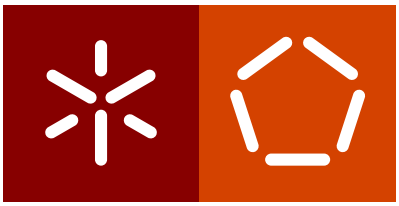
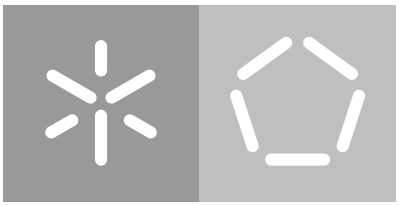


Universidade do Minho
Escola de Engenharia
Departamento de Informática
CEB - Centro de Engenharia Biológica
Chr.Hansen A/S

José Miguel Gonçalves Dias

Reconstructing the metabolic network of Lactobacillus helveticus on a genome-wide scale

September 2017



Universidade do Minho
Escola de Engenharia
Departamento de Informática
CEB - Centro de Engenharia Biológica
Chr.Hansen A/S

José Miguel Gonçalves Dias

Reconstructing the metabolic network of Lactobacillus helveticus on a genome-wide scale

Thesis dissertation
Master Degree in Bioinformatics

Supervisors
Isabel Cristina Almeida Pereira Rocha
Oscar Manuel Lima Dias
Ahmad Adel Zeidan

September 2017

'Winter is Coming'
George R. R. Martin
'I did not come this far to only come this far'
Uncertain author

ACKNOWLEDGMENTS

This thesis represents the pinnacle of many years of work and my evolution and growth as person and professional. To achieve it I cannot forget all the people who accompanied me in this stage. First of all, I want to acknowledge my supervisors for all the help provided. To Doctor Isabel Rocha for all the knowledge transmitted and for pulling up my interest in the Systems Biology field. A special thanks to Doctor Oscar Dias for being available at almost any hour, maintaining proximity even during my time abroad. Another special word to Doctor Ahmad Zeidan for being an infinite source of knowledge in so many diverse fields from microbiology to bioinformatics and for making possible my stay in Denmark. I sincerely hope they substitute the acidic coffee in the future.

I want to also thank Doctor Ana Rute Neves and Doctor Patrick Derkx for receiving me in Chr. Hansen facilities and making available all the necessary tools for this project. I will bring bombocas in my next visit! Maiken and Paula, thank you for the time spent helping me in the lab work and the patience to try the get understand and improv the results. Another word to all the people I met in Denmark who helped me to improve professionally and as a person and also to have a good time: Alfonso, Julia, Lisandra, Martin, Patricia, Shashank, Sonali, Thomas, thank you all. I want also to thank my colleagues in the Byosistems department: António, Bruna, Fernando e Pedro. A special word to Fernando who accompanied me in this experience abroad and helped me in all the moments. Thank you for all the help in script development and all the hours and good time and talking outside work. Wish you all the success! Another thanks to all my hometown friends that despite my absence to maintain always the feeling that everything remains the same.

A very big acknowledgement to Sara who has been with me in every single moment in the last few years, sharing the better and worst moments no matter the distance. You believe in me in every second and make me better every day Thank you for all the trust and strength you transmit me. Thank you also for the painful review of this thesis.

Finally, a big THANK YOU to my family. Without your support all my achievements would not be possible. Thank you for all the foundations you offered me and for transmitting me the principles of Education, Honesty and Hard Work.

OBRIGADO !

ABSTRACT

The constant growth of high-throughput data generation and omics approaches require informatics support and (semi) automated processes to be developed. With increasing number of sequenced genomes available, metabolic engineering processes will allow a rational alteration of the genetic architecture to achieve specific phenotypes. These alterations will allow to generate and optimize features of some organisms with economic and health interest.

Lactobacillus helveticus is an important industrial lactic-acid bacterium being used in the production of several types of cheese. The metabolic activities of the bacterium contribute to the cheese flavour and reduce bitterness. *Lb. helveticus* is a growing body of literature on the health-promoting properties of its various strains and generally accepted as probiotic for its anti-mutagenic, immunomodulatory and anti-diarrheal effects.

The aim of this project was to reconstruct a genome-scale metabolic network of *Lb. helveticus* CNRZ32, based on its genome sequence annotation as well as known biochemical and physiological characteristics. The generated model contained 790 reactions, 894 metabolites and 1687 genes. The growth rate predicted by the model on sugar was comparable to the reported in literature.

This model provides the basis for a constraint-based mathematical model capable of simulating the phenotype of the organism under different growth conditions and guiding in-depth physiological studies and hypothesis generation.

Keywords: metabolic network, *Lactobacillus helveticus*, *Metabolic Models Reconstruction using Genome-Scale Information (merlin)*, computational biology; enzymes; transporters; TRIAGE; COBRA; OptFlux

RESUMO

O crescimento constante do volume de dados de alto rendimento gerados e de abordagens ómicas urgem de desenvolvimento de suporte informático e processos (semi) automatizados. O aumento do número de genomas sequenciados disponíveis, os processos de engenharia metabólica permitirá uma alteração racional da arquitetura genética para alcançar fenótipos específicos. Estas alterações irão permitir gerar e otimizar características de organismos com interesse económico e de saúde. *Lactobacillus helveticus* é uma bactéria lática com importância para o uso industrial e utilizada na produção de vários tipos de queijo. A atividade metabólicas da bactéria contribui para o sabor do queijo e para a redução da sua acidez. *Lb. Helveticus* é geralmente aceite como probiótico, com um crescente volume de literatura sobre as suas propriedades que contribuem positivamente para a saúde em várias das suass estirpes, assim como os seus efeitos antimutagénicos, imunomoduladores e antidiarreicos.

O objetivo deste projeto é gerar uma reconstrução da rede metabólica à escala geneomica de *Lb. helveticus* CNRZ32 baseado na anotação de sequência do genoma, bem como das suas características bioquímicas e fisiológicas. O modelo gerado continha 790 reações, 894 reações e 1687 genes. A taxa de crescimento prevista pelo modelo sobre o açúcar é comparável ao relatado na literatura.

A reconstrução deste modelo serve como base para a rescontrução de modelo matemático baseado em restrições capaz de simular o fenótipo do organismo sob diferentes condições de crescimento e orientar estudos fisiológicos em profundidade e geração de hipóteses.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

A

ADP Adenosine diphosphate. 1, 51

AMP Adenosine monophosphate. 1, 50

ATP Adenosine triphosphate. 1, 22, 39, 50, 51, 55

B

BiGG Biochemical, Genetic and Genomic knowledge base. 1, 16, 17, 38, 48, 58

BKM-react BRENDA–KEGG–MetaCyc reactions. 1, 15

BLAST Basic Local Alignment Search Tool. xi, 1, 14, 18, 30–32

BRENDA BRaunschweig ENzyme DAtabase. 1, 14, 15, 20, 30, 32, 38

BSA Bovine serum albumin. 1, 40

C

CBS Center for Biological Sequence Analysis. 1, 16, 17

CDM chemically defined medium. 1, 40, 96

CDS Coding Sequences. 1, 14, 30

ChEBI Chemical Entities of Biological Interest. 1, 13, 15

COBRApy COnstraint-Based Reconstruction and Analysis for Python. ix, 1, 42, 52, 59

CoReCo Comparative ReConstruction of genome-scale metabolic networks. 1, 12

D

DI Departamento de Informática. 1

DNA Deoxyribonucleic acid. 1, 8, 9, 24, 39, 41, 49, 51

DOE Department of Energy. 1, 15, 16

E

EC Enzyme Commission. xiii, 1, 12, 16, 18, 32, 33, 37, 44, 46, 54

EECG Enzyme Encoding Candidate Genes. ix, 1, 14, 18, 30–33, 44

EPS Exopolyssacharide. viii, 1, 25, 50, 51, 55

ExPASy Expert Protein Analysis System. 1, 15, 17, 32

F

FAD Flavin adenine dinucleotide. 1, 54

FAME Flux Analysis and Modeling Environment. 1, 12

FBA Flux Balance Analysis. viii, 1, 14, 23, 42

FMN Flavin mononucleotide. 1, 54

G

g grams. 1, 42, 51

GOLD Genomes OnLine Database. 1, 16, 17

GPR Gene-Protein-Reaction. 1, 12, 18

GRAS Generally Recognized as Safe. 1

GSMM Genome Scale Metabolic Models. 1, 11–14, 16, 18, 20–22, 59, 60

GSMN Genome Scale Metabolic Network. 1, 10, 21

H

HMM hidden Markov model. 1, 35

HMMER biosequence analysis using profile hidden Markov models. 1, 18

I

IUBMB Union of Biochemistry and Molecular Biology. 1, 14, 18

IUPAC International Union of Pure and Applied Chemistry. 1, 3, 15

J

JGI Joint Genome Institute. 1, 15–17

K

KEGG Kyoto Encyclopedia of Genes and Genomes. 1, 12, 13, 15, 17, 18, 20, 37–39, 45, 54

L

LAB Lactic Acid Bacteria. 1, 24, 26, 27, 54, 55

LB Lower Bound. 1, 30, 32, 34, 35, 38, 44, 53

LCG *Lactobacillus* Core Genome. 1

LTA Lipoteichoic acids. 1, 25, 29, 49, 50, 55

M

MEMOSys MEtabolic MOdel research and development System. 1, 12

merlin Metabolic Models Reconstruction using Genome-Scale Information. vi, viii, xi, 1, 2, 11–14, 16–18, 20, 30–33, 35–39, 42–45, 48–50, 52, 59

MetaNetX Automated Model Construction and Genome Annotation for Large-Scale Metabolic Networks. 1, 16, 38

mg milligram. 1, 40, 41

MIB Métodos de Investigação para a Bioinformática. 1

mL milliliter. 1, 40–42

mRNA messenger RNA. 1, 7, 9, 49

MYSQL Structured Query Language. 1, 16, 30

N

NAD Nicotinamide adenine dinucleotide. 1, 50, 54–56

NADP Nicotinamide adenine dinucleotide phosphate. 1, 50, 54

NCBI National Center for Biotechnology Information. 1, 12, 14, 17, 30, 49

NC-IUBMB Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. 1, 15

nm nanometer. 1, 41, 42

NPV Negative Predictive Value. 1, 34

O

OD Optical Density. 1, 40

ORF Open Reading Frames. 1, 18

P

PATRIC Pathosystems Resources Integration Center. 1, 15, 17

PBS Phosphate-buffered saline. 1, 40, 41

PDB Brookhaven Protein Data Bank. 1, 14

PEP-PTS phosphoenolpyruvate dependent phosphotransferase system. 1, 26

pFBA parsimonious Flux Balance Analysis. viii, 1, 23, 52

PGDB pathway/genome databases. 1, 15

PIR Protein Information Resource. 1, 14

PPV Positive Predictive Value. 1, 33

PRF Protein Research Foundation. 1, 14

PRPP Phosphoribosyl pyrophosphate. 1, 55, 56

Q

QoS Quality of Service. 1

R

RAST Rapid Annotation using Subsystem Technology. 1, 16

RAVEN Reconstruction, Analysis and Visualization of Me tabolic Networks. 1, 12

RNA Ribonucleic acid. 1, 7, 39, 42, 49, 51

rRNA ribosomal Ribonucleic acid. 1, 49

S

SBC Stockholm Bioinformatics Center. 1, 16, 17

SBML Systems Biology Markup Language. 1, 12, 13, 17, 42, 43, 52

SDF structure-data file. 1, 15

SOA Service Oriented Architecture. 1

T

TA Teichoic acids. 1, 29, 49, 50, 54

TC Transporter Classification. 1, 15, 16, 19

TCA Tricarboxylic acid cycle. 1, 54

TCDB Transporter Classification Database. 1, 18, 19, 35, 44

TCG Transporter Candidate Gene. 1, 13

TE Tris-Ethylenediaminetetraacetic acid. 1, 41

TMHMM Transmembrane Helices Prediction using hidden Markov models. 1, 16

TRIAGE Transport Reactions Annotation and Generation. viii, xi, 1, 13, 35, 36, 56, 59

tRNA transfer RNA. 1, 49, 54

U

UB Upper Bound. 1, 30, 32–35, 38, 44

UM Universidade do Minho. 1

UniProtKB Universal Protein Resource Knowledgebase. 1, 15

1

INTRODUCTION

This document describes the thesis developed in the context of the dissertation for the Masters in Bioinformatics.

1.1 CONTEXT AND MOTIVATION

Lactobacillus helveticus is an important industrial lactic-acid bacterium being used in the production of several types of cheese. The metabolic activities of the bacterium contribute to the cheese flavour and can help reduce bitterness. This organism belongs to the *Lactobacillus delbrueckii* phylogenetic group and is able to grow at high temperatures, to produce high quantities of lactic acid in milk and to express a complexity of strong proteolytic enzymes. Therefore, *Lb. helveticus* has an increasing economic impact in industrial dairy fermentations [8]. In addition, is a growing body of literature on the health-promoting properties of *Lb. helveticus*, such as their anti-mutagenic, immunomodulatory and anti-diarrheal effects. It is a Generally Recognized as Safe (GRAS) organism and considered as probiotic. Furthermore, due to its abilities to survive gastrointestinal transit, adhere to epithelial cells and antagonize pathogens, *Lb. helveticus* seems to have some effect against diseases such as intestinal inflammation and cancer. [9]. In this project, the genome of *Lb. helveticus* was functionally annotated using state-of-the-art semi-automated tool Metabolic Models Reconstruction using Genome-Scale Information (*merlin*) [3]. This tool is an application in continuous development based on Java™created to semi-automatically help in the reconstruction of genome-scale metabolic models for any organism with fully sequenced genome. It provides automated steps for reconstruction process, integrating diverse web servers functionalities and information. The functional annotation was then used to generate a draft metabolic network reconstruction. This draft was subjected to manual curation and refinement using literature as well as experimental data on the organism metabolism and physiology. Furthermore, the reconstruction was converted into a mathematical model and constraint-based methods were used to analyze the model and fill its gaps. Finally, wet-lab experiments were conducted at Chr.Hansen facilities in Hørsholm, for validating the metabolic model.

1.2 OBJECTIVES

The main goal of this work is the development of a genome scale metabolic network of the lactic-acid bacterium *Lactobacillus helveticus* CNRZ32. In detail, it is aimed to perform the reconstruction of the genome-scale metabolic model using *merlin*'s approach. *merlin* allows obtaining an up-to-date, high-quality functional annotation of a representative *Lb. helveticus* genome and a semi-automated generation of a draft network reconstruction. *merlin* also facilitates the conversion of the network into a model via the semi-automatic generation of an equation representing the drains of biomolecules to the biomass and other constraints, which were then included in the model. Finally, the metabolic model should be complemented with experimental data, obtained from laboratory experiments and understanding and the usage of basic microbiological techniques for the cultivation and phenotypic characterization of bacterial cells.

1.3 THESIS OUTLINE

This thesis is organized in five different chapters. The subjects are introduced on this chapter and final conclusions and remarks on chapter 5. The scientific research is covered in the other chapters.

Chapter 2 contains an insight of systems biology evolution and current status, as well the process involved in a metabolic annotation and model reconstruction. It also contains the bacteria description, taxonomy, interest context and general features of some main components.

Chapter 3 goes through the used materials, softwares and techniques. It is separated in two main sections. The first is regarding functional annotation, where the genome and transporter annotation methods are described. The second describes the tools and methods leading to the final model reconstruction.

Chapter 4 follows the same distribution as the previous chapter. Along the chapter the obtained results and troubleshooting processes are described. The chapter finishes with the simulations results and a description of main pathways constituting the model and a final metabolic map.

It should be taken also in consideration that along the thesis it can be used different terms which should be considered to have the same meaning:

- lactate and lactic acid
- D-lactate and (R)-lactate
- L-lactate and (S)-lactate
- drains and exchange reactions

- sugar and carbohydrates
- glutamic acid and glutamate
- aspartic acid and aspartate

Amino acids will be referred by their nomenclature and symbols in agreement with International Union of Pure and Applied Chemistry (IUPAC) convention [10].

2

STATE-OF-THE-ART

2.1 BACKGROUND/PROBLEM ANALYSIS AT SYSTEMS-LEVEL

Since more than 60 years ago there have been several efforts to recreate and model metabolisms. With his work in cybernetics, Wiener created the first draft of what is known today as a network [11]. Most of the work was on phenomenological analysis of physiological processes. Before him, biochemical approaches were also tested, and although restricted to steady-state flow, these have been successfully used to explore system-level properties of the biological metabolism [12]. General systems theory has also been previously performed [13]. Systems biology has therefore been built on multiple efforts with distinct approaches, but all sharing the same vision: comprehend the living species as a whole.

The evolution of research down to a molecular level allowed to apply dogmatic principles to systems biology [14]. The completion and publication of the *Haemophilus influenzae* genome sequence in 1995 became a turning point in the history of biological research [15]. It marked a metamorphosis from a data-poor discipline into a data rich one together with other high-throughput experimental technologies. Advanced automation techniques in genome sequencing protocols allowed the number of fully sequenced organisms to increase rapidly in the past few years. This exponential shift in bioinformatics and genomic fields, demanding new tools to enable high-throughput generation of functioning genome-scale metabolic models. The big challenge now is to deal and interpret all this large-scale data produced and integrate it with the fundamentals in biology in order to generate good quality models with information about the whole system. However, there are significant precautions to take when dealing with the big datasets produced by the modern post-genomic era. For instance, technological platforms, both hardware and software, are available for several omics data types analysis, but some of these are prone to introducing technical artifacts [16]. This can bias in the data, as it creates sample differences with no evident biological cause. Moreover, data is not always represented in a standardized or uniform manner, complicating cross-experiment comparisons [17]. Data quality, context and lab-to-lab variations represent another important hurdle that must be overcome in genome-scale science.

With a functional model, it is possible to identify poorly annotated regions of metabolic network and predict minimum culture conditions. Phenotypes predictions can be gathered together with the experimental data to be validated [18]. The social and economic impact of such projects is expected to be high as many of those organisms have important industrial applications or represent important human pathogens [19].

2.2 SYSTEMS BIOLOGY

A system-level understanding requires changing the study focus. Although identifying and understanding genes and proteins functions is important, it is necessary to look to the problem in a wider perspective. The identification of all genes and proteins just provides a catalog of individual components. It is necessary to know how to assemble all its parts in order to form and know the system's structure and its dynamics [20].

Many high-throughput experimental technologies have been developed in the last few years years and it is likely the speed and potential of this new tools will continue to increase. These developments led to a change in scientific thinking and now it is becoming universally accepted that cells should be viewed as systems. Understanding complex biological systems requires the integration of experimental and computational research. Computational biology, through pragmatic modelling and theoretical exploration, involves the development and application of data-analyses and computational simulation techniques for the study of the biological system. This system-bases approach provides a powerful foundation to address critical scientific questions head-on.

Covert describes the reality of a biological systems as consisting in a large numbers of functional diversity and frequently multifunctional [21]. Systems biology is defined as sets of elements that interact selectively and non-linearly. It rather attains a coherent and simpler behavior instead of a complex behaviorism. It is also claimed that the system-level understanding of a biological system can be derived into three key properties:

1. System dynamics: a system behavior over time and under various conditions can be understood through metabolic analysis, sensitivity analysis, dynamic analysis methods such as phase portrait and bifurcation analysis, and by identifying essential mechanisms underlying specific behaviors. Bifurcation analysis traces time-varying change(s) in the state of the system in a multidimensional space where each dimension represents a particular concentration of the biochemical factor involved.
2. The control method: mechanisms that systematically control the state of the cell can be modulated to minimize malfunctions and provide potential therapeutic targets for treatment of disease.

3. The design method: strategies to modify and construct biological systems having desired properties can be devised based on definite design principles and simulations, instead of blind trial-and-error.

The system-wide genome, transcriptome, proteome and fluxome experiments allow to understand the hidden layers of the relationships between all the different levels. Computational systems biology abilities the design and manipulation of robust models, contributing to practical innovations in medicine, drug discovery and engineerings and therefore, to the understanding of life.

2.2.1 *Omics Data*

Systems biology methods accumulate a vast array of information to generate hypotheses and discover new cellular relationships. Since the assembly of the first complete genome using Sanger capillary sequencing in 1977, scientific and industrial developments have improved the so called Omics technologies [22]. A combination of these technologies provides important proof of biochemical predictions and creates new opportunities for understanding cellular functional architecture [23]. The integration of data from distinct Omics technologies allows to identify unexpected regulations modes in cellular metabolism. The understanding of the cell functioning and reciprocal relationships in different levels, namely the metabolite and enzyme concentration levels, is allowed with the data integration.

2.2.1.1 *Genomics*

Genomics is defined as the study of the whole genome sequence and the information contained therein, aiming the collective characterization and quantification of each gene. It is the most mature of the different Omics fields. The raw sequence data allows performing quantitative and comparative genomic studies, contributing to the construction of the tree of life [24]. Having the full sequence of the genome *per se* is not enough to provide all the answers. The human genome sequencing project, for instance, was widely expected to bring a huge revolution toward understanding human evolution, the causation of diseas and the interplay between the environment and heredity [25]. But without interdisciplinarity and complement of other omics, this field by itself is not enough for the comprehension of the living world.

2.2.1.2 *Proteomics*

Proteomics is the study of the function of all expressed proteins. Holds promise for an unbiased, systematic discovery route. The term proteome was first attributed to describe the set of proteins encoded by the genome. Despite the levels of complexity and dynamic

ranges and the difficulty to measure and analyse body fluids, the proteome are a rich source of potential biomarkers [26].
Tremendous progress has been made in the past few years, now evoking the set of all protein isoforms and modifications.in any given cell. Progresses in the 'post-genomic' era allowed to generate large-scale data sets for protein–protein interactions, organelle composition, protein activity patterns and protein profiles, for instance in cancer patients [27].

2.2.1.3 *Transcriptomics*

The transcriptome is the complete set and their quantity of transcripts in a cell, for a specific developmental stage or physiological condition. The comprehension of the transcriptome is essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues and also for understanding development and disease.
Transcriptomics provides powerful tools for understanding gene structures and Ribonucleic acid (RNA)-based regulation in any organism. In contrast to the genomics approach, transcriptomics provide a bird's-eye view of selected phenomena in all genes simultaneously. The key aims of transcriptomics are catalog all species of transcript, including messenger RNA (mRNA), non-coding RNA and small RNA, determine the transcriptional structure of genes, in terms of their start sites, 5' and 3' ends, splicing patterns and post-transcriptional modifications and quantify the changing expression levels of each transcript during development and under different conditions [28]. Although whole-transcriptome studies have been highly productive in eukaryotes for more than a decade, the transcriptomes of bacteria and *Archaea* have been largely overlooked until recently [29].

2.2.1.4 *Metabolomics*

Metabolomics regards the naturally-occurring, low molecular weight organic endogenous metabolites within a cell, tissue or biofluid [30]. Metabolites detection is carried by either nuclear magnetic resonance or mass spectrometry. Metabolomics, when used as a translational research tool, can provide a link between the laboratory and clinic. This happens because metabolic and molecular imaging technologies particularly, such as positron emission tomography and magnetic resonance spectroscopic imaging, enable the discrimination of non-invasive metabolic markers *in vivo* [31].

Understanding such gene-to-metabolite networks in primary and secondary metabolism through integration of transcriptomics and metabolomics can lead to identification of gene function and subsequent improvement of production of useful compounds [32].

2.2.1.5 *Localizomics*

Localizomics can be defined as a field aiming to identify the sub- cellular location of all proteins in the cell, which can provide key insights into the cellular function of the individual proteins as well as their probable interacting partner. Generally localizomics compared with other omics data types requires extraordinary efforts. Nonetheless, experimental efforts have generated a genome-wide resource of individual promoter constructs. Moreover computational techniques are also allowing for the *in silico* prediction of protein localization in eukaryotes [1].

2.2.1.6 *Other Omics*

Colquhoun describes the omics trend as excessive [33]. Still, adding to the better described above, some other deserve the reference:

- In metagenomics, the genomic Deoxyribonucleic acid (DNA) of a microbial community is recovered from the environment and sequenced [34], unlike traditional techniques for studying a prokaryotic species that rely on the ability to grow it in a pure culture. The rationale for this is that most of prokaryotic species cannot be readily grown in laboratory conditions [35]. Metagenomics has been successfully used to assess species diversity in the soil [36], ocean [37] and other niches [38].
- Lipidomics objectives to identify and classify the complete inventory of lipids and their associated interacting factors within the cell [39];
- Glycomics aim to do the same for carbohydrates and glycans. However, these methods are in their infancy and relatively few data sets have been generated so far. Therefore, data-integration efforts using this data type remain on the horizon [40].

Integrating all these omics data, will allow building robust, good quality system-based models.

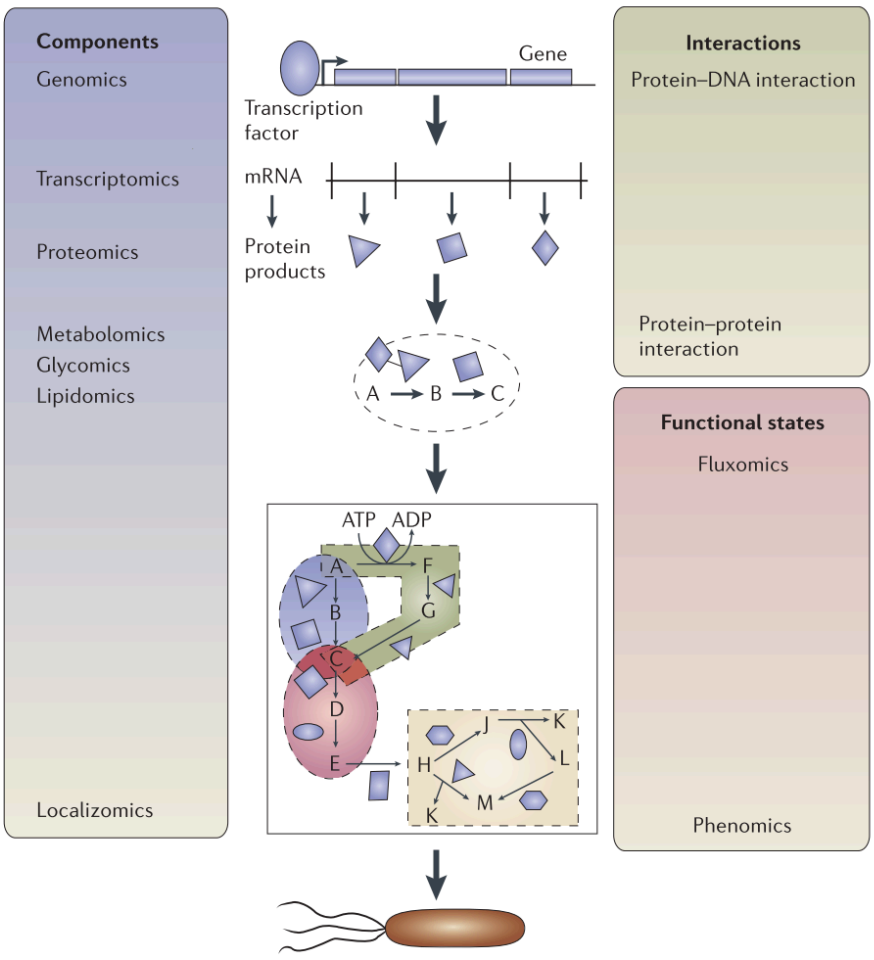


Figure 1.
Omics data in the comprehensive descriptions of components and interactions within the cell [1].

Data is generally classified into three categories: components, interactions and functional-states (Fig1). Components data detail the molecular content of the cell or system, interactions data specify links between molecular components, and functional-states data provide an integrated readout of all omics data types by revealing the overall cellular phenotype. The central pathway traces the biological information flow from the genome to the ultimate cellular phenotype and the available omics data types that are used to describe these processes are indicated in the adjacent boxes. DNA (genomics) is first transcribed to mRNA (transcriptomics) then translated into proteins (proteomics), which can catalyze reactions acting on metabolites (metabolomics), glycoproteins and oligosaccharides (glycomics), and various lipids (lipidomics). Many of these components can be tagged and localized within the cell (localizomics). The processes that are responsible for generating and modifying these cellular components are generally dictated by molecular interactions, for example by protein–DNA interactions in the case of transcription, and protein–protein interactions in translational

processes as well as enzymatic reactions. Ultimately, the metabolic pathways comprise integrated networks, or flux maps (fluxomics), which dictate the cellular behavior, or phenotype (phenomics).

2.3 CONSTRAINT-BASED METABOLIC MODELING

Covert makes an analogy between simulation of traffic conditions in a typical city and simulation of a microbial cell using systems analysis [21]. For both simulations, the first step is to generate a list of all the relevant components (e.g. roads or gene products) of the system. Then, the integration of these components must be determined and specified. In addition, some qualitative predictions are made about the performance of the system. Finally, mathematical modeling is used to quantitatively analyze the system as it responds to a number of environmental factors or a change in the network.
Constraint-based models of metabolisms are a widely used framework for predicting flux distributions in genome-scale biochemical networks. These network reconstructions contain all the known metabolic reactions and the genes encoding each enzyme in an organism but they are absent of regulatory information.

The number of published methods for integration of transcriptomic data into constraint-based models has been rapidly increasing due to the speeding-up of amounts of high-throughput data and the better understanding of information in different omic levels create conditions to (re)construct metabolic models [41, 42].

2.3.1 *Genome-wide scale Modeling Models (GSMM)*

Genome-scale metabolic reconstructions and their analysis with constraint-based modeling techniques have gained enormous momentum [43]. Genome Scale Metabolic Network (GSMN) can be defined as the set of biological reactions retrieved from the enzymes encoded in the target organism's genome [44]. Following the complete sequencing of a genome, this is the next step to take in account in systems biology analysis. These kind of models are built bottom-up from the genes to the enzymes encoding those genes. These reconstructions contain large amounts of structured and pertinent information providing bases to biochemical understanding in specific target organisms.

Afterwards, a mathematical conversion is necessary to facilitate the computational biology studies. Thus, these models should be able to foretell the prototypical behavior of a cell, an organism, or an individual. After this *in silico* analysis the work is directed towards to the best predicted output[3].

