T.R.

GEBZE TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

GENOME-SCALE METABOLIC NETWORK RECONSTRUCTION AND CONSTRAINT-BASED ANALYSIS OF SELECTED DISEASE-ASSOCIATED BACTERIA: *KLEBSIELLA PNEUMONIAE HS11286* AND *PREVOTELLA COPRI DSM 18205*

BETÜL BAZ

A THESIS SUBMITTED FOR THE DEGREE OF

MASTER OF SCIENCE

DEPARTMENT OF BIOENGINEERING

BIOINFORMATICS AND SYSTEMS BIOLOGY PROGRAMME

GEBZE

2020

T.R.

GEBZE TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

GENOME-SCALE METABOLIC NETWORK RECONSTRUCTION AND CONSTRAINT-BASED ANALYSIS OF SELECTED DISEASE-ASSOCIATED BACTERIA: *KLEBSIELLA PNEUMONIAE HS11286 AND PREVOTELLA COPRI DSM 18205*

BETÜL BAZ

A THESIS SUBMITTED FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF BIOENGINEERING

THESIS SUPERVISOR

ASSOC. PROF. DR. TUNAHAN ÇAKIR

GEBZE

2020

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

HASTALIKLARLA İLİŞKİLİ BAKTERİLER İÇİN GENOM-ÖLÇEKLİ METABOLİK AĞ MODELLERİNİN İNŞAASI VE KISIT-TABANLI ANALİZİ: *KLEBSIELLA PNEUMONIAE HS11286* VE *PREVOTELLA COPRI DSM 18205* ÖRNEKLERİ

BETÜL BAZ
YÜKSEK LİSANS TEZİ
BİYOMÜHENDİSLİK ANABİLİM DALI

DANIŞMANI
DOÇ. DR. TUNAHAN ÇAKIR

GEBZE
2020

GEBZE TEKNİK ÜNİVERSİTESİ	**YÜKSEK LİSANS JÜRİ ONAY FORMU**

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 17/07/2020 tarih ve 34/2020 sayılı kararıyla oluşturulan jüri tarafından 27/07/2020 tarihinde tez savunma sınavı yapılan Betül BAZ'ın tez çalışması Biyomühendislik Anabilim Dalı Biyoinformatik ve Sistem Biyolojisi Programında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

JÜRİ

ÜYE
(TEZ DANIŞMANI) : Doç. Dr. Tunahan ÇAKIR

ÜYE : Prof. Dr. Adil MARDİNOĞLU

ÜYE : Dr. Öğr. Üyesi Saliha DURMUŞ

ONAY

Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun
…..…/…..…/..…… tarih ve ……/…….. sayılı kararı.

İMZA/MÜHÜR

SUMMARY

Klebsiella pneumoniae HS11286 is a multi-drug resistant bacteria causing mortality-related infections. *Prevotella copri DSM 18205* is a beneficial bacterial strain inhabited in human gut flora, and importantly, the low abundance of *P.copri* in gut microbiota causes many diseases. The representation of metabolic interactions within bacteria elucidates the molecular mechanisms of the diseases and consequently provides design of novel treatments. The genome-scale metabolic network models (GEMs) have great importance for this purpose. In this study, a GEM for *Klebsiella pneumoniae HS11286*, called *i*Kp1481, was reconstructed and it contains 2649 reactions, 2116 metabolites and 1481 genes. In addition, a GEM for *Prevotella copri DSM 18205*, called *i*Pc621, was generated and it contains 1775 reactions, 1373 metabolites and 621 genes. The GEM reconstruction has four phases: draft reconstruction, manual curation, conversion into mathematical model, and model validation. In the draft model reconstruction stage, high-quality GEMs of genetically close organisms to the target organisms as template models were used. As a further study, the draft models were extensively curated, and validated with the experimental data to have more reliable demonstration of organism's metabolism. *i*Kp1481 was validated with the growth simulations performed via Flux Balance Analysis method on different carbon sources in aerobic condition, and the gene deletion analysis. *i*Pc621 was validated with *in silico* growth predictions on D-glucose minimal medium in anaerobic conditions, and the gene essentiality analysis. As a result, the final GEMs are functional and can demonstrate the metabolic behavior of the target organisms to a certain level.

Key Words: Metabolic Network, Genome-scale Model Reconstruction, *Klebsiella pneumoniae, Prevotella copri*.

ÖZET

Klebsiella pneumoniae HS11286 suşu ölümcül enfeksiyonlara neden olan çoklu ilaca direnç gösteren bir bakteridir. *Prevotella copri DSM 18205*, insan bağırsak florasında yaşayan yararlı bir bakteridir ve en önemlisi, insan bağırsak florasında *P.copri* organizmasının düşük seviyede olması birçok hastalığa neden olur. Bakterilerdeki metabolik etkileşimlerin gösterimi, hastalıkların moleküler mekanizmalarını aydınlatır ve bu sayede yeni tedavi yöntemlerinin geliştirilmesini sağlar. Genom ölçeğinde metabolik ağ modelleri (GEM'ler) bu amaç için büyük öneme sahiptir. Bu çalışmada, *i*Kp1481 olarak adlandırılan *Klebsiella pneumoniae HS11286* suşu için bir GEM oluşturuldu ve bu metabolik model 2649 reaksiyon, 2116 metabolit ve 1481 gen içermektedir. Buna ek olarak, *Prevotella copri DSM 18205* için *i*Pc621 adı verilen bir GEM üretildi ve bu metabolik model 1775 reaksiyon, 1373 metabolit ve 621 gen içermektedir. Genom bilgisinden yola çıkarak GEM oluşturmanın dört aşaması vardır: taslak metabolik model oluşturulması, manuel iyileştirme, matematiksel modele dönüşüm ve modelin doğrulanması. Taslak model oluşturulması aşamasında, şablon model olarak, hedef organizmalara genetik olarak yakınlığı olan organizmalara ait yüksek kaliteli genom ölçekli metabolik modeller kullanılmıştır. Daha ileri bir çalışma olarak ise, oluşturulan taslak metabolik modeller kapsamlı bir şekilde iyileştirilmiştir ve iyileştirilmiş modelin ilgili organizmanın metabolizmasının güvenilir bir gösterimine sahip olduğu deneysel verilerle doğrulanmıştır. Bu amaçla, *i*Kp1481 için aerobik koşulda farklı karbon kaynakları kullanılarak Akı Dengesi Analizi yoluyla büyüme simülasyonları ve gen delesyon analizleri gerçekleştirilerek modelin validasyonu sağlandı. *i*Pc621 için ise, anaerobik koşullarda D-glikoz minimal ortamında in siliko büyüme tahminleri ile ve esansiyel gen analizi ile doğrulama analizleri gerçekleştirildi. Sonuç olarak, nihai GEM'ler çalışır durumdadır ve hedef organizmaların metabolik davranışını belirli ölçüde yansıtabilmektedir.

Anahtar kelimeler: Metabolik Ağ, Genom-ölçekli Model İnşaası, *Klebsiella pneumoniae, Prevotella copri.*

ACKNOWLEDGMENTS

I would like to sincerely thank to my supervisor, Assoc. Dr. Tunahan Çakır, for his endless patience, excellent guidance and encouragement. It was a privilege for me to study under his supervision during my master's degree. I very much appreciate that my dear supervisor brought me valuable experiences that cannot be learned from the books and what he taught me. Also, I would like to thank the other members of my thesis committee, Assist. Prof. Saliha Durmuş and Prof. Adil Mardinoğlu for their comments and suggestions regarding my thesis.

I deeply thank the members of Computational Systems Biology Group at Gebze Technical University for providing me with a peaceful working environment and their very nice collaborations.

I would like to offer my special thanks to my family members Selma Baz, Mehmet Baz, Büşra Baz and especially my big brother, Özkul Baz for their support, courage and making me feel like they are always by me.

I acknowledge that my thesis study was financially supported through a grant by TUBITAK (Project Code: 316S005).

TABLE of CONTENTS

	Page
SUMMARY	v
ÖZET	vi
ACKNOWLEDGMENTS	vii
TABLE of CONTENTS	viii
LIST of ABBREVIATIONS and ACRONYMS	x
LIST of FIGURES	xi
LIST of TABLES	xii
1. INTRODUCTION	1
2. BACKGROUND INFORMATION	3
2.1. Bacteria	3
2.2. Bacterial growth	4
2.3. Bacterial Metabolism	7
2.3.1. Carbon Metabolism	8
2.3.2. Energy Metabolism	13
2.4. Klebsiella pneumoniae subsp. pneumoniae HS11286	16
2.5. Prevotella copri DSM 18205	18
2.6. Genome-scale Metabolic Network Model	21
2.7. Applications of Genome-scale Metabolic Network Models	21
2.8. Genome-scale Metabolic Modelling Process	22
2.8.1. Stage 1: Draft Reconstruction	23
2.8.2. Stage 2: Manual Curation / Refinement	25
2.8.3. Stage 3: Conversion into a Mathematical Model	27
2.8.4. Stage 4: Model Validation	28
2.9. Reconstruction Tools & Softwares	29
2.10. Constraint-Based Analysis of Metabolic Networks	31
2.10.1. Flux Balance Analysis	31
2.10.2. Minimization of Metabolic Adjustment	35

3. GENOME-SCALE METABOLIC MODEL RECONSTRUCTION FOR BACTERIA 37

3.1. Draft Metabolic Model Reconstruction 39

3.2. Model Refinements Based on Manual Curation 41

3.4. Metabolic Model Validation 46

4. GENOME-SCALE METABOLIC MODEL RECONSTRUCTION FOR KLEBSIELLA PNEUMONIAE SUBSP. PNEUMONIAE HS11286 48

4.1. Draft Model for Klebsiella pneumoniae subsp. pneumoniae HS11286 48

4.2. Manual Evaluation of the Draft Model for Klebsiella pneumoniae subsp. pneumoniae HS11286 51

4.2.1. Integration of KEGG-Based Metabolic Network Model 53

4.3. Validation of Model for Klebsiella pneumoniae subsp. pneumoniae HS11286 57

4.3.1. Growth Simulation on Different Carbon Sources 57

4.3.2. Single Gene Deletion Simulations 60

5. GENOME-SCALE METABOLIC MODEL RECONSTRUCTION FOR PREVOTELLA COPRI DSM 18205 62

5.1. Draft Model for Prevotella copri DSM 18205 62

5.2. Manual Evaluation of the Draft Model for Prevotella copri DSM 18205 64

5.2.1. Integration of CarveMe Draft Model 66

5.3. Validation of Model for Prevotella copri DSM 18205 69

5.3.1. Fermentation Product Simulation 69

5.3.2. Gene Essentiality Simulations 73

6. CONCLUSIONS 75

REFERENCES 77

BIOGRAPHY 86

APPENDICES 87

LIST of ABBREVIATIONS and ACRONYMS

Abbreviations and Acronyms		**Explanations**
ACP	:	Acyl-carrier protein
ATP	:	Adenosine Triphosphate
BLASTP	:	Protein Basic Local Alignment Search Tool
COBRA	:	COnstraint-Based Reconstruction and Analysis
ED	:	Entner-Doudoroff
EMP	:	Embden-Meyerhof-Parnas
FBA	:	Flux balance analysis
GAM	:	Growth Associated Maintenance
GEM	:	Genome-scale Metabolic Network Model
GPR	:	Gene-Protein-Reaction
HMM	:	Hidden Markov Model
KEGG	:	Kyoto Encyclopedia of Genes and Genomes
KO	:	KEGG Orthology
LP	:	Linear Programming
MOMA	:	The Method of Minimization of Metabolic Adjustment
NADP	:	Nicotinamide adenine dinucleotide phosphate
NCBI	:	National Center for Biotechnology Information
NGAM	:	Non-growth Associated Maintenance
PD	:	Parkinson’s Disease
PPP	:	Pentose Phosphate Pathway
QP	:	Quadratic Programing
S	:	Stoichiometric Matrix
SCFA	:	Short Chain Fatty Acids
SBML	:	Systems Biology Markup Language
TCA	:	Tricarboxylic Acid
WT	:	Wild Type

LIST of FIGURES

Figure No:		Page
2.1:	Different bacterial cell shapes.	4
2.2:	The plot of logarithm of viable cell concentration versus time, which shows the bacterial growth curve.	5
2.3:	The summary of cell metabolism.	8
2.4:	The Embden Meyerhof Parnas Pathway.	10
2.5:	The two routes of Pentose Phosphate pathway. (A) Oxidative Pentose Phosphate Pathway; (B) Non-oxidative Pentose Phosphate Pathway; (C) Glycolysis.	12
2.6:	The dehydratase and subsequent aldolase reaction in EDP.	13
2.7:	The substrate level phosphorylation in glycolysis.	14
2.8:	The tricarboxylic acid (TCA) cycle.	16
2.9:	The genome-scale metabolic network reconstruction process in four phases.	23
2.10:	Representation of Gene-protein-reaction (GPR) rules.	25
2.11:	A representation of SBML model structure.	28
2.12:	A toy metabolic network.	32
2.13:	A representation of Stoichiometric matrix for the toy metabolic network in Figure 2.12.	33
2.14:	The flux distribution space by constraints-based analysis simulation.	34
2.15:	The solution space of minimization of metabolic adjustment.	35
3.1:	The illustration of the overview of the GEM reconstruction process.	38
3.2:	The illustration of gene deletion analysis via metabolic network simulations.	47
5.1:	Intermediate flow rates of P. copri's central carbon metabolism. The numbers represent mmol of the related compound per g DW. GAP, glyceraldehyde; PEP, phosphoenolpyruvate; OAA, oxaloacetate; Fdred, reduced ferredoxin; Pi, phosphate.	70

LIST of TABLES

Table No:		Page
2.1:	A comparison of software & tools used in GEM reconstructions.	29
4.1:	The Constraints for The Different Carbon Sources.	58
4.2:	Growth Phenotype Comparison for Different Carbon Sources.	58
4.3:	Gene Essentiality Analysis Results.	61
5.1:	The Anaerobic Growth Simulation Results Under Modified Minimal Medium.	72
5.2:	In silico simulation results predicted by iPc621 Model.	73

1. INTRODUCTION

There are thousands of bacterial strains on earth, and a number of them are associated with diseases. Therefore, a closer inspection of disease-associated bacteria is crucial for maintaining and improving the quality of human health.

The global increase in bacteria with multiple drug resistance is giving birth to an urgent problem for public health [Christaki et al., 2019]. Carbapenemase-producing Enterobacteriaceae (CR-CPE) strain *Klebsiella pneumoniae HS11286* is an important gram-negative bacterial pathogen including multidrug-resistance plasmids. Since *K.pneumoniae HS11286* strain causes opportunistic and severe hospital-associated infectious diseases resulting in death, the investigation of metabolism has great importance to develop crucial treatments [Liu et al., 2012].

The human intestinal microbial community is a high-profile research area to explore the connectivity of the host organism and bacterial flora and the metabolism of bacteria living here [Hamer et al., 2012]. Recent studies have reported a significant alteration in the bacterial populations of the gastrointestinal microbiota in patients with various diseases. *Prevotella* is a member of Bacteroidetes, which is one of the two important phyla in human microbiota. *Prevotella copri DSM 18205* is a gram-negative and beneficial anaerobic bacteria inhabited in human microbiota [Hayashi et al., 2007]. The diminished content of *Prevotella copri* is significantly linked with many diseases such as rheumatoid arthritis and neurodegenerative disorders like Parkinson's disease, Alzheimer's disease, multiple sclerosis [Roy Sarkar and Banerjee, 2019], [Alpizar-Rodriguez et al., 2019]. The bacteria produces short chain fatty acids (SCFAs) that promote disease progression by fermentation, and, therefore, SCFAs are important potential biomarkers [Kumar et al., 2018].

The reconstruction of genome-scale metabolic network models (GEMs) is a promising computational biology method for the investigation of the metabolic activities of bacteria [Shoaie et al., 2013]. For that reason, the reconstruction of high-quality GEMs for *Klebsiella pneumoniae HS11286* and *Prevotella copri DSM 18205* is aimed in this study. The GEM reconstruction and its further processing with constraint-based analysis

can provide an increased understanding of the mechanisms behind complex diseases, and a holistic insight on the functional capacity of bacteria.

Chapter 2 includes background information about bacterial cells, bacterial growth, and the carbon and energy metabolisms of bacteria. In addition, this section contains brief information about the two bacteria studied in this study, which are *K.pneumoniae HS11286* and *Prevotella copri DSM 18205* strains. Chapter 2 also covers information about genome-scale metabolic models, the short introduction to GEM modelling and the constraint-based approaches. The next chapter explains the methodology followed in this study for the reconstruction of GEMs. The reconstruction and the validation of GEM for *K.pneumoniae HS11286* strain was described in Chapter 4. Chapter 5 focuses on how the genome-scale metabolic network model for *Prevotella copri DSM 18205* was reconstructed and validated. The main conclusion of this study was mentioned in Chapter 6 with the recommendations about the improvement of the newly reconstructed GEMs for the disease-related bacteria on further studies.

2. BACKGROUND INFORMATION

2.1. Bacteria

Bacteria are microscopic, unicellular organisms that live in a great variety of environments, from Earth's biota, in sediments, to human digestive systems [Whitman et al., 1998], [Fredrickson et al., 2004], [Sears, 2005]. They are one of the two domains of prokaryotes that lack membrane-bound nuclei and any membrane-bound organelles. Unlike eukaryotes, in prokaryotes, all water-soluble intracellular components, DNA, protein, and metabolites are compartmented within cytoplasm rather than in separate subcellular organelles. The genetic material of bacteria is typically in the form of a freely floating circular chromosome of DNA. Additionally, they possess multiple circular or linear chromosomes [Stanier and van Niel, 1962], [Thanbichler et al., 2005]. Bacterial cytoskeleton proteins and peptidoglycan synthesis enzymes are responsible for bacterial cell shape [Cabeen and Jacobs-Wagner, 2005]. Bacteria can have various cell shapes, for instance, spherical (coccus), rod-shaped (bacillus), spiral shape (spirillum) cells as shown in Figure 2.1 [Henze et al., 2008].

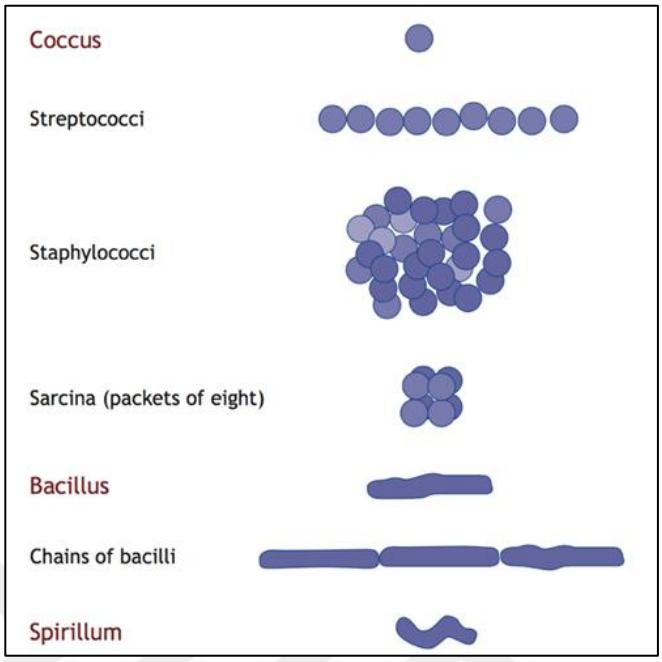


Figure 2.1: Different bacterial cell shapes.

There are two major classes of bacteria based on the differential gram staining procedure contributing to bacterial identification and taxonomic division, which was developed by Hans Christian Gram in 1884. They are Gram-negative and Gram-positive bacteria, and they have different types of cell walls [Coico and Wiley, 2005], [Moyes et al., 2009], [Riedel et al., 2019]. Bacterial species can also be classified distinctively according to their oxygen requirement for growth. Aerobic bacteria require oxygen to grow while anaerobic bacteria can grow without oxygen. Some bacteria grow facultatively, in the presence or absence of oxygen, or micro-aerobically, in the presence of a limited oxygen pressure less than atmospheric pressure, because high levels of oxygen can inhibit growth [Riedel et al., 2019].

2.2. Bacterial growth

Bacterial growth occurs as a result of the division of a bacterial cell into two genetically identical daughter cells by binary fission while some group of bacteria reproduces by budding and by multiple fission. Consequently, bacteria increase the

number of cells over time and can create a population. The time bacteria spend to double itself is called generation time or doubling time. Then, growth yield can be calculated in terms of cell mass, which is the number of viable cells per unit volume of culture or the dry weight of cells per unit volume of culture [Kim and Gadd, 2008], [Riedel et al., 2019]. The growth rate is a measurement of biomass produced per hour in the unit of grams. The specific growth rate (μ) at steady-state condition can be represented by Equation 2.1.

$$\mu = \frac{1}{X}\frac{dX}{dt} \tag{2.1}$$

where X is biomass concentration (g/L) and t is time. The amount of growth in a particular period or the amount of time required for a specified amount of growth can be calculated by using this equation [Riedel et al., 2019].

Batch culture is a technique used to grow bacteria with finite resources in a closed system. This process consists of four distinguishable growth phases: lag, exponential, stationary, and death phases. A typical bacterial growth curve is shown in Figure 2.2 [Riedel et al., 2019].

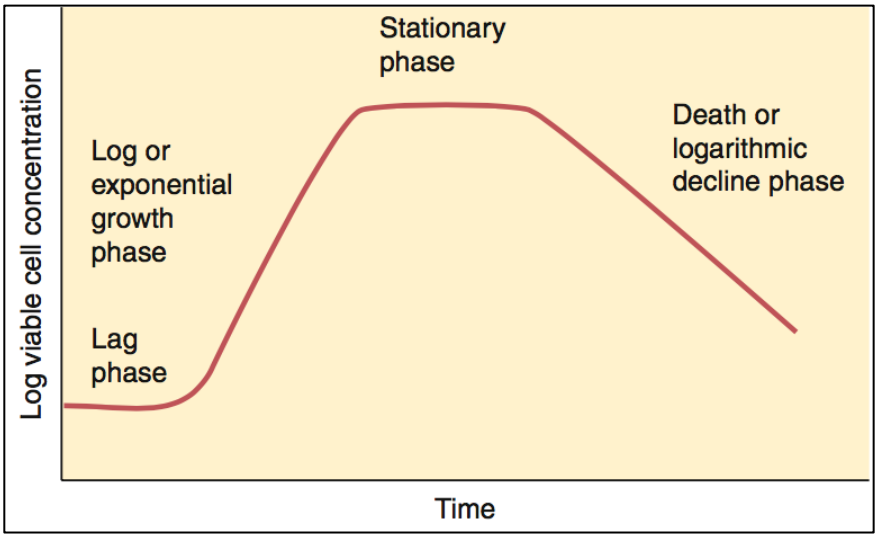


Figure 2.2: The plot of logarithm of viable cell concentration versus time, which shows the bacterial growth curve.

There is a temporary period in which bacteria do not immediately start to proliferate since there is an unfavorable condition. They try to adapt to new media or to adjust themselves for the change in the environment. They start to synthesize enzymes required for growth. They also need essential intermediates to be formed and accumulated until they reach the concentration where bacteria are capable of growth. This period is represented by the lag phase [Riedel et al., 2019], [Bertrand, 2019]. Duration of lag phase can differ depending on multiple criteria such as if bacteria have been pre-adapted to harsh conditions, or which strategy they choose to optimize their metabolism [Bertrand, 2019].

The exponential phase is a period where cells replicate. Doubling of cells continues at a constant rate if there is no limitation. However, exponential growth stops at a point where nutrients are depleted or cellular wastes increase and inhibit the growth. For instance, oxygen is a limitation of growth factors for aerobic organisms. The slope of the line that represents the exponential phase in Figure 2.2 determines the number of divisions per cell per unit time [Riedel et al., 2019].

The growth is limited after a certain time by the exhaustion of the source of nourishment and/or rise in toxicity of the environment, so viable cell count stays stable. The stationary phase is characterized by the equivalence of the rate of cell loss due to death and the rate of new cell generation through growth and replication [Riedel et al., 2019]. As a consequence, during the stationary phase, there is a constant portion of the logarithmic growth curve. When the cells stop actively dividing and switch to the resting state, the mutagenesis process causing DNA damage increases evidently [Bridges, 1998].

