367, 2014.

- [17] Gephi. The Open Graph Viz Platform. <https://gephi.org/>. [Online; accessed 23 September-2016].
- [18] Emanuel Gonçalves, Joachim Bucher, Anke Ryll, Jens Niklas, Klaus Mauch, Stefan Klamt, Miguel Rocha, and Julio Saez-Rodriguez. Bridging the layers: towards integration of signal transduction, regulation and metabolism into mathematical models. *Molecular BioSystems*, 9(7):1576–1583, 2013.
- [19] Koji Hayashi, Naoki Morooka, Yoshihiro Yamamoto, Katsutoshi Fujita, Katsumi Isono, Sunju Choi, Eiichi Ohtsubo, Tomoya Baba, Barry L Wanner, Hirotada Mor, and Takashi Horiuchi. Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. *Mol Syst Biol.*, 2, February 2006.
- [20] Christopher S Henry, Matthew DeJongh, Aaron A Best, Paul M Frybarger, Ben Linsay, and Rick L Stevens. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology*, 28(9):977–982, 2010.
- [21] K Höffner, SM Harwood, and PI Barton. A reliable simulator for dynamic flux balance analysis. *Biotechnology and bioengineering*, 110(3):792–802, 2013.
- [22] jhive. jhive - A Java GUI for Hive Plots. <https://www.bcgsc.ca/wiki/display/jhive/home>. [Online; accessed 23 September-2016].
- [23] L. J Kahl and D. Endy. A survey of enabling technologies in synthetic biology. *Journal of Biological Engineering*, 7(1):13, May 2013.
- [24] Andrew L Kau, Philip P Ahern, Nicholas W Griffin, Andrew L Goodman, and Jeffrey I Gordon. Human nutrition, the gut microbiome and the immune system. *Nature*, 474(7351):327–336, 2011.

- [25] KBase. Department of Energy Systems Biology Knowledgebase (KBase). <http://kbase.us>. [Online; accessed 22-September-2016].
- [26] KEGG. Kyoto encyclopedia of genes and genomes (KEGG). <http://www.genome.jp/kegg/>. [Online; accessed 22-September-2016].
- [27] Martin Krzywinski, Inanc Birol, Steven JM Jones, and Marco A Marra. Hive plots: a rational approach to visualizing networks. *Briefings in bioinformatics*, 13(5):627–644, 2012.
- [28] Wendell A. Lim. The promise of optogenetics in cell biology: interrogating molecular circuits in space and time. *Nature Reviews*, 11:393–403, 2010.
- [29] Diego Barcena Menendez, Vivek Raj Senthivel, and Mark Isalan. Sender-receiver systems and applying information theory for quantitative synthetic biology. *Curr Opin Biotechnol*, 31C:101–107, March 2015.
- [30] Christian M. Metallo and Matthew G. Vander Heiden. Understanding metabolic regulation and its influence on cell physiology. *Molecular Cell*, 49:388–398, February 2013.
- [31] Reza Mosayebi, Hamidreza Arjmandi, Amin Gohari, Masoumeh Nasiri-Kenari, and Urbashi Mitra. Receivers for diffusion-based molecular communication: Exploiting memory and sampling rate. *IEEE Journal on Selected Areas in Communications*, 32(12):2368–2380, December 2014.
- [32] Chris J. Myers. *Engineering genetic circuits*. Chapman & Hall/CRC, Mathematical and Computational Biology Series, 2009.
- [33] D. L. Nelson and M. M. Cox. *Lehninger Principles of Biochemistry*, chapter 12.2, pages 425–429. W. H. Freeman, 2005.

- [34] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–248, 2010.
- [35] Stephen Payne and Lingchong You. Engineered cell-cell communication and its applications. *Adv Biochem Eng Biotechnol*, 146:97–121, 2014.
- [36] M. Pierobon and I. F. Akyildiz. Capacity of a diffusion-based molecular communication system with channel memory and molecular noise. *IEEE Transactions on Information Theory*, 59(2):942–954, February 2013.
- [37] Massimiliano Pierobon, Myra B. Cohen, Nicole Buan, and Christine Kelley. SCIM: Sampling, Characterization, Inference and Modeling of Biological Consortia. Technical Report TR-UNL-CSE-2015-0002, University of Nebraska-Lincoln, 2015.
- [38] Massimiliano Pierobon, Zahmeeth Sakkaff, Jennie L Catlett, and Nicole R Buan. Mutual information upper bound of molecular communication based on cell metabolism. In *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–6. IEEE, 2016.
- [39] Alex Rhee, Raymond Cheong, and Andre Levchenko. The application of information theory to biochemical signaling systems. *Phys Biol.*, 9(4):045011, August 2012.
- [40] Moisés Santillán. Bistable behavior in a model of the lac operon in escherichia coli with variable growth rate. *Biophysical journal*, 94(6):2065–2081, 2008.
- [41] Anna Sheydina, Ruth Y Eberhardt, Daniel J Rigden, Yuanyuan Chang, Zhanwen Li, Christian C Zmasek, Herbert L Axelrod, and Adam Godzik. Structural genomics analysis of uncharacterized protein families overrepresented in human

- gut bacteria identifies a novel glycoside hydrolase. *BMC bioinformatics*, 15(1):1, 2014.
- [42] Orkun S Soyer, Marcel Salathe, and Sebastian Bonhoeffer. Signal transduction networks: topology, response and biochemical processes. *Journal of Theoretical Biology*, 238(2):416–425, 2006.
- [43] K.V. Srinivas, A.W. Eckford, and R.S. Adve. Molecular communication in fluid media: The additive inverse gaussian noise channel. *IEEE Transactions on Information Theory*, 58(7):4678–4692, 2012.
- [44] Gašper Tkačik, Curtis G Callan Jr, and William Bialek. Information capacity of genetic regulatory elements. *Physical Review E*, 78(1):011910, 2008.
- [45] Jared E Toettcher, Christopher A Voigt, Orion D Weiner, and Wendell A Lim. The promise of optogenetics in cell biology: interrogating molecular circuits in space and time. *Nature Methods*, 8:35–38, 2011.
- [46] Alejandro F Villaverde and Julio R Banga. Reverse engineering and identification in systems biology: strategies, perspectives and challenges. *Journal of The Royal Society Interface*, 11(91):20130505, 2014.
- [47] M Zhou, Y-H Chung, KA Beauchemin, L Holtshausen, M Oba, TA McAllister, and LL Guan. Relationship between rumen methanogens and methane production in dairy cows fed diets supplemented with a feed enzyme additive. *Journal of applied microbiology*, 111(5):1148–1158, 2011.