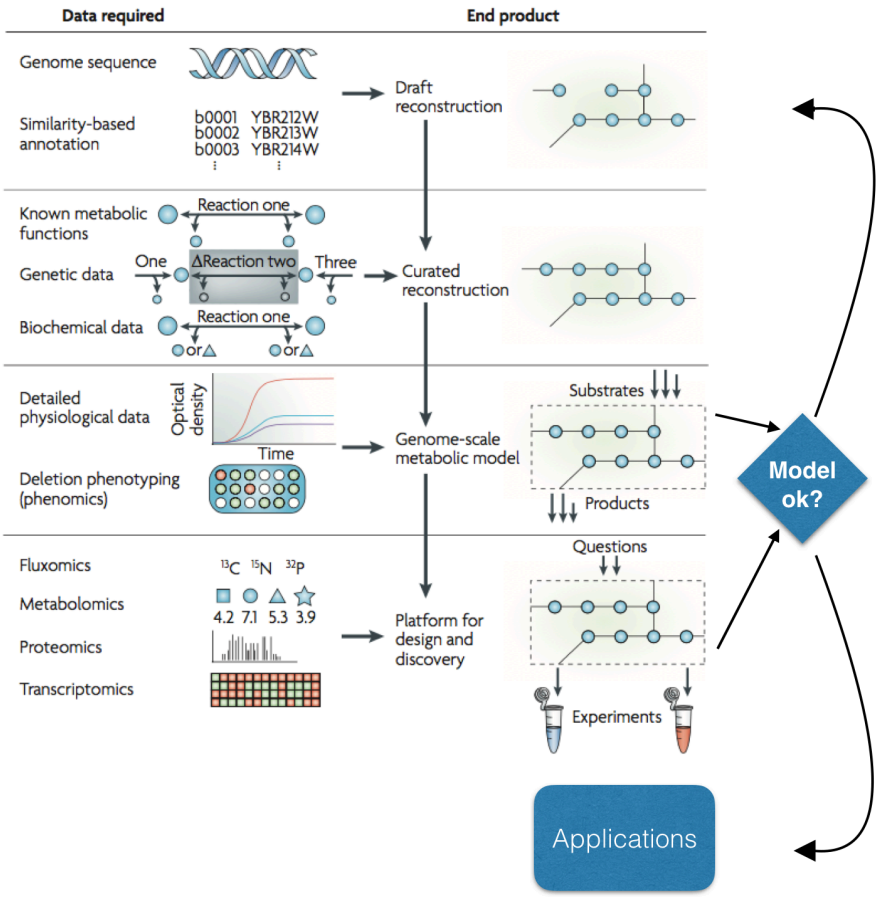


Figure 2.
Adapted scheme of the phases and data used to generate a metabolic reconstruction [2].

As shown in figure 2 the reconstruction of genome-scale metabolic models can be divided into four consecutive major phases. These phases are the draft reconstruction, curated reconstruction and Genome Scale Metabolic Models (GSMM) which together will produce a platform for design and discovery. An additional characteristic of the reconstruction process is the iterative refinement of reconstruction content that is driven by experimental data from the three later phases. If after all the stages performed, the model is not properly working a reverse engineering process is performed, revisiting previous steps.

2.3.2 *merlin*

merlin is an open-source application, distributed under the GNU General Public License at `http://www.MERLIN-sysbio.org`. *merlin*'s methodology provides an automated genome-wide functional annotation assigned with a numeric confidence level score, based on the taxonomy and frequency within the similar sequences to each one of the genes, according to

Eqs. (1) to (3). With minimum user interaction, it establishes a comparison between biological sequences from the organism being studied with all of the National Center for Biotechnology Information (NCBI) databases. Furthermore, Gene-Protein-Reaction (GPR) associations are automatically generated and included in the model. With the 'Draw in Browser' option, *merlin* aids the user in the gap filling process by showing enzymes and reactions annotated directly in the selected Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway browser. Also, it allows compartmentation of the model predicting the organelle localisation of the proteins encoded in the genome. Finally, *merlin* converts the genomic data to draft metabolic models reconstructions in the Systems Biology Markup Language (SBML) standard format, allowing a preliminary view of the biochemical network [3].

$$score_f = \frac{\sum_{i=1}^{n} v_i}{n} \tag{1}$$

if Enzyme Commission (EC) number exists v_i is 1. Otherwise is 0.

$$score_t = \frac{\sum_{i=1}^{n} (v_i \times t_i) x \times penalty_score}{Max_Taxonomy \times min(\sum_{i=1}^{n} v_i, n_Homologies)} \tag{2}$$

$$score_{annotation} = \alpha.score_{frequency} + (1 - \alpha).score_{taxonomy} \tag{3}$$

Although there are another platforms with similar individual tools for metabolic engineering, *merlin* has the advantage to combine all platforms in a unique and completely capable software. This tools are listed in Table 8. *merlin* is the only available tool to our knowledge that provides an integrated framework for the reconstruction of GSMM for both prokaryotes and eukaryotes that retrieves enzymatic, transport and localisation information. Other frameworks have some of the *merlin* capabilities but none of them can gather all of the features in one platform. Namely, Flux Analysis and Modeling Environment (FAME) [45], MEtabolic MOdel research and development System (MEMOSys) [46], MicrobesFlux [47] and the Pathway Tools [48] do not allow metabolic (re)annotations. Comparative ReConstruction of genome-scale metabolic networks (CoReCo) [49] and Reconstruction, Analysis and Visualization of Me tabolic Networks (RAVEN) toolbox are only able to perform genome-wide functional annotation. ModelSEED [50] and RAVEN [51] do not perform transports annotation and obligate the users' data to be shared to SEED's web server. ModelSEED does not support eukaryotic GSMM as well. A more visual comparison of this tools is available in Figure 3.

Software	FAME	MEMOSys	MicrobesFlux	Pathway Tools	CoReCo	RAVEN	Model SEED	*merlin*
Enzymes annotation					•	•	•	•
Transporters annotation				•			•	•
Compartments prediction	*i*	*i*				•	•	•
Biomass reaction	*ii*		*ii*		*ii*	*ii*	•	*ii*
Export to SBML	•	•	•		•	•	•	•
Runs locally		•		•	•	•		•
Requires commercial software						•		
Graphical interface for manual curation				•				•
Pathways visualisation	•		•	•		•	•	•
Gene-Protein-Reaction rules							•	•
Highlight metabolic dead-ends	•		•	•		•		•
Reactions stoichiometry validation		*iii*	•	*iii*	•		*iii*	•
Prokaryotic models	•	•	•	•	•	•	•	•
Eukaryotic models				•	•	•		•

[*i*] Allow to manually assign compartments to reactions (*merlin,* RAVEN and Model SEED automatically predict reactions localisation).
[*ii*] Biomass reaction inserted manually (Model SEED - Biomass reaction automatically generated).
[*iii*] Model SEED and Pathways tools use their own metabolic databases. MicrobesFlux checks for new reactions.
* SBML - Systems Biology Markup Language

Figure 3.
A comparison between *merlin* capabilities on genome-scale models reconstruction when compared to other tools [3]

2.3.2.1 *Annotation of Transporter Systems and Transport Reactions Annotation and Generation (TRIAGE)*

TRIAGE is a tool based on the identification and classification of genes that encode transmembrane proteins. It allows to identify metabolites transported by each transmembrane protein and its transporter family. In the reconstruction of GSMMs transport reactions are added as a complementary element for the model, but usually without association to specific genes. This is a big limitation when considering changes prone to happen in the system, such as gene deletions. TRIAGE is a novel approach for genome-wide transporter functional annotation and appeared as a response to the lack of good transporters annotation. Increasing the compartments indicates compounds needing to reach enzymes and as such have to cross-cell or organelle-specific membranes for reactions to happen. Transport reactions are built considering the metabolites annotated in the TCDB records identified as similar to the TCG) in the target genome. The transporter candidates' layer (dynamic layer) is organism specific and is connected to the shared layer of the database, the transport reactions layer (static layer), by three connections. It allow Transporter Candidate Gene (TCG)s to be assigned with a TC family, a range of metabolites to be transported and a direction for such transport. The metabolites used to construct transport reactions are retrieved from TCDB records. KEGG, Chemical Entities of Biological Interest (ChEBI) and semantics SBML [52] 2.0 for collecting additional data. Uniprot was used to retrieve phylogenetic data in order to assign the transport reactions to the candidate gene. These reactions can be directly integrated with GSMMs since all metabolites involved have KEGG and/or identifiers [53].

2.3.3 *Principles and methods in constraint-based metabolic modeling*

In the last few years a myriad of computational modelings of cellular metabolism in biotechnology have been complemented [54]. From the different mathematical formalisms proposed for computational modeling of cellular metabolism in biotechnology, kinetic and constraint-based models are among the most widely adopted ones [55]. Constraint-based models describe the range of steady-state flux distributions of a metabolic network, using a Flux Balance Analysis (FBA) approach [56].

Despite of requiring a big amount of experimental data for determining the rate laws and kinetic parameters of biochemical reactions, constraint-based modeling mainly demands knowledge of the stoichiometry of the metabolic network. This information can be obtained from annotated genome sequences and metabolic pathway databases.

The simplicity and scalability of FBA, coupled with the advances in genome sequencing, led to an explosion in the number of GSMM currently available [57].

2.3.3.1 *Online Databases*

Online databases access is mandatory to retrieve and collect the necessary data sources for the GSMM construction. Different online sources and respective tools work in synesthesia with *merlin*, allowing the Enzyme Encoding Candidate Genes (EECG) validation, collection of data or literature research (Table 1).

NCBI is a repository of several databases that provides analysis, visualization, and retrieval resources for biomedical, genomic, and other biological data. Basic Local Alignment Search Tool (BLAST) [58] allowed a similarity search performed with *merlin* used all non-redundant sequences (including GenBank coding sequences translations, RefSeq Proteins, Brookhaven Protein Data Bank (PDB), SwissProt, Protein Information Resource (PIR), Protein Research Foundation (PRF) databases (nrDB) available in the NCBI [59] databases to find any protein sequence similar to the target organism.

The Entrez Protein `http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein` database is a collection of sequences from several sources, including GenBank Coding Sequences (CDS) translations, RefSeq Proteins, SwissProt, PIR, PRF, and PDB [59].

The UniProtKB/Swiss-Prot `http://www.UniProt.org/` database is a manually curated protein sequences database which provides annotations with minimal redundancy and high level of integration with other databases [60].

BRaunschweig ENzyme DAtabase (BRENDA) `http://www.brenda-enzymes.info/` provides enzyme functional data obtained directly from literature by professional curators. This database was used to confirm the information gathered in the previous two databases, thus being the third reference database selected for this work. TCDB `http://www.tcdb.org/` details a comprehensive classification system, approved by the Union of Biochemistry and

Molecular Biology (IUBMB), for membrane transporter proteins known as the Transporter Classification (TC) system. (Universal Protein Resource Knowledgebase (UniProtKB)) is the central hub for the collection of accurate, rich, and consistent functional information on proteins. It consists of two sections: a section containing manually annotated records with information extracted from literature and computational analysis (referred to as UniProtKB/Swiss-Prot) and a section with computationally analyzed records waiting full manual annotation (UniProtKB/TrEMBL)

MetaCyc is a database of nonredundant metabolic pathways. MetaCyc is curated from the scientific literature and contains pathways involved in primary and secondary metabolism and associated compounds, enzymes and genes.

KEGG is an online public repository that is, currently, the most extensive combined collection of information on genes, metabolites, reactions, and pathways. KEGG contains genomic and metabolic data.

Expert Protein Analysis System (ExPASy) is the Swiss Institute of Bioinformatics Resource Portal in different areas of life sciences including systems biology. Furthermore, ExPASy is one of the main bioinformatics resources for proteomics in the world.

BRENDA–KEGG–MetaCyc reactions (BKM-react) is an integrated and non redundant database containing known enzyme-catalyzed and spontaneous biological reactions collected from BRENDA, KEGG, and MetaCyc by aligning substrates and products.

BioCyc is a collection of pathway/genome databases (PGDB)s. Each PGDB in the BioCyc collection describes the genome and metabolic pathways of a single organism. These PGDBs contain additional features, including transport systems and gap fillers. Also, the BioCyc website contains tools for the visualization and analysis of the PGDBs.

ChEBI is a freely available dictionary of molecular entities focused on small chemical compounds stored in a relational database. ChEBI incorporates an ontological classification, whereby the relationships between molecular entities or classes of entities and their parents and/or children are specified. ChEBI provides its chemical structures and additional data in structure-data file (SDF) format. It uses nomenclature, symbolism and terminology endorsed by the following international scientific bodies: IUPAC and Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB).

Pathosystems Resources Integration Center (PATRIC)[61] allows genome assembly, genome annotation and provides a protein family sorter, a comparative pathway tool and genome metadata.

The Department of Energy (DOE) Joint Genome Institute (JGI) [62] is a national user facility with massive-scale DNA sequencing and analysis capabilities dedicated to advancing genomics for bioenergy and environmental applications. Beyond generating tens of trillions of DNA bases annually, the Institute develops and maintains data management systems and specialized analytical capabilities to manage and interpret complex genomic data sets. The

JGI Genome Portal `http://genome.jgi.doe.gov` provides a unified access point to all JGI genomic databases and analytical tools. Genomes OnLine Database (GOLD)[63] is a web-based resource for comprehensive information regarding genome and metagenome sequencing projects, and their associated metadata, around the world. Since 2011, the GOLD database has been run by the DOE-JGI.

The Center for Biological Sequence Analysis (CBS) at the Technical University of Denmark was formed in 1993, and conducts basic research in the field of bioinformatics and systems biology. CBS has a highly multi-disciplinary profile (molecular biologists, biochemists, medical doctors, physicists and computer scientists). CBS has produced a large number of computational methods, which are offered to others via WWW servers, as is for instance the Transmembrane Helices Prediction using hidden Markov models (TMHMM).

Stockholm Bioinformatics Center (SBC) provides the tool *Phobius* that provides combined transmembrane topology and signal peptide predictor [64].

Biochemical, Genetic and Genomic knowledge base (BiGG) contains high-quality, manually-curated genome-scale metabolic models containing information for metabolites and reactions. Users can browse and visualize models. BiGG Models connects genome-scale models to genome annotations and external databases [65]. ModelSEED is a source for the reconstruction, exploration, comparison, and analysis of metabolic models. ModelSEED is based in Rapid Annotation using Subsystem Technology (RAST) fully-automated service for annotating complete or nearly complete bacterial and archaeal genomes. RAST provides high quality genome annotations for these genomes across the whole phylogenetic tree. Automated Model Construction and Genome Annotation for Large-Scale Metabolic Networks (MetaNetX) is an online platform for accessing, analyzing and manipulating genome-scale metabolic networks and biochemical pathways. It integrates a great variety of data sources and tools and provides a single identifier to every single metabolic reaction as well as the existing aliases for each reaction over several databases [66, 67, 68]. Structured Query Language (MYSQL) is an open-source relational database management system in which *merlin* is supported.

2.3.3.2 *Genome Annotation*

Genome annotation can be defined as the process of identifying and labeling all the relevant features on a genome sequence [69]. It is the first step of a genome-sale metabolic reconstruction. It is absolutely critical this stage to provide a good quality annotation, as it will henceforward constitute the basis of the reconstruction process. This process assigns genes with functions, providing unique identifiers, such as the EC and TC numbers, to the reconstruction [70, 71]. Genes encoding enzymes or transport systems are labeled metabolic genes, and are mandatory for the development of the GSMMs. Due to its importance, the re-annotation of the genome is encouraged to assure quality and reliability of the gene functional assignments. An example of the iterativity of the process is the *E. coli* metabolic network

BiGG	BiGG Models, metabolites and reactions
NCBI	National Center for Biotechnology Information With the tools: The first Basic Local Alignment Search Tool (BLAST) RefSeq Proteins Entrez Proteins PubMed
EMBL-EBI/ SBI	European Bioinformatics Institute/ SIB Swiss Institute of Bioinformatics with the tools: The UniProtKB/Swiss-Prot Universal Protein Resource SwissProt
BRENDA	The Comprehensive Enzyme Information System BRaunschweig ENzyme DAtabase
JGI	Joint Genome Institute and tools: GOLD Genomes OnLine Database IMG/M Integrated Microbial & Microbiome Samples
KEGG	Kyoto Encyclopedia of Genes and Genomes
ExPASy	SIB Bioinformatics Resource Portal
KEGG	Kyoto Encyclopedia of Genes and Genomes
BioCyc	Pathway/Genome Database Collection
PATRIC	Pathosystems Resources Integration Center
CBS	Center for Biological Sequence Analysis
SBC	Stockholm Bioinformatics Center, *Phobius*
ChEBI	Chemical Entities of Biological Interest database
Semantincs SBML 2.0	Systems Biology Markup Language with the tool: TransMembrane Helices prediction based on a Hidden Markov Model (TMHMM)
TCDB	Transport Classification DataBase

Table 1.

Online Databases and respective tools.
This databases work with *merlin* to retrieve and process information.

Top Level codes	
EC 1	Oxireductases
EC 2	Transferases
EC 3	Hydrolasess
EC 4	Lyases
EC 5	Isomerases
EC 6	Ligases

Table 2.

EC number classification and organization.
The EECG are classified by their function (e.g. Oxidoreductases Transferases, Hydrolases, Lyases, Isomerases or Ligases), following the Enzyme Nomenclature, in which the first level number is associated with the enzyme function (adapted from International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes [70].

process studies, going trough a series of expansions and refinement [72]- [79]. This involves looking to specific data, such as gene or Open Reading Frames (ORF) names, product names, and, if available, EC numbers. Other genes involved in regulatory control or signaling are not included in GSMMs but may be useful for later integration in the model.

For the annotation *merlin* utilizes two different tools (BLAST [58] and biosequence analysis using profile hidden Markov models (HMMER) [80]) to perform the (re) annotation of genomes [3, 81]. The similarity search results are then evaluated and an automatic annotation of the genome is presented, as the tool assigns annotations to each gene of the target organism [82].

2.3.3.3 *Assembling the metabolic network*

In this stage, the biochemical reactions are identified and collected to build the backbone of the network. The reactions catalyzed by enzymes and transport systems encoded in the annotated genome are used.

Genes, Proteins, and Reactions

The association between annotated genes, proteins, and reactions (the GPR associations) is performed by searching biological databases (table 1) with the protein names, EC numbers, or other identifiers (e.g., KEGG reaction number) to which the reaction was associated [83]. TCDB is the only transport protein classification database adopted by the IUBMB (Table 3). The TCDB can be accessed for retrieving the metabolites and type of transport supported by a given carrier protein . These transport reactions should also be added to the draft network [82].

Spontaneous Reactions

Top Level codes	
TC 1	Channels/Pores
TC 2	Electrochemical Potential-driven Transporters
TC 3	Primary Active Transporters
TC 4	Group Translocators
TC 5	Transmembrane Electron Carriers
TC 8	Accessory Factors Involved in Transport
TC 9	Incompletely Characterized Transport Systems

Table 3.
Organization of the transporters classified on TCDB (top levels.)

From these, sub-levels are defined, going from types, to superfamilies and families of transporters. The TC number represent protein which promote metabolites relocation. These follow a classification based on the Transporter Classification Database (TCDB),a freely accessible reference database for transport protein research, which provides structural, functional, mechanistic, evolutionary and disease/medical information about transporters. The TC code contains five elements, separated by four dots (#.*.#.#.#). The left most number represents one of the seven main divisions to which the transporters may belong to, namely, channels/-pores, electrochemical potential-driven, primary active, group translocators, transmembrane electron carriers, accessory factors involved in transport, and incompletely characterized transport systems. The second element is a letter and the remainder elements are numbers. Each element to the right of the main class restricts the classification of the transporter. The TCDB can be accessed for retrieving the metabolites and type of transport supported by a given carrier protein. The ones classified in TC9 group are still not completely characterized and with the information and new studies performed will the transporters be likely to move the other categories. These transport reactions should also be added to the draft network (adapted from `http://www.tcdb.org/browse.php` on June 30th, 2016) [82].

The next step is adding spontaneous reactions that do not require enzymatic catalysis to the network. These reactions can be found in published literature or in a few online data sources, such as KEGG and are included in *merlin*.

Stoichiometry

After collecting the set of reactions, their stoichiometry should be revised. The reaction's stoichiometry provides information regarding quantities of reactants consumed and products formed [84]. Information on this step may be found in databases such as BRENDA KEGG, MetaCyc reactions or BKM-react.

Stoichiometric models have been used to estimate the metabolic flux distribution under given circumstances in the cell at some given moment (metabolic flux analysis), to predict it on the basis of some optimality hypothesis (flux balance analysis), and as tools for the structural analysis of metabolism providing information about systemic characteristics of the cell under investigation (network-based pathway analysis) [85].

Localization/Compartmentation

The compartmentation of the reactions in the cell may induce the regulation of an enzyme function. The localization of an enzyme inside or outside a determined compartment determines the organelle wherein the enzyme will operate. For instance, similar reactions with the same metabolites and stoichiometry, but taking place in different compartments need to be distinguished, as these are considered distinct reactions [86].

The compartamentation evolves as organisms become more complex, for instance:

- Prokaryotes: compartments are typically limited to the cytosol, periplasmic space, and extracellular space.
- Fungi and other eukaryotes: the reactions can occur in various compartments including Golgi apparatus, lysosome, mitochondrion, endoplasmic reticulum, or glyoxysome.
- Higher eukaryotes: it may be further necessary to differentiate between tissues [82].

Hence, each metabolite should include in its name an identifier reflecting its localization. Otherwise reactions and metabolites are usually assigned to the cytosol.

This area is still in development, but in continuous improvement. For instance, *S. cerivisae*'s first GSMM reconstruction accounted for 3 compartments, the second 8 different locations, and the consensus 15 compartments [87, 88, 89].

Manual Curation

All automated processes provide only the basics towards a reconstruction of the metabolic network from a sequenced genome. Unfortunately, despite being very useful and process-accelerating, automated methods ate still fallible and produce incomplete or inappropriate reconstructions [90, 91, 92].

Manual curation is therefore a requirement and includes:

(i) Inspection of the annotation present in the source databases in order to solve incorrect entries [92].

(ii) Resolving inconsistencies between protein and function identifiers in different databases. For instance, due to these inconsistencies, different annotations published for the small genome of *Mycoplasma genitalium* deviated for 8% the gene product [93].

(iii) Addition of new and/or organism-specific reactions or pathways that are absent in the queried databases.

(iv) Judging the correctness of the coupling between query sequence and the sequence in the resource database. The homology and profile-based methods do not always yield a correct coupling [94, 95].

(v) Evaluation of the coupling between the function identifiers and the retrieved reactions . The use of unspecific functional identifiers (like incomplete EC numbers) could lead to false reaction associations, which have to be checked manually [96].

Phylogeny, gene context and high-throughput data

The curation should reconsider all individual proteins within the context of the initial reconstruction. Comparative genomics can be applied to generate additional data to support or reconsider the functional attributes of individual proteins [97]. This might involve the analysis of phylogeny [98], gene fusions [99], gene order [100], co-occurrence [101], regulatory motifs [102] or experimental evidence. Instead of a single indicator, several aspects, should be weighed when performing this assessment.

Pathway analysis: filling gaps and completing the network

The process of reconstructing a metabolic model is never really finished. Therefore, a few minimal quality requirements should be fulfilled. First, the metabolic capabilities represented by the reconstructed network should be consistent with the physiology of the organism. Furthermore, when the reconstruction is used to produce a genome-scale metabolic model with the purpose of yield and flux predictions, its reactions should be elementary balanced and essential pathways should be reviewed [103]. Moreover, unbalanced reactions or gaps in pathways should also be reviewed [104].

Finally, after these aspects are weighed, a debugged GSMN is obtained, which is then converted into a mathematical computational GSMM in the next stage.

2.3.4 *Converting the Metabolic Network to a Stoichiometric Model and Validation*

$$S * v = 0 \tag{4}$$

Stoichiometric modeling avoids difficulties in the development of kinetic models, such as the lack of intracellular experimental measurements. Thus, allows to explore the knowledge

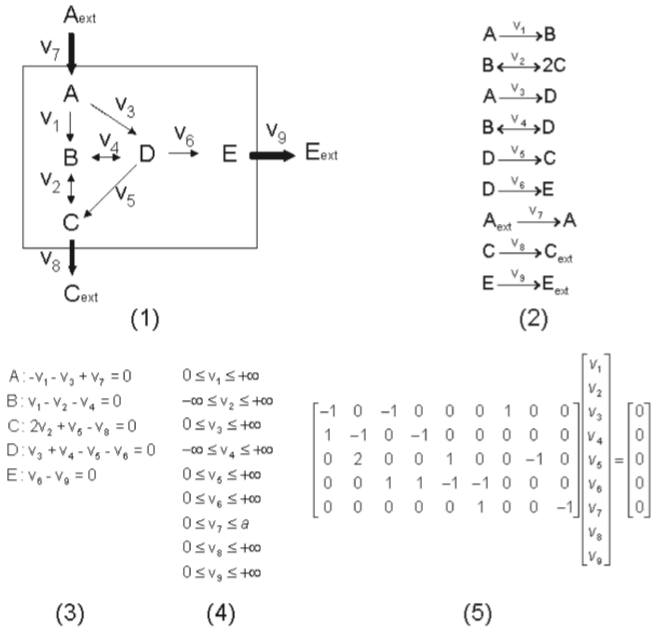


Figure 4.
Example of a metabolic network with five metabolites (A to E) and 9 fluxes (v1 to v9) [4]. The reaction scheme is shown in (1), where the boundaries of the system are also outlined. Fluxes v7 to v9 represent exchange fluxes of both, metabolic substrate (A) and products (C and E). Reversible reactions are shown by double arrows, and irreversible reactions are indicated with a forward arrow. The stoichiometry of the network is represented in panel (2). Panel (3) shows the steady-state mass balances, and panel (4) illustrates the constraints around the flux values (a represents the maximum uptake rate for the consumption of the substrate A). Note that a flux value can be negative for reversible reactions with unconstrained fluxes. Panel (5) shows the representation of the mass balances in matrix format.

about the structure of cell metabolism, without having to consider the intracellular kinetic processes. When the metabolic network is complemented with the biomass equation and the nongrowth Adenosine triphosphate (ATP) requirements, the set of reactions can be represented in the form of a stoichiometric matrix (S) and its flux vectors (v) Eq. (4). In this matrix the columns represent the reactions and rows the metabolites. The classic principles of chemical engineering can be used to construct the matrix that represents the dynamic behavior of the metabolite concentration, by performing dynamic mass balances with ordinary differential equation.

2.3.4.1 *Biomass formation abstraction*

Before converting the network to a GSMM, the biomass formation equation should be included in the reactions set [82]. The biomass equation represents the cell macromolecular

composition and the building blocks used to generate those molecules. Hence, this reaction denotes a drain of biomolecules (e.g., amino acids, nucleotides) into the biomass.

2.3.4.2 *FBA*

Once the mathematical representation of the model is created, it can be used to predict the behavior of the target organism and compare it to experimental data.

The core feature of this representation is a tabulation, in the form of a numerical matrix, of the stoichiometric coefficients of each reaction. These stoichiometries impose constraints on the flow of metabolites through the network. Constraints are represented in two ways, as equations that balance reaction inputs and outputs and as inequalities that impose bounds on the system. The stoichiometric matrix imposes flux balance constraints on the system, ensuring that the total amount of any compound being produced must be equal to the total amount being consumed at steady state. Every reaction can also be limited by upper and lower bounds, which define the maximum and minimum allowable fluxes of the reactions.

These balances and boundaries define the space of allowable flux distributions of a system and the rates at which every metabolite is consumed or produced by each reaction.[56, 105].

2.3.4.3 *parsimonious Flux Balance Analysis (pFBA)*

There is an underlying assumption that when cells are growing exponentially it occurs selection for the fastest growers. Among the fastest growers, there will be a fitness advantage to cells using the least amount of enzyme as they can process the growth substrate the most rapidly and efficiently. The flux parsimony tries to emulate this behavior by minimizing the total material flow required to achieve an objective. pFBA changes the objective to the minimum total flux objective [106].

2.3.4.4 *Data assessment*

Independently of the methodology used to validate the model, it should be thoroughly inspected to find all possible errors. The fact is that if the model does not comply with *in vivo* data, further debugging must be performed. Data sources should be queried subsequently and the reactions set and stoichiometric matrix corrected. A validation with experimental data should be performed, and when not not in conformation with the experimental data, the process of reconstruction should be repeated. The final step is revisiting decisions taken in the manual curation step, in which wrong conclusions may have been inferred [87].

2.4 LACTIC ACID BACTERIA (LAB)

Lactic Acid Bacteria (LAB) are a group of Gram-positive, non-spore forming, anaerobic bacteria which excrete lactic acid or lactate as the main fermentation product into the medium when supplied with suitable carbohydrates [107].

Present in the human body and environment, can colonize the mouth and the nasopharyngal mucosa (oral *streptococci* the gut and intestine (*bifidobacteria*, *enterococci*, some *lactobacilli*) and the mucosa of the vagina (specific *lactobacilli*) [108]. Although LAB are only just a small portion of the total gastrointestinal microbial community, they are predominant microbiota in the small intestine and considered and essential to its protection [109, 110]. These microorganisms are prominent in fermentation of organic matter of various animal and plants niches containing sufficient levels of mono and dissaccharides, playing a play a key role in the production of fermented foods and beverages [111]. The first scientific exploration of lactic fermentations started with the isolation and chemical characterization of lactic acid from fermented milk by Carl Wilhelmscheele (1780). It was followed by reports from Pasteur (1857) which destroyed the theory of spontaneous generation and Lister (1873) who obtained the first bacterial pure culture. Wilhelm Storch and German Weigmann were the first to isolate the LAB from spontaneously fermented milk and cream that are responsible for sour milk and cheese fermentation. Lactic acid bacteria also contributed to the field of genetics, biochemistry and molecular biology in Griffith's work (1928) and later in the DNA work by Avery, MacLeonard and McCarthy (1944). In biochemistry and physiology, LAB allowed to perform quantitative determinations of vitamins by Snell (1952) [112]. The ability to fundamentally understand the genotype–phenotype relationship began to change in the mid 1990s, on completion of the first bacterial genome-sequencing projects in 1995 [113].