The death phase shows a decline in cell growth. Cell death occurs at a drastic rate because of disadvantageous conditions in the cell culture such as lack of nutrients. Thereafter, a few cells can survive for a time even up to a year at the cost of nutrients released from dying and lysing cells, and this indicates the cell turnover [Riedel et al., 2019]. Through taking all of these into account, interpretation of growth in batch culture provides basic insight into the genetics and physiology of bacterial replication.

All cells in an organism needs nutritional elements such as carbon, nitrogen, sulfur, phosphorus, various inorganic salts (e.g., potassium, magnesium, sodium, calcium, and iron) as a source to build cell materials and for the activity of enzymes and transport systems [Kim and Gadd, 2008], [Gottschalk, 1986]. Carbon is the most abundant element

required by bacteria. Bacteria can utilize many different organic and inorganic compounds as a carbon source. The prokaryotes are widely classified according to the carbon sources: organotrophs (heterotrophs) use organic compounds as their carbon source while lithotrophs (autotrophs) use CO_2 [Kim and Gadd, 2008], [Nealson, 1999], [Henze et al., 2008].

Bacterial strains can be cultivated in different culture media containing glucose as a sole organic compound and mineral salts as inorganic compounds under laboratory conditions. On the other hand, there are prokaryotic species that cannot be cultured [Kim and Gadd, 2008]. Carbon sources are utilized by bacteria and contribute to amino acids and other constituents that form cell biomass [Wang et al., 2019]. In the laboratory, there are various types of bacterial culture media depending on the requirements of the experiments or treatments. In general, these are solid growth media, agar plates that are used to isolate the pure culture of the bacterial strain, and liquid growth media, a nutrient broth that is used to measure the bacterial growth. Most bacteria require nutritionally rich media, such as Luria-Bertani (Lennox) medium (LB), yeast extract-tryptone (YT) medium, to grow [Kram and Finkel, 2015], [Riedel et al., 2019]. Undefined medium, also called complex medium, contains all the elements needed for bacterial growth and, minimal medium consists of an adequate amount of nutrients to promote cell growth. These media types are not selective. There are also selective media that intend to remove unrelated bacteria from the environment by including inhibitory agents [Riedel et al., 2019]. Hence, bacteria can show different phenotypes and therefore have strategies to survive and to increase their cell yield regarding medium composition [Kram and Finkel, 2015].

2.3. Bacterial Metabolism

Metabolism refers to all biochemical transformations that take place in a cell or an organism. Metabolism is divided into energy-generating reactions, called catabolism, and energy-consuming reactions, called anabolism (Figure 2.3) [Henze et al., 2008]. Bacterial metabolism studies concentrate on the chemical heterogeneity of oxidations of substrates and break down reactions [Kim and Gadd, 2008].

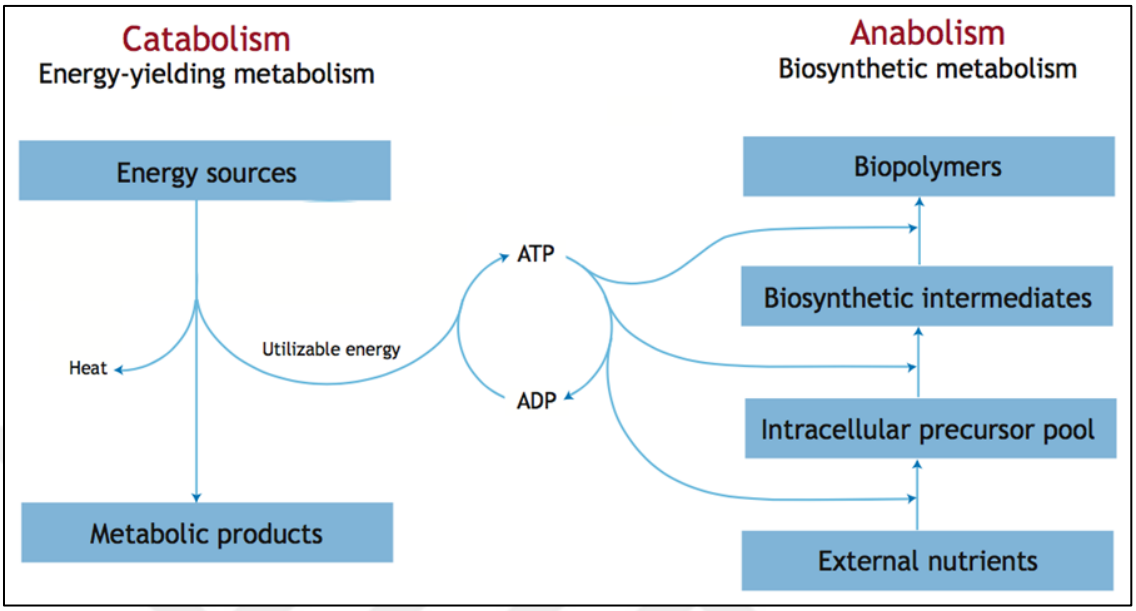


Figure 2.3: The summary of cell metabolism.

Heterotrophic metabolism is the combination of oxidation-reduction transformations of organic compounds to generate ATP and inorganic compounds that are needed by the bacterial cell for biosynthetic or dissimilatory reactions [Kim and Gadd, 2008], [Riedel et al., 2019]. Bacteria show a great variety of metabolic traits. The metabolic capability of the bacterial cell to grow, function, reproduce and nourish in a favorable chemical environment such as medium growth culture and related chemical transformations are extensively studied in the scope of bacterial metabolism.

2.3.1. Carbon Metabolism

Carbohydrate catabolism is processed by mechanisms of breaking down of large molecules and deriving energy. Glucose is the main carbon source of most bacteria and utilized through the glycolysis. The common type of glycolysis is the Embden-Meyerhof-Parnas (EMP) pathway, and it is the oxygen-independent metabolic pathway [Kim and Gadd, 2008]. Differently from eukaryotic organisms, most bacteria make use of alternative pathways when they metabolize carbon sources. These pathways are the

pentose phosphate pathway (PPP) and the Entner-Doudoroff pathway (ED) [Munoz-Elias and McKinney, 2006], [Kim and Gadd, 2008]. Glucose-6-phosphate, fructose-6-phosphate, 3-phosphoglycerate, phosphoenolpyruvate (PEP), acetyl-CoA, propionyl-CoA, oxaloacetate, and α-ketoglutarate are the vital biosynthetic precursors included in these pathways [Munoz-Elias and McKinney, 2006]. Certain pathways vary among the prokaryotes according to their purpose, advantage, competitiveness to a substrate or genetic origin [Munoz-Elias and McKinney, 2006].

In *Escherichia coli*, about 70% of glucose is metabolized through the Embden-Meyerhof-Parnas pathway and the pathway branches out to PPP to profit from the remaining 30% because EMP pathway cannot afford all the biosynthesis cost [Kim and Gadd, 2008]. Glucose is utilized and converted into two molecules of pyruvate via a metabolic reaction cascade by releasing net yield of two ATP and seven biosynthetic intermediates in EMP, which is shown in Figure 2.4 [Riedel et al., 2019].

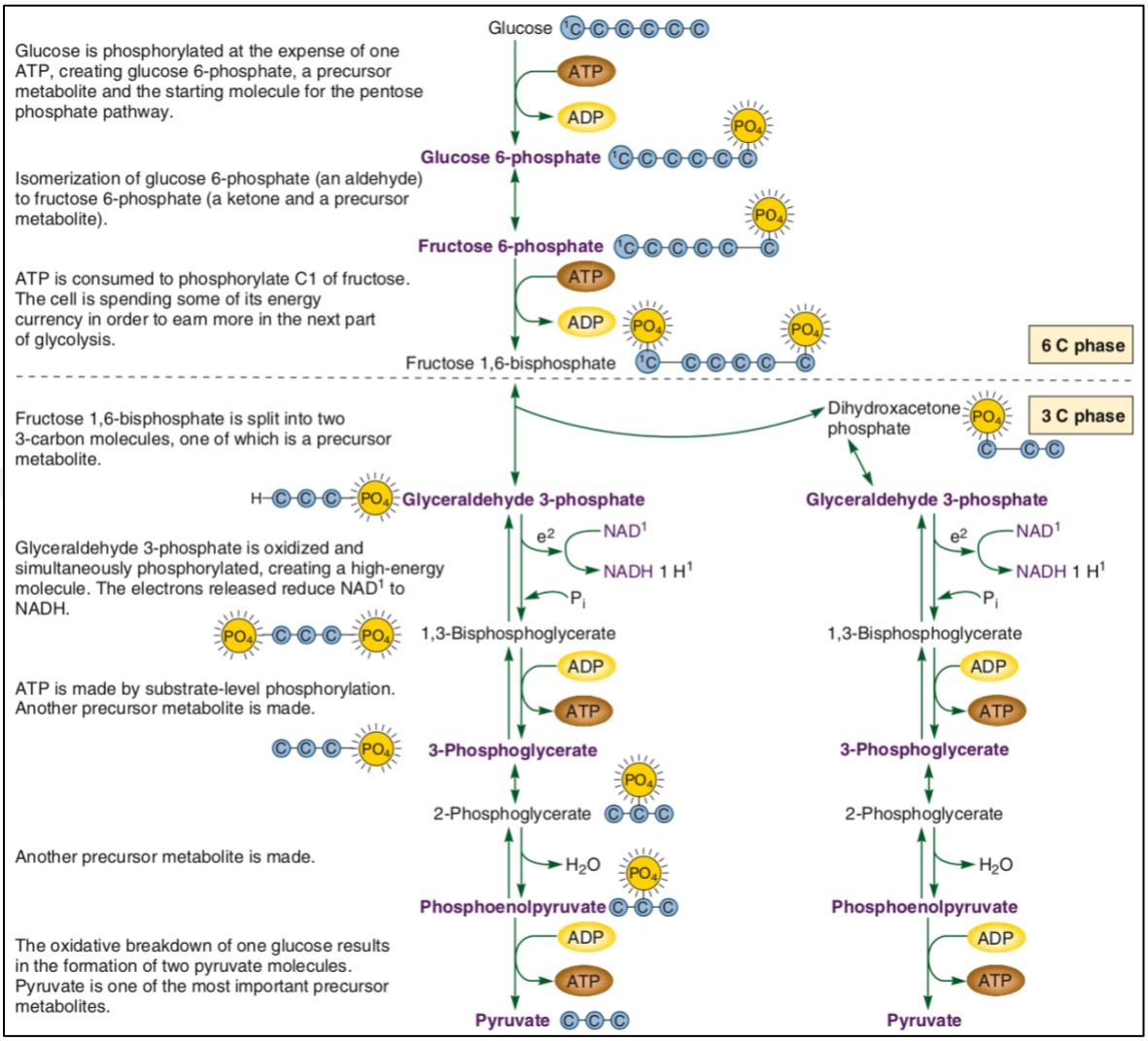


Figure 2.4: The Embden Meyerhof Parnas Pathway.

Glucose is first phosphorylated with ATP cost resulting in glucose 6- phosphate, a precursor metabolite and the starting molecule of PP pathway. The next consumption of ATP occurs during the phosphorylation of fructose 6-phosphate to create fructose 1,6-biphosphate by 6-phosphofructokinase enzyme (PFK). In the lower part of the EMP pathway, in a reaction catalyzed by glyceraldehyde dehydrogenase (GAPDH), the triose phosphate molecule is phosphorylated and oxidized via reduction of NAD to NADH. Consequently, two phosphorylation reactions consume four ATP molecules and lead to two pyruvate molecules [Kim and Gadd, 2008], [Riedel et al., 2019].

EMP pathway is important and bacteria use this pathway for different purposes. Dihydroxyacetone phosphate is a precursor for the lipid synthesis, glyceraldehyde-3-phosphate has a role in the production of vitamin B6, and pyruvate is a precursor for the biosynthesis of certain amino acids, which are all generated in EMP pathway. [Riedel et al., 2019].

Pentose phosphate pathway, also called hexose monophosphate pathway (HMP), has consecutive reactions for the oxidation of glucose-6-phosphate to pentose phosphates [Conway and Cohen, 2015]. The HMP pathway produces pentose-5-phosphate, erythrose-4-phosphate, and NADPH as supporting constituent for the biosynthetic metabolism. A major part of NADPH production in the cell is originated in this pathway [Kim and Gadd, 2008]. PPP is generally divided into preliminary oxidative and subsequent non-oxidative routes (Figure 2.5) [Wamelink et al., 2008], [Conway and Cohen, 2015]. NADPH is produced via the oxidation of glucose-6-phosphate (G-6P) by G-6P dehydrogenase (G6PDH, encoded by zwf gene), and it plays a role as a reducing power for cellular biosynthesis in the oxidative branch of the pathway [Papagianni, 2012]. D-ribose-5-phosphate, sedoheptulose-7-phosphate, and erythrose-4-phosphate are generated in the PP pathway and they act as precursors for the biosynthesis of amino acids and nucleic acids [Conway and Cohen, 2015]. The synthesis of ribulose-5P, which has a role in the riboflavin synthesis and NADPH regeneration, also occurs in the pathway. In addition to that, for some bacteria, HMP is crucial for lysine biosynthesis [Papagianni, 2012]. PPP becomes involved in the entire glucose oxidation and utilization of pentose sugars when bacteria do not have functional EMP, EDP or TCA cycle pathways. This means that bacteria are not able to complete the route to have appropriate resource for biosynthetic precursors [Kim and Gadd, 2008]. Even though PPP does not use O2 or ATP, it is an important pathway for metabolism and have important intermediates. NADPH protects cells against oxidative damage in hypoxia and in malignant cells [Wamelink et al., 2008].

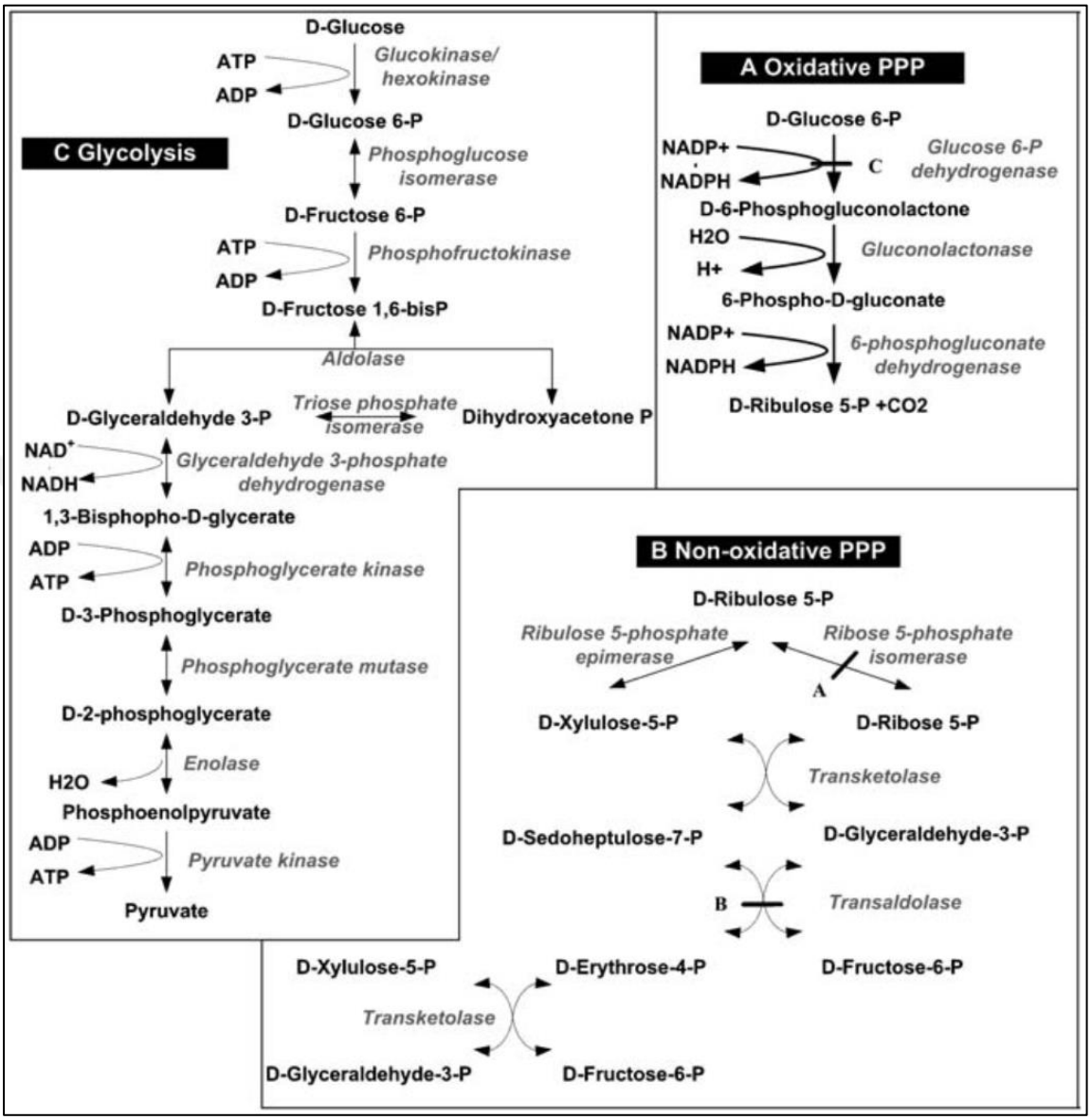


Figure 2.5: The two routes of Pentose Phosphate pathway. (A) Oxidative Pentose Phosphate Pathway; (B) Non-oxidative Pentose Phosphate Pathway; (C) Glycolysis.

The difference of the Entner-Doudoroff pathway from other pathways in central carbon metabolism is the conversion of 6-phosphogluconate into 2-keto-3-deoxy-6-phosphogluconate by dehydratase and the subsequent aldolase reaction, which generates pyruvate and glyceraldehyde 3-phosphate, as shown in Figure 2.6 [Riedel et al., 2019]. Overall, ED pathway respectively has lower energy yield from glucose [Riedel et al., 2019].

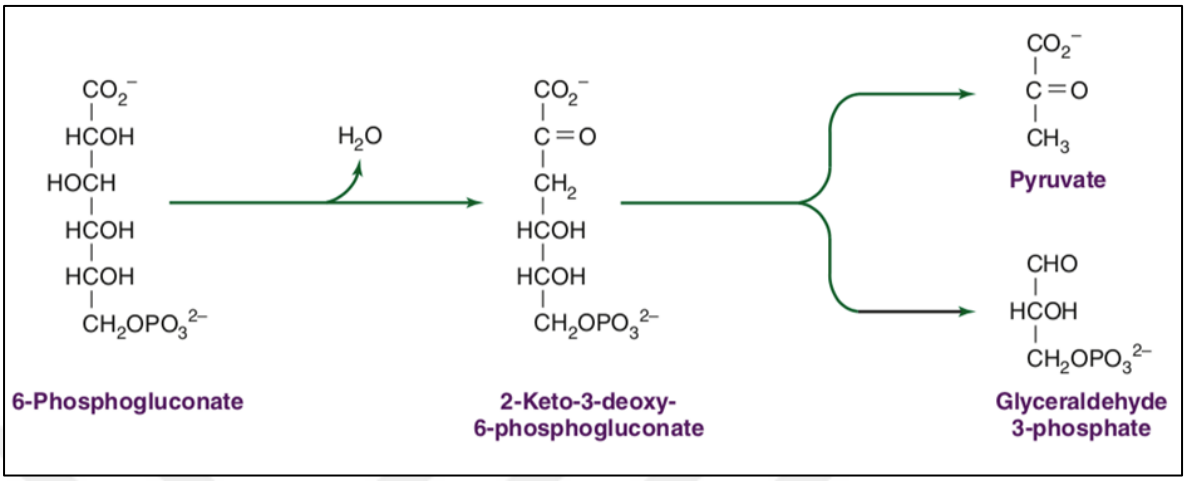


Figure 2.6: The dehydratase and subsequent aldolase reaction in EDP.

The Entner-Doudoroff pathway shows a predominant catabolic metabolism during growth on glucose in some bacteria with inactive EMP [Fuhrer et al., 2005]. Even if glucose metabolism follows EMP and PPP instead of the ED pathway in *E. coli*, the Entner– Doudoroff pathway is essential for the catabolism of aldonic sugars and pentoses and for the intestinal colonization when *E.coli* grows on mucus. In addition, in some cases, gluconate metabolism through the ED pathway is more favorable as a carbon-degrading mechanism, with a considerably lower carbon flow rate in glycolysis and PP pathways [Munoz-Elias and McKinney, 2006].

2.3.2. Energy Metabolism

Bacteria require metabolic energy to synthesize macromolecules they need and to maintain cellular homeostasis. There are three main mechanisms to gain energy: from fermentation, respiration, and photosynthesis. ATP and NADH are the energy carrier molecules, and they are produced in different pathways at a different yield [Riedel et al., 2019].

Anaerobic bacteria use exclusively fermentation to generate their energy since there is a deficiency of respiration or photosynthesis. The fermentation process can be described as substrate phosphorylation that leads to the generation of ATP by the direct ligation of

a phosphoryl group to ADP from a phosphorylated compound/substrate (Figure 2.7) [Riedel et al., 2019].

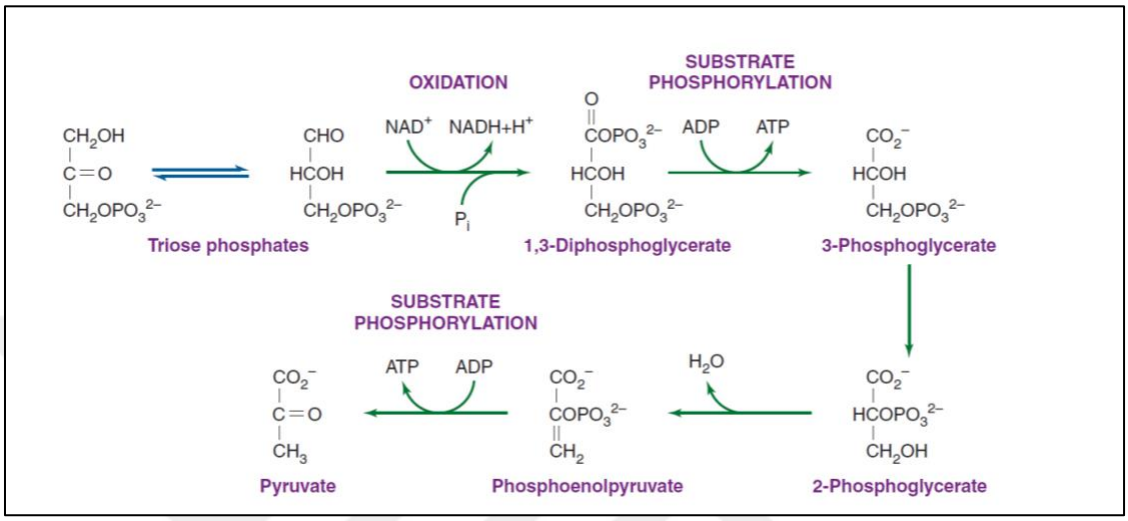


Figure 2.7: The substrate level phosphorylation in glycolysis.

Acids such as lactic acid, formic acid, succinic acid and/or gases such as CO_2 and H2 are the end products of carbohydrate fermentation [Bisen et al., 2012]. Glucose fermentation via EMP pathway can lead to two alternative routes. Two molecules of lactate formation is one of the routes . Pyruvate acts as the electron acceptor here, and it is reduced to lactate, yielding a net gain of two pyrophosphate bonds in ATP. In the second route, pyruvate decarboxylation occurs and lead to two molecules of ethanol and two molecules of CO_2 production. [Riedel et al., 2019].

Homolactic fermentation: glucose → 2 pyruvate → 2 lactate + 2 ATP

Ethanol fermentation: glucose → 2 pyruvate → 2 ethanol + 2 CO_2

There are many other fermentation processes that yield various end products and energy. The branch of acetate formation has higher energy yield than the branch of butyrate production.

glucose → 2 acetate + 4 H2 + 2 CO_2 + 4 ATP

glucose → butyrate + 2 H2 + 2 CO_2 + 3 ATP

Butyrate production is performed by the obligate anaerobe bacteria present in gut microbiome, and the formation of butanoic acid is important since it provides energy for colonocytes [Ward, 2015].

Facultative anaerobes can perform aerobic respiration using O2 as a final electron acceptor and also can take advantage of alternative acceptors such as nitrate, fumarate, sulfate, trimethylamine oxide or dimethylsulphoxide as electron acceptors in anaerobic respiration. In aerobic conditions, respiratory chain is activated through cytochrome oxidases as a function of oxygen level. ATP generation by aerobic metabolism, in principle, is also possible by the oxidation of acetyl-CoA, which acts as the first reaction in the tricarboxylic acid (TCA) cycle, which is also called Krebs cycle or citric acid cycle (Figure 2.8) [Ward, 2015], [Riedel et al., 2019]. Pyruvate, a metabolic product of glycolysis, is oxidatively decarboxylated by pyruvate dehydrogenase complex to acetyl-CoA, which is also responsible for lipid synthesis. As the second step in TCA cycle, citrate synthesis occurs from acetyl-CoA and oxaloacetate. Later, aconitase catalyzes conversion of citrate to isocitrate, which is oxidized to α-ketoglutaric acid (2-oxoglutarate) by isocitrate dehydrogenase with reduction of NAD to NADH. α-ketoglutarate is converted into succinate by oxoglutarate dehydrogenase. Succinate is transformed into fumarate by reducing FAD to FADH via succinate dehydrogenase. Malate formation occurs by the hydration of fumarate by fumarase enzyme. The last step of TCA cycle is conversion of malate into oxaloacetate by malate dehydrogenase with reduced NAD. Oxaloacetate synthesis is balanced with the synthesis of acetyl-CoA, the initiating metabolite for TCA cycle. Guanosine triphosphate (GTP) is not synthesized in TCA cycle by the prokaryotes due to lack of mitochondria in their cells [Kim and Gadd, 2008].

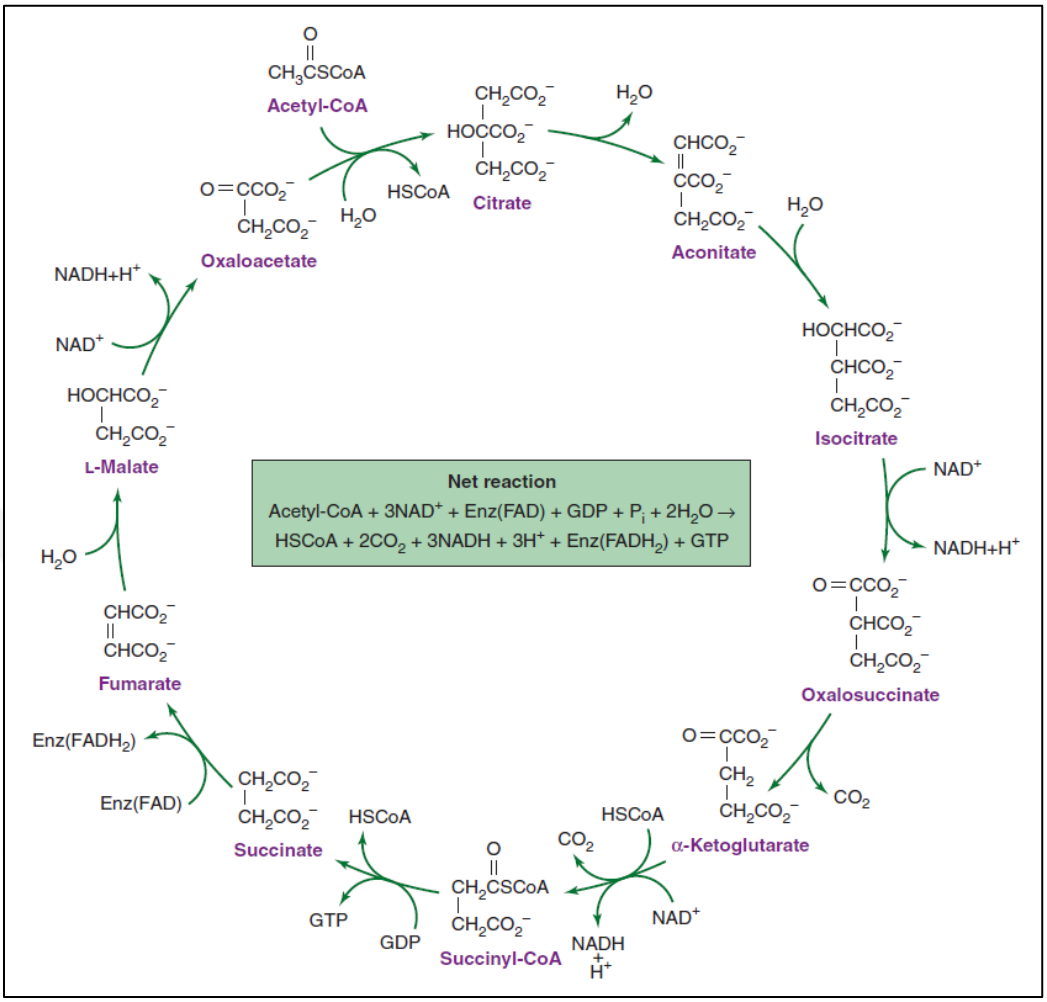


Figure 2.8: The tricarboxylic acid (TCA) cycle.

Escherichia coli diverges in terms of the TCA cycle process. The bacteria uses pyruvate oxidoreductase in order to convert pyruvate into acetyl-CoA and 2-oxoglutarate oxidoreductase for succinate formation from 2-oxoglutarate rather than using pyruvate dehydrogenase and 2-oxoglutarate dehydrogenase respectively. Therefore, TCA cycle is incomplete in this organism [Ward, 2015].

2.4. *Klebsiella pneumoniae subsp. pneumoniae HS11286*

Global rising in the levels of multidrug resistance among bacteria has been an emerging problem of public health. These bacteria have a considerable impact on the

mortality associated infectious diseases [Ramos-Castaneda et al., 2018], [Christaki et al., 2019]. The multidrug-resistant *K. pneumoniae* strain HS11286 was isolated by taking human sputum samples from patients in 2011 at Huashan Hospital, Shanghai, China. Its whole genome was sequenced by Liu and his colleagues and the genome size of the bacteria is about 5.3 Mb. They found that *K. pneumoniae* HS11286 contains seven circular replicons, one chromosome and six plasmids including three multidrug resistance plasmids. Of the multidrug resistance plasmids, pKPHS1 codes for extended-spectrum betalactamase (ESBLs), pKPHS2 carries the blaTEM-1 gene and the carbapenemase gene blaKPC-2, and pKPHS3 has 13 important resistance determinants, such as tetG, cat, sul1, dfra12 [Liu et al., 2012]. This strain possesses carbapenemase, an enzyme which suppresses the activation of an antibiotic agent, carbapenem, as a last hope resource. Therefore, it is called Carbapenem-Resistant or Carbapenemase-Producing Enterobacteriaceae (CR-CPE) [Bi et al., 2015].

There are two mechanisms for the bacterial multidrug resistance. One is that bacteria can have multiple genes that have coding regions for the drug resistance. Another mechanism takes place when there is an elevation in the expression of genes that are responsible for multidrug efflux pumps, enzymatic inactivation or other related mechanisms [Nikaido, 2009]. This pathogenic strain is a representative ST11 lineage and pan-drug-resistant, meaning that it has resistance to all available classes of antimicrobials [Falagas et al., 2008], [Bi et al., 2015]. Furthermore, *K. pneumoniae* strains possess an extremely plastic genome, which refers to the ability to adapt to different conditions with mobile genetic elements such as plasmids, transposons, genomic islands and so on [Capy, 1998], [Bi et al., 2015]. Therefore, bacteria can easily transfer their plasmids carrying the antibiotic resistant genes, and/or virulence-associated genes can spread within species by the conjugation, which is an important mechanism of gene transfer via direct physical contact, and so, their control for resistance is really difficult [Liu et al., 2012], [Christaki et al., 2019]. The reason of the antimicrobial persistence can be highly adequate plasmid transfer rate resulting in favorable plasmid maintenance [Christaki et al., 2019]. All these drug resistance mechanisms make it very challenging to find an efficient therapy for the infectious diseases with low survival rate.

K.pneumoniae HS11286 is an important bacterial pathogen, commonly associated with opportunistic and hospital-acquired infections with limited treatments [Liu et al., 2012], [Campos et al., 2016]. K. *pneumoniae* causes the infections of the lower digestive tract, lower respiratory tract, wounds-surgical sites, bloodstream, and urinary tract [Campos et al., 2016].

Klebsiella species can be present in various environments, from surface water, soil, plants to mucosal surfaces of humans. The bacteria are gram-negative and rod-shaped. Since they are facultative anaerobes, they have capability to utilize both fermentative and respiratory metabolisms [De Jesus et al., 2015]. They reproduce specifically in mucoid colonies on carbohydrate rich media and can be cultured on nutrient agar, tryptic casein soy agar, bromocresol purple lactose agar, Drigalski agar, MacConkey agar, eosin-methylene blue (EMB) agar and bromothymol blue (BTB) agar. They can be identified and differentiated within species based upon the biochemical reactions in different carbon source utilization. L-arabinose, D-arabitol, D-cellobiose, citrate D-fructose, D-galactose, D-glucose, 2-ketogluconate, maltose, D-mannitol, D-melibiose, D- raffinose, D-trehalose and D-xylose, lactose and D-sorbitol can be utilized as a carbon source by all *Klebsiella* strains [Brenner et al., 2005]. *K. pneumoniae* can grow on glycerol fermentatively under the favor of the enzymes involved in the glycerol dissimilation pathway [De Jesus et al., 2015]. A conventional fermentation of glucose in most *Klebsiella* strain yields acid and gases in addition to acetoin and 2,3-butanediol as a major end product, according to Voges-Proskauer test [Drancourt et al., 2001], [Brenner et al., 2005].

2.5. *Prevotella copri DSM 18205*

The human microbiota that belongs to ‘reference man being between 20-30 years of age, weighing 70 kg, is 170 cm in height’, consists of 39 trillion bacterial cells. These bacteria predominantly inhabit colon (large intestine), dental plaque, saliva, and lower small intestine [Sender et al., 2016]. The term “human microbiome” refers to the genome collection of the microbial community in human, and it is an important research area to reveal the physiology and metabolism of the host organism and the bacterial flora. Microbial phylogenetic diversity analysis is performed mainly with the conserved 16S

ribosomal RNA genes in order to determine the classification of human microbiome. Firmicutes, Bacteroidetes, and Actinobacteria are the main phyla in healthy human gut microbiota. [Hamer et al., 2012]. There is a vast majority of anaerobic bacteria compared to aerobic bacteria in human gut flora. The obligate anaerobic bacteria e.g. Bacteroidetes reach the highest concentration at adulthood while humans at birth contain the facultative anaerobic bacteria e.g. E.coli at most [Sears, 2005]. *Provetella* is one of the two domain genera of Bacteroidetes in human microbiota [Ley, 2016]. *P.copri* is generally more abundant in human gut even if *Provetella* species are mainly isolated from oral cavity. *Prevotella copri* DSM 18205 strain is gram-negative, obligate anaerobe, non-pigmented, non-spore-forming, non-motile, and rod shaped bacteria [Hayashi et al., 2007].

Gut flora provides many benefits on human health such as polysaccharide utilization and nutrient release, enhancement of the fat storage, contribution to mucosal homeostasis and repair capacity, and induction of mucosal glucose transporters, selected proteins of innate immunity and villous capillary formation [Sears, 2005]. They altogether act like a bioreactor and produce short chain fatty acids (SCFAs), certain amino acids and vitamins by fermenting the nutrients into beneficial metabolites [Kumar et al., 2018]. There are two main types of colonic microbial fermentation. Carbohydrate fermentation, called saccharolytic fermentation, produces SCFA as a fermentative metabolite while the fermentation of proteins, called proteolytic fermentation, yields p-cresol, phenol, ammonia, or H_2S.

Some of the end products show protective action against the colonic epithelium and others can be proinflammatory or pro-carcinogenic metabolites. These fermentative bacteria can be distinguished from each other by targeting their specific end products [Hamer et al., 2012]. *Prevotella copri* DSM 18205 characteristically can perform D-mannose and L-rhamnose fermentation [Hayashi et al., 2007]. The gene repertoire of the bacteria indicates that it contains genes that contribute to glucose tolerance through the host glycan degradation as a gut fermenter [Ley, 2016]. This strain has a major role in the cellular fatty acids synthesis and also possesses dimethyl acetals and predominant menaquinones. Acid production by *P.copri* is possible through the utilization of glucose, lactose, sucrose, maltose, raffinose, salicin, D-xylose, L-arabinose, D-cellobiose and L-rhamnose [Hayashi et al., 2007]. The genomic profile of *P.copri* shows that it can carry

out the route of EMP pathway. Nevertheless, the bacteria is not capable of activating the oxidative branch of the pentose phosphate pathway, and it does not have the necessary enzymes for KDPG aldolase and 6-phosphogluconate dehydratase activities, which are a part of the Entner-Doudoroff pathway. The metabolic profile analysis, e.g. enzyme activity measurements, demonstrates that succinate formation from fumarate, acetate and formic acid production from degraded pyruvate are the products of glucose fermentation (glycolysis) [Franke and Deppenmeier, 2018].

The human gut microbiota contributes positively to human health and development. However, the alteration in gut microbiota, called dysbiosis, is responsible for various diseases including diabetes, inflammatory bowel disease, atopic diseases, cancer, obesity, malnutrition (stunting and wasting), and neurodegenerative disorders such as Parkinson's disease, Alzheimer's disease, multiple sclerosis and amyotrophic lateral sclerosis [Hamer et al., 2012], [Kumar et al., 2018], [Roy Sarkar and Banerjee, 2019].

The relation of plant-rich and high carbohydrate diet with *Prevotella* is commonly noted by many studies and, therefore, they are proposed as beneficial microbes [Ley, 2016]. Furthermore, they are importantly linked to high-fat diet and consequently to the obesity with insulin resistance [Precup and Vodnar, 2019]. On the other hand, *Prevotella* species are mainly linked to HIV and chronic inflammatory diseases, and they also increase the susceptibility to colitis [Ley, 2016]. It is reported that there is a relatively high abundance of *P. copri* populations in the human stool samples of the rheumatoid arthritis (RA) patients. Therefore, *P. copri* is associated with a higher risk for the pathogenesis of RA by affecting the immune cell activation [Pianta et al., 2017], [Alpizar-Rodriguez et al., 2019]. Many studies have reported that *P. copri* demonstrates low abundance in human gut microbiota of Parkinson’s disease (PD) patients compared to healthy controls. The bacteria generates neurotoxic products, and, consequently, an immune response is triggered in the brain by microglial cells. The activation of microglial cells and motor symptoms are associated with inflammation-induced misfolding of a-Syn proteins and intestinal permeability, promoting PD progression [Keshavarzian et al., 2015], [Scheperjans et al., 2015], [Bedarf et al., 2017], [Petrov et al., 2017].

2.6. Genome-scale Metabolic Network Model

In silico genome-scale metabolic network models (GEMs) are mathematical representation of biochemical transformations (reactions) and associated molecular components (enzymes, substrates, and products) that refer to the metabolic capabilities of a metabolism. This modeling approach covers all the metabolic pathways that occur in a cell. Modeling approaches in order to simulate cell behavior need manual efforts and are labor-intensive but they facilitate the understanding of a complex system. Further, the metabolic model validation and analysis provide an insight on the functional capacity of the organism in question such as growth yield, production of a metabolite, network robustness, and gene essentiality [Arakawa et al., 2006], [Durot et al., 2009]. Until February 2019, GEMs have been reconstructed for 6239 organisms (5897 bacteria, 127 archaea, and 215 eukaryotes) [Gu et al., 2019]. The first genome-scale metabolic model, iJE296, was reconstructed in 1999 for *Haemophilus influenzae* Rd KW20 strain, and the model contains 488 metabolic reactions operating on 343 metabolites driven by 296 genes [Edwards and Palsson, 1999], [Gu et al., 2019]. *In silico* metabolic genotype for Escherichia coli, a model organism, is the second reconstructed bacterial metabolic network model in literature. Its GEM is called *i*JE660 and consists of 627 reactions catalyzed by 660 genes [Edwards and Palsson, 2000]. Since then, many reconstructions for the common laboratory strain *Escherichia coli* K-12 MG1655 have been periodically released based on new approaches, curated genomic and biochemical knowledge. An updated last version of the model, named *i*ML1515, accounts for 1515 genes, 2719 metabolic reactions, and 1192 unique metabolites [Monk et al., 2017].

2.7. Applications of Genome-scale Metabolic Network Models

It is important to demonstrate metabolic interactions within bacteria to elucidate molecular mechanisms and consequently design novel treatments for diseases. Computational systems biology approaches can give further insights by constructing data-driven networks. Hence, the application scope of GEMs has expanded considerably [Gu et al., 2019].

The bacterial GEMs are commonly used for metabolic engineering strategies. For example, *E.coli* GEM can be used to improve fermentative production of bio-based industrial chemicals and fuels from alternative carbon sources [Mienda, 2017]. They can be also used in food and nutrients industry in order to improve the yield of fermentation by-products [Xu et al., 2013]. The effects of the modification on the bacterial metabolism by gene deletions can be predicted in silico by using GEMs, leading to a list of candidate gene deletions that can improve the industrial production [Navid, 2011].

The genome-scale metabolic models of microbes are important also for pharmaceutical purposes. They can be used for the improvement of the production of antibiotics such as penicillin, cephalosporin, and tetracycline by natural or engineered microbes [Xu et al., 2013]. In addition to that, GEMs of pathogenic bacteria can be practiced for the systems biology based analysis of bacterial metabolism about the infection process as well as host–pathogen interactions. This can guide the identification of important pathways or metabolites or genes that can be targeted for drug development [Navid, 2011], [Cesur et al., 2020], [Gu et al., 2019].

GEMs can be also utilized to elucidate metabolic mechanisms in cells under disease conditions, and to suggest effective therapeutic targets. They can be used for many human diseases such as cancer, neurodegenerative diseases, or diabetes [Gu et al., 2019].

2.8. Genome-scale Metabolic Modelling Process

High-throughput sequencing technologies led to huge datasets, and systems biology tools are needed for explicating such data. A highly detailed protocol is present for *in silico* genome-scale metabolic network reconstruction for organisms [Thiele and Palsson, 2010]. The quality of a metabolic network model reconstruction depends on having a complete genome with an adequate annotation, following the reconstruction procedure in detail and having sufficient available data about the target organism. High-quality GEMs can be built based on four stages: draft reconstruction, manual refinement process, metabolic model conversion into a mathematical model and the validation of the model [Thiele and Palsson, 2010], [Haggart et al., 2011]. The process is summarized in Figure 2.9 [Feist et al., 2009].

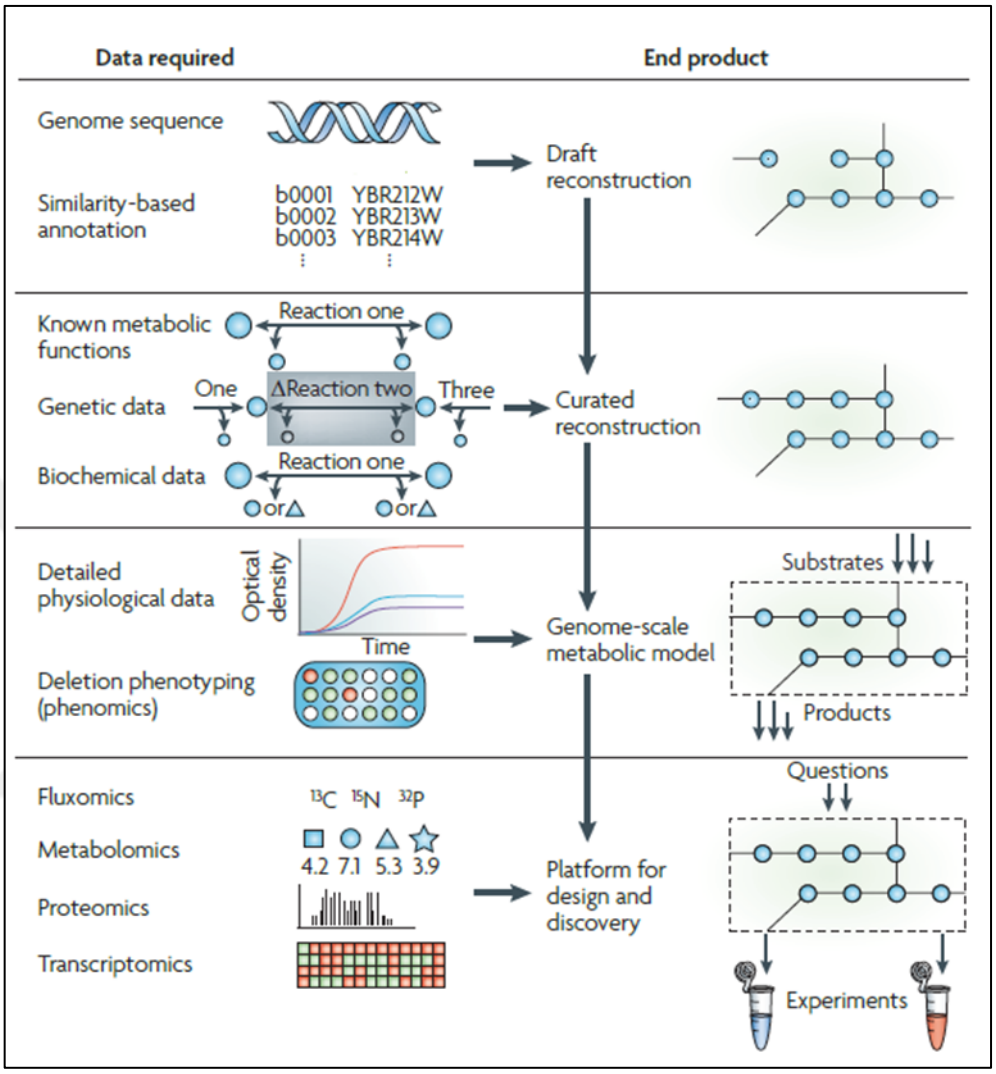


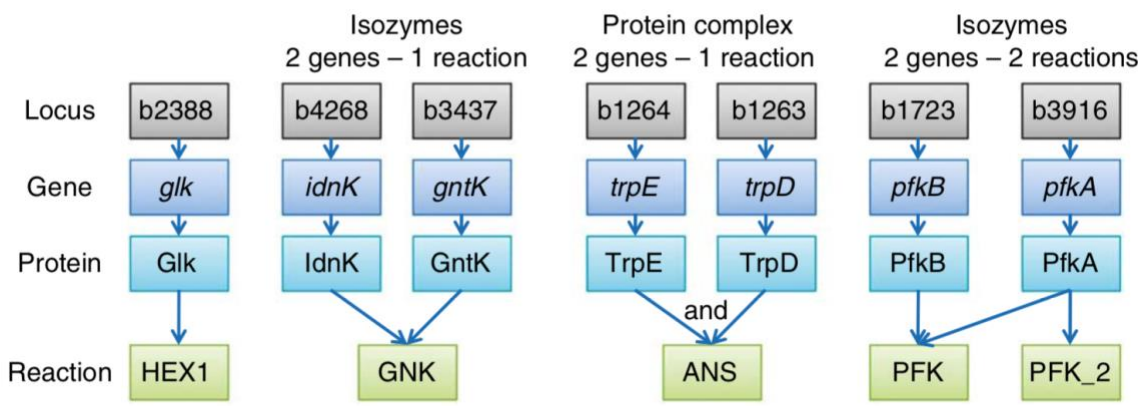
Figure 2.9: The genome-scale metabolic network reconstruction process in four phases.