2.4.1 *Lactobacillus helveticus*

Lactobacilli demand carbohydrates, protein breakdown products, vitamins, and usually a total absence or low oxygen tension. *Lactobacillus helveticus* is a homofermentative thermophilic rod-shaped LAB with a genome composed by one circular chromosomal sequence with around 2MB and around 1700 coding sequences with biotechnological interest and potential. It has low G+C content and it is acid tolerant [5]. The bacteria includes 163 predicted pseudogenes (excluding transposases) and 356 complete or partial insertion sequence (IS) elements. The large number of pseudogenes and IS elements is consistent with a previous report for Lb. helveticus DPC 4571 and supports the hypothesis that this species has experienced significant genome decay. *Lactobacillus helveticus strain CNRZ 32* was the selected organism for the model reconstruction as exists structured knowledgebase on its biochemical, genetic and genomic features. As other other strains, is characterized primarily by the ability to

form various isomers of lactic acid from the fermentation of glucose producing as byproducts (L)-lactate and (D)-lactate.

2.4.1.1 *Taxomony*

Orla Jensen divided the Lactobacilli into the three groups (Thermobacteria, Streptobacteria and Betabacteria) based on growth temperature and biochemical reactions [114]. Although those three groups have been replaced for different classifications, the three names are still in common use. London classified Lactobacillus as part of a phylogenetic cluster with close relations to the genera Streptococculs, Pediococcus and Leuconostoc [115]. This microorganism belongs to the family of *Lactobacillaceae* and to the *Lactobacillus delbrueckii* phylogenetic group that is characterized by being able to grow at high temperatures, to produce high quantities of lactic acid in milk and to express a complexity of strong proteolytic enzymes. In 2008 it was confirmed the proximity between *Lb. helveticus*, *L. acidophilus* and *L. delbrueckii* [116] (Fig.9).

2.4.1.2 *Cell wall*

Gram-positive bacteria have as main component of their cell wall peptidoglycan. It is mainly composed of alternated two sugar chains attached to a chain of amino acids. The cross-linking of this structures create a rigid 3-D structure offering stability to the cell [117]. The wall includes anionic polymers such as teichoic acid which are cross-linked to the N-acetylmuramyl and residue of the peptidoglycan, N-acetyl glucosamine and lesser amounts of membrane bound Lipoteichoic acids (LTA), neutral carbohydrates, and proteins (Fig.10a). LTA with a poly glycerol-phosphate main chain represents the most common type of membrane-anchored anionic polymer The type of peptidoglycan structure can be also a complement for taxonomic classification as it can vary in each specie and strand [118].*Lb. helveticus* belongs to the subtype A4α species [5, 119] (Fig 10b).

2.4.1.3 *Exopolyssacharide (EPS) production*

EPS are long-chain polyssaccharides consisting of branched, repeating units of glucose, galactose and rhamnose, in different ratios. They are secreted to the surrounding of cells during growth, not remaining attached permanently to cell surface [122]. They have a major role in the manufacturing fermented dairy products in Northen, Eastern Europe and Asia such as yogurt, drinking yogurt, cheese, fermented cream, milk based desserts. EPS may act both as texturizers and stabilizers, firstly increasing the viscosity of a final product, and secondly by binding hydration water and interacting with other milk constituents. They can decrease syneresis and contribute to the texture, mouth-feel, taste perception and stability of the final product. The use of EPS avoid the use of aditives which is attractive for the consumer and as consequence for the dairy market of this products [123].

2.4.1.4 *Amino acids auxotrophies and transport system*

Lb. helveticus usually grows in rich environments and has evolutionarily lost the capacity to synthesize most of the amino acids by itself being distinguished by selective gene loss [116]. The bacteria possesses a proteolytic system to release amino acids from the milk protein, casein.

It presents auxotrophies to 14 amino acids, being able to synthesize only alanine, cysteine, lysine and serine. These are called prototrophic amino acids. Glutamine is obtained from glutamate and glycine from serine [124, 125]. So, in order to obtain the remaining essential amino acids, *Lb. helveticus* needs to uptake them from the medium surrounding it using a myriad of mechanisms. They can also vary depending of the availability of energy or sugar in the organism. The transport systems of *Lb. helveticus* are similar to the ones in *L. Lactis* [126]. Cysteine, leucine, isoleucine, valine, threonine, lysine, aspartic acid, glutamic acid, tryptophan, tyrosine, arginine, and histidine are actively transported when glucose is available [127]. Leucine, isoleucine, valine, threonine, and lysine are transported by a proton motive force coupled system by hybrid membranes. *Lb. helveticus* has also a gene encoding a proton motive force coupled di and tripeptide transporter with better efficiency for Pro-Ala, Phe-Val and Leu-Val dipeptides [128]. Methionine has usually a low concentration in milk, so it is liberated from casein by the proteases and uptaked in tripeptides [129]. The amino acids aspartate, glutamate, histidine, arginine, and tyrosine are most likely transported by primary ATP-driven systems . Five secondary amino acid transporters (branched amino acids, alanine and threonine, serine and threonine, and lysine transporters) and glutamine-glutamate and asparagine-aspartate have ATP coupled systems [130]. Histidine, tyrosine and arginine can also be transported by a special class of secondary transport systems such as as precursor/product antiport systems [131].

2.4.1.5 *Sugar uptake and metabolism*

There are multiple sugar uptake systems reported for LAB. Uptake of mono and dissacharides can be performed by phosphoenolpyruvate dependent phosphotransferase system (PEP-PTS), symport, permease or antiport systems. It have been reported mechanisms for fructose galactose, glucose, lactose or sacarose. *Lactobacillus helveticus* can uptake glucose trough PEP-PTS. It also possesses a permease uptake system for lactose (encoded by the gene lhe1439, lacS) and galactose. It is capable of metabolize both galactose and glucose moieties of lactose by the Leloir pathway and does not accumulate free galactose in the external medium [132]. Hence, it exhibits glucokinase and phospho-β-galactosidase and β-galactosidase activity [133]. *Lb. helveticus* possess a *gal* gene cluster consituted of *gal*K, *gal*T and *gal*M plus *lac*L *lac*M and *gal*E, possibly constituting an operan [134]. Ganzle also suggests the presence of possible genes for internal metabolism of oligossacharides such maltose and maltodextrins, despite the transport mechanism being limited or non-occurring [135]. Thus, in *Lb. helveticus*,

as in other organisms, the *gal*E product might be involved in preparation of carbohydrate residues for incorporation into complex polymers, such as exopolysaccharides [136].

2.4.1.6 *Health, Economical, Industrial and Scientific Interest*

Lb. helveticus can be used in the manufacture of dairy products such as acidophilus milk, yogurt, buttermilk, and cheeses. It has also commercially importance in the processing of meats (sausage, cured hams), alcoholic beverages (beer, fortified spirits), and vegetables (pickles, and sauerkraut) *Lb. helveticus* is generally recognized as safe having probiotic features. Probiotics are defined as living microorganisms that, upon ingestion in certain numbers, exert health benefits and have potential applications for conditions such as gastro-intestinal infections and certain bowel disorders [137]. Furthermore, its behavior both in batch [138, 139] and continuous [140]-[142] lactic acid fermentation has been extensively studied [143]. LAB are nutritionally fastidious and cannot synthesize several essential amino acids necessary for growth. Therefore, the lack of the ability to biosynthesize these amino acids is compensated through the expression of a complex proteolytic enzyme system that provides essential amino acids via hydrolysis of casein. Besides providing essential amino acids, these proteolytic enzymes are involved in the development of flavor in fermented dairy products [144]. This bacteria is used extensively as a starter or adjunct culture for manufacturing swiss type and aged italian cheese and fermented milk [145, 121]. It is the dominant microflora of the natural whey starters used for Parmigiano Reggiano cheese making [146]. *Lb. helveticus* CNRZ32 has the ability to reduce bitterness and accelerate the development of cheese flavor [147]. The genomes of a number of *lactobacilli* have been determined [148]-[153], creating good basis for the reconstruction of metabolic models. This background coupled with the industrial interest in the production of a wide range of fermented milk, meat, and plant and health-benefiting products and supplements reveals the potential and importance of the metabolic model reconstruction of *Lb. helveticus*.

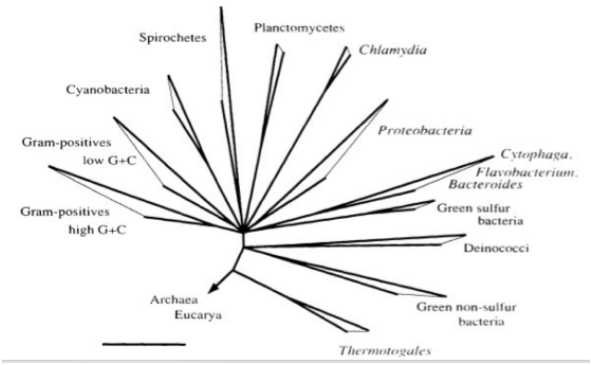


Figure 5.
Phylogenetic tree of bacteria

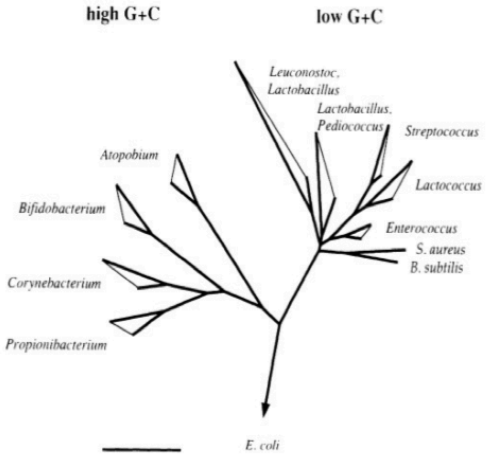


Figure 6.
Gram-positive bacteria

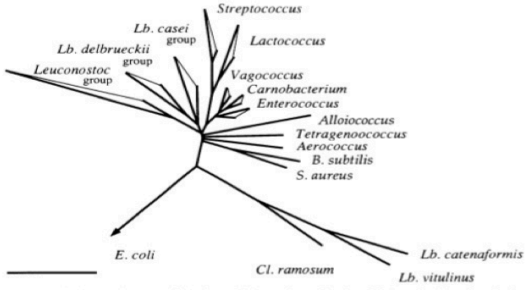


Figure 7.
lactic acid bacteria and related Species

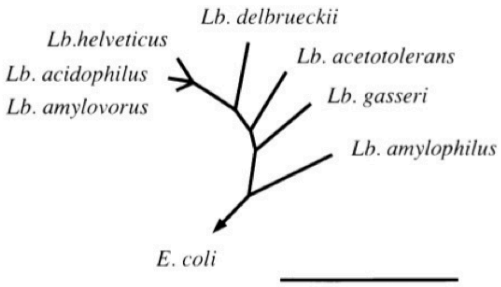
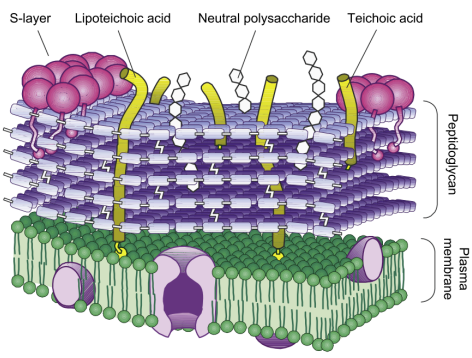
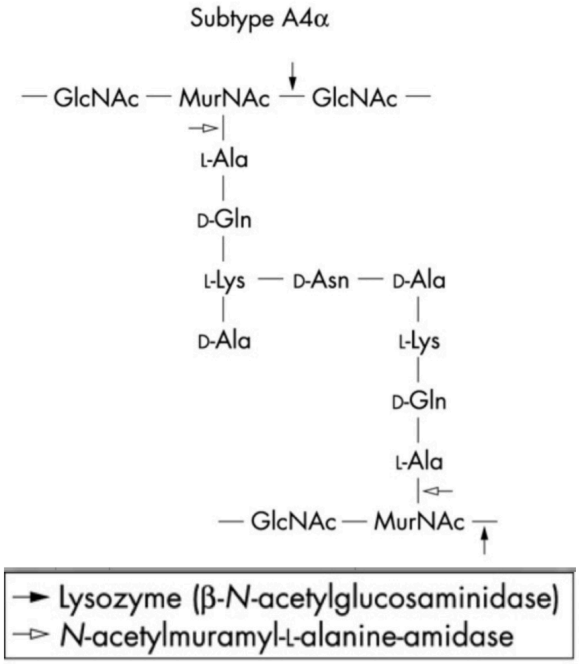


Figure 8.
L. delrueckii group

Figure 9.
Taxonomic grouping in pylogenetic trees of *lactobacilli* in different contexts [5].
Based upon 16s rRNA sequence comparison. Each bar represents 10% of expected sequence divergence



(a)
Schematic representation of cell wall [120]



(b)
Peptidoglycan structure [6]

Figure 10.
General constitution of a gram positive bacteria cell wall and peptidoglycan structure [6].
Covering the plasma membrane there is a peptidoglycan structure embedded with LTA and Teichoic acids (TA) [121]. Peptidoglycan contains the sugar unit of N-acetylglucosamine (GlcNAc) and N-acetylmuramyl(MurNAc) with an aminoacid chain constituted of L-Alanine D-Glutamate, L-Lysine and D-Alanine. Each repetitive peptidoglycan unit is connect by a bond of D-Asparagine

3

MATERIALS AND METHODS

3.1 FUNCTIONAL ANNOTATION

3.1.1 *merlin interface and integration for the annotation process*

merlin has a friendly-usage window interface with tools and options available by mouse-clicking. It has a menu bar in the top and a clipboard on the left side. Different views will be render for each selected option on the clipboard (Fig. 11). The process starts by creating a new project, connecting it to a MYSQL database, choosing the project name and the Taxonomy ID of the target organism. It is also required a FASTA file with the genomic CDS from NCBI Assembly. Afterwards, a BLAST is performed with personalized options for the user (fig. 12).

3.1.2 *Enzymes annotation*

After the rendered BLAST results it is necessary an analysis of the alpha-value and subsequent score annotation. The alpha-value is defined in *merlin* by the user and it controls the weight given to the frequency comparatively to taxonomy (See Equation 3). This choice will allow now to define a threshold. Above this defined threshold, the Upper Bound (UB), all the EECG can be automatically accepted and annotated. An inferior level is also necessary to be set. All the EECG with the score level below this threshold will be rejected [154]. This inferior threshold, the Lower Bound (LB) will be the one defined by the user in *merlin*. Even so, a manual curation is still necessary, in the range of the threshold in order to reduce as maximum as possible the number false-positives (FP) annotated. This task is facilitated by *merlin* fast two-clicks access to Uniprot and BRENDA information on each EECG.

Another defined parameters are the β-value, that defines the weight of the taxonomy and is fixed in 0.15 and the Minimum number of homologies, defined as 3.

After this analysis the re-annotation is performed leading the process of reconstructing a genome-scale metabolic model to the next steps.

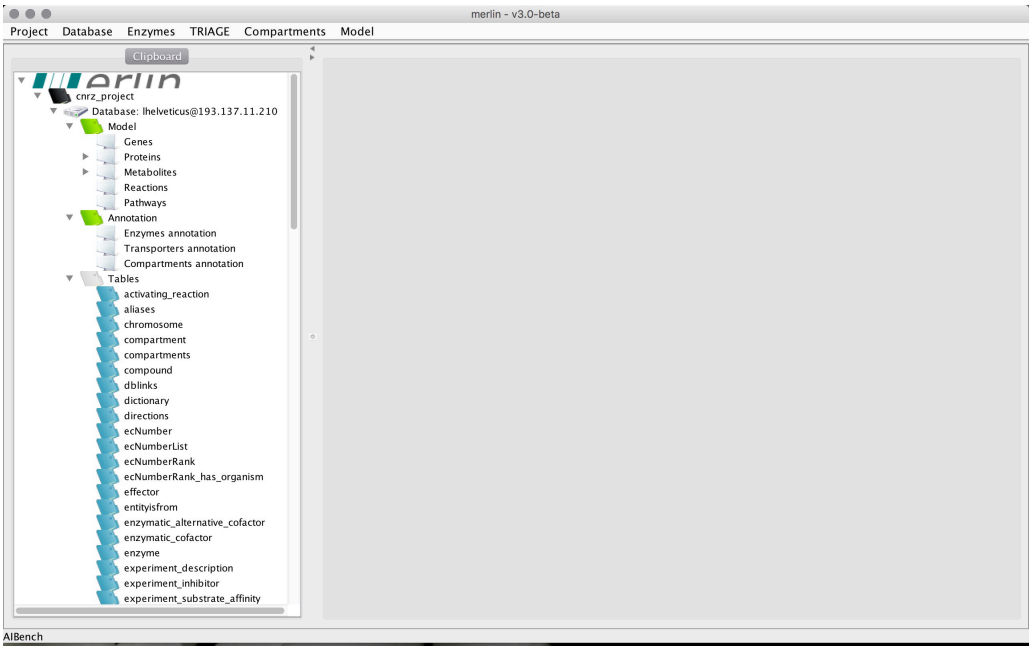


Figure 11.

Screenshot of *merlin* v.3.0 beta interface

On the top bar, multiple options are available for integrate, edit or remove data in the software. On the left is present a clipboard. The main levels on the clipboard represent the created project (cnrz_project), the hosting database (lhelveticus 193.173.11.210), and then Model, Annotation and Tables, which one with their own sublevels. Clicking in any of this levels or sublevels will render a view in the right side (grey when nothing is selected).

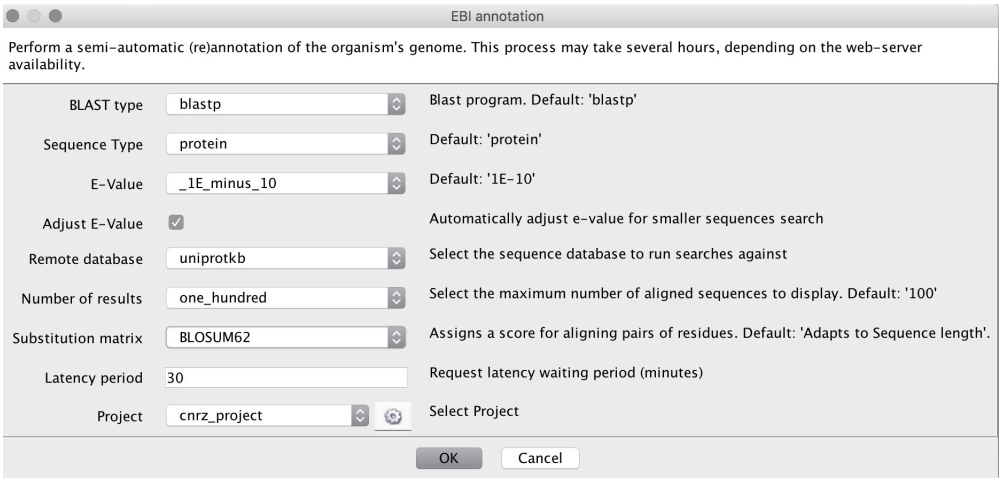


Figure 12.

Parameters definition for BLAST performance

In this example, 'blastp' and 'protein' options are selected to obtain the EECG. The E-value represents the minimum accepted E-value for an annotated enzyme be accepted. UniprotKB was the chosen remote database which blastp was performed against. Finally, 100 results were maximum number of aligned sequences to display, and chosen substitution matrix was BLOSUM62

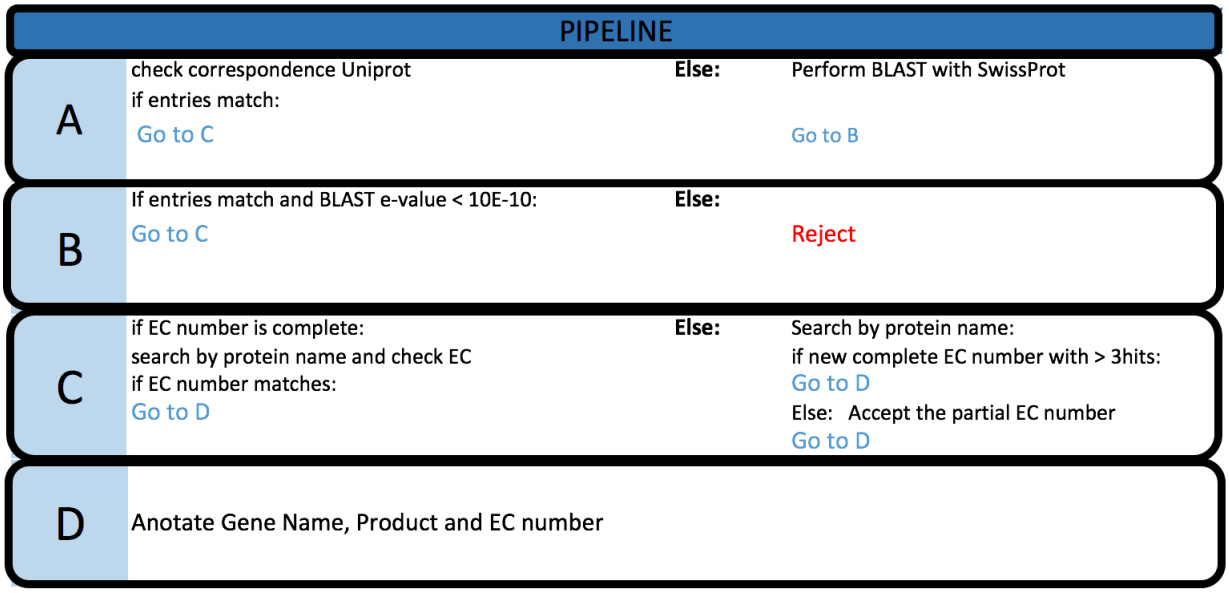


Figure 13.
Annotation pipeline for the assignment of enzymatic functions.
In step A, the locus-tag on each EECG was verified on Uniprot. In case of no correspondence, a BLAST with the SwissProt Database was performed, advancing to step C if conditions are fulfilled. In step C the EC number was checked and revised on BRENDA If the EC number was complete and function confirmed, then the gene was annotated. If the EC number is incomplete, alternatives with complete EC were tried to be found. If there were less than 3 homologies, they were dumped and the partial EC was annotated. If there found at least 3 homologies with the complete EC, the revision made before was performed again and then, again if was an entry with the accomplished requirements, then it was annotated. Also research was made in the databases listed in Table 1. For instance, a search on ExPASy was performed to confirm the product name as there were several EECG that had alternative accepted names.

3.1.2.1 *EECG Annotation Curation Pipeline*

The manual curation on the enzyme annotation was performed for the EECG which annotation confidence score was between the determined UB and LB thresholds. This curation follows the described work-flow described in Fig. 13.

The definition of the UB and LB consists in a stage with multiple steps:

1. Definition of True (T) / False (F), Positive (P) / Negative (N). It has in consideration the pipeline (Fig. 13) and the annotation performed by *merlin*.
   -TP: if there is correspondence between the pipeline workflow and annotation;
   -FP: Exists an EC but in the pipeline it is considered non-metabolic or determines a different EC;
   -FN: lower score than threshold but by the work-flow an EC number is attributed;

0	EECG NAME	EC NUMBER	score	0.2	score	0.3	score	(...)	0.8	score	
	EECG 1	3.4.19.12	0	TN	0.01	TN	0.01		TN		0.01
	EECG 2	3.6.4.13	0.08	TN	0.07	TN	0.06		TN		0.05
	EECG 3	3.1.3.5	0.08	TN	0.07	TN	0.06		TN		0.05
	EECG 4	3.1.3.-	0.08	TN	0.07	TN	0.06		TN		0.05
	EECG 5	3.4.-.-	0.82	TN	0.09	TN	0.08		TN		0.07
0.1											
	EECG 6	2.3.1.189	0.16	TN	0.14	TN	0.12		TN		0.1
	EECG 7	3.6.3.-	0.16	TN	0.14	TN	0.12		TN		0.1
	EECG 8	3.4.-.-	0.16	TN	0.14	TN	0.12		TN		0.1
	EECG 9	1.1.1.49	0.16	FN	0.14	FN	0.13		FN		0.11
	EECG 10	2.4.2.31	0.23	TN	0.21	TN	0.18		TN		0.16
(...)											
1											
	AGQ22965.1	2.3.1.-	0.21	TP	0.23	TN	0.28		TN		0.24
	AGQ23919.1	2.1.1.-	0.22	TP	0.34	TN	0.3		TN		0.26
	AGQ24190.1	1.8.4.14	0.22	TP	0.34	TN	0.3		TN		0.26
	AGQ23524.1	3.5.4.5	0.25	TP	0.39	FN	0.34		FN		0.3
	AGQ23965.1	3.4.11.2	0.27	TP	0.39	FN	0.35		FN		0.31

Figure 14.

Example of annotation analysis appearance

In the first column will be the threshold values, varying from 0 to 1. In the second column the EECG randomly chosen (5 genes between each 0.1 interval of threshold(In the following column will be the confidence score. (The genes were chosen with an alpha-value of 0.5) Afterwards, each column will have respectively the classification of T/F, P/N for each alpha-value from 0.2 until 0.8 and the next column the corresponding confidence score.

-TN: lower score than the threshold and non-metabolic gene.

So all the genes with EC attributed by *merlin* will be considered as positive.

2. Annotation analysis: There are selected randomly fifty EECG, five between each 0.1 confidence score gap from 0 until 1. Afterwards, each gene is evaluated in each alpha-value level from 0.2 until 0.8 (gaps of 0.1). The classification will be TP, FP, TN or FN under the conditions described above. This procedure is better described in Fig. 14

3. Confusion matrix: it is built counting the number of genes in each classification (Fig. 15).

4. Choice of alpha-value and upper and lower bounds. Using the values obtained in the confusion matrix, another table is built considering different metrics:
   - *Accuracy*: allows to choose the most adequate alpha-value (Equation 5).

$$Accuracy = \frac{\sum_{i=1}^{n}(TP + TN)}{\sum_{i=1}^{n}(Total\_Population)} \tag{5}$$

   - *Precision/ Positive Predictive Value (PPV)*: from this calculation is defined the UB. The Precision is calculated for each cell. The biggest value of precision on each column

Classification		T	F	T	F	T	F	T	F	T	F	T	F	T	F
		0.2		0.3		0.4		0.5		0.6		0.7		0.8	
P	0	35	15	35	15	33	17	31	19	30	20	30	20	30	20
N		0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0.1	35	10	35	10	33	12	31	14	35	15	30	15	30	15
N		5	0	5	0	5	0	5	0	0	0	5	0	5	0
P	0.2	34	6	34	6	32	8	30	10	29	11	29	11	29	11
N		10	0	9	1	9	1	9	1	9	1	9	1	9	1
P	0.3	27	8	27	8	27	8	28	7	27	8	27	8	27	8
N		12	3	12	3	12	3	12	3	12	3	12	3	12	3
P	0.4	27	3	27	3	27	3	28	2	27	3	27	3	27	3
N		17	3	17	3	17	3	17	3	17	3	17	3	17	3
P	0.5	24	1	24	1	24	1	24	1	24	1	24	1	24	1
N		20	5	20	5	20	5	20	5	20	5	20	5	20	5
P	0.6	12	0	12	0	12	0	20	0	12	0	12	0	12	0
N		28	10	28	10	28	10	20	10	28	10	28	10	28	10
P	0.7	15	0	15	0	15	0	15	0	15	0	15	0	15	0
N		20	15	20	15	20	15	20	15	20	15	20	15	20	15
P	0.8	13	0	13	0	13	0	13	0	13	0	13	0	13	0
N		23	14	23	14	23	14	26	11	30	7	30	7	33	4
P	0.9	8	0	8	0	8	0	8	0	8	0	8	0	8	0
N		20	22	20	22	20	22	20	22	20	22	20	22	20	22
P	1	3	0	3	0	3	0	3	0	3	0	3	0	3	0
N		20	27	20	27	20	27	20	27	20	27	20	27	23	24

Figure 15.

Example of a confusion matrix construction

After the annotation analysis, this confusion matrix is built counting the number of occurrences of T,F,P,N. It will be the basis for the matrix calculations.

is chosen, which will correspond to the UB for the corresponding alpha-value of the column (Equation 6).

$$PPV = \frac{\sum_{i=1}^{n}(TP}{\sum_{i=1}^{n}(Test\_outcome\_positive(TP + FP))} \tag{6}$$

- *Negative Predictive Value (NPV)*: allows to determine the LB. In each column corresponding to each alpha-value, the one with biggest value will correspond to the LB for that alpha (Equation 7).

$$NPV = \frac{\sum_{i=1}^{n}(TN}{\sum_{i=1}^{n}(Test\_outcome\_negative(TN + FN))} \tag{7}$$

- *'y-value' and 'z-value'*: after the previous calculations to decide the correct alpha and threshold values, it was used a metric. A "y-value" is computed, equal to the number of genes that would need to be curated (genes with confidence score between the lower and upper bounds) for each alpha-value, divided by the total number of genes. After, a "z-value" is computed which will be equal to the accuracy divided by the "y-value".

So, the alpha-value corresponding to the biggest "z-value" was chosen and therefore the respective LB and UB determined for that alpha-value.

$$y - value = \frac{\sum_{i=1}^{n}(curated\_genes)}{(\sum_{i=1}^{n}(Total\_number\_of\_genes)} \tag{8}$$

$$z - value = \frac{Accuracy}{y - value} \tag{9}$$

3.1.3 *Transporter Proteins Annotation*

The annotation was performed with TRIAGE tool, included in *merlin*. TRIAGE works with *Phobius* tool to retrieve the transporter proteins [64]. *Phobius* identifies transmembrane protein topology and signal peptide predictor based on a hidden Markov model (HMM) that models the different sequence regions of a signal peptide and the different regions of a transmembrane protein in a series of inter-connected states.

3.1.3.1 *Transporter Proteins Annotation Curation Pipeline*

After the TRIAGE running in *merlin*, there are still transport proteins that are not annotated, so they have to be manually added to the model (Fig. 16). The process is made by filling an internal database in which 5 major columns are fulfilled, as they are the parameters *merlin* will use in its smith-waterman alignment algorithm performed in TRIAGE. This parameters are direction, metabolite, reversibility(T/F), reacting metabolites and equation. The curation is performed by reviewing the genes one by one, retrieving information from different levels: TCDB structure, Uniprot description, family, subfamily and super family descriptions and literature.