2.8.1. Stage 1: Draft Reconstruction

Draft reconstruction process holds two main constituents: an annotated genome and biochemical databases. The annotated genome is important for giving information about the features and functionality of a gene. The annotated genomes can be obtained from resources such as BASys (Bacterial Annotation System), Genomes Online Database (GOLD), Integrated Microbial Genomes, Ensembl Genomes, NCBI Entrez Gene, Uniprot, UCSC Genome Bioinformatics, and Rapid Annotations using Subsystems

Technology (RAST) server [Thiele and Palsson, 2010], [Haggart et al., 2011]. The draft reconstruction can be performed in an automatized manner. The information of the enzymes, proteins with a catalytic function, encoded by genes present in bacteria is provided by the genome annotation. Gene-protein-reaction (GPR) associations indicate how the gene products collaborate for active enzymes, protein complexes or isoenzymes in order to catalyze reactions (Figure 2.10) [Feist et al., 2009], [Thiele and Palsson, 2010]. Each enzyme then is assigned to a set of biochemical reactions obtained from biochemical databases such as KEGG, BRENDA, PubChem, Metacyc, Biocyc, Transport Classification Database (TCDB), PSORT. Therefore, draft genome-scale metabolic network is constructed basically by obtaining candidate gene list and the associated reactions [Thiele and Palsson, 2010]. The draft model contains substrates and products that participates in reaction(s), the stoichiometric coefficients for each metabolite, reaction reversibility, and compartment information that shows where the reaction occurs in the cell [Durot et al., 2009].

Alternatively, when generating a draft model, genetic information from a phylogenetically close organism can be utilized by homology/orthology predictions [Wang et al., 2018], [Machado et al., 2018]. The draft reconstruction process can be automated by reconstruction-specific bioinformatic tools such as ModelSEED [Henry et al., 2010], Raven Toolbox [Wang et al., 2018], PathwayTools [Karp et al., 2016], CarveMe [Machado et al., 2018] and so on [Gu et al., 2019]. However, a draft model, as the name suggests, is still not fully completed. There may be inadequate or missing information about the metabolism, pathways or GPR rules of the organism in question. In addition to that, incompatibility and erroneous information within the model makes the model less accurate and non-functional. Hence, the draft model needs further manual curation [Thiele and Palsson, 2010].



Reaction abbreviation	Reaction name	E. C.number	GPR
HEX1	Hexokinase (D-glucose:ATP)	2.7.1.1	(b2388)
GNK	Gluconokinase	2.7.1.12	(b3437) or (b4268)
ANS	Anthranilate synthase	4.1.3.27	(b1264) and (b1263)
PFK	Phosphofructokinase	2.7.1.11	(b1723) or (b3916)
PFK_2	Phosphofructokinase (2)	2.7.1.11	(b3916)

Figure 2.10: Representation of Gene-protein-reaction (GPR) rules.

2.8.2. Stage 2: Manual Curation / Refinement

Draft model can be automatically retrieved from databases in the form of a set of candidate metabolic genes and biochemical reactions encoded on a genome. But, it may still lack certain organism-specific features since it is based on genes common in other phylogenetically related organisms [Feist et al., 2009]. Each reaction in the draft model should be checked for its necessity for the model, for accurate reaction stoichiometry, reaction directionality, localization, GPR associations and the role in metabolite production, usage and recycling [Haggart et al., 2011].Therefore, the second stage of the reconstruction process, the manual curation and refinement of the metabolic network, is the most time-consuming and laborious task. It needs comprehensive organism-specific literature review and manual evaluation [Thiele and Palsson, 2010].

The draft model has some metabolic gaps because of incompleteness. The primary reason behind this is the fact that not all the reactions are associated with a gene or protein [Feist et al., 2009]. Therefore, some crucial reactions with no gene/protein association are

not added to the draft model in the first step of the reconstruction. Identification of the candidate reactions for gap-filling is an important step in the reconstruction process, with the intent of having a fully functional model. In an attempt to figure out the gaps in the model, some approaches have arisen. The first introduction of automatic gap-filling process is an optimization-based algorithm built upon resolving the disagreement of the computational and experimental results through the prediction of missing reactions [Reed et al., 2006]. Some other developed tools or algorithms to predict missing reactions are *GapFind* or *GapFill* procedure with bottom-up approaches [Satish Kumar et al., 2007], *fastGapFill*, a COBRA toolbox built-in function [Thiele et al., 2014], *fillGaps*, a RAVEN toolbox built-in function [Wang et al., 2018], and Meneco, a topology-based method [Prigent et al., 2017]. The elimination of metabolic gaps ensures the improvement of draft reconstructions by fixing missing information from metabolic databases and/or genome annotations [Feist et al., 2009].

In addition to gap filling step, the following steps are strongly recommended in the refinement of a draft genome-scale metabolic network:

- Reactions containing generic terms such as protein, DNA, or electron acceptor should not be included.
- The charge balance of the metabolites based on the pH of the metabolic network requires manual refinements.
- The directionality of metabolic reactions can be reversible or irreversible. In general, if there is no available information, reactions can be assumed to be reversible. However, too many reversible reactions in the model can cause futile cycles. Non-accurate assignment of the reaction direction has considerable impact on the model simulations.
- The localization of proteins and reactions is another important step. There are many tools and databases for the cellular localization prediction based on nucleotide or amino acid sequences such as WoLF PSORT [Horton et al., 2007], ClubSub-P [Paramasivam and Linke, 2011], and PSORTb 3.0 [Yu et al., 2010]. PSORTb was specifically developed for bacterial localization prediction. Since bacteria do not have cellular compartments like eukaryotes, the reactions can be localized mainly in three compartments; cytosol, periplasm or extracellular matrix.

- Reaction stoichiometry is necessary to balance the mass of every compounds in the model.
- In prokaryotes, cytoplasm creates a separation between intracellular and extracellular environments, which limits the transport of dissolved components. Transport/exchange reactions are important for connecting extracellular and intracellular environments.
- The refinement of GPR associations is also crucial. One important point is that the reactions named as 'orphan' are not assigned to any representative genes and require both computational and experimental approaches to be confirmed.
- Biomass reaction is a non-enzymatic reaction that represents cellular growth requirements with the partial contributions of the biomass constituents. Assessment of organism-specific biomass composition, when possible, and adding a biomass reaction are crucial steps.
- Growth and non-growth associated ATP maintenance reactions (GAM/NGAM) should be determined and added. The NGAM reaction is ATP hydrolysis reaction that represents the energy for cell maintenance while the GAM reaction corresponds to the energy required for the reproduction of a cell e.g. macromolecular synthesis [Thiele and Palsson, 2010].

2.8.3. Stage 3: Conversion into a Mathematical Model

The mathematical representation of a metabolic network model is a necessary step to perform metabolic network analysis. The modeling approach is also known as constraint-based analysis. Here, the relationship between genotype and phenotype is described mathematically and computationally based on the constraints of the biochemical system [Heirendt et al., 2019]. Therefore, GEM can be converted into a mathematical model by using the popular COBRA toolbox or COBRApy. COBRA Toolbox v.3.0 is with a package developed in MATLAB while COBRApy is a python-based package for constraints-based modeling of biological networks [Heirendt et al., 2019], [Ebrahim et al., 2013]. Systems Biology Markup Language (SBML) is the commonly used standard format to be able to publish and distribute the GEMs [Hamilton and Reed, 2014]. COBRA

Toolbox transforms the input SBML model into a structure (Figure 2.11), and the information in the separated fields is stored in submatrices [Thiele and Palsson, 2010].

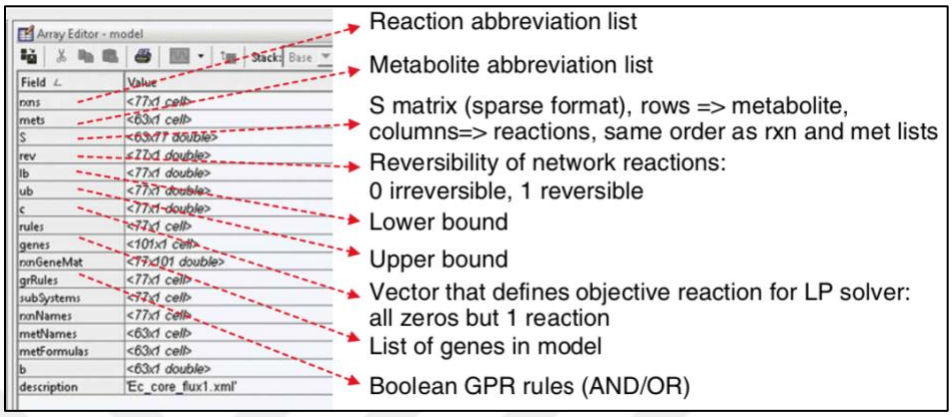


Figure 2.11: A representation of SBML model structure.

2.8.4. Stage 4: Model Validation

The final stage of GEM reconstruction includes network evaluation and validation using experimental data. In this stage, additional gap-filling could be performed for debugging the metabolic network. The identification of metabolic dead-ends, metabolites that cannot be either consumed or produced, or blocked reactions that cannot carry any flux in any simulation conditions can be the first step for network evaluation. These metabolites and reactions highlight metabolic gaps [Thiele and Palsson, 2010], [Hamilton and Reed, 2014]. Testing synthesis of each separate biomass precursor or by-product secretion is the part of the debugging process, and they can be the reason behind blocked reactions [Hamilton and Reed, 2014]. *In silico* quantitative predictions are made with GEMs using the mathematical representation (Stage 3) for the validation of GEMs. The growth rate and/or single-gene deletion phenotype predictions are among the commonly used simulations in the validation process. Afterward, the computational predictions are compared against experimental data. If there is inconsistency between the phenotypes of the modeled organism and the mathematical model, the iteration of steps in stage 2 and stage 4 may be necessary. Therefore, model improvement with a repeated iteration cycle

is necessary to have a fully functional genome scale metabolic model [Thiele and Palsson, 2010], [Hamilton and Reed, 2014].

2.9. Reconstruction Tools & Softwares

Genome-scale metabolic models (GEMs) are a powerful tool in System Biology due to their prediction ability. Since GEMs are becoming popular, several bioinformatic tools/softwares have been developed for GEM reconstructions. There are hundreds of models reconstructed by using these tools for notable microorganisms from important human pathogens to the industrially relevant species [Faria et al., 2018], [Mendoza et al., 2019]. The automated reconstruction tools facilitate the reconstruction process by accelerating draft reconstruction, and supporting and/or providing advanced guide for manual curation and gap filling processes [Mendoza et al., 2019]. Table 2.1 summarizes the comparison of the recent and commonly used GEM reconstruction tools.

Table 2.1: A comparison of software & tools used in GEM reconstructions.

FEATURES	**ModelSEED**	**merlin**	**RAVEN Toolbox 2.0**	**CarveMe**	**MetaDraft**
Input	Genome sequence	Genome sequence or Taxonomy ID	Annotated genome or template model(s)	Universal model, template model	Annotated genome or template model(s)
Programming Language	Web-based	Java	MATLAB	Python	Python
Graphical User Interface (GUI)	Stand-alone website	Stand-alone interface	Command line	Command line	Stand-alone interface

Table 2.1: Table of continued.

Reference Database	SEED	KEGG, MetaCyc, UniprotKB, TCDB	KEGG, MetaCyc	BiGG database	BiGG database
Mapping Method	RAST Annotation	BLAST, HMMER	Bidirectional BLASTP, Diamond	Diamond, eggNOG-mapper	Autograph (Inparanoid)
Gap-filling function	Yes	No	Yes	Yes	No
Assignment of sub-cellular localization	No	Yes	Yes	Yes	No
Simulation support	Yes	No	Yes	Yes	No
Output	SBML, Excel	SBML	SBML, Excel	SBML	SBML
References	[Henry et al., 2010]	[Dias et al., 2015]	[Wang et al., 2018]	[Machado et al., 2018]	[Hanemaaijer et al., 2017]

ModelSEED is the first bioinformatic tool to generate draft models. It has the ability of automating the gap filling step based on medium growth, and it allows to perform flux balance analysis and phenotype simulations. It performs genome annotation itself via RAST system [Faria et al., 2018], [Mendoza et al., 2019]. ModelSEED is user-friendly and the fastest-automated tool for generating GEM. Therefore, some studies like AGORA reconstruction for microbial community used ModelSEED for automatic reconstruction of 773 members of the human gut microbiome [Magnusdottir et al., 2017].

merlin (metabolic models reconstruction using genome-scale information) takes nucleotide and amino acid sequence or taxonomy ID as input. *merlin* is a powerful tool characterized by the annotation. It contains several tools for enzyme and transporter annotation and compartmentation prediction [Dias et al., 2015], [Mendoza et al., 2019]. *merlin* and ModelSEED support a basic visualization of metabolic networks, drawing active routes on top of KEGG maps [Faria et al., 2018].

Raven Toolbox version 2.0 (Reconstruction, Analysis and Visualization of Metabolic Networks) provides a semi-automated refinement process with many MATLAB functions. Since the source-code of the functions is open, users can modify/improve its codes [Wang et al., 2018], [Mendoza et al., 2019].

CarveMe was introduced as a unique top-down approach for GEM generation from BIGG-based manually curated universal template reaction set. The tool has its own gap-filling algorithm that principally includes reactions with higher genetic evidence [Mendoza et al., 2019].

MetaDraft uses CBMPy, a python-based software package [Hanemaaijer et al., 2017], [Olivier, 2018]. The tool incorporates BiGG models as an internal database, or users can load specific template models. The resulting GEM has the latest features of current SBML standards. It has functions to make manual curation steps semi-automatic [Mendoza et al., 2019].

As a consequence, there are various GEM reconstruction tools available in the literature. However, they have advantages and vulnerabilities in different tasks, and none of the tools has performed better than others in all categories.

2.10. Constraint-Based Analysis of Metabolic Networks

2.10.1. Flux Balance Analysis

Genome-scale metabolic networks with the constraint-based approach is a powerful tool for the system biology of metabolism in order to better understand and predict the behavior of a biochemical network under various environmental and genetic disturbances [Haggart et al., 2011]. Constraint-based approaches do not require kinetic information.

Hence, their flux rate prediction is based on steady-state conditions. The approach is basically the prediction of phenotypes from the genotype of an organism [Lewis et al., 2012], [Bordbar et al., 2014].

Flux Balance Analysis (FBA) is the first and widely used mathematical approach to study genome-scale metabolic networks. FBA analysis depends on the mass balance equations around each intracellular metabolite in the metabolic network at steady state conditions. Firstly, FBA needs a mathematical representation of the network, and stoichiometric matrix, S, provides the need in the form of a numerical matrix via the stoichiometric coefficients of the metabolites participating in a reaction [Orth et al., 2011]. Stoichiometric matrix, S, is the matrix representation of mass balance equations at steady state. The toy example below explains how S is constructed.

A toy metabolic network represented in Figure 2.12 consists of A,B,C,D,E,F, and G metabolites where A,B, and C are the metabolites that are taken up and passed through cell boundary [Lewis et al., 2012]. The six metabolites in the network are catalyzed by the reactions represented by V_1, V_2, V_3, V_4, V_5, b_a, b_b, b_c, and b_G, where b_a, b_b, b_c, and b_G are the exchange reactions. There are both reversible and irreversible reactions in the network. V_4 and V_5 are reversible while others are irreversible reactions.

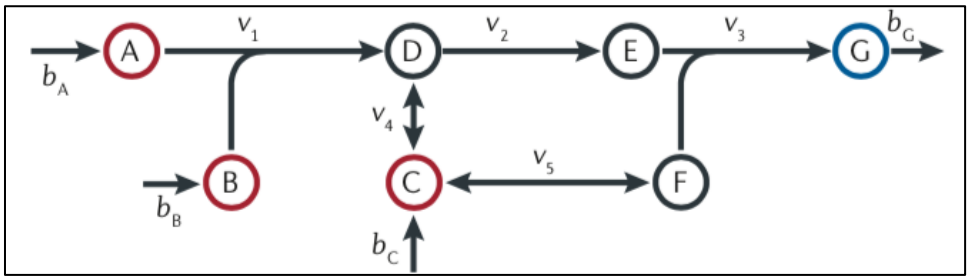


Figure 2.12: A toy metabolic network.

The network can be converted into a mathematical model in the form of a stoichiometric matrix. The rows represent the balanced metabolites and the columns represent the reactions in the toy network. The S matrix is obtained by multiplying the coefficients of the metabolites in the associated reaction by -1 if these coefficients are consumed and 1 if produced. These are showed in Figure 2.13 [Lewis et al., 2012].

	v_1	v_2	v_3	v_4	v_5	b_A	b_B	b_C	b_G
A	-1	0	0	0	0	1	0	0	0
B	-1	0	0	0	0	0	1	0	0
C	0	0	0	-1	-1	0	0	1	0
D	1	-1	0	1	0	0	0	0	0
E	0	1	-1	0	0	0	0	0	0
F	0	0	-1	0	1	0	0	0	0
G	0	0	1	0	0	0	0	0	-1

Figure 2.13: A representation of Stoichiometric matrix for the toy metabolic network in Figure 2.12.

FBA approach is based on some constraints. The stoichiometry in the matrix indicates the flux (mass) balance constraints on the system since the total rates of reactions consuming a metabolite must equal to the total rates of reactions producing that metabolite for intracellular metabolites. Each reaction flux is described by upper and lower bounds that characterize the maximum and minimum possible flux rate. Reaction rates can be constrained by the measurement constraints, which are experimentally measured fluxes [Orth et al., 2011]. The thermodynamic constraint that indicate the kinetic limitations of a particular enzyme or membrane transporter is defined according to reaction reversibility. The upper bounds (ub) of the reversible or irreversible reactions can be infinite, mostly defined as 1000 for practical reason. The lower bounds (lb) of reversible reactions can take minus infinite or -1000, indicating reverse direction, while the lower bounds of irreversible reactions are assigned to zero [Haggart et al., 2011].

The vector of the flux rates is represented by v, which has a length of n, the number of reactions in the network. All metabolite concentrations are represented by the vector x, with length m, the number of metabolites. The flux balance of a metabolite at a given time is calculated by the derivative (dx/dt). The system of mass balance equations at steady

state must equal to zero with no accumulation for the flow of metabolites, as given in equations (Eqn 2.2 and Eqn 2.3) [Orth et al., 2011].

$$\frac{dx}{dt} = S \times v \tag{2.2}$$

$$S \times v = 0 \tag{2.3}$$

After these constraints are set, objective function, a mathematical representation of cellular objective, is needed to be defined. If the objective function is the maximization of growth rate, it means that the aim of the cells of the organism of interest is maximum biomass production, equivalent to maximizing the rate of the conversion of metabolites into biomass components such as proteins, lipids and nucleic acids. The rate of uptake metabolites are defined in units of per gram dry weight of the cell per hour (gDW^{-1} hr^{-1}), which results in a predicted exponential growth rate in units of hr^{-1}. *In silico* prediction of maximum growth rate in order to simulate biomass production is a linear optimization problem and can be calculated by many computational linear programming (LP) interfaces present in COBRA toolbox [Orth et al., 2011], [Heirendt et al., 2019]. The calculated flux distributions show a set of reaction fluxes at steady-state condition, which satisfy the mass balance, measurement and reversibility constraints [Lewis et al., 2012].

The flux solution space is defined by designating certain assumptions (constraints). The final predicted flux distribution is a single optimal point in the solution space obtained by solving an optimization problem (Figure 2.14) [Xu et al., 2013].

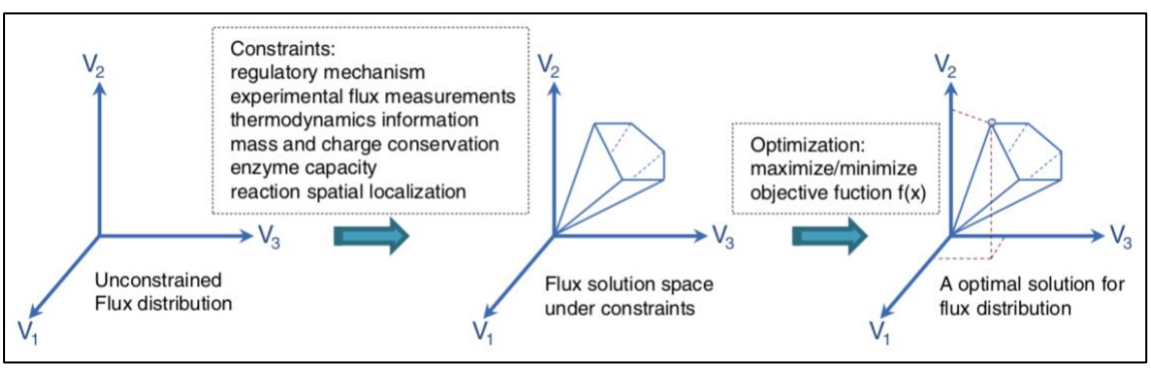


Figure 2.14: The flux distribution space by constraints-based analysis simulation.

FBA approach is usually performed in order to identify essential genes for growth, to discover novel drug targets and to predict the maximum yield of targeted byproducts based on different objective functions [Haggart et al., 2011].

2.10.2. Minimization of Metabolic Adjustment

Constraint-based analysis provides a prediction from genotypes to phenotypes, leading to precise insights into the biological mechanisms. Further, Minimization of Metabolic Adjustment (MOMA) is established for the simulation of genetically disturbed organisms in terms of metabolic flux prediction. In such organisms, there is a deviation in their flux rates compared to wild type (WT). MOMA hypothesizes that there should be minimum change between the genetically perturbed, knock-out (KO), flux distribution and WT flux distribution due to the metabolic adaptation of the organisms. That is, the perturbed organisms try to keep the change in their metabolism at minimum rather than maximizing their growth rates. Therefore, MOMA performs distance minimization in flux space to predict metabolic phenotypes of such organisms (Figure 2.15) [Segre et al., 2002].

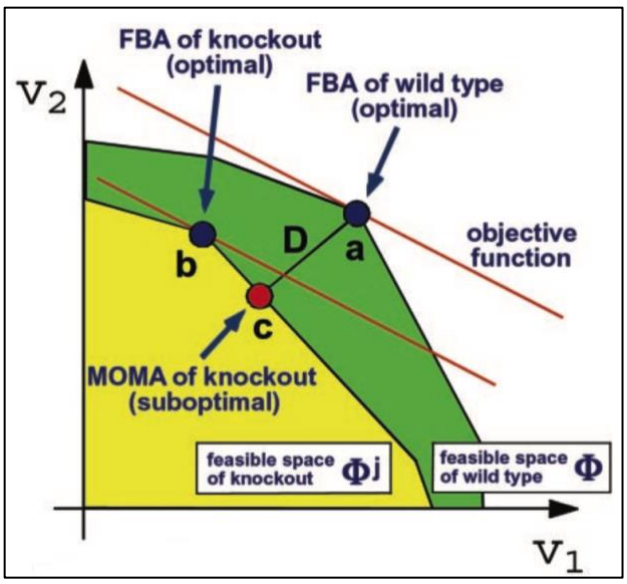


Figure 2.15: The solution space of minimization of metabolic adjustment.

The new predicted flux vector (v) is a suboptimal solution between WT (v_{FBA}, Eqn 2.4) and KO flux distributions. Thereby, since it is a minimal Euclidean distance problem, the algorithm uses quadratic programming (QP) to predict the flux distribution of the reaction set (Eqn 2.5). MOMA is still based on the standard FBA mass balance constraints and other FBA constraints (reversibility and measurement), with only difference being the choice of objective function. Also, all KO gene-associated reaction fluxes (v_k, Eqn 2.6) are made equal to zero in MOMA approach [Haggart et al., 2011].

$$D(v^{FBA}, v) = \sqrt{\sum_{j=1}^{N} (v_j^{FBA} - v_j)^2} \qquad \forall j \in N \tag{2.4}$$

$$\min(v^{FBA} - v)^T \, (v^{FBA} - v) \tag{2.5}$$

$$v_k = 0, \qquad \forall k \in A \tag{2.6}$$

MOMA also exhibits more accurate results than standard/classical growth maximization simulation result for lethal gene prediction [Segre et al., 2002].

3. GENOME-SCALE METABOLIC MODEL RECONSTRUCTION FOR BACTERIA

Genome sequencing leads to huge datasets, which need to be analyzed to extract useful information. Genome annotation provides an understanding of the genomic data. It is also valuable to illustrate metabolic interactions within the organism at genome-scale. Genome-scale metabolic models are powerful tools for *in silico* simulations of the bacterial metabolic networks. Systems biology approaches, bioinformatic tools and databases are used for this purpose, and they provide an insight on the functional capacity of the organism.

In this study, genome-scale metabolic network models for *Klebsiella pneumoniae subsp. pneumoniae HS11286* and *Prevotella copri DSM 18205* were reconstructed by using RAVEN Toolbox 2.0, the MATLAB-based reconstruction tool, in order to investigate bacterial metabolisms and clarify the mechanisms of the diseases that these bacteria are responsible for. It is a good strategy to combine a semi-automatic tool with manual curation to get more accurate GEMs. The main steps of GEM reconstruction are (i) draft metabolic model reconstruction based on protein similarity via BLASTP between the target organism and a genetically similar species for which a genome-scale metabolic reconstruction is available, (ii) manual evaluation of the draft model with the application of gap-filling strategies, and (iii) the metabolic model validation by performing the constraint-based approaches. In addition, if obtained model needs further manual curation, repetition of those steps is required in order to get a final model. The GEM reconstruction process is summarized in Figure 3.1.

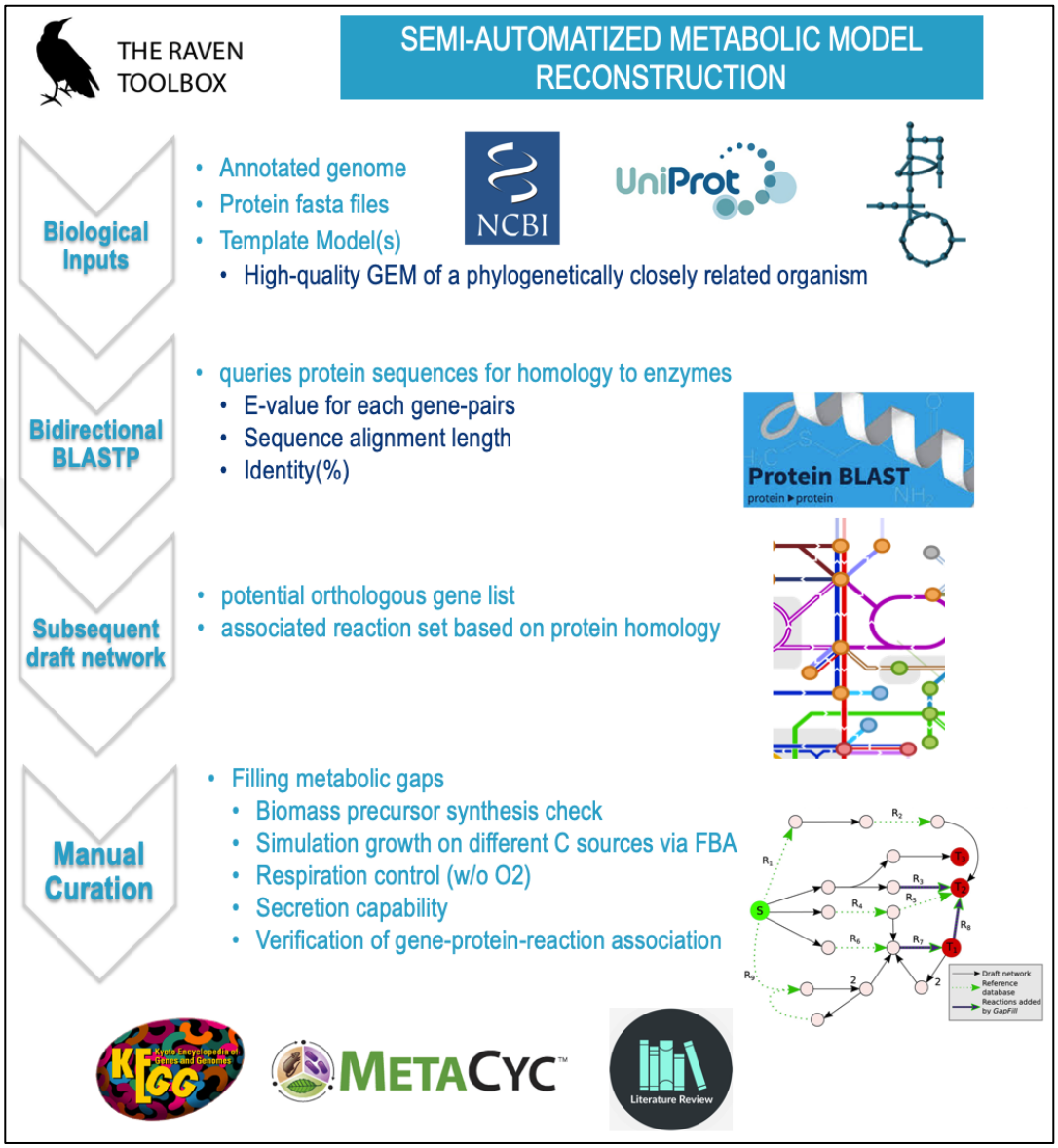


Figure 3.1: The illustration of the overview of the GEM reconstruction process.

The draft metabolic models for these two bacteria were created automatically by using RAVEN toolbox, which can produce a draft model in SBML format. These draft models were imported into MATLAB with COBRA Toolbox, a package in MATLAB. All other implementations to the models were performed in MATLAB (version R2018b) by utilizing the functions from RAVEN toolbox v.2.0 and COBRA Toolbox v.3.0. Gurobi 8.1.1 was used as a powerful mathematical optimization problem solver in simulations,

with both linear programming and quadratic programming features [Gurobi Optimization, 2018].

3.1. Draft Metabolic Model Reconstruction

The draft metabolic model was reconstructed by providing two inputs: (i) the annotated genome of the organism for which metabolic model reconstruction will be performed and (ii) an existing curated genome-scale metabolic network model of a genetically similar species, referred also as template model. The first step to create a draft network is to import the template model into RAVEN toolbox and confirm that the model is functional. The template models used in this study to reconstruct draft models for the two bacteria were double checked whether they have duplicate reactions, metabolites or genes, and if any duplications were found, they were eliminated. In addition, *grRules* of template models had be standardized based on the required format from RAVEN. If there were any missing fields in the model format, new model fields were integrated into the template model. For example, the 'metComps' field, which indicates in which compartments of the cell the metabolites are present in the model, is necessary for the functions we used in this study, and therefore it was created for the *i*AH991_norm model, a GEM for *Bacteroides thetaiotaomicron VPI-5482*.

Protein sequence homology is the similarity in protein sequences due to shared ancestry between a pair of genes in different taxa. The draft model is reconstructed based on this homology between the organisms. Bi-directional BLASTP is a method used to find and assign homolog genes by searching all proteins in an organism against all the proteins in the other organism in both ways. After successful parsing of the models, bi-directional BLASTP was carried out via '*getBlast*', a built-in function of RAVEN toolbox, between the reference (template) organism and organism of interest. This function was run by supplying the organism genomes as FASTA files with protein sequences, which were obtained from Uniprot-Proteome database. At this point, protein FASTA files must be modified such that gene ID formats should be named in the same manner with the gene ID format in the template model. Thus, all gene IDs in the FASTA files and the template model were matched before applying '*getBlast*' function. The only filtering criteria of this

function is the removal of hits with E-value higher than 10^{-5}. This function provides a Blast structure as an output. The Blast structure consists of various homology measurements such as E-value, alignment length, percentage of identity, and bit score.

In the next step, the draft model was obtained by running '*getModelFromHomology*' function in RAVEN. This function gets many inputs including the Blast structure obtained with '*getBlast*' function and the template model. The draft models were generated by applying cut-offs to the following parameters: the maximum E-value, the minimum alignment length, and the minimum identity percentage. E-value designates the number of matches expected by chance in a BLAST-based comparison of the sequences of a protein pair, and it represents the significance of the alignment. Its default value is 10^{-30}, and a lower E-value indicates a better quality in the search. Alignment length is the length of the aligned regions measured in base-pairs (bp). The alignment includes gaps and mismatches, and the shorter sequences may have a higher probability of aligning randomly to the target sequences. Therefore, the alignment length should be specified long enough to get a high percent of identity. The default value of the cut-off for alignment length is 200 bp. We also took into account the protein lengths when determining this cut-off. Since the short proteins will have a short alignment length, we tried to not to miss them. But if it is a long protein, when we select a very short alignment length, the homology between the proteins will be low, and this will prevent accuracy. Therefore, the alignment length should effectively cover both cases. Also, since we do not only evaluate similarity results over the alignment length, the percentage of identity would be supportive to balance these two situations. Thus, we used identity percentage as the third parameter in '*getModelFromHomology*' function. It is a number that describes how similar the query sequence is to the target sequence (the percent of how many residues in both query and target sequences are identical). Identical residues conserve similar chemical properties. The higher the percent identity is, the more significant the match, and its default value is 40 %. When we determine the minimum percent identity, we took into account the genetic similarity of the template and target organisms. Since bidirectional BLASTP was performed, all gene pairs that have acceptable BLASTP results by meeting the mentioned criteria in both directions were considered for further analysis, and so, the homolog gene pairs were chosen for the metabolic model reconstruction. Thus, the draft metabolic

network was created by retrieving possible set of reactions from the template model based on protein sequence homology.

3.2. Model Refinements Based on Manual Curation

The obtained draft models were unable to produce biomass, and they lacked many important metabolic reactions such as transport reactions, exchange reactions, ATP maintenance reactions. There are many automatized or manual ways of carrying out the processes that are mentioned in the second stage of the genome-scale metabolic reconstruction process (Section 2.8.2). In the model refinement process, laborious manual efforts with extensive literature review were performed. The reactions that are not associated with genes cannot be included in the draft model via protein homology. Therefore, these reactions were added to the draft model manually. The biomass reaction, exchange reactions, transport reactions, non-growth associated ATP maintenance reaction (NGAM), growth-associated ATP maintenance reaction (GAM) which was added through the biomass reaction formula, and ATP synthesis reaction were firstly added into the draft model. The reactions in the draft model came from an already published high-quality GEM (the template model), thus, they have accurate reaction stoichiometry, reaction directionality, localization, and GPR associations. Nevertheless, the reactions to be added should still be checked in terms of their necessity to the model when we consider the metabolism of the organism of interest.

The main step of the manual curation process is the gap filling process. First, the function, '*biomassPrecursorCheck*', was used because the biomass reaction rate (growth rate) of the draft model was zero. This function runs several FBA simulations to identify whether each of the biomass precursors, the molecules used in the biomass reaction, can be synthesized in the model. The identification of biomass precursors that cannot be synthesized by the draft model gives clues on the reason behind having zero flux through biomass formation. One major reason is blocked reactions, the reactions that cannot carry flux in any condition of a simulation. Those blocked reactions are mostly due to gaps in the metabolic network in a pathway, preventing carbon flow, and hence the synthesis of

certain biomass metabolites. RAVEN built-in function, '*fillGaps*', was run in order to find out the candidate reactions to solve the blocked reactions and to fill the metabolic gaps.

The idea behind the automatic gap-filling function is that it fills gaps with the smallest set of reactions from template model by targeting the functionality and the connectivity of the metabolic network. In addition, the gap-filling function basically added active reactions that are required in order synthesize the biomass constituents and therefore to have a non-zero flux through the biomass reaction, and also saved the reactions from being blocked in the model. At the end of this step, a functional metabolic model is obtained with nonzero growth rate. However, the consistency of the predicted fluxes by the model needs to be checked separately.

For a further study, different strategies were performed in order to expand the functional capacity of the models. For this purpose, in this study, a new KEGG-based network model was generated for *Klebsiella pneumoniae subsp. pneumoniae HS11286* and it was combined with the model we have built. For *Prevotella copri DSM 18205*, a template model for this organism was integrated with the model we reconstructed.

3.3. KEGG-Based Genome-Scale Draft Metabolic Network Model Reconstruction

The orthologous genes are the genes with homologous sequences between different species. These genes have high probability to perform similar functions. Thus, these genes might be involved in the same metabolisms by catalyzing same reactions [Koonin, 2005, Troyanskaya et al., 2010]. Therefore, using this information, we can obtain new reactions through the ortholog genes of the target organism using sources other than the template genome-scale metabolic network reconstruction. These reactions and the new genes associated with them are also important because they increase the genome coverage of the model we created.

KEGG GENES database is a collection of completely sequenced genomes. KEGG orthology (KO) is a database including identifiers that define molecular functions represented in KEGG PATHWAY database, and orthologs of experimentally characterized genes or proteins present in KEGG GENES database [Kanehisa et al., 2016].

There is a KO hierarchy starting with a gene and its corresponding KO terms and continue with its associated pathways in KEGG. KO is basically used to connect pathway and genomic information in KEGG database. Furthermore, K number, also called KO identifier, is assigned to experimentally characterized genes and proteins in specific organisms. K number assignment is used to find orthologs in other organisms based on the similarity of amino acid sequences in the KEGG. The sequence similarity search is performed for query amino acid sequences against genes in KEGG based on orthology, and with regards to the outcome of the search, the most appropriate K numbers are assigned to query sequences. In principle, a KO corresponds to a highly similar single sequence group but also one KO may consist of multiple sequence similarity groups.

The ortholog annotation leads to the automatic reconstruction of the corresponding metabolic network. KEGG pathway maps are created as networks of K number nodes, and this connectivity provides the automatic metabolic network reconstruction by pathway identification [Kanehisa et al., 2016, Mao et al., 2005].

HMMs (Hidden Markov Models) is a computational algorithm that can be used to describe the evolution of observable events due to internal factors that are not directly observable. Characteristically, biological sequences consist of sub-structures with different functions, and different functional regions generally show different statistical properties. For instance, proteins are generally made up of more than one domain that is conserved and functional region of a protein. Using this method, when a new protein is given, the constitutions of domains and their locations in the amino acid sequences can be estimated, and the family of the new protein can also be identified [Yoon, 2009]. The method starts with a single target sequence as a seed and iteratively builds a hidden Markov model (HMM) from the sequence and thus creates a model library [Karplus et al., 1998].

'*getKEGGModelForOrganism*' function in RAVEN basically reconstructs a genome-scale draft metabolic model based on protein homology search against KEGG orthology specific HMMs. There are many inputs for the function and they may be confusing. The input criteria in this study were determined by careful thinking and after many trials. '*getKEGGModelForOrganism*' function first downloads the HMMs archive file containing pre-trained HMMs, which were already trained on the selected sequences

by choosing the percent of sequence similarity as a threshold according to the given input. The given input can be like ‘prok100_kegg91’ where 100 represents 100 % sequence similarity and kegg91 represents the fitting KEGG version. The function retrieves the input data in a file. The file contains four subfolders, called keggdb, fasta, aligned and hmms (Hidden Markov Models).

The function imports KEGG database files including KEGG reactions, metabolites and genes in RAVEN Toolbox and constructs a generic KEGG metabolic network model. The files are stored in keggdb subfolder. The generic model is then basically reduced in size based on criteria given as input to the function. The organism ID used in KEGG was given as another input to the function. By using the organism ID, the genes included in the generic KEGG model was reduced. In other words, only the genes of *K.pneumoniae HS11286* organism in the KEGG database remained in the model. keepSpontaneous, keepUndefinedStoich, keepIncomplete, and keepGeneral are four inputs that are used as criteria for the reactions in the global KEGG model. They represent reactions respectively that are not associated with any genes, have undefined stoichiometry, have been labelled as ‘incomplete’, ‘erroneous’, and are on the generic form like ‘an aldehyde <=> an alcohol’, and therefore unsuited for modelling purposes. maxPhylDist (maximum phylogenetic distance) is the input parameter of the function used to filter KO-specific protein sets and to select which sequences should be used for constructing HMMs. The default value of the threshold is Inf, which means that all sequences will be used. The final information stored in keggdb folder was the imported KEGG phylogenetic distance matrix in RAVEN, and the function created the phylogenetic distance matrix, which was constrained according to the maxPhylDist threshold.

Next, the function checked genes with KO IDs in the global KEGG model whether these KO IDs are present in the generated sub-folders. Then, the function generated protein multi-FASTA files for each KO that is missing in these sub-files from KO datasets obtained based on protein sequence similarity analysis in KEGG database. Then, the multi-FASTA files were stored in the fasta folder.

Thereafter, the function clustered KO-specific protein FASTA files based on sequence identity threshold (seqIdentity). Then, it performed multi-sequence alignment

for KO-specific protein sets (or multi-FASTA files). The result was stored in the aligned folder.

In this function, HMMs are trained in order to predict protein function by identifying significantly similar protein sequences via the detection of homologs and consequently provide homology-based inference of knowledge. Each HMM in the HMMs archive has a KO ID. The function checked HMMs in the hmms folder to see whether HMMs for each KOs in the global KEGG model are present in this sub-folder. The required KO-specific HMMs were trained for each of the aligned KO-specific FASTA files by the function, and these HMMs were stored in the hmms folder.

The protein FASTA file of the target organism was supplied as one of the inputs to the function. The function queried the sequences belonging to the target organism via the provided FASTA file and performed an HMM search for KO-specific HMMs containing the sequences of the target organism. The significance score, the input named as cut-off, is used to filter out KO hits obtained as a result of the HMM search. Its default value is 10^{-50}. The results with KO hit lower than the cut-off was stored in a matrix called 'koGeneMat'. The parameters, minScoreRatioKO and minScoreRatioG, are used to remove genes with KO associations below these scores from the 'koGeneMat' matrix. minScoreRatioKO is a score that is used to prune KOs that were assigned to many genes and where some are clearly a better fit. The function ignores genes in a KO if their score is lower than the score ratio (log (score)/log (best score in KO)). The default value of minScoreRatioKO is 0.3. minScoreRatioG is a score that is used to prevent the allocation of a gene clearly belonging to a KO to KOs with lower scores. The function keeps the gene that is only assigned to KOs where their scores are equal and/or higher than the score ratio (log (score)/log (best score)) for that gene. The default value of minScoreRatioG is 0.8. The global KEGG model already has the KO-reaction associations and the KOs are converted to genes by using the pre-processed 'koGeneMat' matrix. Thus, the final KEGG model have had the gene annotations. The function removes reactions that do not have associations with these KOs from the model. Also, the reactions that have no longer GPR associations after HMM search are removed. Consequently, the reconstruction of genome-scale draft KEGG model is completed. Later, the draft KEGG model is merged with the

template-based reconstructed model to increase the genome coverage of the reconstruction.

3.4. Metabolic Model Validation

After manual network evaluation of the models, the next step is the validation of the models against experimental data. The metabolic models need to mimic the organism's metabolism at best. This step is crucial for assessing the accuracy and reliability of the produced models. At this stage, different simulations were performed for the models of *Klebsiella pneumoniae subsp. pneumoniae HS11286* and *Prevotella copri DSM 18205*.

Growth phenotypes in different carbon sources were predicted by performing simulations on *K. pneumoniae HS11286* model. The constraints and objective function were determined first, and flux rates were obtained via FBA. The consistency of the results with the experimental data was verified. On the other hand, *in silico* predictions for the fermentation products of *Prevotella copri DSM 18205* were calculated in specific simulation conditions. It was tested whether the model reflects the experimental conditions.

Finally, for both models, single gene deletion analysis was carried out to elucidate the gene essentiality of the organisms. The COBRA toolbox function '*singleGeneDeletion*' was used for the essential metabolic gene predictions based on MOMA approach. The function basically deletes a gene in the model and thereby the associated reaction(s). The removal of the related reaction(s) can create metabolic gaps or disturbs connectivity of the metabolic network so, it can prevent the biomass production (Figure 3.2). The genes of such reactions are predicted to be essential genes. Thus, the analysis can give information about the crucial genes that play an essential role in the metabolism.

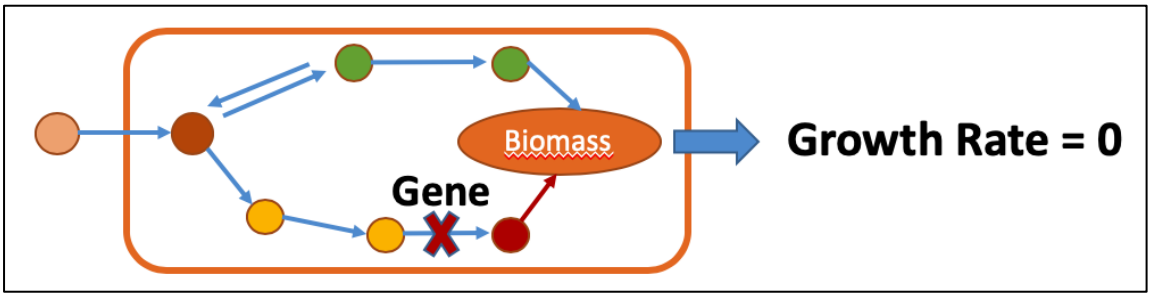


Figure 3.2: The illustration of gene deletion analysis via metabolic network simulations.

4. GENOME-SCALE METABOLIC MODEL RECONSTRUCTION FOR *KLEBSIELLA PNEUMONIAE SUBSP. PNEUMONIAE HS11286*

4.1. Draft Model for *Klebsiella pneumoniae subsp. pneumoniae HS11286*

Within the scope of the draft reconstruction, three inputs were provided: (i) the protein information of *K.pneumoniae HS11286*, (ii) the protein information of the reference/template organism *K.pneumoniae KPPR1*, and (iii) the recently reconstructed high-quality genome-scale metabolic network model of the reference organism, called *i*Kp1289 [Henry et al., 2017]. KPPR1 strain (a rifampin-resistant derivative of ATCC 43816 strain), a human *K. pneumoniae* isolate, is a highly virulent strain with several infection types [Henry et al., 2017]. The reference organism was chosen due to its high genetic similarity with *K.pneumoniae HS11286* strain and also since its GEM is a good template to build a draft model from it. *i*Kp1289 model in SBML format can be obtained from the Department of Energy Systems Biology Knowledgebase (KBase) Narrative interface [Arkin et al., 2018] and the excel format of the model is provided in its article [Henry et al., 2017]. *i*Kp1289 model consists of 2474 reactions, 2000 metabolites and 1289 associated genes.

In order to make the template model ready for use, the model was checked by '*checkCobraModelUnique*' function in COBRA toolbox. The function checks the model if there is any metabolite names or reaction names that are non-unique. Thus, the function found duplicates in the template model and provided them as names. Therefore, these repetitive reactions and metabolites should be removed by applying extra steps. For repetitive reactions, '*checkDuplicateRxn*' function in COBRA toolbox was used because it finds duplicate reactions and removes them from the given model and provides a new model as one of the outputs. However, we have benefited by using the indices of the duplicate reactions, another output from the function. The model field (grRules), which was missing in the model and required for the function, created and added to the model. The gene-protein-reaction associations (model.grRules) were generated by using

‘*generateGrRules*’ function in COBRA toolbox that uses the rules and genes fields present in the model to be consistent with GPR rules. 26 GPR rules in the ‘*grRules’* of *i*Kp1289 model were manually curated by removing ‘Unknown’ GPR relations such as ‘(Unknown and Unknown and VK055_4748 and VK055_2303 and VK055_4749)’. Also, the template model lacked reversibility vector for the reactions (model.rev), so it was created and added manually. ‘*checkDuplicateRxn*’ function in COBRA toolbox was then used to eliminate identical reactions, which resulted in the identification of 69 duplicate reactions in *i*Kp1289. However, two reactions can be identical but they can have different gene associations and/or different reversibility information. Such reactions were kept in the model. Therefore, 32 identical reactions out of 69 reactions were removed from the model by using ‘*removeReactions*’ function in RAVEN Toolbox. The duplicate metabolite names were detected in the model and their related rows in S matrix were also checked manually. Then, detected 200 duplicate metabolites were removed from *i*Kp1289 model. There were also duplicates in compartments field in the model. Therefore, compartments (model.comps) and compartment names (model.compNames) were made unique and the names were standardized as Cytosol (c0), Periplasm (p0), Extracellular (e0). The modified *i*Kp1289 model contains 2442 reactions, 1800 metabolites, and 1289 genes. To confirm that the model is functional after the modification, growth simulation on D-glucose media was performed using FBA.