3.1.3.2 *Transport Reactions Creation and Integration*

After completing the manual annotation of the missing metabolites, the TRIAGE database information is integrated in *merlin*. The information is linked to the genes data to create the transport reactions and finally integrated to the model.

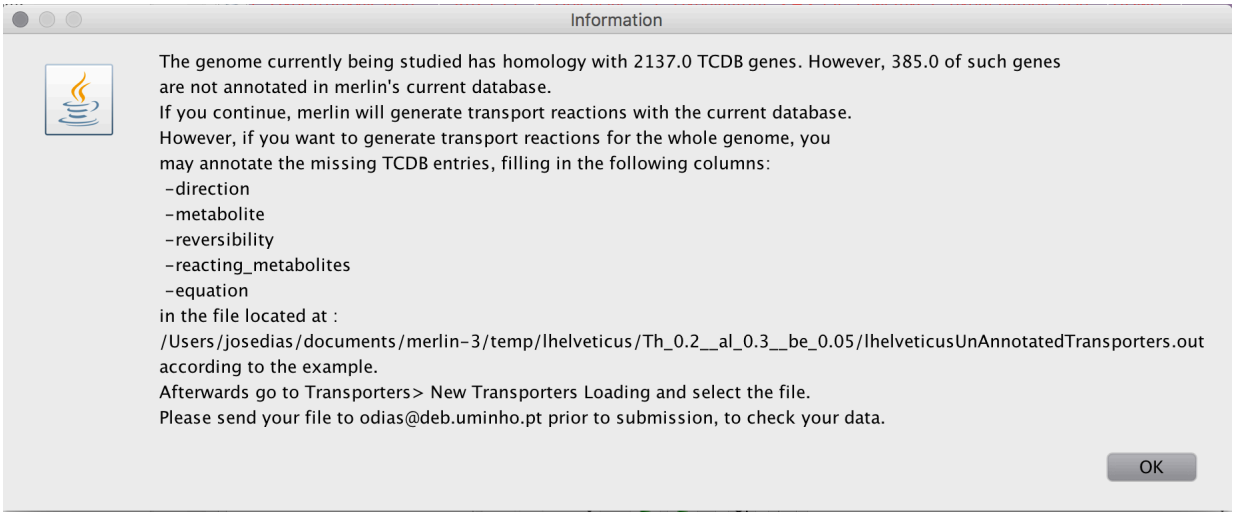


Figure 16.

merlin output after TRIAGE performance.

It contains information on the total number of homologies found (2137) and the number of genes that are still not in *merlin* database (385). For those, a manual annotation should be performed, fillings the fields of direction, metabolite, reversibility, reacting metabolites and equation

3.2 DRAFT NETWORK RECONSTRUCTION

After the annotation and curation of enzymes and transporters, the Enzyme Annotation is committed and integrated to the model (clickable options on 'Enzyme Annotation view'). Then with the 'Load Metabolic Data' tool a list of all the pathways and reactions with homology to the KEGG database is generated. In this view, it is possible to select each pathway individually and visualize the active reactions using the 'Draw in Browser' option. It connects to the KEGG maps opening it in a new browser window. When used together with the 'unconnected reactions' tool, the active reactions will appear on the pathway map with associated colors.

3.2.1 *Pathways and Reactions Curation*

A pathway by pathway analysis was performed. For each pathway was performed the following workflow:

- Identifying the putative routes within the pathway;
- Screenshot of annotation view of *Lb. helveticus* CNRZ 32;
- Screenshot of KEGG reference organism model annotation view;
- Describe differences between annotated model and reference one and relevant notes;

3.2.1.1 *Unconnected reactions*

This feature available in *merlin* allows to differentiate the reactions in a color scale by their interconnection as described below:

- **Green** EC numbers are integrated in the model. Generally recognized as the main active enzyme to a given reaction.
- **Blue** EC numbers are integrated in the model. Generally recognized as secondary enzyme, often associated with another pathways, to a given reaction.
- **Cyan** EC numbers are integrated in the model, but connected to a dead end, i. e., the next connected EC number is not in the model.
- **Red** EC numbers integrated in the model, but represent a dead end reaction, i. e. their products are end results of the pathways.
- **Colorless** EC number are absent in the model

3.2.1.2 *Directionality and Reversibility of Reactions*

As the reactions are obtained from KEGG database, they are established as reversible by default. To try to approximate the model to the constraints of the real world it is necessary to identify the irreversible reactions and constrain them as unidirectional. To fulfill this purpose, *merlin* automated tool 'correct reversibility' was used. This tool was constructed based on previous studies combining BRENDA results with ones in KEGG [155, 156]. Reactions are turned irreversible settling the flag 'reversibility' to '0'. Initially the UB and LB are settled in thousands, simulating a hypothetical unlimited flux (-99999 to 10000). To settle the direction of the reaction the lower and upper bounds are adjusted. Therefore, if a reaction is supposed to occur in the direction of the products, the LB is changed to '0' and in the case of a reaction occurring in the reactants direction it is the UB that is changed to '0'. However, the process is still liable to errors. To try to correct most of the possible errors a complementary process was developed combining knowledge obtained from MetaNetX, BiGG and modelSEED databases, which are manually curated. An in-house python script was written to correlate the reactions in the model (KEGG based) with the aforementioned databases (unpublished data kindly provided by Ahmad Zeidan). The final output was an excel file with the single identifier of MetaNetX for every single reaction aligned with the respective aliases of the other three databases by identifier, equation, directionality and reversibility. When information for a reaction was just partial or not in agreement between the different databases it was corrected following the trustworthy scale:

1. BiGG
2. modelSEED
3. KEGG

3.2.1.3 *Redundant and active pathways curation*

Firstly, general and complex pathways were removed. This removal process is performed in *merlin* on the 'Reactions' view. For the remaining pathways their removal was performed after a comparative analysis with the following criteria:

- no reference model found on KEGG (no *Lb. Helveticus CNRZ32* or other strains found);
- few active reactions on the pathway and the the main product of the pathway is not produced. For instance, the 'Tyrosine metabolism' has just two active enzymes, both redundant to other pathways and the tyrosine metabolite itself is not being produced. In this case, the tyrosine metabolism pathway would be removed;
- redundant reactions. If all the reactions on the pathway are linked to other pathways, then the pathway is removed.

After the removal of the redundant pathways a new and deeper pathway-by-pathway analysis was performed. For each ongoing active pathway it was described the putative routes and taken screenshots of the pathway in reconstructed network and the KEGG reference organism annotation views. All the differences and potential gaps were evaluated and notes were taken for further evaluation.

3.2.1.4 *Unbalanced reactions*

In order to have a running model it is necessary to all reactions to be balanced. For this task, *merlin* has a built-in tool retrieving the reactions with the wrong mass balance. On the reactions view, this unbalanced reactions appear in bold to be easiyl identified for further analysis. It also calculates the difference between the components in reactants and products resulting in the missing elements to obtain a balanced reaction. The unbalanced reactions could then be analyzed and corrected.

3.2.2 *Biomass Equation*

For the biomass equation construction *merlin* has a built-in tool to create a draft version for the equation. Named 'e-biomass equation', it allows to create semi-automatically the reactions that lumped together will correspond to the biomass production. The user provides FASTA files with genomic, proteic and the different RNA (tRNA, mRNA, rRNA) sequences and indicates the content amount (from 0 to 1) of each one of the main macromolecules constituting the cells. These macromolecules are proteins, carbohydrates, DNA, RNA and cofactors, generating the respective 'e-protein', 'e-carbohydrates', 'e-DNA', 'e-RNA', 'e-cofactors' and finally the 'e-Biomass' reactions. It is important to mention that this process only provides in an assisted starting point for the Biomass equation construction. It is necessary to add, remove or adapt the equations for the model under construction. For that, literature data on the composition of different macromolecules in Lb. helveticus or closely related organisms were used to refine the biomass equation. Experiments were also conducted to determine the macromolecular composition of *Lb. helveticus* CNRZ 32 cells.

3.2.2.1 *Energy Requirements*

Energy is necessary for the cell metabolism. This energy is available in the form of ATP to the organism. For the model construction it is necessary to take in consideration two types of energy. The energy associated to growth which should be included in the main biomass equation and the maintenance energy or non-growth associated energy. The latter considers the energy spent by the cell in other functions other than growing new cell material [157].

3.2.3 *Experimental Determination of the Macromolecular Composition of Biomass*

a) Medium and Growth Conditions

Lactobacillus helveticus CNRZ32 cells were inoculated in a chemically defined medium (CDM) supplemented with 2% of glucose. The medium contained as main elements glucose Trace Elements Basal Solution (80.2%), Magnesium Chloride (1%), Calcium Chloride (1%), Cysteine Hydrochloride (1%), Urea (1%), Amino Acid Stock Solution (3.8%), Vitamin Basal Solution (1%) and Bases Solution (1%). The medium components are fully described in the Support Material. An exponential growing culture was selected and incubated over night at 40 °C. Afterwards, 200 milliliter (mL) of CDM were inoculated at the 0.05 - 0.1 value of Optical Density (OD)600 and then incubated up to the exponential phase, commonly achieved round OD600 of 1. When the exponential phase was achieved, samples of 10 mL were collected and split according to its destination. Samples used for macromolecular content determination were centrifuged and its supernatant kept at −20 °C to be used as control. The pellets were washed with cold Phosphate-buffered saline (PBS). Samples for the cell dry weight determination were used immediately and the remaining were kept at −20 °C until further use.

b) Cell Dry Weight

The cell dry weight (CDW) of *L. helevteicus* CNRZ32 was determined in triplicate using 10-mL aliquots of cell suspension, harvested at mid-log phase. For that, 0.22 µM membrane filters were pre-dried in a microwave oven at 350 W for 4 min. The filters were cooled down in a desiccator before being weighed on an analytical balance. Then, by using the vacuum filtration assembly, the cell suspensions were filtered and washed three times with equal volumes of MiliQ water. The filters were then dried in the microwave oven at 350 W for 8 min, cooled down in the desiccator and weighed again. Later, the difference between the initial and final weight was calculated. The cell dry weight was achieved by dividing the obtained result by the filtered volume.

c) Protein

The total protein content of the cells was determined by the Biuret method, as described by Herbert et al. [158]. The previously frozen pellets were resuspended in MiliQ water and subsequently washed from any trace of growth medium. The volume of 0.6 mL washed cell suspension was transfered to a 1.5mL tube. Simultaneously, standard protein solutions were prepared using 2 milligram (mg)/mL Bovine serum albumin (BSA) and demineralized water with different concentrations: 0.25 mg/mL; 0.5 mg/mL; 1 mg/mL; 2 mg/mL. Also a blank of demineralized water was prepared. The standards and blank were treated henceforth as

with the cell suspensions. After the addition of sodium hydroxide, the samples were placed in boiling water bath and then cooled down in ice. Posteriorly, copper sulfate was added to the suspension, mixed and incubated at room temperature. Finally, the suspension was centrifuged and the supernatant collected to measure the absorbance in the spectrophotometer at 555 nanometer (nm). The standards were used to plot the absorbance values against the known concentrations and determine the regression line. This allowed to estimate the sample concentration.

d) Carbohydrates

Total carbohydrate content (capsular polysaccharide and free sugars inside the cell in addition to sugar residues in peptidoglycan and lipoteicoic and teichoic acids) of Lb. helveticus cells was determined according to the phenol-sulfuric acid method [158]. A sample of 1 mL frozen cells formerly suspend in MiliQ water in a thick glass tube. At the same time, glucose standards were prepared in parallel for the following concentrations: 0.25 mg/mL; 0.5 mg/mL; 1 mg/mL; 2 mg/mL. A blank was prepared with demineralized water. The standards and blank were treated henceforth as with the cell suspensions. It was subsequently added to the suspension 5 % phenol and sulfuric acid, and mixed immediately. Thereafter the samples were incubated at room temperature and afterwards placed in a water bath at 25 °C. Then, 1 mL out of the suspension was measured in the spectrophotometer at 488 nm. Finally, the carbohydrates content was estimated as with the protein content.

e) DNA

To determine the DNA content, it was used the Invitrogen Easy-DNA™kit for genomic DNA extraction (Support Material). The protocol started with the suspension of the samples in PBS. Mutanolysin was added and the samples incubated at 37 °C. It was added the kit solution A and the resulting suspension suffered a vortex before another incubation, this time at 65 °C. The solution B was added next and immediately suffered another vortex until the resulting precipitate was dissolved. Chloroform was added to the suspension. Subsequently, the suspension suffered vortex and centrifugation. The upper phase was collected into a tube followed by the addition of 100% ethanol. To the mixture it was applied other centrifugation. Prior to that, the samples were incubated on ice. The 100% ethanol was removed as ethanol at 80% concentration was added. Posteriorly, the samples were inverted five times and centrifuged. Lastly, the ethanol was removed and if necessary other centrifugation is performed to remove residual ethanol. The pellet was dried at 37 °C and resuspended on Tris-Ethylenediaminetetraacetic acid (TE) buffer plus RNase. The samples were incubated at 37 °C before measuring the DNA concentration using the Qubit.

f) RNA

The total cellular RNA content was quantified by the KOH/UV method [159]. The cells were washed three times with perchloric acid and afterwards digested by potassium hydroxide at 37 °C with mixing intervals every 10 minutes. Posterior to that the suspension was cooled down and neutralized with perchloric acid. The suspension was centrifuged whereupon the supernatant was collected. As so, the pellet was washed twice with perchloric acid and once again the supernatant were collected. To reach a final volume of 15 mL, the supernatant were pooled. Then, the remaining potassium perchlorate was removed by centrifugation. The resulting samples were measured at the absorbance of 260 nm was measured using the spectrophotometer. Later, the Beer-Lambert law was firstly used to calculate the final concentration of the samples while the 340 grams (g)/mol molecular weight was used hereafter to calculate the RNA content.

3.2.4 *Validation and Simulation*

After Biomass structure being built a validation of the model is necessary. In order to do it, two distinct softwares were used: COBRAPY and OptFlux. The first is orientated to find errors in the model and the latter to perform FBA simulations.

3.2.4.1 *COnstraint-Based Reconstruction and Analysis for Python (COBRApy)*

COBRA for Python (COBRAPY) is a Python package that provides support for basic COBRA methods. The openCOBRA Project is a community effort to promote constraints-based research through the distribution of freely available software (available in `http://opencobra.github.io/COBRAPY/`). COBRApy implies three fundamental concepts: the presence of physicochemical constraints, mathematical description of evolutionary selective pressures genome-scale perspective of cell metabolism accounting all gene products in a cell [160]. COBRA methods can be used in metabolic networks of prokaryotes and eukaryotes with an integration framework for the multiomics data used in systems biology. It provides access to commonly used COBRA methods, such as flux balance analysis, flux variability analysis, and gene deletion analyses. COBRApy serves as an enabling framework for which the community can develop and contribute application specific modules [161]. A SBML file (*merlin* output can be loaded and methods powered by the COBRA package can be extended in python scripts to analyze the model. The developed script goes reaction by reaction following the workflow:

1. The reactants metabolites are retrieved from the reaction under analysis;
2. A single sink reaction is iteratively created to each metabolite;
3. The new sink reaction is settled as objective function;

4. The model is simulated to the objective function 'maximize';
5. The output flux is recorded;
6. If the metabolite tested is biomass precursor entitled as macromolecular entity, the process is repeated regarding to the synthesis reaction of the macromolecular entity.

The running code generates an excel file with the fluxes for each of the precursors. When the flux was zero or close to zero it meant that the metabolite was not being produced in the model.

3.2.4.2 *OptFlux*

OptFlux is an open-source software platform for *in silico* metabolic engineering (available from `http://www.optflux.org`). It appears as a response for the needs of the scientific community. A bunch of methods have been already proposed before for the phenotype simulation of microorganisms under different environmental and genetic conditions. Although they were restricted to expert researchers and bioinformaticians. OptFlux brings a powerful tool for metabolic engineering working with genome-scale models in a user-friendly environment.

OptFlux is able to perform strain optimization, being the first metabolic engineering computational tool to provide algorithms and simulated annealing metaheuristics to reach targets given a user-defined objective function. It also allows the use of stoichiometric metabolic models for phenotype simulation of both wild-type and mutant organisms, using the methods of Flux Balance Analysis, minimization of metabolic adjustment or regulatory on/off minimization of metabolic flux changes, metabolic flux analysis, computing the admissible flux space given a set of measured fluxes, and pathway analysis through the calculation of Elementary Flux Modes. The software supports importing/exporting to several flat file formats and it is compatible with the SBML standard. [162].

OptFlux has already been used in published in testing metabolic engineering predictions testing different carbon sources uptake experiments [163, 164]. A SBML file (export format of models in *merlin* output can be loaded and then tested in the different modules of OptFlux. Predictions of cell behavior for different environmental conditions, carbon sources and reactions and genes Knock-out can be performed, looking for the best possible outputs to validate later on the wet-lab.

4

RESULTS AND DISCUSSION

4.1 FUNCTIONAL ANNOTATION

4.1.1 *EECG Annotation Results*

After uploading the file with the coding sequences for *Lactobacillus helveticus CNRZ32* and run the BLAST tool, the α-values and thresholds were evaluated. The values obtained were 0.8 for α-value, and 0.2 and 0.5 for the LB and UB respectively.

The curation process had an impact on the number of total genes, as there are less 145 genes when compared to before the evaluation. This is explained by the defined threshold levels. Before the curation the threshold was defined at 0, for it would be possible to make an evaluation of all metabolic genes. After the metrics evaluation it was established of the alpha-value of 0.8 and the threshold of 0.2 as LB. It was also defined 0.5 as the UB, which all the genes with confidence score above it are considered as high-confidence genes and accepted as correctly annotated. The genes with confidence score between the LB and the UB were manually evaluated and curated giving resulting in differences their groups distribution and in 30 new complete EC numbers annotated s and add extra information to three more, despite not completely fulfill them (Tables 4 and 5).

4.1.2 *Transporters Annotation Results*

A total of 2137 TCDB homologies were found for the organism genome (Fig. 16). From those, 385 were still not annotated in *merlin* database and needed to be manually curated. All this data was integrated in *merlin* and 232 transport reactions were created and integrated to the model.

Table 4.
Differences of distribution of EC numbers before and after the manual curation.

	after	*before*	*dif*
EC numbers	575	720	-145
EC incomplete	19%	28%	-10%
EC complete	81%	72%	9%
Gene name	19%	17%	2
Oxidoreductases	10%	10%	0%
Transferases	33%	33%	0%
Hydrolases	34%	38%	-4%
Lyases	5%	4%	1%
Isomerases	7%	6%	1%
Ligases	12%	9%	2%

4.2 DRAFT MODEL RECONSTRUCTION

4.2.1 *Pathways and Reactions Curation*

4.2.1.1 *Directionality and Reversibility of Reactions*

Followed the automatic correction in *merlin*, the script developed revealed that a total of 131 did not reveal agreement between all the databases. From those, 49 were switched from reversible to irreversible and 36 the reverse process. The remaining were kept unaltered as they resulted from parsing errors on the script or from the differences in the approach of mass and charge balance between modelSEED and KEGG. The changes performed were confirmed with literature support when available.

4.2.1.2 *Unbalanced Reactions*

The balance of the reactions was evaluated. From a total of 939 reactions 217 were unbalanced. Although not all those reactions had to be corrected. Seventy-three of those reactions were exchange reactions. In this reactions only the reactants side of equation has metabolites. Another 20 of those unbalanced reactions belonged to the 'Amynoacyl-tRNA biosynthesis' pathway. In this pathway each reaction is unbalanced, but the sum of all of them results in a balanced matrix, simulating the protein building and elongation. Nine reactions belonged to the biomass pathway. Each of this reactions were unbalanced as the artificial 'e-components' do not have chemical formula or structure. Still, the final 'e-biomass' equation balanced. This lead to a total of 115 unbalanced reactions that needed to be manually updated. From those 87 were non-associated to pathways and therefore were left to later evaluation. These reactions would be later reviewed when assured that they would be necessary or not in the model. A final number of potential 28 reactions to balance remained. The reactions

Table 5.

Completed EC numbers.

Gene	Incomplete EC	New annotation	Protein Name
AGQ22686.1	1.1.1.-	1.1.1.274	2,5-didehydrogluconate reductase
AGQ22711.1	2.7.3.-	2.7.13.3	Histidine kinase
AGQ22837.1	3.1.3.-	3.1.3.48	Protein-tyrosine phosphatase
AGQ22860.1	1.-.-.-	1.16.1.1	Mercury II reductase
AGQ22862.1	2.3.1.-	2.3.1.183	Phosphinothricin N-acetyltransferase
AGQ22878.1	4.2.1.-	4.2.1.59	(3R)-hydroxymyristoyl-[acyl carrier protein] dehydratase
AGQ22898.1	2.7.7.-	2.7.7.80	Molybdopterin-synthase adenylyltransferase
AGQ22931.1	1.7.-.-	1.6.5.2	Possible NAD(P)H dehydrogenase (Quinone)
AGQ22932.1	1.7.-.-	1.7.1.6	Possible NAD(P)H dehydrogenase (Quinone)
AGQ22946.1	2.7.3.-	2.7.13.3	Histidine kinase
AGQ22987.1	2.4.1.-	2.7.8.6	Undecaprenyl-phosphate galactosephosphotransferase
AGQ23025.1	5.4.2.-	5.4.2.1	Phosphoglycerate mutase
AGQ23087.1	5.4.2.-	5.4.2.1	Phosphoglycerate mutase
AGQ23095.1	2.4.1.-	2.4.1.187	UDP-N-acetyl-D-mannosamine transferase
AGQ23096.1	2.4.-.-	2.4.1.-	Glycosyltransferase
AGQ23099.1	2.7.8.-	2.7.8.12	Putative CDP-glycerol:glycerophosphate glycerophosphotransferase
AGQ23110.1	2.7.1.-	2.7.1.107	Diacylglycerol kinase
AGQ23147.1	3.2.-.-	3.2.2.1	Nucleoside hydrolase
AGQ23167.1	3.4.11.-	3.4.11.7	Glutamyl aminopeptidase
AGQ23179.1	2.6.1.-	2.6.1.83	LL-diaminopimelate aminotransferase
AGQ23208.1	3.6.3.-	3.6.3.10	H+-K+-exchanging ATPase
AGQ23282.1	2.7.3.-	2.7.13.3	Histidine kinase
AGQ23300.1	2.1.1.-	1.3.3.4	Protoporphyrinogen oxidase
AGQ23335.1	2.4.-.-	2.4.1.-	Bactoprenol glucosyl transferase
AGQ23367.1	5.4.2.-	5.4.2.1	Phosphoglycerate mutase
AGQ23376.1	2.1.1.-	2.1.1.171	Methyltransferase
AGQ23596.1	3.6.3.-	3.6.3.21	Polar amino acid transport system ATP-binding protein
AGQ23846.1	3.6.3.-	3.6.3.21	Polar-amino-acid-transporting ATPase
AGQ23914.1	1.1.1.-	1.1.1.27	L-2-hydroxyisocaproate dehydrogenase
AGQ23923.1	2.7.3.-	2.7.13.3	Histidine kinase
AGQ24041.1	2.-.-.-	2.4.-.-	Glycosyltransferase
AGQ24137.1	5.4.2.-	5.4.2.1	Phosphoglycerate mutase
AGQ24254.1	3.4.22.-	3.4.22.40	Aminopeptidase C

Table 6.

Removed pathways.

Metabolic pathways	Limonene and pinene degradation
Microbial metabolism in diverse environments	Lysine Degradation
Acarbose and validamycin biosynthesis	Metabolism of xenobiotics by cytochrome P450
Aflatoxin biosynthesis	Methane metabolism
Arginine and proline metabolism	Monobactam Biosynthesis
Arginine biosynthesis	Monoterpenoid biosynthesis
Ascorbate and aldarate metabolism	Nitrogen metabolism
Aminobenzoate degradation	Nitrotoluene degradation
Benzoate degradation	Penicillin and cephalosporin biosynthesis
Betalyn Biosynthesis (not able to draw on KEGG)	Pentose and glucoronate interconversions
Biosynthesis of type II polyketide products	Phenylalanine metabolism
Biosynthesis of unsaturated fatty acids	Phenylalanine, tyrosine and tryptophan biosynthesis
Biosynthesis of vancomycin group antibiotics	Porphyrin and chlorophyll metabolism
Biotine Metabolism	Propanoate metabolism
Bisphenol degradation	Retinol metabolism
Butanoate metabolism	Selenocompound metabolism
Caprolactam degradation	Sphingolipid metabolism
Carbon fixation in photosynthetic organisms	Steroid degradation
Carbon fixation pathways in prokaryotes	Streptomycin biosynthesis
Chloroalkane and chloroalkene degradation	Styrene degradation
Chlorocyclohexane and chlorobenzene degradation	Synthesis and degradation of ketone bodies
Cyanoamino acid metabolism	Taurine and hypotaurine metabolism
Drug metabolism - other enzymes	Tetracycline biosynthesis
Drug metabolism - cytochrome P450	Tropane, piperidine and pyridine alkaloid biosynthesis
Fatty acid elongation	Tryptophan metabolism
Glyoxylate and dicarboxylate metabolism	Tyrosine metabolism
Glycosaminoglycan degradation	Valine, leucine and isoleucine biosynthesis
Glycosphingolipid biosynthesis - ganglio series	Valine, leucine and isoleucine degradation
Glycosphingolipid biosynthesis - globo and isoglobo series	Various types of N-glycan biosynthesis
Histidine Metabolism	Xylene degradation
Indole alkaloid biosynthesis	Zeatin biosynthesis
Inositol phosphate metabolism	α-Linolenic acid metabolism
Isoquinoline alkaloid biosynthesis	

containing polymers such as starch and dextrin were removed and the remained corrected after confirming the agreement with BiGG and modelSEED databases. Mostly proton balance corrections were performed.

The next step was the 'Unconnected reactions' tool in *merlin*. It was not performed the reaction removal in order to keep the model closest to the reality, i.e., isolated reactions or reactions leading to partial pathways with no evident connection or interoperability with other pathways were kept for further studies. The tool was mainly used for identify dead ends. One reaction follows the rule '$\mathbf{A + B = C + D}$' This 'C' and 'D' metabolites cannot be 'lost'. They had to either be used as reactants for further reactions or excreted by the metabolism. When none of this conditions is respected it is created a dead end that has to be analyzed and corrected.

4.2.1.3 *Redundant and active pathways curation*

A total of 65 pathways were considered redundant and therefore removed from a total of the initially 106 generated in *merlin* (Table 6). Removed the redundant pathways, a total of 36 pathways were maintained and analyzed one-by-one. The main pathways which are responsible for the carbon sources uptake and metabolism and building blocks from main precursor metabolites constituting the cell components are briefly described later on section 4.4.

4.2.2 *Biomass Equation*

To fill the different components content in *merlin*'s "e-Biomass equation" tool it was followed experimentally validated values by Santos [165]. These were 0.53 in protein, 0.02 in DNA, 0.08 in RNA, 0.125 in lipids and 0.2 in carbohydrates. Beside these components, it also adds a residual value for the called 'e-cofactors'. The information for these fields had to be uploaded with different FASTA files for each one of components: genome nucleotide sequence, genome amino acid sequence, genome transfer RNA (tRNA) sequence, genome mRNA sequence and genome ribosomal Ribonucleic acid (rRNA) sequence. These files were obtained from the NCBI Assembly database lastly accessed in March, 7h 2017 in :`https://www.ncbi.nlm.nih.gov/assembly/GCF_000422165.1`. Although another adjustments were still necessary. For having the most accurate and specific structure and components amounts as possible, other published models were taken in consideration. The models were available for *L.lactis* [166] and *B.subtillis* [167]. In these models, the cellular composition of the organism is distributed differently from *merlin*. In the *L.lactis* model it is taken in consideration the LTA, peptidoglycan and polysacharides. In the *B.subtillis* is described the structure and weight of each component of the LTA and the cell wall (peptidoglycan and TA. So, from the first e-biomass equation it was removed the 'e-carbohydrates' entity and created the 'e-peptidoglycan', 'e-EPS', 'e-teichoic acids' and 'e-lipoteichoic-acids' The weight of each one of these biomass precursors was distributed and adapted from *merlin* starting point as described below.

For the nucleic acids equations, 'e-DNA' and 'e-RNA' no changes were performed and their precursors and content percentage were kept as generated in the semi-automatic tool. For the 'e-Protein' equation, the precursors were kept as generated by *merlin* which estimates the codon usage from the amino acids sequence. The amino acids were considered in their activated form already associated with tRNA. The weight of protein would be later updated as new entities for the biomass equation were created. So, from the original 0.53 of protein content it was removed the amount corresponding to the proteic components in peptidoglycan, teichoic acids and lipoteichoic acids.

For the lipids components , it was necessary to create a *de novo* equation as *merlin* does not create it automatically. Literature review revealed that the lipid content for *Lb. helveticus* was based in cardiolipin, phosphatidylglycerol and phosphatic acid. The sum of their contents in the study was normalized to 1 after 7% of unidentified phospholipids were suppressed [168]. Plus, it was necessary to estimate the fatty acid composition as they are part of lipids constitution. Due to absence of specific information for the CNRZ32 strain, the average fatty acid composition was calculated with data obtained from different studies in other *Lb. helveticus* strains (Table 15 in Support Material). From this fatty acid profile it was still removed the odd number chain fatty acids as in the *Lb. helveticus* CNRZ32 was not

possible to observe the synthesis of 2-Methylbutanoyl-CoA or Propanoyl-CoA, precursors for this kind of fatty acids. This fatty acid component was virtually created inside *merlin* with the reaction 'R-fatty-acid:

0.16 tetradecanoic acid + 0.20 hexadecanoic acid + 0.03 (9Z)-Hexadecenoic acid + 0.03 octadecanoic acid + 0.58 cis-9-Ocatadecanoic acid = 1.0 Fatty acid'

For the 'e-carbohydrate' equation it was used a different approach. As a gram-positive bacteria, *Lb. helveticus* cell wall contains peptidoglycan, TA and LTA as main components. Plus, produces EPS. Being all important constituents of the organism, independent entities were created for each one of them. The content initially corresponding to the total 'e-carbohydrate' was distributed with the corresponding weight of the sugar constituents for the new entities: 'e-peptidoglycan', 'e-TA', 'e-LTA' and 'e-EPS'.