After successful parsing of the modified *i*Kp1289 model, gene ID matching was performed between the protein FASTA files of *K.pneumoniae HS11286* and *K.pneumoniae KPPR1* and the template model (*i*Kp1289). The gene IDs in the FASTA file of *K.pneumoniae KPPR1* was matched and changed according to the gene IDs in the *i*Kp1289 model. The gene ID format in the FASTA file of *K.pneumoniae HS11286* was also changed in the same manner with the gene ID format in the *i*Kp1289 model. After that, bi-directional BLASTP was applied via ‘*getBlast*’ function and the Blast structure with homology scores was obtained. Later, the standardization of ‘*grRules*’ is needed by ‘*getModelFromHomology*’ function. Therefore, they were rearranged to standardize them according to the following specialized ‘*grRules*’ modification guidelines in RAVEN:

- There should be no overall rules containing brackets.
- The rules containing only enzyme complexes are enclosed into brackets.

- ' and ' & ' or ' strings are strictly set to lowercases.
- If the reaction is catalyzed by possible different protein complexes, all the possible combinations should be explicitly written.

In the final step, cut-offs for '*getModelFromHomology*' function were specified carefully by considering also other studies similar to our study to generate the draft model. In the genome-scale metabolic model reconstruction for *K .pneumoniae MGH 78578*, the following cut-offs were used in homology search using BLASTP between *K. pneumoniae MGH 78578* and *E. coli*: E-value of 10^{-5}, 35% amino acid sequence identity and match lengths of at least 70% of length of both query and subject [Liao et al., 2011]. In another article, the protein sequences of *S. coelicolor* A3(2) in the metabolic model iMK1208 were used as the reference and its sequence similarities in protein level were compared with the protein coding sequences of 31 *Streptomyces* strains using BLAST at the threshold of 40% identity percentage, 50% of query protein length aligned and E-value of 10^{-5} [Wang et al., 2017]. Our maximum E-value was chosen as 10^{-10}, the threshold filters out less significant matches. Since the genetic similarity between the target and reference organisms is very high (they are different strains of the same organism), the minimum identity percentage was chosen as 80%. The mentioned articles give us an idea of what percentage of the query protein length the aligned region should cover. The minimum alignment length was set to 100 considering both the minimum and average lengths of the query and target protein sequences, which are about 35 bp and 280 bp, respectively. As a result, the draft model was created as an output of '*getModelFromHomology*' function and it contains 1928 reactions, 1721 metabolites and 1238 genes. Thereafter, 46 unused genes were determined and removed from the model because they were not associated with any reaction, leading to 1192 genes left in the draft model. The 46 genes not associated with any reactions were added to the model by '*getModelFromHomology*' function while generating draft model. The problem is most probably due to a bug in this RAVEN function. Such unexpected results from available reconstruction toolboxes/functions is a disadvantage of using automatization in the reconstruction of a draft model, mostly because these functions were not probably tested by independent users in detail, and, hence, they are not bug/error-free. Therefore, it is always recommended to perform manual checks in every step of the automated reconstruction workflow.

4.2. Manual Evaluation of the Draft Model for *Klebsiella pneumoniae subsp. pneumoniae HS11286*

The generated draft model was improved by a manual curation process to have a complete functional model. The biomass formation reaction was obtained from the reference model (*i*Kp1289) and the same reaction was added to the draft model since they were genetically highly similar organisms. While the exchange reactions provide the balance between the transitions at the cell boundary, ATP reactions are important for energy requirements within the cell. Addition of biomass reaction also ensured the addition of the growth associated ATP maintenance requirement (GAM) necessary for cell replication, since GAM is represented as a term in the biomass reaction. Therefore, besides the biomass reaction, the addition of (i) 297 exchange reactions, (ii) ATP hydrolysis reaction, which is the non-growth associated ATP requirements for the maintenance of the cell, and (iii) ATP synthesis reaction from the template model was the first step.

Then, the growth simulation on D-glucose media in aerobic condition was performed using FBA by using the objective function to maximize growth in order to see if the current model can grow *in silico*. However, the biomass reaction rate of the model was found to be zero in the FBA simulation. Hence, the function '*biomassPrecursorCheck*' was used to check if biomass precursors are able to be synthesized by the metabolic network. This is an important check because there will be no *in silico* cell growth if any of the biomass components cannot be produced. The function adds demand reaction for each biomass metabolite, sets the objective function to maximize its production, and optimizes the objective function. This function identified 18 biomass precursor compounds that could not be synthesized and, hence, caused to get zero flux through biomass formation.

Later, '*fillGaps*' function of RAVEN Toolbox was run in order to find out the candidate reactions whose inclusion into the draft model can make nonzero biomass formation rate possible. *i*Kp1289 model was used again as a template model for this function. The criteria defined within the function is to fill the gaps of the draft model by

the smallest set of reactions from the template model. Beforehand, the vector (model.metComps) is required by the function, and for *i*Kp1289 model, the vector was defined. It indicates to which compartment each metabolite belongs. Consequently, 175 reactions were provided by this function. However, 120 of these reactions were already present in the model and most of them were exchange reactions in the draft model. When providing a new reaction, the RAVEN function adds the model ID to the end of the reaction name if that reaction is already available in the draft model, to avoid any conflict between the reaction names. The reason why the function adds already available reactions to the model is unknown to us, and could be a bug in the function. The remaining 55 reactions were added to the model manually. Some of the reactions provided by the function have GPR rules, but the genes present in these rules belong to *K.pneumoniae KPPR1* organism. Therefore, the homology scores of the 28 associated genes in GPR rules of the 55 reactions were manually checked and matched with our target organism. At this point, some tolerance was introduced to the threshold values as we needed these reactions to achieve *in silico* growth. There was no need to tolerate E-value in all genes, they were already below the chosen threshold. For the gene identity score, the threshold of 40% was used. For the alignment length, if the gene had score below the 100 bp, the determined threshold value, for the decision was made case-by-case based on the gene length. Namely, when a gene had alignment length shorter than 100 bp, if the length of the gene is short and its identity score was also high, this was considered as a good match. Thus, 8 genes were eliminated and 20 genes were added to the model via GPR rules of 55 reactions. The GPR rules of the reactions associated with the eliminated genes were left blank. In addition, if the reaction has GPR rules consisting of co-occurrence of multiple genes, that is, if the rule contains a protein complex, and not all the genes in the rule meet our specified cut-off criteria, such GPR rules for the reactions were removed. Together with the new reactions that were added from the *i*Kp1289 model for gap-filling, 50 transport reactions that are necessary to provide intracellular connectivity and transitions between compartments have been added to our model from the template model.

Apart from the reactions that have been added to the draft model so far, there were reactions associated with the genes that could not come into the draft model because they could not exceed the threshold value used for BLASTP as a result of the homology search.

The genes that have a role in these reactions were actually paired via BLASTP with the genes in the reference organism but they could not be selected due to the chosen cut-off values. Especially, the threshold value for the gene identity parameter, which was 80%, was stringently high since the gene identity scores used in the mentioned articles in Section 4.1 were 35-40 %. Therefore, the reactions associated with these genes, which still have acceptable matches because they have gene identity scores higher than 40 %, can be considered to be added to our model even if their genes did not really pass the 80% threshold value we determined. The reactions that cannot be included in the draft model because of exceeding the threshold value used for BLASTP were manually checked and questioned as to whether the reaction was anyhow important and needed for the metabolism of the organism, and whether the associated gene had acceptable threshold values, for example, the gene identity score above 40%. According to the criteria, 22 reactions that have gene associations were selected and added into our model. These 22 reactions spanned reactions for acetyl-transferase mechanism, Ubiquinol reactions that participates in the mitochondrial electron transport chain, pyruvate formation pathway, among others.

At last, the reconstructed metabolic model for *K.pneumoniae strain HS11286* consisted of 2355 reactions, 1753 metabolites, and 1218 genes. The growth simulation on D-glucose media both in aerobic condition and anaerobic condition was performed using FBA by using the objective function of growth maximization on the final metabolic model. The final version of the reconstructed metabolic model could grow *in silico* with and without oxygen and simulate the respiration phenotype of the bacteria, which is facultative respiration.

4.2.1. Integration of KEGG-Based Metabolic Network Model

In the previous steps, a draft metabolic model was created first using the high-quality metabolic model of an organism that has close genetic similarity with the target organism. In the second step, the same template model was used to make the draft model more complete and functional. However, the reactions we obtained from the template model to reflect the metabolism of the organism are only the common reactions between

the two bacteria. Therefore, the model we created does not have any reactions for the genes that are specific to our target organism, leading to a lower gene coverage as a genome-scale metabolic model in its current form. To include also the genes and reactions specific to our target organism in the reconstructed genome scale metabolic network, databases with comprehensive and reliable information about the genome, chemical properties and functions of a biological system can be used. KEGG is such a comprehensive database, which also includes metabolic reactions of HS11286 strain of *K. pneumoniae*. However, manually searching KEGG for metabolic reactions that belong to *K. pneumoniae HS11286* can require extensive time. Therefore, as a third step in our reconstruction effort, a KEGG-based genome-scale metabolic network was automatically generated to increase the gene coverage of the model and include specific reactions to our target organism that do not exist in the reference strain used as the template. Therefore, RAVEN Toolbox was used to reconstruct a KEGG-based metabolic network of the organism automatically by using '*getKEGGModelForOrganism*' function.

The reconstruction of KEGG-based metabolic network is mainly based on the generation of a metabolic network via orthologous genes and their associated reactions. The '*getKEGGModelForOrganism*' function in RAVEN makes the assignment of our query sequences of *K.pneumoniae HS11286* to ortholog genes in the KEGG GENES database as a result of homology search, and it leads to the reconstruction of the corresponding KEGG-based metabolic network through the KO hierarchy.

The function first downloaded the HMMs archive file containing pre-trained HMMs that were trained on only prokaryotic sequences with % 100 sequence similarity as a threshold by choosing the input 'prok100_kegg91'. The function retrieves the input data in a file with four subfolders, called keggdb, fasta, aligned and hmms. Keggdb folder contains the global KEGG metabolic network including KEGG reactions, metabolites and genes in RAVEN. The other input, the organism ID, which is 'kpm', was used to reduce the global KEGG model with respect to the genes of *K.pneumoniae HS11286* organism in the KEGG database. keepSpontaneous, keepUndefinedStoich, keepIncomplete, and keepGeneral are four inputs that are used to reduce the KEGG model in terms of reactions (also please see Section 4.3). 'false' was chosen for these four inputs, meaning that the function did not keep these types of reactions in the global model. The last information

stored in keggdb folder was the imported KEGG phylogenetic distance matrix in RAVEN. In our case, maxPhylDist (maximum phylogenetic distance) was chosen as -1 in order to have HMMs that are trained on only sequences from the same domain. The other input that is provided to the function was the protein FASTA file of *K.pneumoniae HS11286* strain. The function questioned the sequences that belong to *K.pneumoniae HS11286* strain via the supplied FASTA file and conducted an HMM search for KO-specific HMMs containing sequences of the target organism.

The significance score threshold, called cut-off, filters out KO hits obtained from the HMM search and we used its default value of 10^{-50}. A matrix called 'koGeneMat' was created from the results obtained with a lower KO hits than the cut-off. minScoreRatioKO and minScoreRatioG parameters were used to remove genes with KO relationships below these scores from the 'koGeneMat' matrix. For both thresholds, lower is less strict. Hence, these two scores were defined higher than the default value in order to have the best gene-KO associations in the KEGG model. minScoreRatioKO was defined as 0.7 (default value: 0.3), thus the function ignored the genes in a KO lower than the score ratio (log(score)/log(best score in KO)). minScoreRatioG was defined as 0.95 (default value: 0.8). Thus, the function kept the genes that are only assigned to KOs whose scores are equal and/or higher than the score ratio (log (score)/log (best score)) for that gene.

'*getKEGGModelForOrganism*' function reconstructs a genome-scale draft metabolic model based primarily on protein homology search against KEGG orthology-specific HMMs. The homology results of the genes with significance scores below the cut-off, which is defined with E-values and refer to the values lower than 10^{-50}, are summarized into KO-gene occurrence matrix with E-values in intersections as 'koGeneMat'. The two inputs, minScoreRatioKO and minScoreRatioG, were used to prune the undesirable KO-gene association in this matrix. The function constructs the KEGG-based model based on the pre-processed 'koGeneMat' matrix. Since KO-reaction relationships are already stored in the global KEGG model, KOs are converted to query genes by using the 'koGeneMat' matrix. Thus, the final draft KEGG model contains only those reactions that are associated with KOs from 'koGeneMat'. Consequently, the reconstructed KEGG model contains 1589 reactions, 1618 metabolites and 1196 genes.

The created genome-scale KEGG model is aimed to be integrated into the template-based reconstructed model for *K.pneumoniae HS11286* created in Section 4.2. The metabolite names in the KEGG model were matched and changed to the same format as the metabolic names in the template-based reconstructed model in order to facilitate this step and make it more accurate. The metabolic IDs in the template-based model, which are in the ModelSEED ID format, and KEGG metabolic IDs were matched by using the ModelSEED database. The used data in this stage is available for download from the GitHub page [Web 1, 2019].

The KEGG model has many reactions that are already available in the genome-scale model for *K.pneumoniae HS11286* generated based on RAVEN Toolbox. To identify reactions specific to the KEGG model, the KEGG model was downsized before integrating these two models. To this aim, the reactions that are common between the two models were first removed from the KEGG model. Since our main goal is to increase gene coverage in the current model, the genes common in both models were searched and 787 genes were found. Then, we continued to shrink the KEGG model by removing the reactions that are associated with these common genes. Thus, the modified small version of KEGG model was created, which contains only the reactions associated with the KEGG model specific genes. At the final step, the small KEGG model, which contains 294 reactions, 548 metabolites and 261 genes, and the RAVEN-based model for *K.pneumoniae HS11286* were combined by using '*mergeModels*', a built-in function in RAVEN. The generated merged model contains 2301 metabolites, 2649 reactions and associated 1479 genes. In the '*mergeModels*' function, when both models have metabolites with the same naming, it changes the metabolite name in the reactions coming from KEGG by adding a model ID at the end in order to avoid conflicts. Since these metabolites are indeed the same, to ensure a proper mass balance in the merged model, these two rows belonging to duplicate metabolites in the S matrix have been updated by merging them. This led to the removal of 185 metabolites from the merged model. Therefore, a more comprehensive and complete model has been generated for *K.pneumoniae HS11286*. The final model contains 2649 reactions, 2116 metabolites and 1481 genes. This means that 263 genes and 294 reactions were added to the RAVEN-

based model from the KEGG model, significantly increasing the genome-scale coverage of the reconstructed model.

4.3. Validation of Model for *Klebsiella pneumoniae subsp. pneumoniae HS11286*

4.3.1. Growth Simulation on Different Carbon Sources

As a validation step of our model, called *i*Kp1481, the growth phenotype of *Klebsiella pneumoniae HS11286* was simulated via FBA. Since we do not have high-throughput experimental assays or experimental growth rates of this organism, the growth simulation results were compared to the experimental growth rates of *Klebsiella pneumoniae MGH 78578*, a metabolically similar strain of *Klebsiella pneumoniae KPPR1* [Liao et al., 2011]. In addition, the simulation results were compared to *in silico* growth simulation results of the template model (iKp1289). The growth simulations on six different carbon sources, acetate, citrate, glucose, glycerol, L-lactate, and L-malate under aerobic conditions were performed by using both our model and *i*Kp1289 model. Carbon D-glucose minimal medium constraints were obtained from the Department of Energy Systems Biology Knowledgebase (KBase) Narrative interface [Arkin et al., 2018]. The metabolites included in the minimal medium are H2O, Sulfate, Phosphate, NH_3, Mn^{2+}, Zn^{2+}, Cu^{2+}, Ca^{2+}, H^+, Cl^-, Co^{2+}, K^+, Ni^{2+}, Mg, Na^+, Fe^{2+}, Fe^{3+}, Molybdate. For all these metabolites, their upper and lower boundaries were set to 100 and -100 respectively.

Additional uptake rates for six different carbon (C) sources and oxygen were obtained from the literature (Table 4.1) [Liao et al., 2011]. The exchange reactions for other metabolites were inactivated by setting their upper-bound and lower-bound values to zero. Thus, all the constraints were introduced to the models. While performing FBA, the objective function was set the maximum growth of the organism in different carbon sources.

Table 4.1: The Constraints for The Different Carbon Sources.

	Experimental Uptake Rates (mmol/gDW/h)	
	C source	**Oxygen**
Acetate	14.291	14.657
Citrate	14.017	21.837
Glucose	10.457	21.744
Glycerol	10.609	13.618
L_Lactate	22.686	21.837
L_Malate	34.572	21.837

*i*YL1228 is the experimentally validated genome-scale metabolic network model of *Klebsiella pneumoniae MGH 78578* strain. The quantitative results of *in silico* growth simulations of the three GEMs, which are *i*Kp1289, *i*Kp1481, and *i*YL1228, and the experimental data of *Klebsiella pneumoniae MGH 78578* strain were provided in Table 4.2 [Liao et al., 2011].

Table 4.2: Growth Phenotype Comparison for Different Carbon Sources.

	Growth Rate (1/h)			
	i*YL1228** ***(In silico)	***i*Kp1289** ***(In silico)***	***i*Kp1481** ***(In silico)***	**Experimental data**
Acetate	0.355	0.432	0.750	0.293
Citrate	0.937	1.072	1.688	0.570
Glucose	1.040	1.140	1.646	1.084
Glycerol	0.599	0.692	0.835	0.804
L_Lactate	0.655	0.884	1.785	0.658
L_Malate	1.053	1.203	2.478	0.834

The results show that *Klebsiella pneumoniae HS11286* is able to grow on these six carbon sources. In other words, the organism has specific enzymes encoded by associated

genes and accordingly, responsible metabolic pathways to utilize these carbon sources. This result is in agreement with the fact that *Klebsiella* strains can inhabit a variety of environments and also utilize multiple carbon sources mentioned in Section 2.4. In addition, it is noteworthy that *Klebsiella pneumoniae HS11286* GEM shows the highest growth in all the conditions. There are many possible explanations for these differences. One possible reason for having higher growth rate than experimental data is that the metabolic capabilities of *Klebsiella pneumoniae MGH 78578* strain differs from both *Klebsiella pneumoniae HS11286* and *Klebsiella pneumoniae KPPR1* strains although they are close organisms. *Klebsiella pneumoniae MGH 78578* strain has low virulence while *Klebsiella pneumoniae HS11286* and *Klebsiella pneumoniae KPPR1* are hypervirulent strains [Henry et al., 2017]. That difference supports the fact that *Klebsiella pneumoniae MGH 78578* strain has a low growth rate than the other two organisms because these hypervirulent strains are highly adapted to a wide variety of hosts in which *K. pneumoniae* can infect and proliferate. Therefore, their capability of nutrient uptake, their catabolism, and cell division are relatively higher. In the meantime, since both *Klebsiella pneumoniae HS11286* and *Klebsiella pneumoniae KPPR1* strains have high genetic similarity with each other, a similar growth simulation results are expected to see. However, to obtain a result higher than expected metabolic capacity suggests that a small genetic difference may lead to a large metabolic difference. Another reason may be that the reconstructed model is unrealistically efficient because the mechanism responsible for gene regulatory factors, like transcription factors, is not taken into account in the current model.

As growth-associated ATP is integrated into the biomass reaction, ATP production is highly correlated with biomass formation. Therefore, the maximum ATP production of *i*Kp1289 and *i*Kp1481 was checked. As a result, they have same ATP generation capacity. In addition, the biomass formula of the models are same, however, the models have small differences in their capacity of biomass precursor metabolite production individually. Based on this, the differences in the maximum production rate of the two amino acids, L-cysteine and L-methionine, which attracted particular attention, suggested that it may be related with the difference in growth results of *i*Kp1289 and *i*Kp1481. Thus, the higher production capacity of some biomass precursors, such as these two amino acids in our

model, may be due to additional alternative pathways involved in the production of these metabolites, which may also result in an increase in the total biomass rate.

Furthermore, FBA-based analysis supposes that the organism grows in an optimal condition due to its optimization based on maximum growth, however in reality this is not the case every time. Therefore, in such cases, the reconciliation of GEM with experimental data is important in terms of improving GEM and accurately represents the organism's metabolism. Overall, we could say that the final model was successful by simulating the growth on different carbon supplements and provided insights on the growth behavior of *Klebsiella pneumoniae HS11286*.

4.3.2. Single Gene Deletion Simulations

The gene essentiality analysis simulations were performed for the organism of interest and the reference organism. The growth simulation was performed on aerobic condition in Carbon D-glucose minimal medium by taking the glucose uptake rate of 10 mmol/gDW/h and the oxygen uptake rate of 20 mmol/gDW/h. The essential genes were predicted computationally via the COBRA Toolbox built-in function, '*singleGeneDeletion*', based on MOMA approach by using models of *Klebsiella pneumoniae HS11286* (*i*Kp1481) and *Klebsiella pneumoniae KPPR1* (*i*Kp1289). After the deletion simulation of each single gene, the predicted growth rate at that condition is examined.

Consequently, 117 genes were identified as essential for growth of *Klebsiella pneumoniae KPPR1*. In addition, 106 genes were essential for growth of *Klebsiella pneumoniae HS11286* while the deletion of 423 genes had no impacts at all on growth according to in silico model predictions. The essential genes of both models were matched by using BLAST results. According to the mapping genes, *i*Kp1481 model has 101 common essential genes with *i*Kp1289. The gene essentiality analysis results represent that when a gene is deleted and thereby the associated reactions are prevented, consequently, related pathways that have roles in growth will be interrupted and so, they cannot utilized. It indicates that the deleted gene is essential for growth. 5 out of 106 computationally essential genes predicted by *i*Kp1481 were present only in *K.*

pneumoniae HS11286. These survival genes are needed for many important metabolic pathways like amino acid metabolism (Table 4.4). The essential genes have important roles in these pathways and they are targetable for infection treatments.

Table 4.3: Gene Essentiality Analysis Results.

Unique Essential Genes	**The Associated Pathways**	**The Associated Reactions**
KPHS_07980	Valine, Leucine, and Isoleucine Metabolism	ACHBS: 2_aceto 2_hydroxybutanoate synthase reaction ACLS: acetolactate synthase reaction Ilvg: Isoleucine biosynthesis step_i
KPHS_07990	Valine, Leucine, and Isoleucine Metabolism	ACHBS: 2_aceto 2_hydroxybutanoate synthase reaction ACLS: acetolactate synthase reaction Ilvg: Isoleucine biosynthesis step_i
KPHS_11560	Inorganic Ion Transport and Metabolism	NH4tpp: ammonia reversible transport reaction
KPHS_24190	Alternate Carbon Metabolism	MAN6PI: mannose_6_phosphate isomerase reaction
KPHS_37480	Amino acid, ala	ALATA_L: L-alanine transaminase

5. GENOME-SCALE METABOLIC MODEL RECONSTRUCTION FOR *PREVOTELLA COPRI DSM 18205*

5.1. Draft Model for *Prevotella copri DSM 18205*

The human host has restrictions on its ability to use certain dietary polysaccharides. Therefore, the human host uses saccharolytic gut inhabitants, such as *Bacteroides thetaiotaomicron*, that ferment inaccessible polysaccharides into the short chain fatty acids (SCFAs), and thus it can consume some dietary polysaccharides in the form of SCFAs [Heinken et al., 2013]. *Bacteroidetes* is one of two dominant phyla inhabited in human microbiota. The order *Bacteroidales* contains *Bacteroidaceae* and *Prevotellaceae* families, to which *Bacteroides thetaiotaomicron* and *Prevotella copri* belongs respectively. The protein information of *Prevotella copri DSM 18205* and the reference organism, *Bacteroides thetaiotaomicron VPI-5482*, and the genome-scale metabolic model of *Bacteroides thetaiotaomicron VPI-5482* (*i*AH991) as a template model were provided as three inputs to perform the draft model reconstruction of *Prevotella copri DSM 18205*.

Since there is no high-quality GEM for any *Prevotella* species in the literature, a model of a genetically close organism with high-quality GEM was used. The GEM for *Bacteroides thetaiotaomicron VPI-5482* was first reconstructed in 2013, and it was called *i*AH991 [Heinken et al., 2013]. It was later updated in 2015 [Heinken and Thiele, 2015]. Afterwards, *i*AH991_norm model, the last updated version, was released with manually curated biomass reaction. Verification of the biomass reaction was done by Chan et. al via fixing three ‘error leading problems a) the reaction generated by automated platforms or adapted from other models without proper verification, b) inaccurate stoichiometric coefficients in the biomass reaction, and c) missing cofactors in macromolecular synthesis reactions [Chan et al., 2017]. *i*AH991_norm model, which is supplied in MAT-file format in its article, includes 1539 reactions, 1182 metabolites and 991 genes [Chan et al., 2017]. 22 *grRules* in the recent manually curated high-quality GEM, *i*AH991_norm model were standardized based on the required format by RAVEN as described in section 4.1 before

running '*getModelFromHomology*' , a built-in function in RAVEN. Also, the model was checked in terms of whether there is any duplication in metabolites, reactions or genes. There are no duplications in any such fields of the model. After that, the growth simulation via FBA was performed by setting objective to maximum growth. As a result, the functionality of the template model was confirmed by getting an optimal solution.

After the successful parsing of the *i*AH991_norm model, the gene ID format in the protein FASTA files of *Prevotella copri* and *Bacteroides thetaiotaomicron* were matched and changed to make both in the same gene ID format as the template model (*i*AH991_norm). The gene ID formatting is important because '*getModelFromHomology*' function creates a draft metabolic network model by ensuring consistency with the genes in the iAH991_norm model and thus accessing reactions related to existing genes without any matching problem. These protein FASTA files of the organisms were then used as inputs to run bi-directional BLASTP through the '*getBlast*' function. As a result, a Blast structure containing protein homology scores was obtained as output. Subsequently, before running '*getModelFromHomology*' function, the three threshold values, which are the maximum E-value, the minimum alignment length and the minimum identity percentage, were determined by considering other similar studies from the literature and the degree of genetic similarity between the reference organism and the organism that we aimed to build a model for. In addition to the two articles that were referred to in Section 4.1 about the cut-off values for BLAST analysis, two other articles that focused on organisms other than bacteria were also considered in choosing cut-offs [Liao et al., 2011], [Wang et al., 2017], [Ye et al., 2015], [Tiukova et al., 2019]. In one of those studies, while reconstructing the genome-scale metabolic model of the oleaginous fungus *Mortierella alpina*, the reactions for the model were chosen based on orthologs shared between *Mortierella alpina* and the three reference organisms as a result of the protein sequence similarity searches using BLAST. Their sufficient similarity scores to ensure the accuracy are E-value of 10^{-30} and the sequence identity of 40% [Ye et al., 2015]. Also, the information on whether they use the criteria of the alignment length was not provided in the article. The other article was the genome-scale model of the basidiomycete red yeast *Rhodotorula toruloides*. When they reconstructed their draft model, the orthologous genes, which were identified via bi-directional BLASTP, were chosen by using the

following cut-offs: E-value of 10^{-20}, 35% amino acid sequence identity and alignment length of 150 bp [Tiukova et al., 2019]. After all the similar studies and the lineage similarity between target and reference organisms were considered, the minimum identity was chosen as 35%. The maximum E-value was chosen as 10^{-10} to get rid of the less significant ortholog gene matches. The minimum alignment length was set to 100 bp. Consequently, the obtained draft model as an output of '*getModelFromHomology*' function for *Prevotella copri DSM 18205*, consists of 795 reactions, 860 metabolites, and 492 genes.

5.2. Manual Evaluation of the Draft Model for *Prevotella copri DSM 18205*

For further analysis, the manual curation process was performed on the generated draft model. First, since the draft model lacks biomass reaction, the biomass reaction in the template model was added to the draft model by removing *Bacteroides thetaiotaomicron* specific metabolites available in the reaction, which are capsule polysaccharide, lipopolysaccharide, ceramide phosphoethanolamine and sphingomyelin. In addition, ACP and apoACP were also removed from the biomass reaction formula to avoid mass-balance based problems. In addition, the growth-associated ATP maintenance was integrated into the biomass reaction. Afterwards, the addition of (i) 278 exchange reactions, (ii) the non-growth associated ATP requirement reaction for the maintenance of the cell, and (iii) ATP synthesis reaction from the template model were completed.

Prevotella copri is an obligate anaerobe bacteria inhabiting the human gut flora (Section 2.5). Therefore, *in silico* growth simulation via FBA was performed in anaerobic environment by using the objective function to maximize growth and without any uptake constraints. That is, the simulations were performed in rich medium. However, there was no *in silico* cell growth, meaning that the rate of biomass reaction of the model was zero. The COBRA toolbox built-in function, '*biomassPrecursorCheck*', was run by setting the objective function to the maximization of the production of biomass precursor metabolites. The function is used to identify the metabolites that cannot be synthesized by

the metabolic network and therefore cause zero biomass reaction flux rate. According to the function output, the synthesis of 59 biomass precursor metabolites was missing.

The blocked reactions, the reactions that cannot provide any carbon flow through them, cause such problems by creating gaps in the metabolic network, and they can prevent the synthesis of metabolites that enable growth. Therefore, the '*fillGaps*' function of the RAVEN Toolbox was run in anaerobic conditions to perform the gap-filling process by using the *i*AH991_norm model. The function supplies the candidate reactions, thanks to which a nonzero biomass formation rate will be achieved if included in the draft model. Addition of these candidate reactions to the draft model can activate blocked reactions, and the metabolites necessary for growth can be synthesized. Consequently, by adding 185 candidate reactions, the '*fillGaps*' function output provides the gap-filled model, which contains 1261 reactions, 1104 metabolites, and 597 genes. However, 22 of those reactions were already present in our current model. The function adds the model ID to the end of the reaction names to prevent any conflict between the reaction names. Hence, 22 reactions were removed from the gap-filled model to eliminate duplication by using '*removeReactions*' function in RAVEN Toolbox. The same is also true for the 24 metabolites in the model. To confirm that these metabolites are duplicate, the paired rows in the S matrix were checked and found to be completely identical. As a result, these duplicate metabolites were removed from the gap-filled model. The reactions and metabolites that are already present in the draft model were probably added again due to an error in the function.

Furthermore, some of the candidate reactions provided by the '*fillGaps'* function have GPR rules, but the genes present in these rules belong to the *Bacteroides thetaiotaomicron VPI-5482* organism. Therefore, 102 genes of the *Bacteroides thetaiotaomicron VPI-5482* associated with the candidate reactions were manually checked. Since the reactions coming from the gap-filling phase will allow the model to grow by making it a functional model, a flexibility was introduced to the previously determined threshold values for the gene homology scores in the associated genes with these reactions. Still, the gene identity score was not dropped below 30%. For the alignment length, if the gene identity value is high, the relevant gene length was checked from the FASTA file. If the gene length is short and the gene identity value is high, these

genes were not eliminated. In this case, the genes of the *Bacteroides thetaiotaomicron VPI-5482* that came from the GPR rules of the candidate reactions were first deleted from the gap-filled model by using the '*removeGenesFromModel*', the built-in COBRA function. Consequently, the GPR rules of 185 candidate reactions were also modified and deleted by the function. Next, if there is an acceptable similarity match between the *Bacteroides thetaiotaomicron* genes and *Prevotella copri* genes, the GPR rules of reactions related to those genes were manually edited and added to the model by the COBRA Toolbox built-in function, '*changeGeneAssociation*' function. This led to the addition of 7 genes to the model associated with the gap-filling reactions. In addition to the new reactions added from the *i*AH991_norm model for gap filling, 81 transport reactions, which serve to provide intracellular transitions between the compartments, were added from the template model. Thus, the current model contains 1320 reactions, 1108 metabolites, and 502 genes. *In silico* growth simulation on D-glucose media in anaerobic condition was performed on the current model via FBA by using the objective function as maximizing growth. As a result, having non-zero growth rate confirmed that the current model can simulate the respiratory behavior of the *Prevotella copri* organism.

5.2.1. Integration of CarveMe Draft Model

In the previous steps, a draft metabolic model for *Prevotella copri DSM 18205* was first created using a high-quality genome-scale metabolic model of an organism that has a close genetic similarity to the target organism, then the same template model (*i*AH991_norm) was used to make the draft model more complete and functional. However, the reactions we obtained from the template model to reflect the metabolism of the target organism were only common reactions between the two bacteria according to the protein homology search. When these two organisms are considered in more detail in terms of genetic similarity, the metabolic reactions we obtain from the template model are not sufficient for the reconstruction of a complete genome-scale metabolic model of the target organism. Therefore, in order to increase the gene coverage and functional capacity of our target organism, in addition to the processes that were performed so far, a different strategy have been applied. To this aim, another draft model created by using the

CarveMe, a bioinformatics tool, for the *Prevotella copri DSM 18205* were integrated into our reconstructed genome-scale metabolic network model. CarveMe is a Python-based bioinformatics tool that reconstructs automatically a model for microbial communities by using the BiGG database as the reference reaction database and the manually selected universal template reaction set [Machado et al., 2018]. The authors performed a fast automated generation of a collection of 5587 microbial metabolic models by CarveMe tool, and the database including the model collection is available at https://github.com/cdanielmachado/embl_gems. The draft template model that was integrated into our model was obtained from the mentioned model collection. The draft CarveMe model for *Prevotella copri DSM 18205* includes 1323 reactions, 1002 metabolites and 498 genes.

Some changes were made to the CarveMe model before integrating the two models. As a first step, the gene ID format in the CarveMe model was matched and changed according to our current model. Thus, 305 common genes were identified between the two models. By removing the common genes and associated reactions from the CarveMe model, it was aimed to prevent the re-addition of reactions already existing in our current model. Before removing common genes, the missing GPR rules field in the CarveMe model was created by using the '*creategrRulesField*', which is the COBRA built-in function. The common genes and the associated reactions were then removed from the template model by using '*removeGenes*' function in RAVEN Toolbox, and the resulting downsized CarveMe model contains 869 metabolites, 823 reactions and 188 genes. In the next step, since our current model did not have a periplasm compartment, the reactions performed in periplasm were modified and this compartment was removed from the model. These reactions were assumed to occur in the cytosol in the CarveMe model. For this purpose, the steps mentioned below were carried out:

- The metabolites in the periplasm compartment in the CarveMe model were identified at first.
- The names of these metabolites were changed to make them associated with the cytosol compartment.
- Subsequently, the reactions that remained the same on both sides of the reaction formula were removed from the model. In other words, transport reactions that play

a role in the transition from periplasm to cytosol or the other way around were deleted. Also, 40 transport reactions which have the metabolites that are transported via ATP-binding cassette transporters (ABC transporters) were found and removed from the model. In addition, the repetitive reactions resulting from the exchange of metabolites between cytosol and periplasm compartments were also removed from the CarveMe model.

- '*removeReactions*' function in RAVEN Toolbox was used in all of the reaction removal steps from the model.
- So far, any metabolites were not removed from the model. But even if the metabolite name was changed, there was still the rows of periplasm metabolites in the S matrix. In order to completely transfer the metabolites in periplasm to cytosol, the periplasm and cytosol rows belonging to the same metabolite in the S matrix were summed up. The row total was assigned to the cytosol row of the relevant metabolite. Then, the rows belonging to the metabolites in periplasm were removed from the model by using '*removeMets*' built-in function of RAVEN Toolbox.
- At last, the unused periplasm compartment was removed from the CarveMe model. Also, the 'metComps' field, which represents which metabolites are placed in which compartment, was updated.

Furthermore, the reaction and metabolite names of the two models, our current model and CarveMe model, were updated to make them compatible. Thus, 182 common reactions were identified and removed from the CarveMe model. The resulting modified CarveMe model contains 538 reactions and 608 metabolites and 118 genes.

After all the pre-treatments to CarveMe model were completed, the downsized CarveMe model was combined with the current template-based reconstructed model for *Prevotella copri DSM 18205* by using the '*mergeModels*' function in RAVEN Toolbox. The merged model contains 1462 metabolites, 1858 reactions and 620 genes. As in the combination of the previously produced model for the *Klebsiella pneumoniae HS11286* and the KEGG-based model (Section 4.2.1), when there were metabolites with the same names in both models, the '*mergeModels*' function in RAVEN Toolbox re-named the metabolite names in the reactions of the CarveMe model by adding the ID of CarveMe model as suffix to avoid conflict in the merged model. Since these metabolites are the

same, these two rows of repetitive metabolites in the S matrix were combined and updated to provide a suitable mass balance in the combined model. In this way, 89 metabolites were removed from the combined model and the metabolites were unified. In addition, although 85 reactions had different names, they were identified as duplicate reactions and removed from the combined model by using the '*removeReactions*' function in RAVEN Toolbox. Almost all of these reactions were the exchange reactions. Thus, a more comprehensive and complete model for *Prevotella copri DSM 18205* was created with the combined model. The final model contains 1773 reactions, 1373 metabolites and 620 genes. Thus, 118 genes and 453 reactions were added from the CarveMe model to the template-based model for *Prevotella copri DSM 18205*. The final reconstructed model is a model with significantly increased genome coverage.

5.3. Validation of Model for *Prevotella copri DSM 18205*

5.3.1. Fermentation Product Simulation

The reconstructed metabolic network model for the *Prevotella copri DSM 12805* was verified by comparing the estimated flux rates with the experimentally measured values to ensure that the model can realistically simulate the metabolism of the organism. *Prevotella copri DSM 12805* strain was analyzed by using bioinformatics and experimental techniques by Franke and Deppenmeier [Franke and Deppenmeier, 2018]. The study provides the essential insight into growth behavior and fundamental characteristics for central carbon and energy metabolism of the organism. Besides the growth yield value, they reported some important experimental data for *Prevotella copri* metabolism including glucose uptake rate and the flux rates for three fermentation products; succinate, acetate, and formate. The measured values obtained from the article were used as constraints or for model validation for the growth phenotype simulations of the organism in this study via FBA by setting objective to the maximization of the growth (Figure 5.1) [Franke and Deppenmeier, 2018].

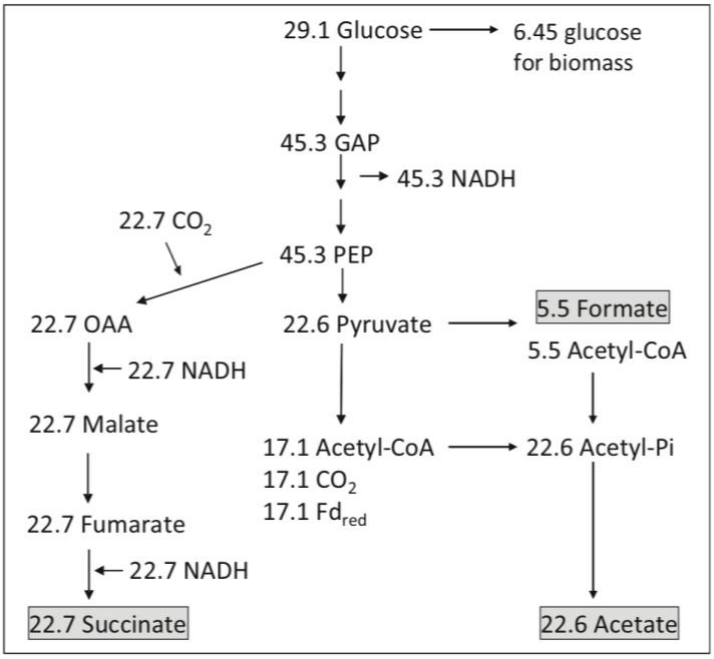


Figure 5.1: Intermediate flow rates of *P. copri*'s central carbon metabolism. The numbers represent mmol of the related compound per g DW. GAP, glyceraldehyde; PEP, phosphoenolpyruvate; OAA, oxaloacetate; Fdred, reduced ferredoxin; Pi, phosphate.

The anaerobic growth simulation was performed on modified Defined Minimal Medium Glucose (DMMG) by using the template-based model for the *Prevotella copri DSM 12805*. Carbon D-glucose medium constraints were defined in our model according to the growth conditions in the reference article [Franke and Deppenmeier, 2018]. The minimal medium has a mineral solution that includes H_2O, K^+, Ca^{2+}, Mn^{2+}, H^+, Cl^-, Co^{2+}, Mg, Na^+, Fe^{2+}, Fe^{3+}, NH_4, Sulfate [Varel and Bryant, 1974]. Varel and his colleagues mentioned that instead of Casitone that is not present in our model, L-Methionine is supportive for growth stimulation [Varel and Bryant, 1974]. Therefore, L-Methionine uptake was defined in the model. The content of the vitamin solution was obtained from Wolin and his colleagues, and the included metabolites are biotin, folic acid, pyridoxine, riboflavin, thiamine, vitamin B12 (cobalamin), nicotinic acid, and pantothenic acid [Wolin et al., 1963]. L-cysteine, protoheme, CO_2, Zn^{2+}, and Phosphate were additionally defined for minimal medium according to the reference article. For all these metabolites, their upper and lower boundaries were set to 100 and -100 respectively.

The experimental D-glucose uptake rate was introduced to the model as 29.1 mmol/gDW. The upper and lower boundaries of the exchange reactions for other metabolites were set to zero. In addition, the measured secretion rates of succinate, acetate, and formate are reported as 22.7, 22.6, and 5.5 mmol/gDW respectively. The reported growth yield in the reference article is 34.4 g DW/mol glucose [Franke and Deppenmeier, 2018].

The reconstructed model could not simulate non-zero biomass flux rate under the growth conditions that were introduced by Franke and Deppenmeier (2018) without utilizing an additional carbon source, especially N-Acetyl-D-glucosamine. Since N-Acetyl-D-glucosamine is used for the synthesis of some biomass precursors, the intermediate reactions required to provide the synthesis of biomass metabolites through the same pathway without requiring an additional carbon source was identified. Therefore, G1PACT reaction, glucosamine-1-phosphate N-acetyltransferase reaction, from *i*AH991 model was added to our model. In addition, the missing transport reaction with GPR rule for Mn+2, a biomass precursor metabolite, was added to our model via CarveMe model.

The citrate synthase (CS) reaction significantly decreases the flow of the route that produces malate from oxaloacetate, fumarate from malate, and ultimately succinate from malate, through the inverse Krebs cycle. Since the CS reaction uses oxaloacetate to produce citrate leading to the initiation of TCA cycle, it was prevented by constraining its rate to zero. In addition, in the reference article, it was stated that *Prevotella copri* has CO_2 dependence for growth in minimal medium. They reported that 5.6 mmol CO_2/g DW was consumed by the organism. Thus, the reported value was introduced to the model as a constraint for CO_2 consumption.

At last, our model, *i*Pc621, could simulate the non-zero growth under the determined growth condition. The obtained result under this simulation condition is higher than the expected growth rate (Table 5.1). Additionally, the model produced 28.6 mmol/gDW acetate and 54.3 mmol/gDW formate, a higher rate than the reported value, which are 22.6 and 5.5 mmol/gDW respectively, while failing to form succinate. Besides, CO_2 consumption has an effect on the production of these two fermentation products.

Table 5.1: The Anaerobic Growth Simulation Results Under Modified Minimal Medium.

	The Experimental Data	*i*Pc621 (*In silico*)
Growth Rate (1/h)	1 [Franke and Deppenmeier, 2018]	5.4

The constraints set in a model significantly affect the flow of reactions and thus the flow of growth. The medium constraints were not exactly known for most of the metabolites for the reported experimental condition, and, therefore, uptake rates were mostly defined as intervals. In addition, improvement of the model with more experimental data of *Prevotella copri DSM 12805* will affect the results positively.

Due to the difference in simulation results, the model was checked for ATP leak. The maximum ATP generation of *i*Pc621 under non-zero growth is 98.8, so there is no ATP leakage in the model. Furthermore, *i*Pc621 does not prefer to produce succinate according to the optimization result aimed at maximum growth under these conditions. However, if the experimentally measured value of succinate was introduced to the model as a constraint to produce it, the model succeeded in succinate formation but this time it could not produce acetate. Therefore, in addition, the exchange reactions were checked in order to control produced and consumed metabolites in the model. Production of glucuronate and xanthine were non-zero in the predicted flux distribution. The production rates of those metabolites were constrained to zero as the pathways associated with production of these metabolites could affect the acetate production capacity of the model and there is no experimental work in literature that reports the production of those metabolites in the studied minimal media. Consequently, *i*Pc621 could simulate growth by producing all three fermentation products (Table 5.2).

Table 5.2: *In silico* simulation results predicted by iPc621 Model.

	Experimental data	iPc621 Model (*in silico*)
Growth rate (1/h)	1	4.1
Succinate (mmol g/DW)	22.7	22.7
Acetate (mmol g/DW)	22.6	51.6
Formate (mmol g/DW)	5.5	15.3

According to the predicted results by iPc621, overall, obtained flux rates are higher than the experimentally measured values. Nevertheless, the model we obtained is a functional model that has managed to grow on glucose, the only carbon source in experimental conditions in the reference article. *i*Pc621 model also produced succinate from fumarate, the format from acetyl-coA and pyruvate, and acetate via acetate kinase reaction in accordance with the article.

5.3.2. Gene Essentiality Simulations

The gene essentiality analysis was performed for *P.copri DSM 12805* organism. The essential genes were predicted by using the draft CarveMe model and *i*Pc621 model for *P.copri DSM 12805*. In gene essentiality analysis, a richer medium environment was used instead of minimal medium as it may be more suitable for the habitat of the organism. The growth simulation was performed at 29.1 mmol/gDW/ h glucose uptake rate with rich medium and anaerobic conditions based on the FBA approach by using the '*singleGeneDeletion*' built-in function of COBRA Toolbox. This function runs several FBA simulations after the deletion of each gene in the model and predicts the growth rate

for each condition. The gene that prevents the growth is considered as the essential gene. Consequently, 66 genes were identified as essential for growth of the organism according to our model, *i*Pc621. In addition, the deletion of 207 genes had no impacts at all on growth according to *in silico* model prediction for *i*Pc621 model. The single gene deletion simulations of our model was compared to the computationally essential genes predicted by draft CarveMe model. According to the results of CarveMe model, 92 genes are essential for the organism while 139 genes do not affect the growth. Furthermore, 33 of 66 essential metabolic genes identified by iPc621 are common with the CarveMe model.

The 66 survival genes are involved in important pathways for the organism's metabolism. Tyrosine, Tryptophan, and Phenylalanine Metabolism, Purine and Pyrimidine Biosynthesis, and Cell Envelope Biosynthesis draw attention. Their further investigations may contribute to the treatments for many diseases for example, obesity with insulin resistance, the rheumatoid arthritis, and Parkinson’s disease that are caused from the bacteria inhabited in human gut (Section 2.5).

6. CONCLUSIONS

The genome-scale metabolic network models have potential for understanding of the growth behavior, nutritional needs, survival genes, and to develop treatments for many diseases. In this study, we developed two GEMs for two different bacteria that are associated with many diseases. The target bacteria to build models for are *Klebsiella pneumonaie HS11286* and *Prevotella copri DSM 12805*. *Klebsiella pneumonaie HS11286* is mainly responsible for mortality associated infectious diseases with multi-drug resistance (Section 2.4). *Prevotella copri DSM 12805* is a beneficial bacterium that live in human microbiota. In addition, when comparing the amount of *P.copri* in the gut microbiota of patients and healthy controls, it is determined to be significantly associated with obesity, HIV, chronic inflammatory diseases, rheumatoid arthritis and Parkinson's disease (Section 2.5).

The draft GEMs were reconstructed by using genomic data and the high-quality GEMs of genetically similar organisms with the target bacteria. The genetic similarity between the reference organism and the target organism affects the size and quality of the draft model we obtained. The genetic similarity between *K.pneumoniae HS11286* and its reference organism, *K.pneumoniae KPPR1*, is much higher than the genetic similarity between *P.copri DSM 12805* and its reference organism, *Bacteroides thetaiotaomicron VPI-5482*. Therefore, the number of reactions and the related genes in the draft model for *K.pneumoniae HS11286* are higher than the draft model we created for *P.copri DSM 12805*. In addition to them, the more reactions the draft model contains from the metabolism of the target organism, the better, and it facilitates the subsequent manual evaluation steps.