For 'e-Peptidoglycan' equation the precursors are UDP-N-acetylmuramate, UDP-N-acetyl-alpha-D-glucosamine, L-Alanine, L-Lysine, D-Glutamate, D-Alanine and L-Asparagine.

For the 'e-teichoic acids' equation the precursors are UDP-N-acetyl-D-mannosamine, UDP-glucose, UDP-N-acetyl-alpha-D-glucosamine, sn-Glycerol 3-phosphate and D-Alanine.

For the 'e-Lipoteichoic acid' equation the selected precursors were UDP-glucose, UDP-N-acetyl-alpha-D-glucosamine, Acyl-CoA, sn-Glycerol 3-phosphate and D-Alanine. Acyl-CoA is important as it is a necessary component for the reactions leading to the lipids synthesis. The presence of Acyl-CoA implied the creation of this entity in s well. This reaction uses the fatty acid entity created before:

'R acyl-CoA: Fatty acid + CoA + ATP = Acyl-CoA + Adenosine monophosphate (AMP) + Diphosphate'

The 'e-EPS' equation was built after creating a profile for the EPS. This profile was constructed with averages of 11 different strands of *Lactobacilli* strains as for our knowledge there is no profile defined yet for CNRZ32 strain [169]- [122]. UDP-Glucose, UDP-alpha-D-Galactose and dTDP-L-Rhamnose were defined as the precursors for the EPS production.

For the precursors of the 'e-cofactors' reaction, a literature review was performed and it was removed heme, as it is related to aerobic organisms, ubiquinone, associated to respiration processes [172] and glutathione, present in cyanobacteria and proteobacteria, and in all mitochondria or chloroplast-bearing eukaryotes [173]. Nicotinamide adenine dinucleotide (NAD), CoA, folate, thiamine and pyridoxal originally generated by *merlin* were kept. Pantothenate, 4-Aminobenzoate, Nicotinamide adenine dinucleotide phosphate (NADP)+, S-Adenosyl-L-methionine, nicotinamide and Biotin were latter added as they shown to be essential to the metabolism of the bacteria. The e-cofactor content amount was estimated as the remaining from all the other components to complete a ratio of 1. For instance, it was just necessary to virtually give a trace amount of NAD+ or folate as a boost to start-up since they are later renewed by the cell metabolism.

Summing up, the final e-Biomass equation included the upward described nine precursors for

Table 7.

Experimental macromolecules content.

CDW-cell dry weight in g/L, STD-Standard Deviation. The macromolecules contents are present in % (g/g). The presented values represent the average from triplicate results in each specie and experiment

	CDW	STD	DNA	RNA	Protein	Carbohydrates
Lb. helveticus	0.35	0.02	0.76	7.25	44.95	2.00
L. bulgaricus	0.60	0.10	0.49	6.48	19.11	1.22
St. termophilus	0.30	0.02	0.66	5.62	52.66	4.50

the bacteria main components plus water and energy. For the growth associated energy it was used a value of 10. It is important to refer that the amino acids are considered already in their activated form and therefore there is already ATP associated to their synthesis. It was also added and equation for the maintenance energy requirements which was fixed in 5 ATP.

The final e-biomass equation simulating the production of 1 mol of biomass was established as follows (note that e-biomass component is just used as a representation):

0.02 e-DNA + 0.08 e-RNA + 0.0617 e-cofactor + 0.034 e-lipids + 0.4863 e-protein + 0.12 e-exopolysaccharides + 0.0531 e-peptidoglycan + 0.08 e-lipoteichoic acids + 0.0649 e-teichoic acids + 10.0 H_2O + 10.0 ATP = 10.0 Adenosine diphosphate (ADP) + 10.0 Orthophosphate + 1.0 e-biomass

4.2.3 *Experimental Determination of the Macromolecular Composition of the biomass*

Calculated the values of the different macromolecules after extract, weight and measure them, values were not completely satisfactory. Experiments were performed in 3 different species and in triplicates. Values on *Lb. helveticus*, *L. Bulgaricus* and *S. thermophilus* were far away from the expected in Carbohydrates and DNA. For *Lb. helveticus* only the RNA values were close to expected (0.72 vs 0.8) as the protein content came up short as well. Even between species the obtained values vary, as for instance in protein content which varied from 19 to 52 %. To this inconsistencies added the non consideration of EPS production (medium is removed in the experiments wiping the EPS from the content). These factors turned inviable to use the measured wet-lab values for the model construction. So, it was decided to keep the combination of values obtained previous studies as described before.

4.3 MODEL VALIDATION AND SIMULATION

4.3.1 *Model troubleshooting and validation*

The biomass equation and model curation processes were followed by the file exporting and validation with COBRApy tool. Briefly, the troubleshooting was carried out through the evaluation of the model's ability to produce each of the biomass precursors. A script was developed to identify the biomass precursors not being produced by the model, i.e., with a flux of zero. Then, a trace back was performed on the reactions leading to precursors formation. For the reactants, it was created sink reactions and set up the objective function. The process was then iteratively repeated until a reactant with flux zero being identified. Furthermore, it was evaluated the reactions balance, the directionality and the reversibility. When necessary, the curation process would go back and forward until validation process was finalized. The model was recognized as operative and functional when all the biomass precursors have fluxes. Finally, the corrections were updated in *merlin* and then exported as an .xml file in the SBML Level 2 version 4 format.

4.3.2 *Simulations*

Obtained an expected functional model, simulations could now be performed. OptFlux (v3.3.5) was used for this task. The first test was performed without any restrictions, working as a control. Growth rate as expected was extremely high (above 1000), consistent with the biomass value obtained with COBRApy. The metabolites production gave already good indications (H_2O, Ammonia, CO_2, (R)-Lactate and (S)-Lactate).

Then, the model was tested with the conditions of growth used in the wet-lab. So, the only carbohydrate made available was glucose with a maximum flux of 10. The obtained growth rate was around the expected (0.26 vs $0.3h^{-1}$) as were the products: H_2O, Ammonia, CO_2 and (R)-Lactate. Although is expected to obtain lactate in its both isoforms, the *in silico* simulation may not contemplate it as it was used a pFBA method, which will try to reduce the number of used enzymes. Therefore, limiting the amount of sugar available it is most likely that it will result in only one of the isoforms to be represented. The bacteria consumed the glucose almost entirely. It was also taken up Uracil and Adenine All the twenty amino acids were taken up in small rates ($<0.1h^{-1}$) such as Hydrogen and cofactors (Riboflavin, Pyridoxal, Biotin, S-Adenosyl-L-methionine, 4-Aminobenzoate and Folate).

Tests with other sugar sources were made. Providing only galactose result in the same output as for glucose, only with the other lactate isoform. Tests with lactose, sucrose and maltose as sugar source failed with no growth obtained for any of them.

Environmental Conditions	Biomass value	Production
no constraints	1106	(H_2O, Ammonia, CO_2, (R)-Lactate and (S)-Lactate)
glucose LB -10	0.26	(H_2O, Ammonia, CO_2 and (R)-Lactate
galactose LB -10	0.26	(H_2O, Ammonia, CO_2 and (S)-Lactate
glycerol LB -10	NaN	None
lactose LB -5	NaN	None
maltose LB -5	NaN	None
sucrose LB -5	NaN	None
adenine omission	0.0	(R)-Lactate
alanine omission	0.0	(R)-Lactate
aspartate omission	0.26	(H_2O, Ammonia, CO_2 and (R)-Lactate
cysteine omission	0.0	(R)-Lactate
glutamine omission	0.26	(H_2O, Ammonia, CO_2 and (R)-Lactate
glycine omission	0.0	(R)-Lactate
isoleucine omission	0.0	(R)-Lactate
lysine omission	0.0	(R)-Lactate
methionine omission	0.0	(R)-Lactate
proline omission	0.0	(R)-Lactate
xantosine omission	0.26	(H_2O, Ammonia, CO_2 and (R)-Lactate

Testing the prototrophic amino acids (alanine, cysteine, lysine and serine) resulted in a change of behavior in the cell. Only the main path from glucose to lactate (represented in the figure 18) had fluxes, with glucose being consumed and lactate produced, but no growth was registered. For glutamine and glycine it was expected that their constraint would not have big impact in the cell performance. This was what happened when glutamine was taken from the medium, but with glycine the output was the same as for the auxotrophic amino acids. All the tests with amino acids and bases restrictions realized were conducted with glucose as sugar source. In table **??** it resumed the performed simulations results.

4.4 METABOLIC NETWORK SUMMARY

Finalized the model, a brief review of main pathways, precursors and respective building blocks was performed. In figure 18 the central carbon metabolism is represented. The process starts with the uptake of glucose and goes through multiple conversions until the lactate production. The schema comprises glycolysis, pentose phosphate and pyruvate metabolism pathways. Each pathway is described below and their respective KEGG maps screenshots are available in the Support Material.

- **Amino acids metabolism and Proteins** The amino acids were considered in the model in their activated form associated with tRNA and represented by the Aminoacyl-tRNA biosynthesis pathway. Most of amino acids pathways are not active and the cell depends of the uptake from the exterior (described in section 2.4.1.4). Even the synthesized amino acids have most of their pathways only partially complete.

- **Aminosugar and Nucleotide Sugar Metabolism** The pathway allows to create metabolites that will be precursors mainly for the peptidoglycan and TA biosynthesizes. The path started Glucose-6-phosphate until the conversions in N-acetyl glucosamine , N-acetylmuramate and UDP-Glucose appears to be complete, containing enzymes to all reactions leading to the formation of the before mentioned metabolites.

- **Citrate Cycle (Tricarboxylic acid cycle (TCA) cycle)** As a homofermentative lactic acid bacteria, *Lb. helveticus* it was expected that the pathway representing the citric acid cycle to be limited. It was observed the presence of only two enzymes metabolizing reactions (EC 4.2.1.2 and 1.3.5.4) converting Malate into Fumarate and the last one into Succinate. The reactions apparently were not functional and their presence in the genome can be explained by the genome decay from other related LAB ancestor species which were able to synthesize these components.

- **Cofactors metabolism** The cofactors necessary for the model to work were distributed in different pathways. The Riboflavin metabolism only had as active enzymes the conversion of Riboflavin in Flavin mononucleotide (FMN) and Flavin adenine dinucleotide (FAD).

  The Nicotinate and Nicotinamide Metabolism pathway was active in a cycle. If Nicotinamide is considered as a starting point, the cycle goes to the below section converting it into Nicotinate and continuing until the formation of NAD. From here the cycle goes back no Nicotinamide. It is also important to refer the presence of the NAD+ kinase enzyme which allows the phosphorylation of NAD into NADP.

  The folate biosynthesis was also well described containing the enzymes and respective reactions since the precursor Guanosine triphosphate until the cycle of regeneration of

Folate. This cycle is better represented in the 'One carbon pool by folate' pathway. The cycle with back-to-back conversions of Dihydrofolate, Tetrahydrofolate and Folate allows also the renewal of NAD and NADP.

- **Fatty Acid Biosynthesis** This pathway contemplates the formation of the multiple fatty acids potentially produced by the organism. It has Acetyl-CoA as precursor for all the reactions forming Malonyl-CoA. Gathered with Acyl-Carrier-Protein (ACP) it will form the starting block for the extension in the different chain sizes of fatty acids.

- **Glycerolipid and Glycerophospholipid Metabolism** These two pathways share the reactions potentially able to convert glycerone phosphate in sn-3-phosphate-glycerol and latter into phosphatidic acid (also known as phosphatidate or 1,2-DiAcyl-sn-glycerol 3-phosphate). The Glycerophospholipid pathway will also construct Cardiolipin and Phosphatidylglycerol. All this components gathered with the Acyl-group gave away by Acyl-CoA form the main structure for the lipids and LTA produced by the organism.

- **Glycolysis** The Glycolysis pathway contains the main backbone of *Lb. helveticus* metabolic network. Again, as a LAB and homofermentative, the metabolism of the organism works basically in being able to produce ATP to grow and maintain main cell functions and recycle the NAD. The pathway compiles the reactions since the sugar uptake (glucose) until its degradation in Pyruvate later converted in Lactate. It is a connection point with the Pentose Phosphate Pathway and the Pyruvate Metabolism.

- **Pentose phosphate pathway** Intimately connected to glycolysis in this pathway important metabolites are synthesized starting from Glucose-6-Phosphate. D-Gluconate-6-phosphate, D-Ribulose-5-phosphate, D-Xylulose-5-Phosphate, Phosphoribosyl pyrophosphate (PRPP) and D-Glyceraldehyde-3-Phosphate. These components are essential to Purine and Pyrimidine Metabolism and Glycolysis pathways. In this model the only source of acetyl-phosphate is also in this pathway.

- **Polyketide Sugar Unit Biosynthesis** It has a small portion active which contains a chain of reactions allowing conversion of Glucose-D-phosphate into dTDP-L-rhamnose, component of the EPS.

- **Purine Metabolism** The 'natural' course expected for the pathway would be to use the PRPP since the cell is capable of synthesize in Pentose phosphate pathway. Although, there is a missing link in the path in the phosphoribosylglycinamide formyltransferase. With no evidence for this enzyme to work, the alternatives would be to purines metabolism initiated by the histidine metabolism, but itis also inactive for this strain. Therefore, the necessary components for producing the purines need to be uptake from the medium. These are the bases Adenine, Guanine, Hypoxanthine and Xanthine.

- **Pyrimidine Metabolism** The Pyrimidine metabolism contrary to Purine Metabolism seems likely to be able to use PRPP for the pyrimidine synthesis. It has also enzymes to synthesize it from bases like Uracil and Cystine. In this context the importance of the proteolytic capacity of the bacteria to obtain all the necessary nutrients from the extracellular medium.

- **Pyruvate Metabolism** The homofermentative nature of *Lactobacillus helveticus* leads to an 'incomplete' pathway as Pyruvate is forwarded to the production of lactic acid and consequent NAD+ regeneration. The citric acid cycle is not active and the acetyl-phosphate is obtained in the reaction R01621 from D-Xylulose 5-phosphate and Orthophosphate. Other byproducts such as acetate are not produced.

- **Transporters pathway and Drains** The transporters pathway is a special pathway which represents the uptake and export of metabolites by the organism. The transport reactions were automatically generated by the TRIAGE tool before explained. It tries to emulate the different processes by which the components pass through the cell membrane. In this reactions the metabolites change between compartments (inside and outside). A total of 73 transport reactions were included in the model, being only 8 of them created manually.

  The Drains pathway is different from all the other as its reactions possess one individual metabolite in their equations present uniquely in the reactants side. It tries to simulate the exchange of components from the external environment to the medium (outside compartment). It is with this drains that the environmental conditions can be simulated. 'Opening' or 'closing' each one of the exchange reactions emulates the possibility of a component to be uptaked or exported by the cell. The model was construct with two different compartments: inside and outside. These compartments represent respectively the interior area of the cell and the exterior environment, separated by the plasma membrane (Fig. 17).

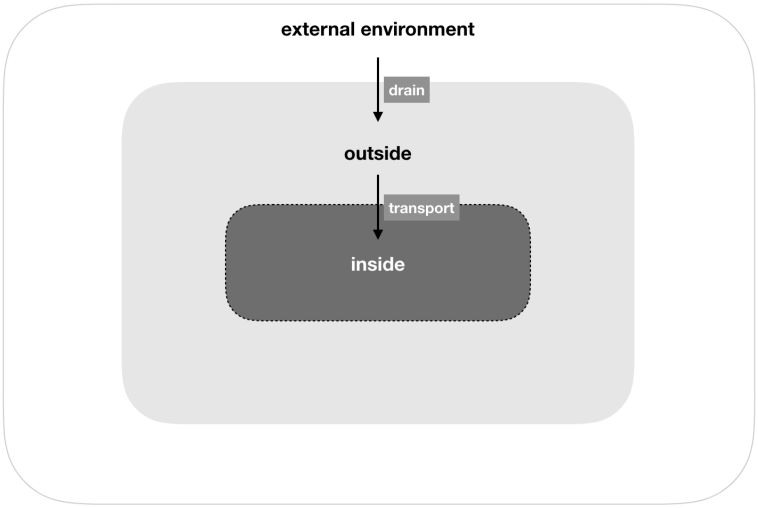


Figure 17.
Schema representing the different compartments and difference between drains (exchange reactions) and transporters.

The drains represent the exchanges between the medium (outside) with the external environment. The transporters represent exchanges between the medium (outside) and the cell (inside)

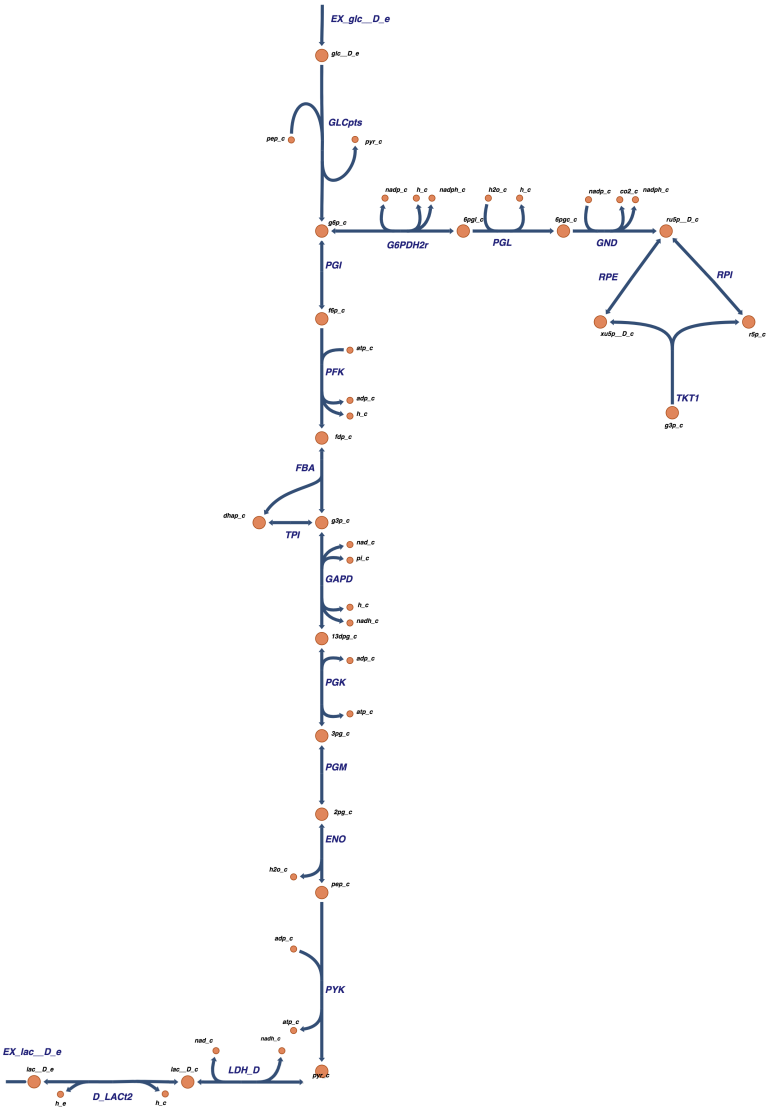


Figure 18.

Central Carbon Metabolism schema

The metabolic map represents the reactions in the model central carbon metabolism from the glucose uptake until the lactate export. Metabolites and reactions id's follow BiGG database nomenclature.

5

SUMMARY AND PROSPECTS FOR FURTHER WORK

Automated genome sequence analysis and annotation has the advantage that the analysis strategy is uniformly applied to all genome sequences against the same database, rendering comparable results. With the speed-up progress in the field of Bioinformatics the (re)construction of the metabolic network on a genome-wide scale has a major role as a support tool for scientific research and developments. It is important to fill the gaps, optimize and standardize the manual stages, trying to create new automated tools and pipelines, and therefore quicken the processes.

Finally, there are several developments that will speed up the reconstruction process and improve its accuracy considerably. This includes efforts to unify nomenclature and to devise physiologically relevant functional classification schemes that enable effective coupling of stored information. This large experimentation stages will allow responding to the bottleneck of rapid data integration, with special attention to GSMM.

In silico approaches proved to be time-saving methods with the multitude of tools provided by *merlin*, COBRApy and OptFlux. Although, when is intended to create a specific and high-quality genome-scale model the manual curation processes are still mandatory. Along the development of the model a manual review of each step revealed that was always space for improvements and corrections, sacrificing the time for a better quality model. During the model reconstruction process innovations were performed such as the alpha-value and threshold choice method. TRIAGE internal database was also extended. An effort of creating cross-references between different knowledge databases was also performed, trying to reduce the hurdle this variations usually bring in science. It was maintained a constant work in synergy with *merlin* developers for future add-ons and bug-fixing. It was also put up to test different softwares which complemented to each other in the attempt of creating the best possible output. All the steps were described as detailed as possible for allowing the work to be reproduced.

As the model was left open, it leaves possibilities of further work in strain improving and optimization. For this particular model, more validation in wet-lab should be performed. A repetition of macromolecule content measuring should be also performed. It is also in perspective the possibility to include an extra compartment, the periplasm to create an even

more accurate and reality approximate model. For the time (and consequently money) saving it will be important to try to look for even more solutions automate the tasks with the minimum sacrifice of accuracy.

Summing up, the intended model reconstruction was accomplished with results believed to be close to reality. All the performed work allowed to create a support model for further studies in *Lb. helveticus* with educational, health, industrial and scientific interest. With hundreds of strains for this specie the existence of a high quality, curated GSMM will ease the task in build future models for those strains and related species.

BIBLIOGRAPHY

[1] A R Joyce and B Ø Palsson. The model organism as a system: Integrating 'omics' data sets. *Nature Reviews Molecular Cell Biology*, 7(3):198–210, 2006.

[2] Adam M. Feist, Markus J Herrgård, Ines Thiele, Jennie L. Reed, and Bernhard Ø Palsson. Reconstruction of biochemical networks in microorganisms. *Nature reviews. Microbiology*, 7(2):129–43, 2009.

[3] Oscar Dias, Miguel Rocha, Eugénio C Ferreira, and Isabel Rocha. Reconstructing genome-scale metabolic models with *merlin*. *Nucleic acids research*, page gkv294, 2015.

[4] Isabel Rocha, Jochen Förster, and Jens Nielsen. Design and Application of Genome-Scale Reconstructed Metabolic Models. *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*, 416:409–431, 2007.

[5] WHN Holzapfel and Brian JB Wood. *The genera of lactic acid bacteria*, volume 2. Springer Science & Business Media, 2012.

[6] E Simelyte, M Rimpiläinen, X Zhang, and P Toivanen. Role of peptidoglycan subtypes in the pathogenesis of bacterial cell wall arthritis. *Annals of the rheumatic diseases*, 62:976–982, 2003.

[7] Wilbert B. Copeland, Bryan A. Bartley, Deepak Chandran, Michal Galdzicki, Kyung H. Kim, Sean C. Sleight, Costas D. Maranas, and Herbert M. Sauro. Computational tools for metabolic engineering. *Metabolic Engineering*, 14(3):270–280, 2012.

[8] X Chen, YQ Song, HY Xu, BLG Menghe, HP Zhang, and ZH Sun. Genetic relationships among *Enterococcus faecalis* isolates from different sources as revealed by multilocus sequence typing. *Journal of dairy science*, 98(8):5183–5193, 2015.

[9] Claude P Champagne, Thomas A Tompkins, Nicole D Buckley, and Julia M Green-Johnson. Effect of fermentation by pure and mixed cultures of *Streptococcus thermophilus* and *Lactobacillus helveticus* on isoflavone and b-vitamin content of a fermented soy beverage. *Food microbiology*, 27(7):968–972, 2010.

[10] JCBN IUPAC-IUBMB. Nomenclature and symbolism for amino acids and peptides. *Pure Appl Chem*, 56(5):595–624, 1983.

[11] Norbert Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*, volume 25. MIT press, 1961.

[12] Walter B Cannon. Physiological regulation of normal states: some tentative postulates concerning biological homeostatics. *Ses Amis, ses Colleges, ses Eleves*, 1926.

[13] Ludwig Von Bertalanffy and John W Sutherland. General systems theory: Foundations, developments, applications. *IEEE Transactions on Systems, Man, and Cybernetics*, 4(6):592–592, 1974.

[14] H Kitano. *Foundations of Systems Biology* . MIT press, 2001.

[15] R D Fleischmann, M D Adams, O White, R A Clayton, E F Kirkness, A R Kerlavage, C J Bult, J F Tomb, B A Dougherty, J M Merrick, K Mckenney, G Sutton, W Fitzhugh, C Fields, J D Gocayne, J Scott, R Shirley, L I Liu, A Glodek, J M Kelley, J F Weidman, C A Phillips, T Spriggs, E Hedblom, M D Cotton, T R Utterback, M C Hanna, D T Nguyen, D M Saudek, R C Brandon, L D Fine, J L Fritchman, J L Fuhrmann, N S M Geoghagen, C L Gnehm, L A Mcdonald, K V Small, C M Fraser, H O Smith, and J C Venter. Whole-Genome Random Sequencing and Assembly of *Haemophilus-Influenzae* Rd. *Science*, 269(5223):496–512, 1995.

[16] Oliver Fiehn, Joachim Kopka, Peter Dörmann, Thomas Altmann, Richard N Trethewey, and Lothar Willmitzer. Metabolite profiling for plant functional genomics. *Nature biotechnology*, 18(11):1157–1161, 2000.

[17] Ute Roessner, Alexander Luedemann, Doreen Brust, Oliver Fiehn, Thomas Linke, Lothar Willmitzer, and Alisdair R Fernie. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *The Plant Cell*, 13(1):11–29, 2001.

[18] Christopher S. Henry, Matthew DeJongh, Aaron A. Best, Paul M. Frybarger, Ben Linsay, and Rick L. Stevens. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology*, 28(9):977–982, 2010.

[19] Isabel Rocha, Jochen Förster, and Jens Nielsen. Design and Application of Genome-Scale Reconstructed Metabolic Models. *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*, 416:409–431, 2007.

[20] H Kitano. Systems biology: A brief overview. *Science (New York, NY)*, 295(5560):1662–1664, 2002.

[21] Markus W. Covert, Christophe H. Schilling, Iman Famili, Jeremy S. Edwards, Igor I. Goryanin, Evgeni Selkov, and Bernhard O. Palsson. Metabolic modeling of microbial strains *in silico*. *Trends in Biochemical Sciences*, 26(3):179–186, 2001.

[22] Frederick Sanger, Gilian M Air, Bart G Barrell, Nigel L Brown, Alan R Coulson, J_ C_ Fiddes, CA Hutchison, Patrick M Slocombe, and Mo Smith. Nucleotide sequence of bacteriophage φx174 dna. *nature*, 265(5596):687–695, 1977.

[23] V. Gopalan and R. Engel-Herbert. Complex oxides: Creative tension in layered crystals. *Nature Materials*, 15(August):928–930, 2016.

[24] Frédéric Delsuc, Henner Brinkmann, and Hervé Philippe. Phylogenomics and the reconstruction of the tree of life. *Nature reviews. Genetics*, 6(5):361–375, 2005.

[25] J C Venter, Myers E W Adams MD Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR,, Kodira C D Zhang Q Zheng XQH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang JH, Miklos GLG, Nelson C, Broder S, Clark AG, Nadeau C, McKusick VA, Zinder, Levine A J N Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S,, Reinert K Mobarry C Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng ZM, Di, Dunn P Francesco V Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge WM, Gong FC, Gu ZP, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke ZX, Ketchum KA, Lai ZW, Lei, Li Z Y YD Li JY, Liang Y, Lin XY, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue BX, Sun, Wang Z Y JT Wang AH, Wang X, Wang J, Wei MH, Wides R, Xiao CL, Yan CH, Yao A, Ye J, Zhan M, Zhang WQ, Zhang HY, Zhao Q, Zheng LS, Zhong F, Zhong WY, Zhu, Zhao S Y SPC Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An HJ, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D,, Center A Carver A Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C,, Hladun S Heiner C Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy, Moy L M Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M,, Strong R Stewart E Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF,, Campbell M J Guigo R Sjolander KV, Karlak B, Kejariwal A, Mi HY, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert, Schwartz R R Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays AD, Dombroski M,, Ely D Donnelly M Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B,

Harris M, Heil J, Henderson S, Hoover J, Jennings, Jordan C D Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu XJ, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell, and Pan S M Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen MY, Wu D, Wu M, Xia A, Zandieh A, Zhu XH. The sequence of the human genome. *Science*, 291(5507):1304, 2001.

[26] Harald Mischak, Rolf Apweiler, Rosamonde E. Banks, Mark Conaway, Joshua Coon, Anna Dominiczak, Jochen H H Ehrich, Danilo Fliser, Mark Girolami, Henning Hermjakob, Denis Hochstrasser, Joachim Jankowski, Bruce A. Julian, Walter Kolch, Ziad A. Massy, Christian Neusuess, Jan Novak, Karlheinz Peter, Kasper Rossing, Joost Schanstra, O. John Semmes, Dan Theodorescu, Visith Thongboonkerd, Eva M. Weissinger, Jennifer E. Van Eyk, and Tadashi Yamamoto. Clinical proteomics: A need to define the field and to begin to set adequate standards. *Proteomics - Clinical Applications*, 1(2):148–156, 2007.

[27] Mike Tyers and Matthias Mann. From genomics to proteomics. *From genomics to proteomics*, 422(March):193–197, 2003.

[28] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, 2009.

[29] Rotem Sorek and Pascale Cossart. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nature reviews. Genetics*, 11(1):9–16, 2010.

[30] Jacob G. Bundy, Matthew P. Davey, and Mark R. Viant. Environmental metabolomics: A critical review and future perspectives. *Metabolomics*, 5(1):3–21, 2009.

[31] Jennifer L. Spratlin, Natalie J. Serkova, and S. Gail Eckhardt. Clinical applications of metabolomics in oncology: A review. *Clinical Cancer Research*, 15(2):431–440, 2009.

[32] Masami Yokota Hirai, Mitsuru Yano, Dayan B Goodenowe, Shigehiko Kanaya, Tomoko Kimura, Motoko Awazuhara, Masanori Arita, Toru Fujiwara, and Kazuki Saito. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America*, 101(27):10205–10, 2004.