The manual curation of the GEMs is an essential step to have functional and accurate model. The two reconstructed models, *i*Kp1481 and *i*Pc621, were curated via gap-filling steps, then the models were integrated with KEGG-based model and draft CarveMe model, respectively, to increase the gene coverage. Finally, *i*Kp1481 and *i*Pc621 were validated by performing growth simulations via FBA and comparison with the experimental data. Also, gene essentiality analyses were accomplished for *Klebsiella*

pneumonaie HS11286 and *Prevotella copri DSM 12805* by using *i*Kp1481 and *i*Pc621 models. It is verified that *i*Kp1481 and *i*Pc621 are functional.

*i*Kp1481 model was successful even though the growth predictions performed on different carbon sources were higher than other models (Section 4.3.1). Still, the reactions in the model can be studied in more detail to understand why the model had a higher growth rate. In addition to the experimental data of genetically highly similar organisms, the comparison of the experimental data of *K.pneumoniae HS11286* strain with the growth simulation results of the model would be more accurate and reliable. For future studies, the growth phenotype estimation on nitrogen, phosphorus and sulfur sources other than carbon source can be performed and the simulation results can be compared with the experimental data of *K.pneumoniae HS11286* strain.

The *i*Pc621 model is accomplished to be compatible with the metabolism of *P.copri*. Particularly, it could reflect the carbon flow although it gives higher biomass flux rate in the growth simulation under the modified DMMG condition. *i*Pc621 model should be supported and improved with more literature survey and manual curation in order to obtain more accurate results for *in silico* environment and more comprehensive metabolic network model.

At last, for both reconstructed model, other omics data such as transcriptome and metabolome data can be integrated to the GEMs in order to improve them and to have more compatible *in silico* predictions with the experimental results and better representation of the organism's metabolism.

REFERENCES

Alpizar-Rodriguez D., Lesker T. R., Gronow A., Gilbert B., Raemy E., Lamacchia C., Gabay C., Finckh A., Strowig T., (2019), "Prevotella copri in individuals at risk for rheumatoid arthritis", Ann Rheum Dis, 78 (5), 590-593.

Arakawa K., Yamada Y., Shinoda K., Nakayama Y., Tomita M., (2006), "GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes", BMC Bioinformatics, 7, 168.

Arkin A. P., Cottingham R. W., Henry C. S., Harris N. L., Stevens R. L., Maslov S., Dehal P., Ware D., Perez F., Canon S., (2018), "KBase: the United States department of energy systems biology knowledgebase", Nature biotechnology, 36 (7), 566.

Bedarf J. R., Hildebrand F., Coelho L. P., Sunagawa S., Bahram M., Goeser F., Bork P., Wullner U., (2017), "Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naive Parkinson's disease patients", Genome Med, 9 (1), 39.

Bertrand R. L., (2019), "Lag Phase Is a Dynamic, Organized, Adaptive, and Evolvable Period That Prepares Bacteria for Cell Division", Journal of bacteriology, 201 (7).

Bi D., Jiang X., Sheng Z. K., Ngmenterebo D., Tai C., Wang M., Deng Z., Rajakumar K., Ou H. Y., (2015), "Mapping the resistance-associated mobilome of a carbapenem-resistant Klebsiella pneumoniae strain reveals insights into factors shaping these regions and facilitates generation of a 'resistance-disarmed' model organism", J Antimicrob Chemother, 70 (10), 2770-2774.

Bisen P. S., Debnath M., Prasad G. B. K. S., (2012), "Microbes : concepts and applications", Edition, Wiley-Blackwell.

Bordbar A., Monk J. M., King Z. A., Palsson B. O., (2014), "Constraint-based models predict metabolic and associated cellular functions", Nat Rev Genet, 15 (2), 107-120.

Brenner D. J., Krieg N. R., Staley J. T., Garrity G. M., (2005), "Bergey's manual of systematic bacteriology. Vol. Two, Part B, Vol. Two, Part B", Edition, Springer.

Bridges B. A., (1998), "The role of DNA damage in stationary phase (`adaptive') mutation", Mutation Research/DNA Repair Mutation Research/DNA Repair, 408 (1), 1-9.

Cabeen M. T., Jacobs-Wagner C., (2005), "Bacterial cell shape", Nat Rev Microbiol, 3 (8), 601-610.

Campos A. C., Albiero J., Ecker A. B., Kuroda C. M., Meirelles L. E., Polato A., Tognim M. C., Wingeter M. A., Teixeira J. J., (2016), "Outbreak of Klebsiella pneumoniae carbapenemase-producing K pneumoniae: A systematic review", Am J Infect Control, 44 (11), 1374-1380.

Capy P., (1998), "A plastic genome", Nature Nature, 396 (6711), 522-523.

Cesur M. F., Siraj B., Uddin R., Durmuş S., Çakır T., (2020), "Network-based metabolism-centered screening of potential drug targets in Klebsiella pneumoniae at genome scale", Frontiers in Cellular and Infection Microbiology, 9, 447.

Chan S. H. J., Cai J., Wang L., Simons-Senftle M. N., Maranas C. D., (2017), "Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models", Bioinformatics, 33 (22), 3603-3609.

Christaki E., Marcou M., Tofarides A., (2019), "Antimicrobial Resistance in Bacteria: Mechanisms, Evolution, and Persistence", J Mol Evol.

Coico R., Wiley I., (2005), "Current protocols in microbiology".

Conway T., Cohen P. S., (2015), "Metabolism and bacterial pathogenesis", Edition, ASM Press.

De Jesus M. B., Ehlers M. M., Dos Santos R. F., Kock M. M. (2015), "Review - Understanding β-lactamase Producing Klebsiella pneumoniae". "Antimicrobial Resistance - An Open Challenge".

Dias O., Rocha M., Ferreira E. C., Rocha I., (2015), "Reconstructing genome-scale metabolic models with merlin", Nucleic Acids Res, 43 (8), 3899-3910.

Drancourt M., Bollet C., Carta A., Rousselier P., (2001), "Phylogenetic analyses of Klebsiella species delineate Klebsiella and Raoultella gen. nov., with description of Raoultella ornithinolytica comb. nov., Raoultella terrigena comb. nov. and Raoultella planticola comb. nov", INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY, 51, 925-932.

Durot M., Bourguignon P. Y., Schachter V., (2009), "Genome-scale models of bacterial metabolism: reconstruction and applications", FEMS Microbiol Rev, 33 (1), 164-190.

Ebrahim A., Lerman J. A., Palsson B. O., Hyduke D. R., (2013), "COBRApy: constraints-based reconstruction and analysis for python", BMC systems biology, 7 (1), 74.

Edwards J. S., Palsson B. O., (1999), "Systems properties of the Haemophilus influenzae Rd metabolic genotype", JOURNAL OF BIOLOGICAL CHEMISTRY, 274 (25), 17410-17416.

Edwards J. S., Palsson B. O., (2000), "The Escherichia coli MG1655 in Silico Metabolic Genotype: Its Definition, Characteristics, and Capabilities", procnatiacadscie Proceedings of the National Academy of Sciences of the United States of America, 97 (10), 5528-5533.

Falagas M. E., Rafailidis P. I., Matthaiou D. K., Virtzili S., Nikita D., Michalopoulos A., (2008), "Pandrug-resistant Klebsiella pneumoniae, Pseudomonas aeruginosa and Acinetobacter baumannii infections: characteristics and outcome in a series of 28 patients", Int J Antimicrob Agents, 32 (5), 450-454.

Faria J. P., Rocha M., Rocha I., Henry C. S., (2018), "Methods for automated genome-scale metabolic model reconstruction", Biochem Soc Trans, 46 (4), 931-936.

Feist A. M., Herrgard M. J., Thiele I., Reed J. L., Palsson B. O., (2009), "Reconstruction of biochemical networks in microorganisms", Nat Rev Microbiol, 7 (2), 129-143.

Franke T., Deppenmeier U., (2018), "Physiology and central carbon metabolism of the gut bacterium Prevotella copri", Mol Microbiol, 109 (4), 528-540.

Fredrickson J. K., Zachara J. M., Balkwill D. L., Kennedy D., Li S. M., Kostandarithes H. M., Daly M. J., Romine M. F., Brockman F. J., (2004), "Geomicrobiology of high-level nuclear waste-contaminated vadose sediments at the hanford site, washington state", Appl Environ Microbiol, 70 (7), 4230-4241.

Fuhrer T., Fischer E., Sauer U., (2005), "Experimental identification and quantification of glucose metabolism in seven bacterial species", J Bacteriol, 187 (5), 1581-1590.

Gottschalk G., (1986), "Bacterial metabolism", Edition, Springer.

Gu C., Kim G. B., Kim W. J., Kim H. U., Lee S. Y., (2019), "Current status and applications of genome-scale metabolic models", Genome Biol, 20 (1), 121.

Gurobi Optimization I., (2018), "Gurobi optimizer reference manual", URL http://www. gurobi. com.

Haggart C. R., Bartell J. A., Saucerman J. J., Papin J. A., (2011), "Whole-genome metabolic network reconstruction and constraint-based modeling", Methods Enzymol, 500, 411-433.

Hamer H. M., De Preter V., Windey K., Verbeke K., (2012), "Functional analysis of colonic bacterial metabolism: relevant to health?", Am J Physiol Gastrointest Liver Physiol, 302 (1), G1-9.

Hamilton J. J., Reed J. L., (2014), "Software platforms to facilitate reconstructing genome-scale metabolic networks", Environ Microbiol, 16 (1), 49-59.

Hanemaaijer M., Olivier B. G., Roling W. F., Bruggeman F. J., Teusink B., (2017), "Model-based quantification of metabolic interactions from dynamic microbial-community data", PLoS One, 12 (3), e0173183.

Hayashi H., Shibata K., Sakamoto M., Tomita S., Benno Y., (2007), "Prevotella copri sp. nov. and Prevotella stercorea sp. nov., isolated from human faeces", Int J Syst Evol Microbiol, 57 (Pt 5), 941-946.

Heinken A., Sahoo S., Fleming R. M., Thiele I., (2013), "Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut", Gut microbes, 4 (1), 28-40.

Heinken A., Thiele I., (2015), "Systematic prediction of health-relevant human-microbial co-metabolism through a computational framework", Gut microbes, 6 (2), 120-130.

Heirendt L., Arreckx S., Pfau T., Mendoza S. N., Richelle A., Heinken A., Haraldsdóttir H. S., Wachowiak J., Palsson B. Ø., Thiele I., Fleming R. M. T., (2019), "Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0", Nature Protocols, 14 (3), 639-702.

Henry C. S., DeJongh M., Best A. A., Frybarger P. M., Linsay B., Stevens R. L., (2010), "High-throughput generation, optimization and analysis of genome-scale metabolic models", Nat Biotechnol, 28 (9), 977-982.

Henry C. S., Rotman E., Lathem W. W., Tyo K. E., Hauser A. R., Mandel M. J., (2017), "Generation and Validation of the iKp1289 Metabolic Model for Klebsiella pneumoniae KPPR1", J Infect Dis, 215 (suppl_1), S37-S43.

Henze M., Loosdrecht M. C. M. v., Ekama G. A., Brdjanovic D., (2008), "Biological wastewater treatment principles, modelling and design".

Horton P., Park K. J., Obayashi T., Fujita N., Harada H., Adams-Collier C. J., Nakai K., (2007), "WoLF PSORT: protein localization predictor", Nucleic Acids Res, 35 (Web Server issue), W585-587.

Kanehisa M., Sato Y., Kawashima M., Furumichi M., Tanabe M., (2016), "KEGG as a reference resource for gene and protein annotation", Nucleic acids research, 44 (D1), D457-D462.

Karp P. D., Latendresse M., Paley S. M., Krummenacker M., Ong Q. D., Billington R., Kothari A., Weaver D., Lee T., Subhraveti P., (2016), "Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology", Briefings in bioinformatics, 17 (5), 877-890.

Karplus K., Barrett C., Hughey R., (1998), "Hidden Markov models for detecting remote protein homologies", Bioinformatics (Oxford, England), 14 (10), 846-856.

Keshavarzian A., Green S. J., Engen P. A., Voigt R. M., Naqib A., Forsyth C. B., Mutlu E., Shannon K. M., (2015), "Colonic bacterial composition in Parkinson's disease", Mov Disord, 30 (10), 1351-1360.

Kim B. H., Gadd G. M., (2008), "Bacterial physiology and metabolism", Edition, Cambridge University Press.

Koonin E. V., (2005), "Orthologs, paralogs, and evolutionary genomics", Annu Rev Genet, 39, 309-338.

Kram K. E., Finkel S. E., (2015), "Rich Medium Composition Affects Escherichia coli Survival, Glycation, and Mutation Frequency during Long-Term Batch Culture", Appl Environ Microbiol, 81 (13), 4442-4450.

Kumar M., Ji B., Babaei P., Das P., Lappa D., Ramakrishnan G., Fox T. E., Haque R., Petri W. A., Backhed F., Nielsen J., (2018), "Gut microbiota dysbiosis is associated with malnutrition and reduced plasma amino acid levels: Lessons from genome-scale metabolic modeling", Metab Eng, 49, 128-142.

Lewis N. E., Nagarajan H., Palsson B. O., (2012), "Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods", Nat Rev Microbiol, 10 (4), 291-305.

Ley R. E., (2016), "Gut microbiota in 2015: Prevotella in the gut: choose carefully", Nat Rev Gastroenterol Hepatol, 13 (2), 69-70.

Liao Y. C., Huang T. W., Chen F. C., Charusanti P., Hong J. S., Chang H. Y., Tsai S. F., Palsson B. O., Hsiung C. A., (2011), "An experimentally validated genome-scale metabolic reconstruction of Klebsiella pneumoniae MGH 78578, iYL1228", J Bacteriol, 193 (7), 1710-1717.

Liu P., Li P., Jiang X., Bi D., Xie Y., Tai C., Deng Z., Rajakumar K., Ou H. Y., (2012), "Complete genome sequence of Klebsiella pneumoniae subsp. pneumoniae HS11286, a multidrug-resistant strain isolated from human sputum", J Bacteriol, 194 (7), 1841-1842.

Machado D., Andrejev S., Tramontano M., Patil K. R., (2018), "Fast automated reconstruction of genome-scale metabolic models for microbial species and communities", Nucleic Acids Res, 46 (15), 7542-7553.

Magnusdottir S., Heinken A., Kutt L., Ravcheev D. A., Bauer E., Noronha A., Greenhalgh K., Jager C., Baginska J., Wilmes P., Fleming R. M., Thiele I., (2017), "Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota", Nat Biotechnol, 35 (1), 81-89.

Mao X., Cai T., Olyarchuk J. G., Wei L., (2005), "Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary", Bioinformatics, 21 (19), 3787-3793.

Mendoza S. N., Olivier B. G., Molenaar D., Teusink B., (2019), "A systematic assessment of current genome-scale metabolic reconstruction tools", Genome Biol, 20 (1), 158.

Mienda B. S., (2017), "Genome-scale metabolic models as platforms for strain design and biological discovery", J Biomol Struct Dyn, 35 (9), 1863-1873.

Monk J. M., Lloyd C. J., Brunk E., Mih N., Sastry A., King Z., Takeuchi R., Nomura W., Zhang Z., Mori H., (2017), "i ML1515, a knowledgebase that computes Escherichia coli traits", Nature biotechnology, 35 (10), 904-908.

Moyes R. B., Reynolds J., Breakwell D. P., (2009), "Differential staining of bacteria: Gram stain", Curr. Protoc. Microbiol. Current Protocols in Microbiology (SUPPL. 15), A.3C.1-A.3C.8.

Munoz-Elias E. J., McKinney J. D., (2006), "Carbon metabolism of intracellular bacteria", Cell Microbiol, 8 (1), 10-22.

Navid A., (2011), "Applications of system-level models of metabolism for analysis of bacterial physiology and identification of new drug targets", Brief Funct Genomics, 10 (6), 354-364.

Nealson K. H., (1999), "Post-Viking Microbiology: New Approaches, New Data, New Insights", Origins of Life and Evolution of the Biosphere, 29 (1), 73-93.

Nikaido H., (2009), "Multidrug resistance in bacteria", Annu Rev Biochem, 78, 119-146.

Olivier B., (2018), https://doi.org/10.5281/zenodo.2398336, (Erişim Tarihi: 14.10.2018).

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

Papagianni M., (2012), "Recent advances in engineering the central carbon metabolism of industrially important bacteria", Microbial Cell Factories, 11 (1), 1-13.

Paramasivam N., Linke D., (2011), "ClubSub-P: cluster-based subcellular localization prediction for Gram-negative bacteria and archaea", Frontiers in microbiology, 2, 218.

Petrov V. A., Saltykova I. V., Zhukova I. A., Alifirova V. M., Zhukova N. G., Dorofeeva Y. B., Tyakht A. V., Kovarsky B. A., Alekseev D. G., Kostryukova E. S., Mironova Y. S., Izhboldina O. P., Nikitina M. A., Perevozchikova T. V., Fait E. A., Babenko V. V., Vakhitova M. T., Govorun V. M., Sazonov A. E., (2017), "Analysis of Gut Microbiota in Patients with Parkinson's Disease", Bull Exp Biol Med, 162 (6), 734-737.

Pianta A., Arvikar S., Strle K., Drouin E. E., Wang Q., Costello C. E., Steere A. C., (2017), "Evidence of the Immune Relevance of Prevotella copri, a Gut Microbe, in Patients With Rheumatoid Arthritis", Arthritis Rheumatol, 69 (5), 964-975.

Precup G., Vodnar D. C., (2019), "Gut Prevotella as a possible biomarker of diet and its eubiotic versus dysbiotic roles: a comprehensive literature review", Br J Nutr, 122 (2), 131-140.

Prigent S., Frioux C., Dittami S. M., Thiele S., Larhlimi A., Collet G., Gutknecht F., Got J., Eveillard D., Bourdon J., Plewniak F., Tonon T., Siegel A., (2017), "Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide Metabolic Networks", PLoS Comput Biol, 13 (1), e1005276.

Ramos-Castaneda J. A., Ruano-Ravina A., Barbosa-Lorenzo R., Paillier-Gonzalez J. E., Saldana-Campos J. C., Salinas D. F., Lemos-Luengas E. V., (2018), "Mortality due to KPC carbapenemase-producing Klebsiella pneumoniae infections: Systematic review and meta-analysis: Mortality due to KPC Klebsiella pneumoniae infections", J Infect, 76 (5), 438-448.

Reed J. L., Patel T. R., Chen K. H., Joyce A. R., Applebee M. K., Herring C. D., Bui O. T., Knight E. M., Fong S. S., Palsson B. O., (2006), "Systems approach to refining genome annotation", Proc Natl Acad Sci U S A, 103 (46), 17480-17484.

Riedel S., Morse S. A., Mietzner T. A., Miller S., (2019), "Jawetz, Melnick & Adelberg's medical microbiology".

Roy Sarkar S., Banerjee S., (2019), "Gut microbiota in neurodegenerative disorders", J Neuroimmunol, 328, 98-104.

Satish Kumar V., Dasika M. S., Maranas C. D., (2007), "Optimization based automated curation of metabolic reconstructions", BMC Bioinformatics, 8, 212.

Scheperjans F., Aho V., Pereira P. A., Koskinen K., Paulin L., Pekkonen E., Haapaniemi E., Kaakkola S., Eerola-Rautio J., Pohja M., Kinnunen E., Murros K., Auvinen P., (2015), "Gut microbiota are related to Parkinson's disease and clinical phenotype", Mov Disord, 30 (3), 350-358.

Sears C. L., (2005), "A dynamic partnership: celebrating our gut flora", Anaerobe, 11 (5), 247-251.

Segre D., Vitkup D., Church G. M., (2002), "Analysis of optimality in natural and perturbed metabolic networks", Proc Natl Acad Sci U S A, 99 (23), 15112-15117.

Sender R., Fuchs S., Milo R., (2016), "Revised estimates for the number of human and bacteria cells in the body", PLoS biology, 14 (8), e1002533.

Shoaie S., Karlsson F., Mardinoglu A., Nookaew I., Bordel S., Nielsen J., (2013), "Understanding the interactions between bacteria in the human gut through metabolic modeling", Scientific reports, 3, 2532.

Stanier R. Y., van Niel C. B., (1962), "The concept of a bacterium", Archiv. Mikrobiol. Archiv für Mikrobiologie, 42 (1), 17-35.

Thanbichler M., Wang S. C., Shapiro L., (2005), "The bacterial nucleoid: a highly organized and dynamic structure", J Cell Biochem, 96 (3), 506-521.

Thiele I., Palsson B. O., (2010), "A protocol for generating a high-quality genome-scale metabolic reconstruction", Nat Protoc, 5 (1), 93-121.

Thiele I., Vlassis N., Fleming R. M., (2014), "fastGapFill: efficient gap filling in metabolic networks", Bioinformatics, 30 (17), 2529-2531.

Tiukova I. A., Prigent S., Nielsen J., Sandgren M., Kerkhoven E. J., (2019), "Genome-scale model of Rhodotorula toruloides metabolism", Biotechnology and bioengineering, 116 (12), 3396-3408.

Troyanskaya O., Fang G., Bhardwaj N., Robilotto R., Gerstein M. B., (2010), "Getting Started in Gene Orthology and Functional Analysis", PLoS Computational Biology, 6 (3). Varel V. H., Bryant M. P., (1974), "Nutritional features of Bacteroides fragilis subsp. fragilis", Applied microbiology, 28 (2), 251-257.

Wamelink M. M., Struys E. A., Jakobs C., (2008), "The biochemistry, metabolism and inherited defects of the pentose phosphate pathway: a review", J Inherit Metab Dis, 31 (6), 703-717.

Wang H., Marcisauskas S., Sanchez B. J., Domenzain I., Hermansson D., Agren R., Nielsen J., Kerkhoven E. J., (2018), "RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor", PLoS Comput Biol, 14 (10), e1006541.

Wang J., Wang C., Song K., Wen J., (2017), "Metabolic network model guided engineering ethylmalonyl-CoA pathway to improve ascomycin production in Streptomyces hygroscopicus var. ascomyceticus", Microb Cell Fact, 16 (1), 169.

Wang X., Xia K., Yang X., Tang C., (2019), "Growth strategy of microbes on mixed carbon sources", Nat Commun, 10 (1), 1279.

Ward B. (2015), "Bacterial energy metabolism". "Molecular Medical Microbiology", Elsevier.

Web 1, (2019), https://github.com/ModelSEED/ModelSEEDDatabase, (Erişim Tarihi: 20/11/2019).

Whitman W. B., Coleman D. C., Wiebe W. J., (1998), "Prokaryotes: The Unseen Majority", procnatiacadscie Proceedings of the National Academy of Sciences of the United States of America, 95 (12), 6578-6583.

Wolin E., Wolin M. J., Wolfe R., (1963), "Formation of methane by bacterial extracts", JOURNAL OF BIOLOGICAL CHEMISTRY, 238 (8), 2882-2886.

Xu C., Liu L., Zhang Z., Jin D., Qiu J., Chen M., (2013), "Genome-scale metabolic model in guiding metabolic engineering of microbial improvement", Appl Microbiol Biotechnol, 97 (2), 519-539.

Ye C., Xu N., Chen H., Chen Y. Q., Chen W., Liu L., (2015), "Reconstruction and analysis of a genome-scale metabolic model of the oleaginous fungus Mortierella alpina", BMC Syst Biol, 9, 1.

Yoon B.-J., (2009), "Hidden Markov models and their applications in biological sequence analysis", Current genomics, 10 (6), 402-415.

Yu N. Y., Wagner J. R., Laird M. R., Melli G., Rey S., Lo R., Dao P., Sahinalp S. C., Ester M., Foster L. J., (2010), "PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes", Bioinformatics, 26 (13), 1608-1615.

BIOGRAPHY

Betül Baz was born in Seyhan, Adana on October 24, 1993. She graduated from Istanbul University Department of Genetics and Bioengineering in 2016. She is an MSc student in the Bioinformatics and Systems Biology Program under Bioengineering Department at Graduate School of Natural and Applied Sciences in Gebze Technical University.

APPENDICES