[33] David Colquhoun. Too many'omics. *The Scientist*, 19(3):8–9, 2005.

[34] Edward F DeLong. The microbial ocean from genomes to biomes. *Nature*, 459(7244):200–206, 2009.

[35] Norman R Pace. A molecular view of microbial diversity and the biosphere. *Science*, 276(5313):734–740, 1997.

[36] Susannah Green Tringe, Christian Von Mering, Arthur Kobayashi, Asaf A Salamov, Kevin Chen, Hwai W Chang, Mircea Podar, Jay M Short, Eric J Mathur, John C Detter, et al. Comparative metagenomics of microbial communities. *Science*, 308(5721):554–557, 2005.

[37] Douglas B Rusch, Aaron L Halpern, Granger Sutton, Karla B Heidelberg, Shannon Williamson, Shibu Yooseph, Dongying Wu, Jonathan A Eisen, Jeff M Hoffman, Karin Remington, et al. The sorcerer ii global ocean sampling expedition: northwest atlantic through eastern tropical pacific. *PLoS Biol*, 5(3):e77, 2007.

[38] Falk Warnecke, Peter Luginbühl, Natalia Ivanova, Majid Ghassemian, Toby H Richardson, Justin T Stege, Michelle Cayouette, Alice C McHardy, Gordana Djordjevic, Nahla Aboushadi, et al. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*, 450(7169):560–565, 2007.

[39] Jens Georg, Björn Voß, Ingeborg Scholz, Jan Mitschke, Annegret Wilde, and Wolfgang R Hess. Evidence for a major role of antisense rnas in cyanobacterial gene regulation. *Molecular Systems Biology*, 5(1):305, 2009.

[40] Jane M Liu, Jonathan Livny, Michael S Lawrence, Marc D Kimball, Matthew K Waldor, and Andrew Camilli. Experimental discovery of srnas in *Vibrio cholerae* by direct cloning, 5s/trna depletion and parallel sequencing. *Nucleic acids research*, 37(6):e46–e46, 2009.

[41] David Eisenberg, Edward M Marcotte, Ioannis Xenarios, and Todd O Yeates. Protein function in the post-genomic era. *Nature*, 405(6788):823–826, 2000.

[42] Daniel Machado and Markus Herrgård. Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. *PLoS Computational Biology*, 10(4), 2014.

[43] Filipe Santos, Joost Boele, and Bas Teusink. A practical guide to genome-scale metabolic models and their analysis. *Methods enzymol*, 500:509–532, 2011.

[44] Oscar Dias and Isabel Rocha. Systems biology in fungi. In *Molecular Biology of Food and Water Borne Mycotoxigenic and Mycotic Fungi*, pages 69–92. CRC Press, 2015.

[45] Joost Boele, Brett G Olivier, and Bas Teusink. Fame, the flux analysis and modeling environment. *BMC systems biology*, 6(1):8, 2012.

[46] Stephan Pabinger, Robert Rader, Rasmus Agren, Jens Nielsen, and Zlatko Trajanoski. Memosys: Bioinformatics platform for genome-scale metabolic models. *BMC systems biology*, 5(1):20, 2011.

[47] Xueyang Feng, You Xu, Yixin Chen, and Yinjie J Tang. Microbesflux: a web platform for drafting metabolic models from the kegg database. *BMC systems biology*, 6(1):94, 2012.

[48] Peter D Karp, Suzanne Paley, and Pedro Romero. The pathway tools software. *Bioinformatics*, 18(suppl 1):S225–S232, 2002.

[49] Esa Pitkänen, Paula Jouhten, Jian Hou, Muhammad Fahad Syed, Peter Blomberg, Jana Kludas, Merja Oja, Liisa Holm, Merja Penttilä, Juho Rousu, et al. Comparative genome-scale reconstruction of gapless metabolic networks for present and ancestral species. *PLoS Comput Biol*, 10(2):e1003465, 2014.

[50] Matthew DeJongh, Kevin Formsma, Paul Boillot, John Gould, Matthew Rycenga, and Aaron Best. Toward the automated generation of genome-scale metabolic networks in the seed. *BMC bioinformatics*, 8(1):139, 2007.

[51] Rasmus Agren, Liming Liu, Saeed Shoaie, Wanwipa Vongsangnak, Intawat Nookaew, and Jens Nielsen. The raven toolbox and its use for generating a genome-scale metabolic model for penicillium chrysogenum. *PLoS Comput Biol*, 9(3):e1002980, 2013.

[52] Michael Hucka, Andrew Finney, Herbert M Sauro, Hamid Bolouri, John C Doyle, Hiroaki Kitano, Adam P Arkin, Benjamin J Bornstein, Dennis Bray, Athel Cornish-Bowden, et al. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[53] Oscar Dias, Daniel Gomes, Paulo Vilaca, Joao Cardoso, Miguel Rocha, Eugenio Ferreira, and Isabel Rocha. Genome-wide semi-automated annotation of transporter systems. *IEEE/ACM transactions on computational biology and bioinformatics/IEEE, ACM*, 2016.

[54] Tae Yong Kim, Seung Bum Sohn, Hyun Uk Kim, and Sang Yup Lee. Strategies for systems-level metabolic engineering. *Biotechnology Journal*, 3(5):612–623, 2008.

[55] Machado D., Costa R.S., Rocha M., Ferreira E.C., Tidor B., and Rocha I. Modeling formalisms in systems biology. *AMB Express*, 1(1):1–14, 2011.

[56] Jeffrey Orth, Ines Thiele, and Bernhard. What is flux balance analysis? *Nature biotechnology*, 28(3):245–248, 2010.

[57] Adam M. Feist, Markus J Herrgård, Ines Thiele, Jennie L. Reed, and Bernhard Ø Palsson. Reconstruction of biochemical networks in microorganisms. *Nature reviews. Microbiology*, 7(2):129–43, 2009.

[58] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[59] McEntyre Jo and J Ostell. The ncbi handbook national library of medicine (us). *National Center for Biotechnology Information. Internet: Bethesda (MD) National Center for Biotechnology Information US*, 2002.

[60] UniProt Consortium et al. Ongoing and future developments at the universal protein resource. *Nucleic acids research*, 39(suppl 1):D214–D219, 2011.

[61] Alice R Wattam, David Abraham, Oral Dalay, Terry L Disz, Timothy Driscoll, Joseph L Gabbard, Joseph J Gillespie, Roger Gough, Deborah Hix, Ronald Kenyon, et al. Patric, the bacterial bioinformatics database and analysis resource. *Nucleic acids research*, page gkt1099, 2013.

[62] Henrik Nordberg, Michael Cantor, Serge Dusheyko, Susan Hua, Alexander Poliakov, Igor Shabalov, Tatyana Smirnova, Igor V Grigoriev, and Inna Dubchak. The genome portal of the department of energy joint genome institute: 2014 updates. *Nucleic acids research*, 42(D1):D26–D31, 2014.

[63] Nikos C Kyrpides. Genomes online database (gold 1.0): a monitor of complete and ongoing genome projects world-wide. *Bioinformatics*, 15(9):773–774, 1999.

[64] Lukas Käll, Anders Krogh, and E. L L Sonnhammer. A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*, 338(5):1027–1036, 2004.

[65] Zachary A. King, Justin Lu, Andreas Dr??ger, Philip Miller, Stephen Federowicz, Joshua A. Lerman, Ali Ebrahim, Bernhard O. Palsson, and Nathan E. Lewis. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research*, 44(D1):D515–D522, 2016.

[66] Thomas Bernard, Alan Bridge, Anne Morgat, Sébastien Moretti, Ioannis Xenarios, and Marco Pagni. Reconciliation of metabolites and biochemical reactions for metabolic networks. *Briefings in bioinformatics*, 15(1):123–135, 2012.

[67] Mathias Ganter, Thomas Bernard, Sébastien Moretti, Joerg Stelling, and Marco Pagni. Metanetx. org: a website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics*, 29(6):815–816, 2013.

[68] Sébastien Moretti, Olivier Martin, T Van Du Tran, Alan Bridge, Anne Morgat, and Marco Pagni. Metanetx/mnxref–reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic acids research*, 44(D1):D523–D526, 2016.

[69] Emily J Richardson and Mick Watson. The automatic annotation of bacterial genomes. *Briefings in bioinformatics*, 14(1):1–12, 2013.

[70] E Webb. International union of biochemistry and molecular biology: Enzyme nomenclature 1992. recommendations of the nomenclature committee of the international union of biochemistry and molecular biology on the nomenclature and classification of enzymes, 1992.

[71] Milton H Saier. A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiology and Molecular Biology Reviews*, 64(2):354–411, 2000.

[72] RA Majewski and MM Domach. Simple constrained-optimization view of acetate overflow in e. coli. *Biotechnology and bioengineering*, 35(7):732–738, 1990.

[73] Amit Varma, Brian W Boesch, and Bernhard O Palsson. Biochemical production capabilities of *Escherichia coli*. *Biotechnology and bioengineering*, 42(1):59–73, 1993.

[74] Amit Varma and Bernhard O Palsson. Metabolic capabilities of *Escherichia coli*: I. synthesis of biosynthetic precursors and cofactors. *Journal of theoretical biology*, 165(4):477–502, 1993.

[75] J Pramanik and JD Keasling. Stoichiometric model of *Escherichia coli* metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnology and bioengineering*, 56(4):398–421, 1997.

[76] J Pramanik and JD Keasling. Effect of *Escherichia coli* biomass composition on central metabolic fluxes predicted by a stoichiometric model. *Biotechnology and bioengineering*, 60(2):230–238, 1998.

[77] Jennifer L Reed and Bernhard Ø Palsson. Thirteen years of building constraint-based *in silico* models of *Escherichia coli*. *Journal of bacteriology*, 185(9):2692–2699, 2003.

[78] Sang Yup Lee, Han Min Woo, Dong-Yup Lee, Hyung Seok Choi, Tae Yong Kim, and Hongseok Yun. Systems-level analysis of genome-scale *in silico* metabolic models using metafluxnet. *Biotechnology and Bioprocess Engineering*, 10(5):425, 2005.

[79] Adam M Feist, Christopher S Henry, Jennifer L Reed, Markus Krummenacker, Andrew R Joyce, Peter D Karp, Linda J Broadbelt, Vassily Hatzimanikatis, and Bernhard Ø Palsson. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular systems biology*, 3(121):121, 2007.

[80] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[81] Oscar Dias, Miguel Rocha, Eugenio C Ferreira, and Isabel Rocha. *merlin*: Metabolic models reconstruction using genome-scale information. *IFAC Proceedings Volumes*, 43(6):120–125, 2010.

[82] L. Zaror. *Mucormycosis. Food and Water Borne Mycotoxigenic and Mycotic Fungi*. CRC Press, 2015.

[83] Christof Francke, Roland J. Siezen, and Bas Teusink. Reconstructing the metabolic network of a bacterium from its genome. *Trends in Microbiology*, 13(11):550–558, 2005.

[84] G J F Smolders, J van der Meij, Mark C M van Loosdrecht, and Joseph J Heijnen. A structured metabolic model for the anaerobic and aerobic stoichiometry of the biological phosphorus removal process. Biotechnology and Bioengineering. *Biotechnology and Bioengineering*, 47(3):277–287, 1995.

[85] Lars Kuepfer. Stoichiometric modelling of microbial metabolism. *Methods in Molecular Biology*, 1191(1):3–18, 2014.

[86] F Walberg. Metabolic compartmentation of glutamate and glutamine: morphological evidence obtained by quantitative immunocytochemistry in rat cerebellum. *Neuroscience*, 46(3):519–534, 1992.

[87] Jochen Förster, Iman Famili, Patrick Fu, Bernhard Ø Palsson, and Jens Nielsen. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome research*, 13(2):244–253, 2003.

[88] Markus J Herrgård, Neil Swainston, Paul Dobson, Warwick B Dunn, K Yalçin Arga, Mikko Arvas, Nils Blüthgen, Simon Borger, Roeland Costenoble, Matthias Heinemann, et al. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature biotechnology*, 26(10):1155–1160, 2008.

[89] Natalie C Duarte, Markus J Herrgård, and Bernhard Ø Palsson. Reconstruction and validation of *Saccharomyces cerevisiae* ind750, a fully compartmentalized genome-scale metabolic model. *Genome research*, 14(7):1298–1309, 2004.

[90] Ioannis Iliopoulos, Sophia Tsoka, Miguel A Andrade, Paul Janssen, Benjamin Audit, Anna Tramontano, Alfonso Valencia, Christophe Leroy, Chris Sander, and Christos A Ouzounis. Open letter Genome sequences and great expectations. *Genome Biology*, 2(1):1–3, 2000.

[91] D Devos and a Valencia. Practical limits of functional prediction. *Proteins*, 41(February):98–107, 2000.

[92] Damien Devos and Alfonso Valencia. Intrinsic errors in genome annotation. *Trends in Genetics*, 17(8):429–431, 2001.

[93] Steven E. Brenner. Errors in genome annotation. *Trends in Genetics*, 15(4):132–133, 1999.

[94] P Bork, T Dandekar, Y Diaz-Lazcoz, F Eisenhaber, M Huynen, and Y Yuan. Predicting function: from genes to genomes and back. *J Mol Biol*, 283(4):707–725, 1998.

[95] John a Gerlt and Patricia C Babbitt. Divergent Evolution of Enzymatic Function: Mechanistically Diverse Superfamilies and Functionally Distinct Suprafamilies. *Annual Review of Biochemistry*, 70(1):209–246, 2001.

[96] S. M J Klaus, Arno Wegkamp, Wilbert Sybesma, Jeroen Hugenholtz, Jesse F. Gregory, and Andrew D. Hanson. A nudix enzyme removes pyrophosphate from dihydroneopterin triphosphate in the folate synthesis pathway of bacteria and plants. *Journal of Biological Chemistry*, 280(7):5274–5280, 2005.

[97] Martijn A. Huynen, Berend Snel, Christian Von Mering, and Peer Bork. Function prediction and protein networks. *Current Opinion in Cell Biology*, 15(2):191–198, 2003.

[98] Florencio Pazos and Alfonso Valencia. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein engineering*, 14(9):609–614, 2001.

[99] Itai Yanai, Adnan Derti, and Charles DeLisi. Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proceedings of the National Academy of Sciences*, 98(14):7940–7945, 2001.

[100] Thomas Dandekar, Berend Snel, Martijn Huynen, and Peer Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in biochemical sciences*, 23(9):324–328, 1998.

[101] Matteo Pellegrini, Edward M Marcotte, Michael J Thompson, David Eisenberg, and Todd O Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8):4285–4288, 1999.

[102] Martha L Bulyk, Abigail M McGuire, Nobuhisa Masuda, and George M Church. A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in *Escherichia coli*. *Genome research*, 14(2):201–208, 2004.

[103] Nathan D Price, Jennifer L Reed, and Bernhard Ø Palsson. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology*, 2(11):886–897, 2004.

[104] Christof Francke, Roland J Siezen, and Bas Teusink. Reconstructing the metabolic network of a bacterium from its genome. *Trends in microbiology*, 13(11):550–558, 2005.

[105] Nathan D Price, Jennifer L Reed, and Bernhard Ø Palsson. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology*, 2(11):886–897, 2004.

[106] Nathan E Lewis, Kim K Hixson, Tom M Conrad, Joshua A Lerman, Pep Charusanti, Ashoka D Polpitiya, Joshua N Adkins, Gunnar Schramm, Samuel O Purvine, Daniel Lopez-Ferrer, et al. Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Molecular systems biology*, 6(1):390, 2010.

[107] Frank J Carr, Don Chill, and Nino Maida. The lactic acid bacteria: a literature survey. *Critical reviews in microbiology*, 28(4):281–370, 2002.

[108] Gerald W Tannock. A Special Fondness for *Lactobacilli*. *Applied and environmental microbiology*, 70(6):3189–3194, 2004.

[109] Hidenori Hayashi, Rei Takahashi, Takahiro Nishi, Mitsuo Sakamoto, and Yoshimi Benno. Molecular analysis of jejunal, ileal, caecal and recto-sigmoidal human colonic microbiota using 16s rrna gene libraries and terminal restriction fragment length polymorphism. *Journal of medical microbiology*, 54(11):1093–1101, 2005.

[110] Hans GHJ Heilig, Erwin G Zoetendal, Elaine E Vaughan, Philippe Marteau, Antoon DL Akkermans, and Willem M de Vos. Molecular diversity of lactobacillus spp. and other lactic acid bacteria in the human intestine as determined by specific amplification of 16s ribosomal dna. *Applied and environmental microbiology*, 68(1):114–123, 2002.

[111] K Makarova, A Slesarev, Y Wolf, A Sorokin, B Mirkin, E Koonin, A Pavlov, N Pavlova, V Karamychev, N Polouchine, V Shakhova, I Grigoriev, Y Lou, D Rohksar, S Lucas, K Huang, D M Goodstein, T Hawkins, V Plengvidhya, D Welker, J Hughes, Y Goh, A Benson, K Baldwin, J.-H Lee, I Díaz-Muñ Iz, B Dosti, V Smeianov, W Wechter, R Barabote, G Lorca, E Altermann, R Barrangou, B Ganesan, Y Xie, H Rawsthorne, D Tamir, C Parker, F Breidt, J Broadbent, R Hutkins, D O 'sullivan, J Steele, G Unlu, M Saier, T Klaenhammer, P Richardson, S Kozyavkin, B Weimer, and D Mills. Comparative genomics of the lactic acid bacteria. *Proceedings of the National Academy of Sciences*, 103(42):15611–15616, 2006.

[112] Michael Teuber. Lactic acid bacteria. *Biotechnology Set, Second Edition*, pages 325–366, 1993.

[113] Miriam Land, Loren Hauser, Se-Ran Jun, Intawat Nookaew, Michael R Leuze, Tae-Hyuk Ahn, Tatiana Karpinets, Ole Lund, Guruprased Kora, Trudy Wassenaar, Suresh

Poudel, David W Ussery, D W Ussery, L Hauser, M R Leuze, T <h Ahn, G Kora, O Lund, T Wassenaar, and S Poudel. Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics*, 15:141–161, 2015.

[114] Sigurd Orla-Jensen. *The lactic acid bacteria*, volume 2. Ejnar Munksgaard, 1942.

[115] Jack London. The ecology and taxonomic status of the *lactobacilli*. *Annual Reviews in Microbiology*, 30(1):279–301, 1976.

[116] Michael Callanan, Pawel Kaleta, John O'Callaghan, Orla O'Sullivan, Kieran Jordan, Olivia McAuliffe, Amaia Sangrador-Vegas, Lydia Slattery, Gerald F. Fitzgerald, Tom Beresford, and R. Paul Ross. Genome sequence of Lactobacillus helveticus, an organism distinguished by selective gene loss and insertion sequence element expansion. *Journal of Bacteriology*, 190(2):727–735, 2008.

[117] Max Firtel, Grant Henderson, and Igor Sokolov. Nanosurgery: Observation of peptidoglycan strands in *Lactobacillus helveticus* cell walls. *Ultramicroscopy*, 101(2-4):105–109, 2004.

[118] Harald Labischinski, G Barnickel, D. Naumann, and P. Keller. Conformational and topological aspects of the three-dimensional architecture of bacterial peptidoglycan. *Annales de l'Institut Pasteur. Microbiology*, 136A(1):45–50, 1985.

[119] K H Schleifer and O Kandler. Peptidoglycan types of bacterial cell walls and their taxonomic implications. *Bacteriological Reviews*, 36(4):407–477, 1972.

[120] Jean Delcour, Thierry Ferain, Marie Deghorain, Emmanuelle Palumbo, and Pascal Hols. The biosynthesis and functionality of the cell-wall of lactic acid bacteria. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, 76(1-4):159–184, 1999.

[121] Evgeny Vinogradov, Florence Valence, Emmanuel Maes, Iva Jebava, Victoria Chuat, Sylvie Lortal, Thierry Grard, Yann Guerardel, and Irina Sadovskaya. Structural studies of the cell wall polysaccharides from three strains of *Lactobacillus helveticus* with different autolytic properties: DPC4571, BROI, and LH1. *Carbohydrate Research*, 379:7–12, 2013.

[122] Andrew Laws, Yucheng Gu, and Valerie Marshall. Biosynthesis, characterisation, and design of bacterial exopolysaccharides from lactic acid bacteria. *Biotechnology advances*, 19(8):597–625, 2001.

[123] Philippe Duboc and Beat Mollet. Applications of exopolysaccharides in the dairy industry. *International Dairy Journal*, 11(9):759–768, 2001.

[124] T Morishita, Y Deguchi, M Yajima, and T Yura. Multiple nutritional requirements of lactobacilli : genetic lesions affecting amino acid biosynthetic Multiple Nutritional Requirements of *Lactobacilli* : Genetic Lesions Affecting Amino Acid Biosynthetic Pathways. *Journal of Bacteriology*, 148(1):64–71, 1981.

[125] Jason K. Christiansen, Joanne E. Hughes, Dennis L. Welker, Beatriz T. Rodríguez, James L. Steele, and Jeff R. Broadbent. Phenotypic and genotypic analysis of amino acid auxotrophy in *Lactobacillus helveticus* CNRZ 32. *Applied and Environmental Microbiology*, 74(2):416–423, 2008.

[126] H Nakajima. Amino acid transport in *Lactobacillus helveticus*. *FEMS Microbiology Letters*, 158(November 1997):249–253, 1998.

[127] B Poolman. Energy transduction in lactic acid bacteria. *FEMS microbiology reviews*, 12(1-3):125–147, 1993.

[128] H Nakajima, a Hagting, E R Kunji, B Poolman, and W N Konings. Cloning and functional expression in *Escherichia coli* of the gene encoding the di- and tripeptide transport protein of Lactobacillus helveticus. *Applied and environmental microbiology*, 63(6):2213–2217, 1997.

[129] Kirsi Savijoki and Airi Palva. Purification and molecular characterization of a tripeptidase (pept) from lactobacillus helveticus. *Applied and environmental microbiology*, 66(2):794–800, 2000.

[130] Herbert J Strobel, James B Russell, AJ Driessen, and Wil N Konings. Transport of amino acids in lactobacillus casei by proton-motive-force-dependent and non-proton-motive-force-dependent mechanisms. *Journal of bacteriology*, 171(1):280–284, 1989.

[131] Berend Tolner, Trees Ubbink-Kok, Bert Poolman, and Wil N Konings. Characterization of the proton/glutamate symport protein of *Bacillus subtilis* and its functional expression in *Escherichia coli*. *Journal of bacteriology*, 177(10):2863–2869, 1995.

[132] BEAT Mollet and NATHALIE Pilloud. Galactose utilization in *Lactobacillus helveticus*: isolation and characterization of the galactokinase (galk) and galactose-1-phosphate uridyl transferase (galt) genes. *Journal of bacteriology*, 173(14):4464–4473, 1991.

[133] M. W. Hickey, A. J. Hillier, and G. R. Jago. Transport and metabolism of lactose, glucose, and galactose in homofermentative *lactobacilli*. *Applied and Environmental Microbiology*, 51(4):825–831, 1986.

[134] B. Grossiord, E.E. E Vaughan, Evert Luesink, and W.M. M de Vos. Genetics of galactose utilisation via the Leloir pathway in lactic acid bacteria. *Le Lait*, 78(1):77–84, 1998.

[135] Michael G Gänzle and Rainer Follador. Metabolism of oligosaccharides and starch in *lactobacilli*: a review. *Frontiers in microbiology*, 3, 2012.

[136] Maria Grazia Fortina, Giovanni Ricci, Diego Mora, Simone Guglielmetti, and Pier Luigi Manachini. Unusual organization for lactose and galactose gene clusters in *Lactobacillus helveticus*. *Applied and environmental microbiology*, 69(6):3238–3243, 2003.

[137] GC O'Sullivan, P Kelly, S O'Halloran, C Collins, JK Collins, C Dunne, and F Shanahan. Probiotics: an emerging therapy. *Current pharmaceutical design*, 11(1):3–10, 2005.

[138] Luigi Chiarini, Luisa Mara, Silvia Tabacchioni, Biotechnology Project, C R E Casaccia, Via Anguillarese, and S Maria Galeria. Influence of growth supplements on lactic acid production in whey ultrafiltrate by *Lactobacillus helveticus*. *Applied Microbiology Biotechnology*, pages 461–464, 1992.

[139] Abdeltif Amrane and Yves Prigent. Differentiation of pH and free lactic acid effects on the various growth and production phases of *Lactobacillus helveticus*. *Journal of Chemical Technology and Biotechnology*, 40(September 1998):33–40, 1999.

[140] Abdeltif Amrane and Yves Prigent. A novel concept of bioreactor: Specialized function two-stage continuous reactor, and its application to lactose conversion into lactic acid. *Journal of biotechnology*, 45:195–203, 1996.

[141] U Kulozik. Physiological aspects of continuous lactic acid fermentations at high dilution rates. *Applied microbiology and biotechnology*, pages 506–510, 1998.

[142] J Ø I Storr and D W Levine. Uptake of lactose and continuous lactic acid fermentation by entrapped non-growing *Lactobacillus helveticus* in whey permeate. *Applied microbiology and biotechnology*, 249:240–249, 1996.

[143] Kari Kylä-Nikkilä, Mervi Hujanen, Matti Leisola, and Airi Palva. Metabolic Engineering of *Lactobacillus helveticus* CNRZ32 for Production of Purel-(+)-Lactic Acid. *Applied and Environmental Microbiology*, 66(9):3835–3841, 2000.

[144] L. Fernandez, T. Bhowmik, and J. L. Steele. Characterization of the Lactobacillus helveticus CNRZ32 pepC gene. *Applied and Environmental Microbiology*, 60(1):333–336, 1994.

[145] Monica Gatti, Carlo Trivisano, Enrico Fabrizi, Erasmo Neviani, Fausto Gardini, Istituto Sperimentale, Lattiero Caseario, and Scienze Statistiche. Biodiversity among *Lactobacillus helveticus* Strains Isolated from Different Natural Whey Starter Cultures as Revealed by Classification Trees. *Applied and environmental microbiology*, 70(1):182–190, 2004.

[146] M Gatti, C Lazzi, L Rossetti, G Mucchetti, and E Neviani. Biodiversity in *Lactobacillus helveticus* strains present in natural whey starter used for Parmigiano Reggiano cheese. *Journal of Applied Microbiology*, pages 463–470, 2003.

[147] HJ Bartels, ME Johnson, and NF Olson. Accelerated ripening of gouda cheese. 1. effect of heat-shocked thermophilic *lactobacilli* and *streptococci* on proteolysis and flavor development. *Milchwissenschaft*, 42(2):83–88, 1987.

[148] R David Pridmore, Bernard Berger, Frank Desiere, David Vilanova, Caroline Barretto, Anne-Cecile Pittet, Marie-Camille Zwahlen, Martine Rouvet, Eric Altermann, Rodolphe Barrangou, et al. The genome sequence of the probiotic intestinal bacterium *Lactobacillus johnsonii* ncc 533. *Proceedings of the National Academy of Sciences of the United States of America*, 101(8):2512–2517, 2004.

[149] Eric Altermann, W Michael Russell, M Andrea Azcarate-Peril, Rodolphe Barrangou, B Logan Buck, Olivia McAuliffe, Nicole Souther, Alleson Dobson, Tri Duong, Michael Callanan, et al. Complete genome sequence of the probiotic lactic acid bacterium lactobacillus acidophilus ncfm. *Proceedings of the National Academy of Sciences of the United States of America*, 102(11):3906–3912, 2005.

[150] Stéphane Chaillou, Marie-Christine Champomier-Vergès, Monique Cornet, Anne-Marie Crutz-Le Coq, Anne-Marie Dudez, Véronique Martin, Sophie Beaufils, Emmanuelle Darbon-Rongère, Robert Bossy, Valentin Loux, et al. The complete genome sequence of the meat-borne lactic acid bacterium lactobacillus sakei 23k. *Nature biotechnology*, 23(12):1527–1533, 2005.

[151] Marcus J Claesson, Yin Li, Sinead Leahy, Carlos Canchaya, Jan Peter van Pijkeren, Ana M Cerdeño-Tárraga, Julian Parkhill, Sarah Flynn, Gerald C O'Sullivan, J Kevin Collins, et al. Multireplicon genome architecture of *Lactobacillus salivarius*. *Proceedings of the National Academy of Sciences*, 103(17):6718–6723, 2006.

[152] Michiel Kleerebezem, Jos Boekhorst, Richard van Kranenburg, Douwe Molenaar, Oscar P Kuipers, Rob Leer, Renato Tarchini, Sander A Peters, Hans M Sandbrink, Mark WEJ Fiers, et al. Complete genome sequence of *Lactobacillus plantarum* wcfs1. *Proceedings of the National Academy of Sciences*, 100(4):1990–1995, 2003.

[153] M Van de Guchte, S Penaud, C Grimaldi, V Barbe, K Bryson, P Nicolas, C Robert, S Oztas, S Mangenot, A Couloux, et al. The complete genome sequence of *Lactobacillus bulgaricus* reveals extensive and ongoing reductive evolution. *Proceedings of the National Academy of Sciences*, 103(24):9274–9279, 2006.

[154] Oscar Dias, Andreas K Gombert, Eugénio C Ferreira, and Isabel Rocha. Genome-wide metabolic (re-) annotation of *Kluyveromyces lactis*. *BMC genomics*, 13(1):1, 2012.

[155] Hongwu Ma and An-Ping Zeng. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2):270–277, 2003.

[156] Michael Stelzer, Jibin Sun, Tom Kamphans, Sándor P Fekete, and An-Ping Zeng. An extended bioreaction database that significantly improves reconstruction and analysis of genome-scale metabolic networks. *Integrative Biology*, 3(11):1071–1086, 2011.

[157] SJ Pirt. The maintenance energy of bacteria in growing cultures. *Proceedings of the Royal Society of London B: Biological Sciences*, 163(991):224–231, 1965.

[158] D Herbert, PJ Phipps, and RE Strange. Chapter iii chemical analysis of microbial cells. *Methods in microbiology*, 5:209–344, 1971.

[159] S Benthin, J Nielsen, and J Villadsen. A simple and reliable method for the determination of cellular rna content. *Biotechnology Techniques*, 5(1):39–42, 1991.

[160] Nathan E Lewis, Harish Nagarajan, and Bernhard O Palsson. Constraining the metabolic genotype-phenotype relationship using a phylogeny of *in silico* methods. *Nature reviews. Microbiology*, 10(4):291–305, 2012.

[161] Ali Ebrahim, Joshua A Lerman, Bernhard O Palsson, and Daniel R Hyduke. Cobrapy: Constraints-based reconstruction and analysis for python. *BMC systems biology*, 7(1):74, 2013.

[162] Isabel Rocha, Paulo Maia, Pedro Evangelista, Paulo Vilaça, Simão Soares, José P Pinto, Jens Nielsen, Kiran R Patil, Eugénio C Ferreira, and Miguel Rocha. Optflux: an open-source software platform for in silico metabolic engineering. *BMC systems biology*, 4(1):45, 2010.

[163] ID Cavalcanti-Montano, CAG Suarez, R Sousa Jr, RC Giordano, EC Ferreira, TC Zangirolami, and I Rocha. Análise dos fluxos metabólicos em *Saccharomyces cerevisiae* a partir de d-xilulose como fonte de carbono utilizando optflux. *SINAFERM 2013-Anais do XIX Simpósio Nacional de Bioprocessos*, (SP 01-05):1–5, 2013.

[164] Bashir Sajo Mienda, Mohd Shahir Shamsir, and Faezah Mohd Salleh. *In silico* metabolic engineering prediction of escherichia coli genome model for production of d-lactic acid from glycerol using the optflux software platform. *International Journal of Computational Bioinformatics and In Silico Modeling*, 3(4):460–5, 2014.

[165] Sophia Santos and Isabel Rocha. Estimation of biomass composition from genomic and transcriptomic information. *Journal of Integrative Bioinformatics,*, 13(2):1–14, 2016.

[166] Ana Paula Oliveira, Jens Nielsen, and Jochen Förster. Modeling *Lactococcus lactis* using a genome-scale flux model. *BMC microbiology*, 5(1):39, 2005.

[167] You Kwan Oh, Bernhard O. Palsson, Sung M. Park, Christophe H. Schilling, and Radhakrishnan Mahadevan. Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data. *Journal of Biological Chemistry*, 282(39):28791–28799, 2007.

[168] F. A. Exterkate, B. J. Otten, H. W. Wassenberg, and J. H. Veerkamp. Comparison of the phospholipid composition of *Bifidobacterium* and *Lactobacillus* strains. *Journal of Bacteriology*, 106(3):824–829, 1971.

[169] Ginka I Frengova, Emilina D Simova, Dora M Beshkova, and Zhelyasko I Simov. Exopolysaccharides produced by lactic acid bacteria of kefir grains. *Zeitschrift für Naturforschung C*, 57(9-10):805–810, 2002.

[170] MI Torino, F Mozzi, and G Font De Valdez. Exopolysaccharide biosynthesis by *Lactobacillus helveticus* atcc 15807. *Applied microbiology and biotechnology*, 68(2):259–265, 2005.

[171] Gerard W Robijn, Dick JC van den Berg, Han Haas, Johannis P Kamerling, and Johannes FG Vliegenthart. Determination of the structure of the exopolysaccharide produced by *Lactobacillus sake*. *Carbohydrate research*, 276(1):117/–136, 1995.

[172] Laurent Aussel, Fabien Pierrel, Laurent Loiseau, Murielle Lombard, Marc Fontecave, and Frédéric Barras. Biosynthesis and physiology of coenzyme Q in bacteria. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1837(7):1004–1011, 2014.

[173] Lluis Masip, Karthik Veeravalli, and George Georgiou. The many faces of glutathione in bacteria. *Antioxidants & redox signaling*, 8(5-6):753–762, 2006.

A

SUPPORT MATERIAL

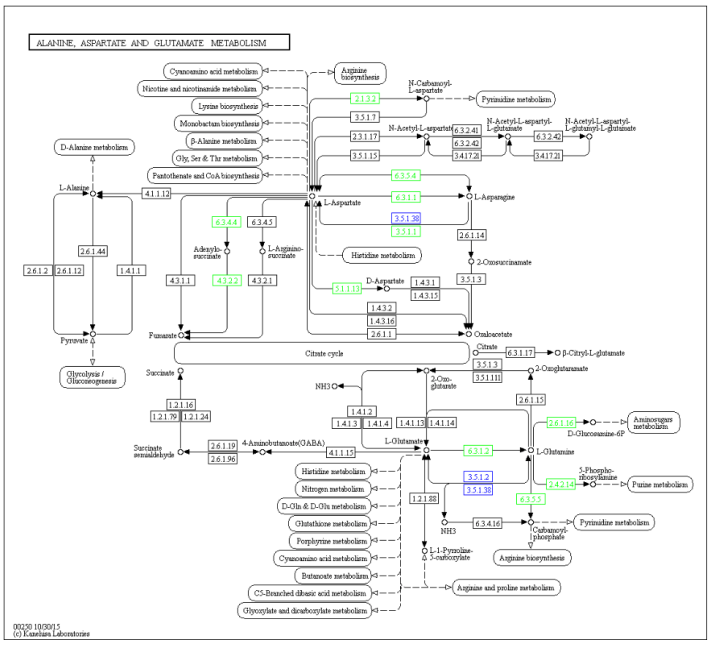


Figure 19.
Screenshot of alanine aspartate and glutamate metabolism

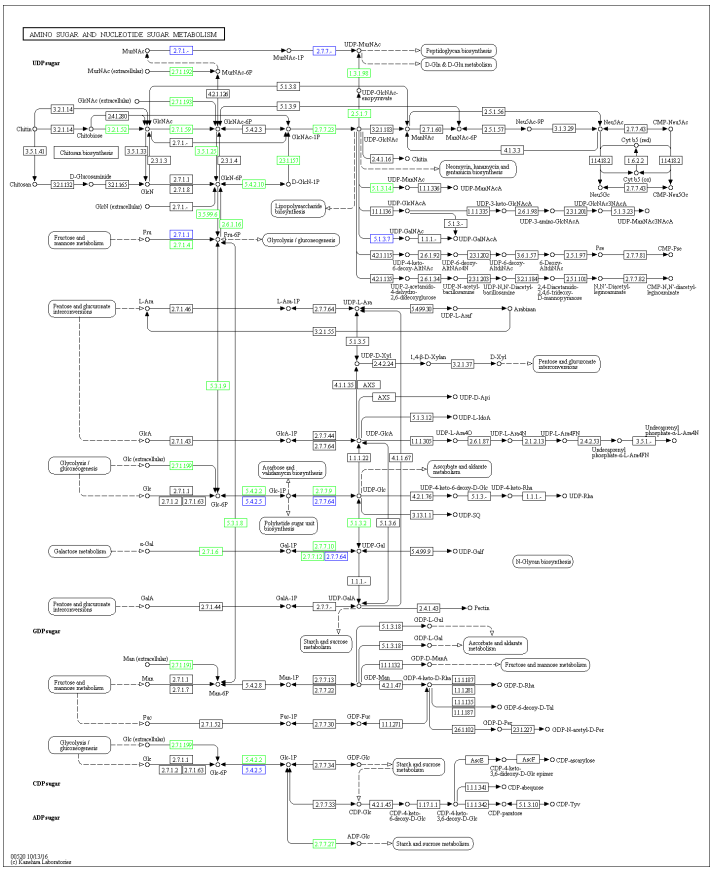


Figure 20.
Screenshot of Amino Sugar and Nucleotide Metabolism

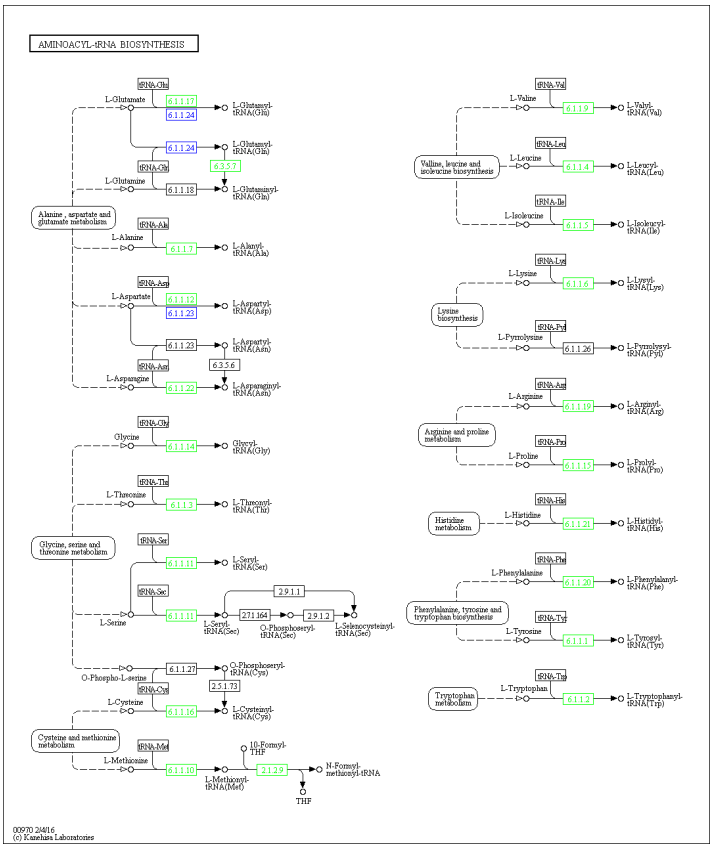


Figure 21.

Screenshot of Aminoacyl-tRNA biosynthesis.

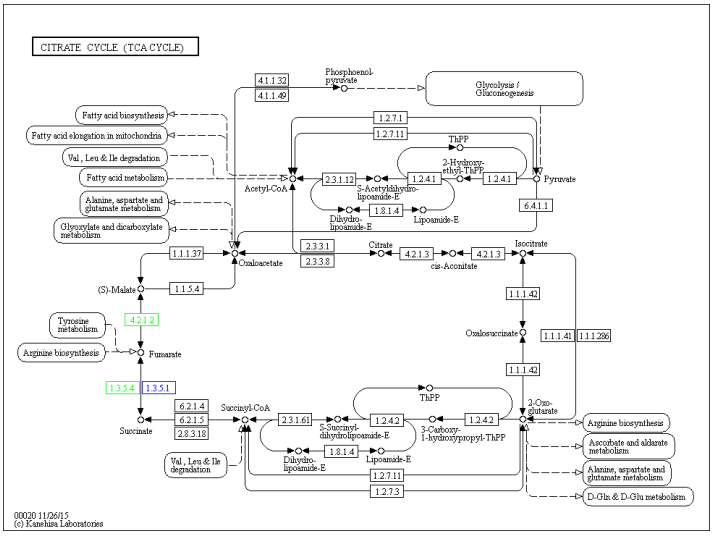


Figure 22.

Screenshot of TCA cycle

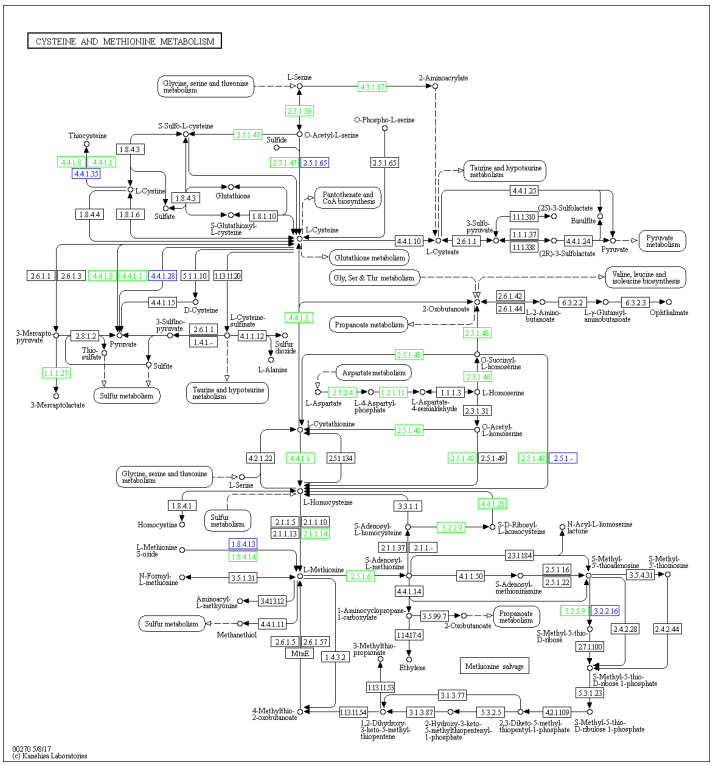


Figure 23.
Screenshot of Cysteine and methionine Metabolism

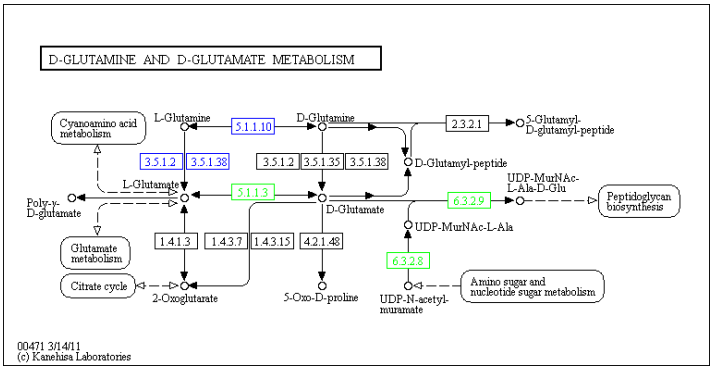


Figure 24.
Screenshot of Glutamine and Glutamate metabolism

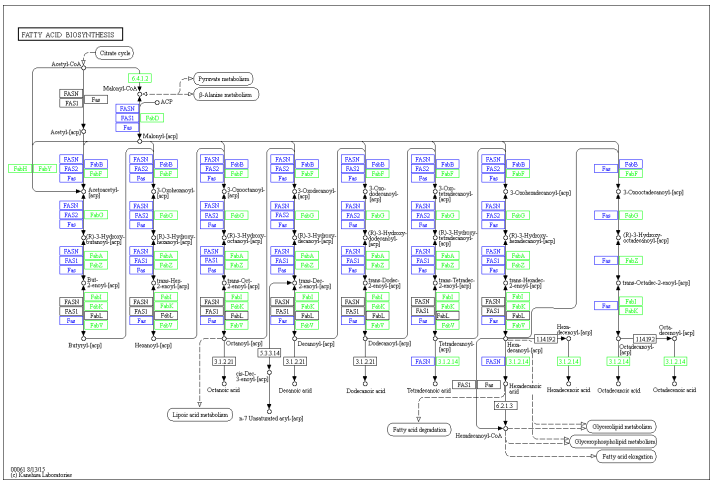


Figure 25.

Screenshot of Fatty acid biosynthesizes

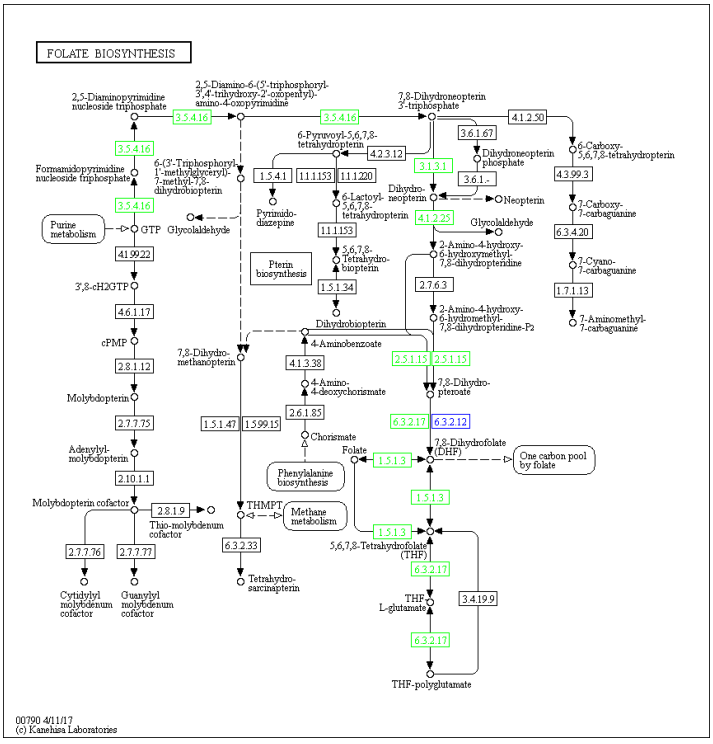


Figure 26.

Screenshot of Folate biosynthesizes

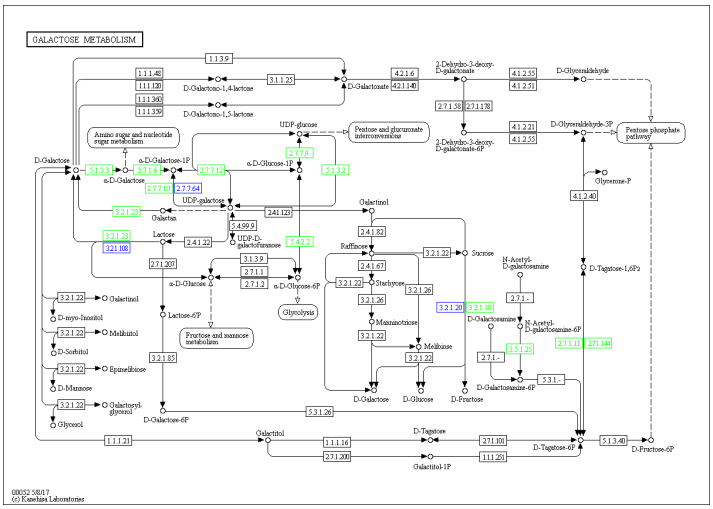


Figure 27.

Screenshot of Galactose metabolism

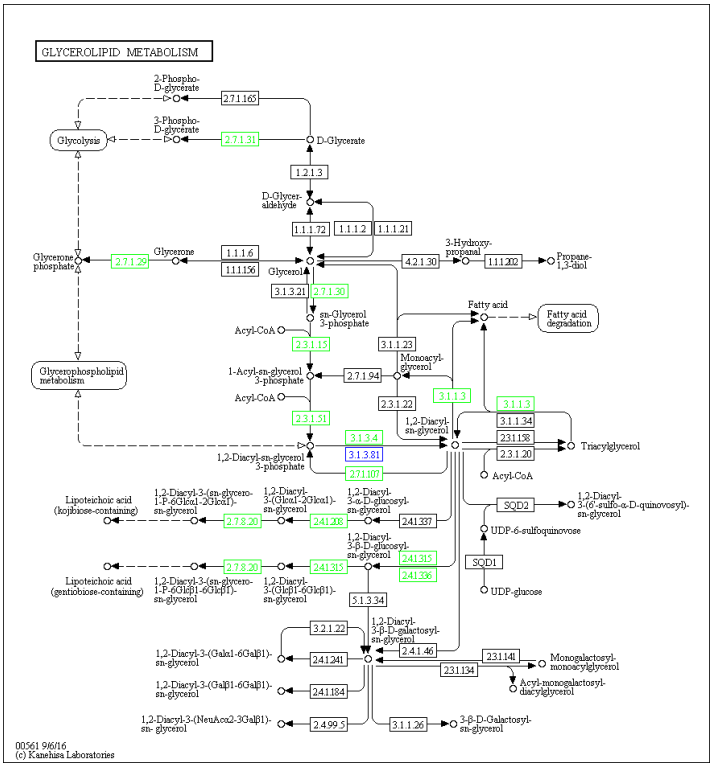


Figure 28.

Screenshot of Glycerolipid metabolism

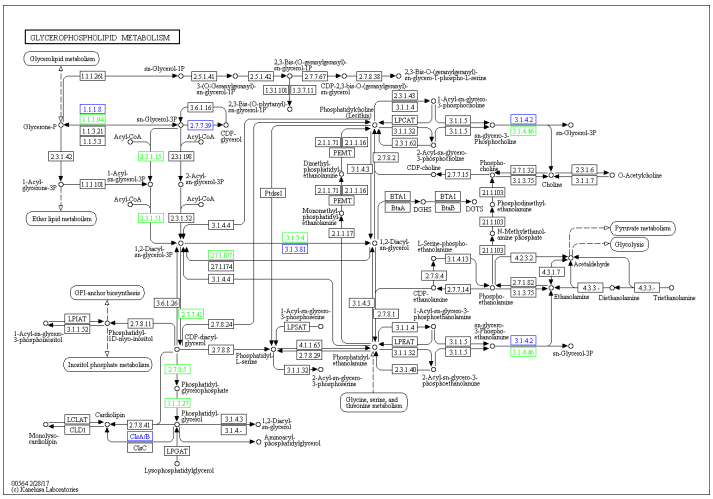


Figure 29.
Screenshot of Glycerophospholipid metabolism

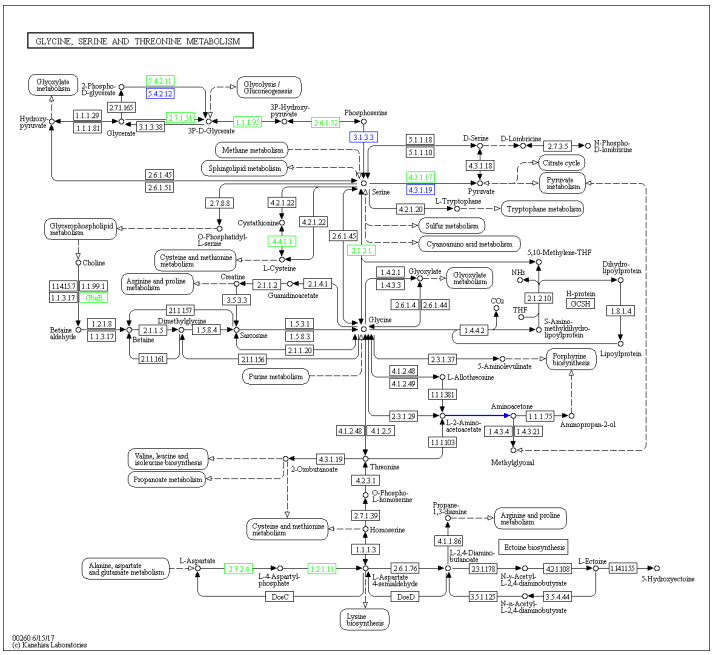


Figure 30.
Screenshot of Glycine, Serine and Threonine metabolism

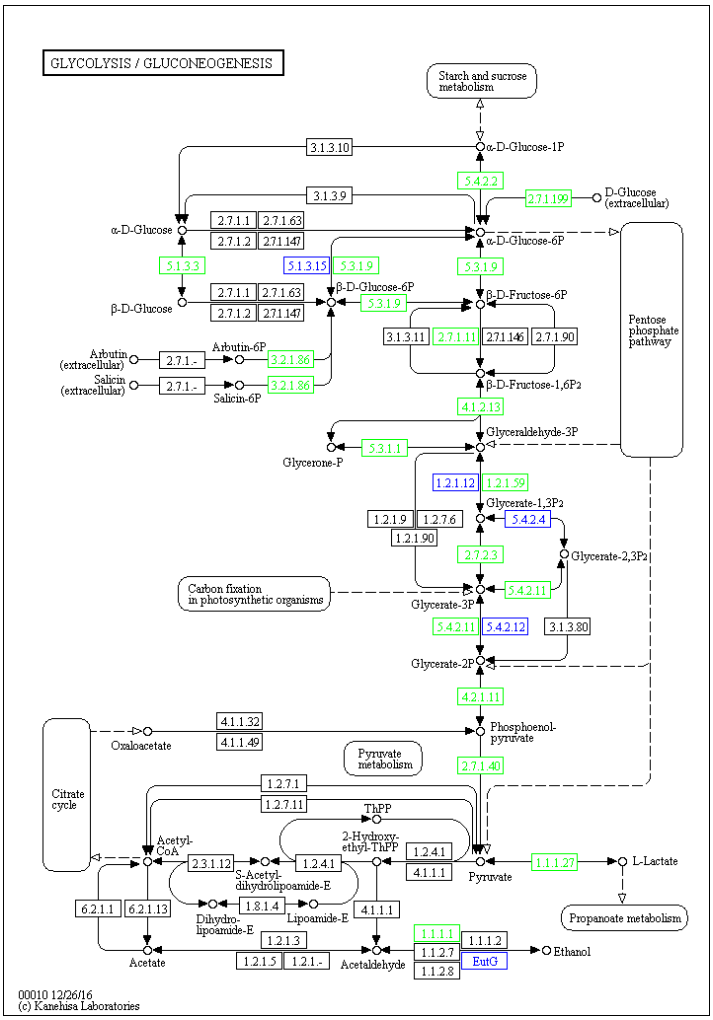


Figure 31.

Screenshot of Glycolysis

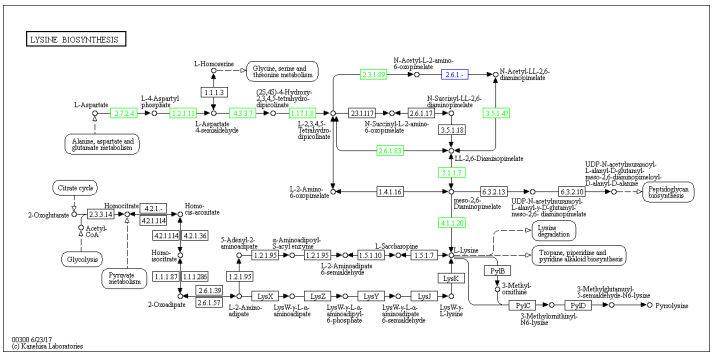


Figure 32.

Screenshot of Lysine biosynthesis

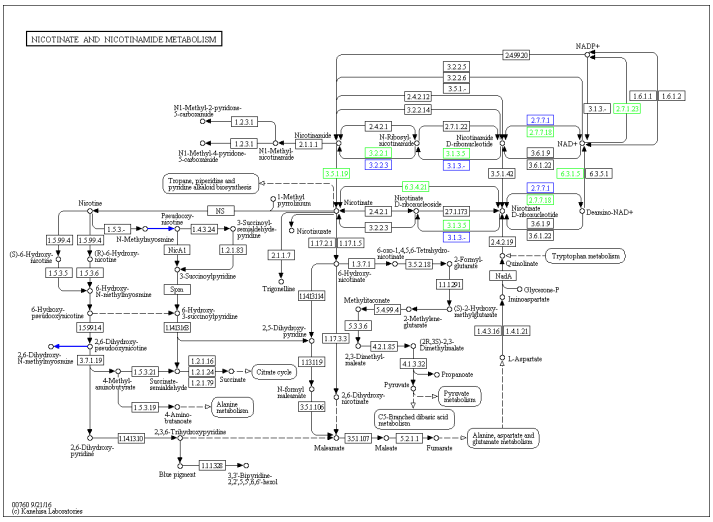


Figure 33.
Screenshot of Nicotinamide and nicotinate metabolism

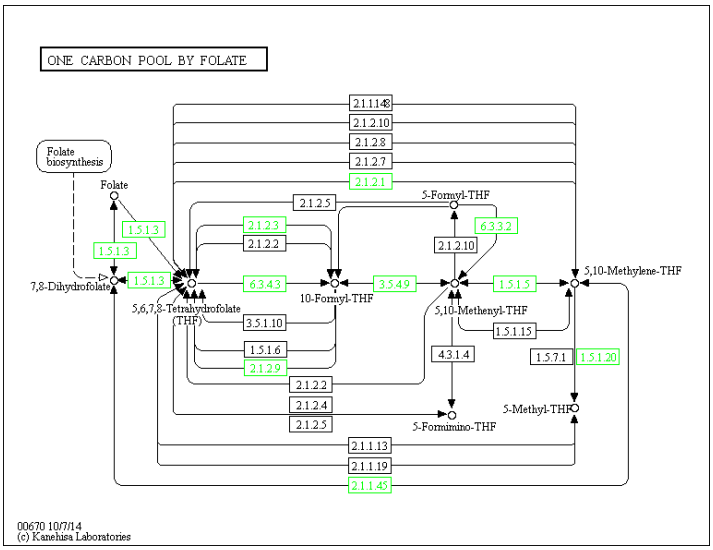


Figure 34.
Screenshot of One carbon pool by folate metabolism

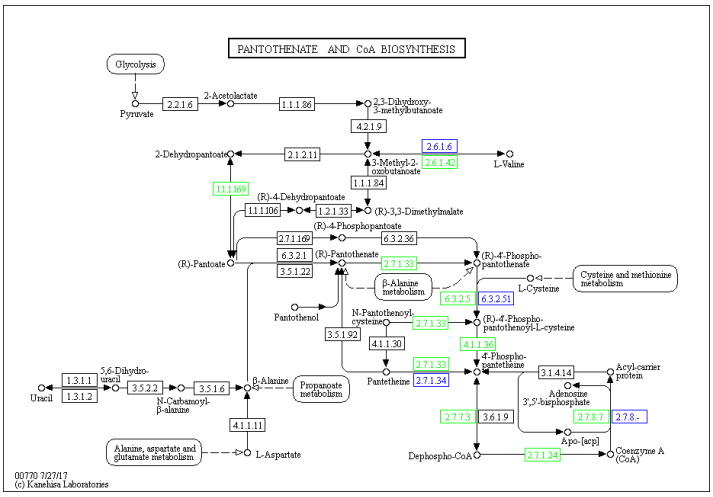


Figure 35.
Screenshot of Pantothenate and CoA biosynthesis pathway

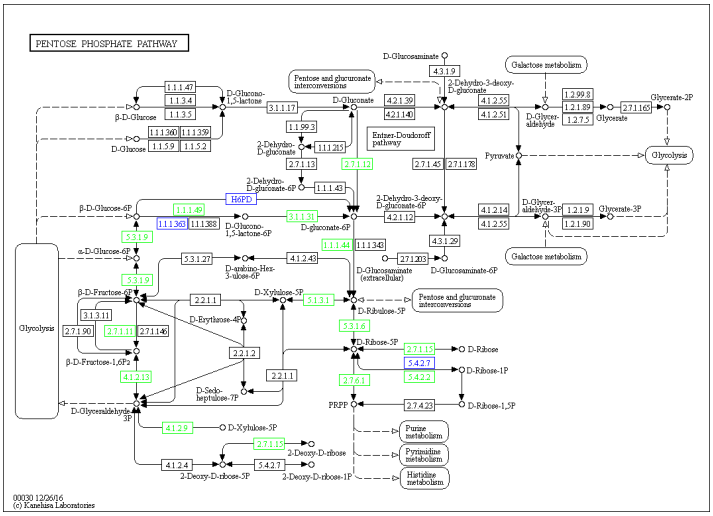


Figure 36.
Screenshot of Pentose phosphate pathway

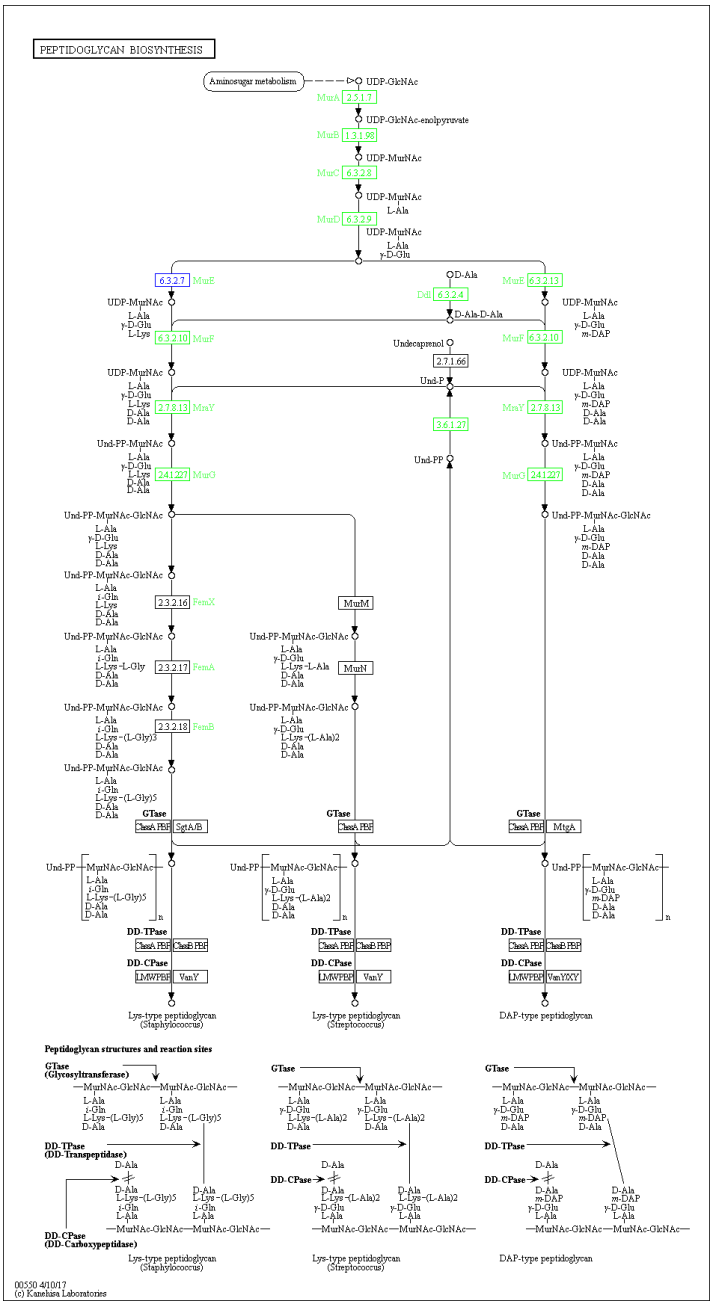


Figure 37.

Screenshot of Peptidoglycan biosynthesis

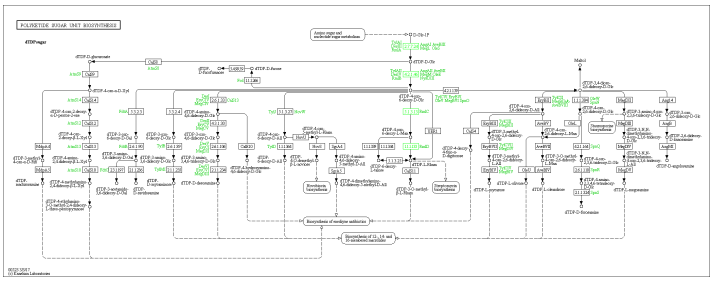


Figure 38.
Screenshot of Polyketide Sugar unit biosynthesis

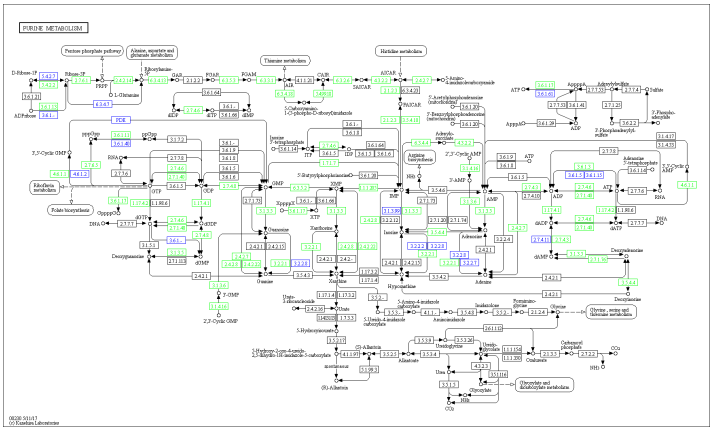


Figure 39.
Screenshot of Purine metabolism

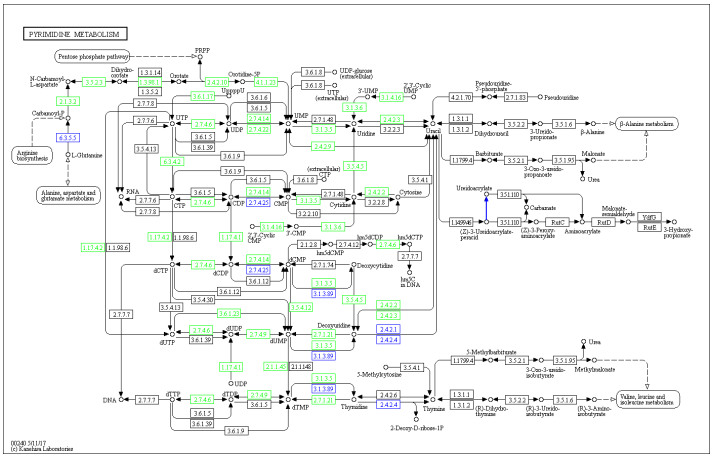


Figure 40.
Screenshot of Pyrimidine metabolism

a

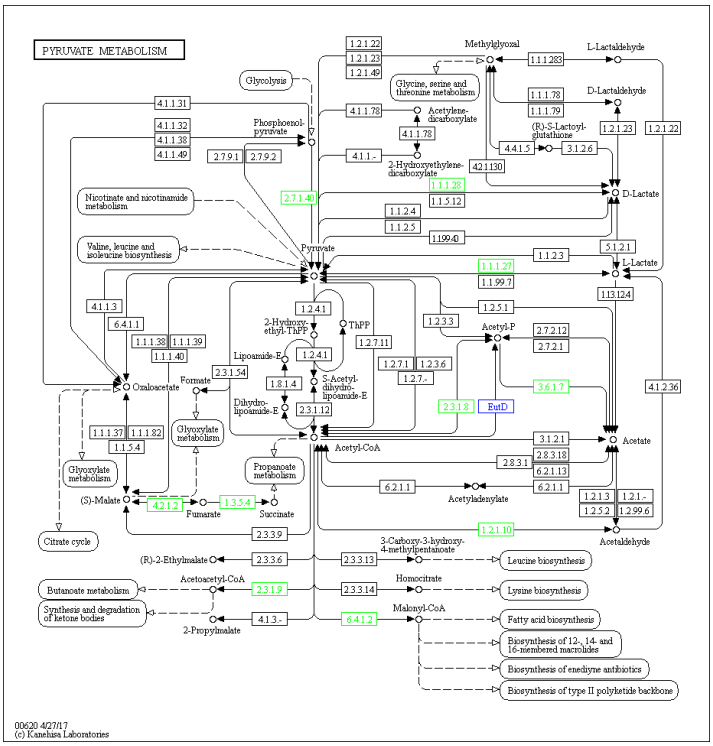


Figure 41.

Screenshot of Pyruvate metabolism

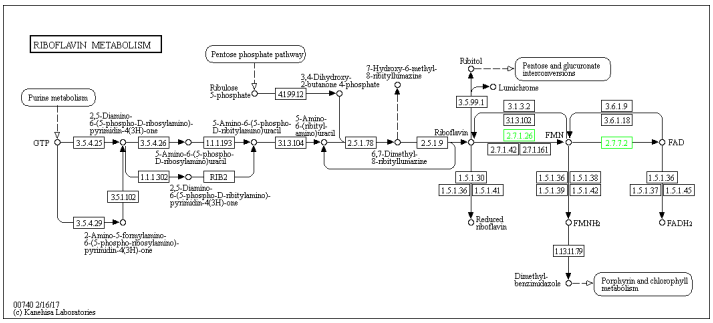


Figure 42.

Screenshot of Riboflavin metabolism

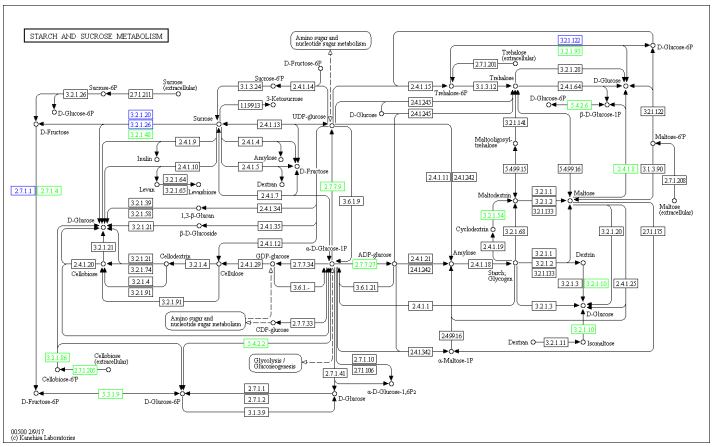


Figure 43.
Screenshot of Starch and sucrose metabolism

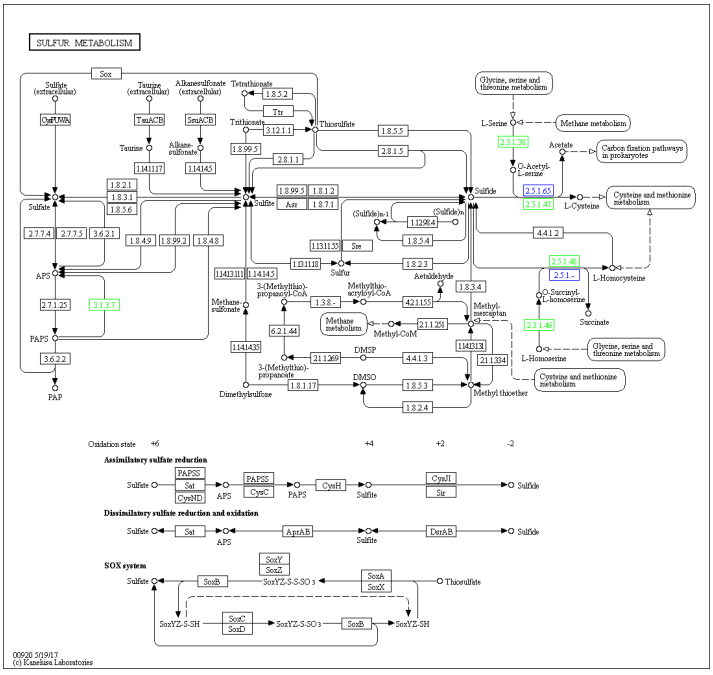


Figure 44.
Screenshot of sulfur metabolism

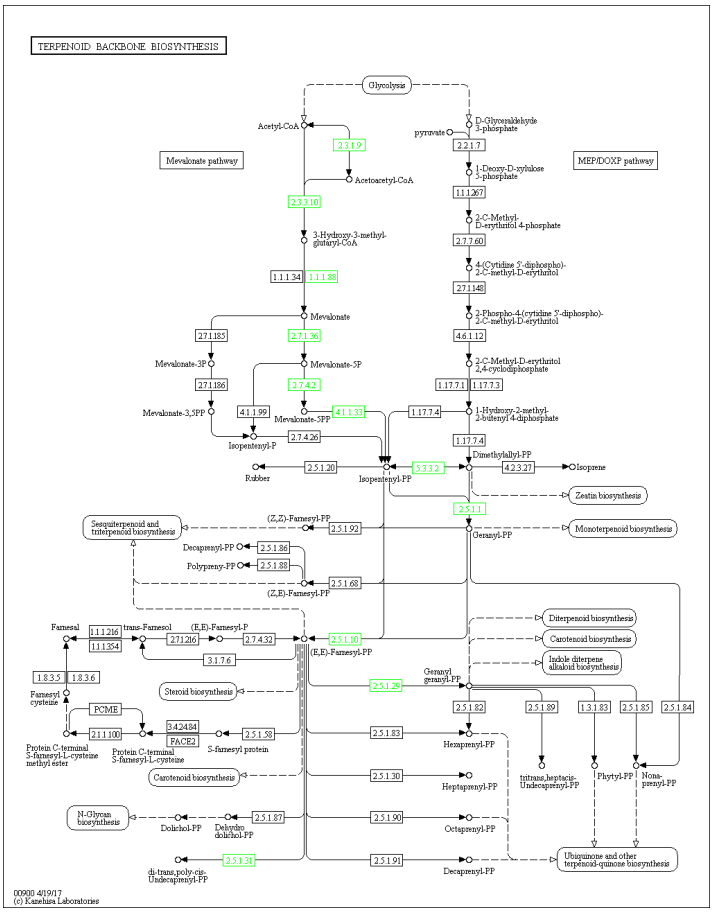


Figure 45.
Screenshot of Terpenoid backbone biosynthesis

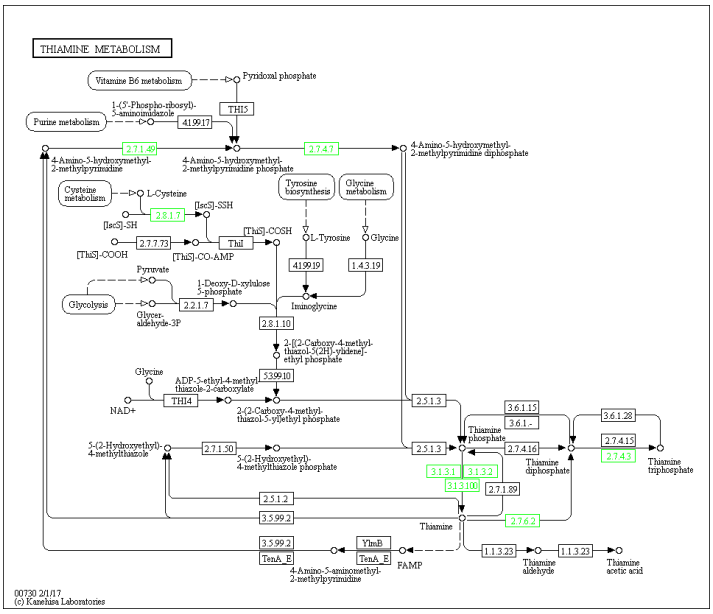


Figure 46.

Screenshot of Thiamine metabolism

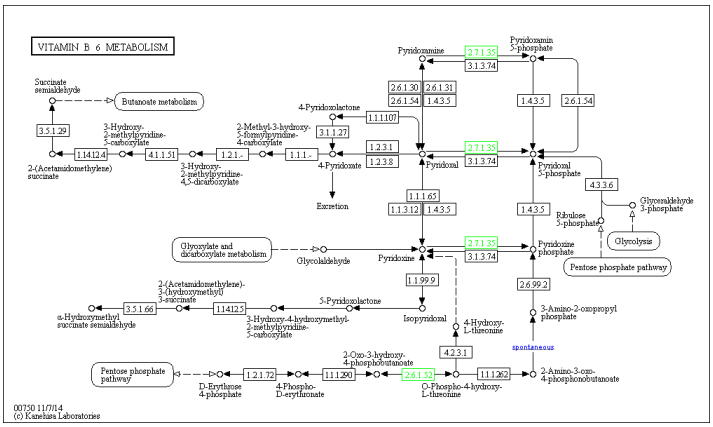


Figure 47.

Screenshot of Vitamin B6 metabolism

Table 8.
Software tools used in metabolic engineering applications [7]

Names	Taks	License	Accesibility
13CFLUX2	MFA	Free non-commercial	UNIX/Linux
A Plasmid Editor (ApE)	DNA visualization, Nucleic acid design	Free	Cross-Platform
Arcadia	Reaction network visualization	GPL	Cross-Platform
BiGG	Metabolic network reconstruction	Free non-commercial	Online
BioMet Toolbox	Constraints-based modeling	Free	Online, Windows
BioModelsDB	Metabolic network reconstruction	Free	Online
BioPax	Annotation	Free	N/A
BioTapestry	Genetic network construction and analysis	Free	Cross-Platform
BLAST	Comparative sequence analysis	Free	Online, Cross-Platform
Cell Illustrator	Reaction network visualization and design	Free, Closed source	Online
CellDesigner	Reaction network visualization and design	Free, Closed source	Cross-Platform
CellNetAnalyzer	Constraint-based modeling, MFA, Network analysis	Free academic, Requires Matlab	Cross-Platform
COBRA 2.0	Constraint-based Modeling, MFA, Network analysis	GNU GPLv3	Cross-Platform
COPASI	Mathematical analysis	Artistic License 2.0	Cross-Platform
Cytoscape	Interaction network visualization	GNU LGPL	Cross-Platform
DNA 2.0 Gene Designer	Codon optimization	Free, Closed source	Cross-Platform
DNAStar	DNA visualization, Nucleic acid design	Academic, Commercial	Cross-Platform
Lasergene	Constraint-based modeling, MFA	GNU GPL	Cross-Platform
FASIMU	MFA	Free academic, Requires Matlab	Cross-Platform
FiatFlux	DNA visualization, Nucleic acid design	Free limited, Academic, Commercial	Cross-Platform
Geneious	DNA visualization, Nucleic acid design	Apache 2.0	Online
GenoCAD	Reaction network visualization	BSD 2	Online
GLAMM	Annotation	Free	N/A
GO	Interaction network visualization	Eclipse Public License	Cross-Platform
GraphViz	Optimize culture conditions	Source code available to academic users	Cross-Platform
GrowMatch	Gene synthesis	Free, Closed source	Online, Windows
HelixWeb DNA Works	Comparative sequence analysis, Annotation	Free, Closed source	Online
IMG	Reaction network visualization and design	BSD 2	Windows
Jdesigner	Metabolic network reconstruction	Free	Online
KAAS	Metabolic network reconstruction	Free web, Licensed download	Online
KEGG Pathway	Metabolic network reconstruction	Free agreement	Online
MetaCyc	Metabolic network reconstruction	Free	Online
MetRxn	Metabolic network reconstruction	Free	Online
ModelSEED	Nucleic acid structure analysis	Free, Open source	Online
NuPack	Reaction network visualization	Free non-commercial, Closed source	Cross-Platform
Omix	MFA	GNU GPL, Requires Matlab	Cross-Platform
OpenFLUX	Constraint-based modeling, MFA, Network analysis	GNU GPLv3	Cross-Platform
OptFlux	Constraints-based modeling	Free, Requires Matlab	Cobra-toobox 2.0
OptKnock	Pathway prospecting	Free	Available by request
OptStrain	Metabolic network model analysis	Free non-commercial	Cross-Platform
PathwayTools	Primer design	Free	Online
PHUSER	Dynamic simulation	BSD 2	Cross-Platform
PySCeS	Nucleic acid design, Expression optimization	Free non-commercial	Online
RBS Calculator	Metabolic network reconstruction	Free	Online
Reactome	Network visualization	Free	N/A
SBGN	Network reconstruction and visualization	Free	N/A
SBML	Annotation	Free	N/A
SBO	Dynamic simulation	BSD 2	Cross-Platform
SBW	Optimize culture conditions	Source code available to academic users	Cross-Platform
SL Finder	Constraint-based modeling, MFA, Network analysis	GNU GPLv2	Cross-Platform
Systems Biology Research Tool Systrip	Interaction network visualization	GNU LGPL	Cross-Platform
TinkerCell	Model visualization and analysis	BSD 2	Cross-Platform
Vanted	Reaction network visualization	GNU GPLv2	Cross-Platform
VectorNTI	DNA visualization, Nucleic acid design	Academic, Commercial	Cross-Platform
Vienna RNA Websuite	Nucleic acid structure analysis	Free, Open source	Online
yEd	Interaction network visualization	Free, Closed source	Cross-Platform

Easy-DNA™Kit For genomic DNA isolation *Samples*

- Suspension or trypsinized cells (103–107 cells)
- E. coli cells (0.5–1.0 mL of an overnight culture, 1×109 cells/mL)
- Mammalian tissues (3.5 mg to 100 mg)

- Fresh plant leaves (50 mg)

Preparation Cells must be pelleted and the medium decanted. Resuspend cell pellet in 200 µŁ1X PBS (Cat. no. 10010-023). This will eliminate the formation of a salt pellet when precipitating DNA. Freeze tissue and plant leaves in liquid nitrogen and pulverize with a mortar and pestle. Place samples in microcentrifuge tubes for processing. Note: Fresh, minced leaves will yield DNA, but not as much and not as high quality as when the fresh leaves are frozen in liquid nitrogen and pulverized.

Before Starting

- Chill 100% and 80% ethanol in a $-20\,^{\circ}\mathrm{C}$ freezer.
- Thaw RNase (if stored at $-20\,^{\circ}\mathrm{C}$) and keep on ice.
- Equilibrate two heat blocks or water baths, one to $37\,^{\circ}\mathrm{C}$ and the other to $65\,^{\circ}\mathrm{C}$.

Isolation of DNA

1. Add 350 µŁSolution A to cell suspension, tissue, or plant parts and vortex in 1 second intervals until evenly dispersed.
2. Incubate at 65°C for 10 minutes.
3. Add 150 µŁSolution B and vortex vigorously until the precipitate moves freely in the tube, and the sample is uniformly viscous (10 seconds–1 minute).
4. Add 500 µŁchloroform and vortex until viscosity decreases and the mixture is homogeneous (10 seconds–1 minute).
5. Centrifuge at maximum speed for 10–20 minutes at$4\,^{\circ}\mathrm{C}$ to separate phases. Transfer the upper phase into a fresh microcentrifuge tube. Proceed to DNA Precipitation.

DNA Precipitation

1. To the DNA solution, add 1 mL of 100% ethanol ($-20\,^{\circ}\mathrm{C}$) and vortex briefly.
2. Incubate tube on ice for 30 minutes.
3. Centrifuge at maximum speed for 10–15 minutes at $4\,^{\circ}\mathrm{C}$. Remove ethanol from the pellet with a drawn-out Pasteur pipette.
4. Add 500 microL of 80% ethanol ($-20\,^{\circ}\mathrm{C}$) and mix by inverting the tube 3–5 times.
5. Centrifuge at maximum speed for 3–5 minutes at $4\,^{\circ}\mathrm{C}$. Save the pellet and remove the 80% ethanol with drawn-out Pasteur pipette.
6. Centrifuge at maximum speed for 2–3 minutes at $4\,^{\circ}\mathrm{C}$. Remove residual ethanol with a pipettor. Let air dry 5 minutes.

7. Resuspend the pellet in 100 microL TE buffer. Add 2 microL of a 2 mg/mL RNase to bring the concentration to 40 micrograms/mL.
8. Incubate at 37 °C for 30 minutes. DNA is ready for further experiments. Store at 4 °C.

Growth Medium

The used CDM was constituted of:

Table 9.

Basal Solution

Reference	Compound	Quantity
Merck 1.05941.0250	$MnSO4 \cdot H2O$	0.028g
Merck 1.06268.1000	Sodium acetate	1g
VWR 271534H	Ammonium citrate	0.6g
Merck 1.05101.1000	K2HPO4	2.5g
Merck 1.04873.1000	KH2PO4	3g
Sigma 6297-250G	NaHCO3	0.42g
Aldrich 380024-5G	Trace elements	1mL

Table 10.

Trace Elements

Reference	Compound	Quantity/L
	HCl (25% 7.7 M)	10mL
Sigma 220299-5G	FeCl2.4H2O	1.5g
Sigma 746355-100G	ZnCl2	70 mg
Sigma M3634-500G	MnCl2. 4H2O	100mg
Sigma C8661-25G	CoCl2.6H2O	190mg
	Distilled water	990mL

Sugar The sugar used for *Lb. helevetivus CNRZ32* cultures was -(+)-Glucose (Merck 1.09342.1000). It were added 10 g of sugar for 50 mL of anoxic water. The final concentration in the medium was 10 g/L.

Table 11.

Amino acid Stock Solutions (4%)

Reference	Compound	Quantity/L
Sigma – A7627-100G	L-Alanine	1g/25ml H2O
Sigma - A8094-25G	L-Arginine	1g/25ml H2O
Sigma - A8381-100G	L-Asparagine(1H2O)	1.13g/22,5ml H2O + 2,5ml 2M NaOH
Sigma – 11189-100G	L-Aspartate	1g/20ml H2O + 5ml 2M NaOH
Sigma – G8415-100G	L-Glutamate	1g/20ml H2O + 5ml 2M NaOH
Sigma – G3126-100G	L-Glutamine	1g/20ml H2O + 5ml 2M NaOH
Sigma – 50046-50G	Glycine	1g/25ml H2O
Sigma – H8125-100G	L-Histidine (1HCl.1H2O)	1.35g/25ml H2O
Sigma – I 2752-25G	L-Isoleucine	1g/20ml H2O + 5ml 2M NaOH
Sigma – 61819-25G	L-Leucine	1g/20ml H2O + 5ml 2M NaOH
Sigma – 62840-25G-F	L -Lysine	1g/25ml H2O
Merck – 1.05707.0025	L-Methionine	1g/25ml H2O + 5ml 2M NaOH
Sigma – 7819-25G	L-Phenylalanine	1g/20ml H2O + 5ml 2M NaOH
Sigma – 81709-10G	L-Proline	1g/25ml H2O
Sigma – 84959-25Gb	L-Serine	1g/25ml H2O
Sigma – T8625-10G	L-Threonine	1g/25ml H2O + 5ml 2M NaOH
Sigma – T8941-25G	L-Tryptophane	1g/20ml H2O + 5ml 2M NaOH
Sigma – T8566-25G	L-Tyrosine	1g/20ml H2O + 5ml 2M NaOH
Sigma – V0513-25G	L-Valine	1g/25ml H2O + 5ml 2M NaOH

Table 12.

Vitamin solution, 100x

Reference	Compound	Quantity (g)
Sigma B4501-10G	Biotin	0.010
Sigma F8798-5G	Folic acid	0.010
Sigma 271748	Pyridoxal.HCl	0.050
VWR chemicals 27414.137	Riboflavin	0,025
Sigma T4625-10G	Thiamine-HCl	0.025
Sigma N0636-100G	Nicotinamide	0.025
Sigma C3607-500mg	Vit B12 Cyanocobalamin	0.025
Sigma A9878-5G	p-Aminobenzoic acid	0.025
Sigma C8731-25G	dl-Ca-pantothenate	0.200
Sigma 62320-5G-F	dl-6,8-Thioctic acid (Lipoic acid)	0.025

Table 13.

Bases Solution (100x)

Reference	Compound	Quantity (g/250mL)
Sigma A8626	Adenine	0.250
Sigma G11950	Guanine	0.250
Sigma U0750	Uracil	0.250
Sigma X0626	Xanthine	0.250

Other Components

MgCl2.6H2O, 2% (100 x)

- 1 g of MgCl2.6H2O for 50 mL of anoxic water. Final concentration in the medium 0.2 g/L (Ref: Merck, 1.05833.0250) Flush the headspace with N2 > 10min as described in EXP-15-AF7901. Autoclave at 121°C for 10-15 minutes. Note: final concentration in the medium is half of that used in CDM BB12

CaCl2.2H2O, 0.5% (100 x)

- 0.25 g of CaCl2.2H2O for 50 mL of anoxic water. Final concentration in the medium 0.05 g/L (Ref: Merck, 1.02382.0500) Flush the headspace with N2 > 10min as described in EXP-15-AF7901. Autoclave at 121°C for 10-15 minutes.

Cysteine-HCl.H2O, 5% (100 x)

- 2,5 g of Cysteine-HCl for 50 mL of anoxic water. Final concentration in the medium 0.5 g/L (Ref: Merck, 1.02839.0100) Flush the headspace with N2 > 10min as described in EXP-15-AF7901. Autoclave at 121°C for 10-15 minutes.

Urea, 1.2% (100 x)

- 0.6 g of Urea for 50 mL. Final concentration in the medium 0.12 g/L (Ref: Sigma, 51456) Filter sterilization (0.22µm) to a serum bottle previous flushed with N2 Keep at 4C.

Table 14.

Final volumes amounts

Values presented for a final solution with 100mL

Component	Volume
Basal solution	80,2 mL
Sugar (20X) – 2%	10 mL
MgCl2.6H2O (100x) – 0,2g/L	1 mL
CaCl2.2H2O (100x) – 0,05g/L	1 mL
Urea (100x) -0,12g/L	1 mL
Vitamin sol. (100x)	1 mL
Amino acids mix (52,6x) – 0,08 g/L	3,8 mL
Bases solution (100x) – 10 mg/L	1 mL
Cysteine.HCl (100x) – 0,05%	1 mL

Table 15.

Fatty acid profile

Fatty acid		Ratio	Chemical Formula	g/mol
tetradecanoic acid (myristic acid)	14:0	0.16	C14H28O2	228.37
hexadecanoic acid (palmitic acid)	16:0	0.20	C16H32O2	256.43
cis-9-Hexadecenoic acid (palmitoleic acid)	16:1	2.51	C16H30O2	254.41
Octadecanoic acid (stearic acid)	18:0	2.51	C18H36O2	284.84
(9Z)-Octadecenoic acid (oleic acid)	18:1	0.58	C18H34O2	282.47
Average fatty acid		1	C17H33O2	269.50