

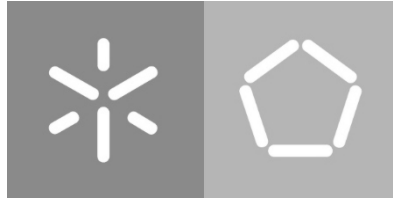
Universidade do Minho

Escola de Engenharia

Departamento de Informática

José Jorge Sampaio Bastos

**Modelling interspecies interactions of
syntrophic communities of
Syntrophobacter fumaroxidans and
*Methanospirillum hungatei***



Universidade do Minho

Escola de Engenharia

Departamento de Informática

José Jorge Sampaio Bastos

**Modelling interspecies interactions of
syntrophic communities of *Syntrophobacter
fumaroxidans* and *Methanospirillum
hungatei***

Thesis Dissertation

Master's Degree in Bioinformatics

Supervisors

Oscar Dias

Alfons Stams

March 19

ACKNOWLEDGEMENTS

It comes to an end one more challenge of my life. During this journey, I found help and support in many people who without them this work would not be concluded. I use this section to thank all of those that in a way or another kept me on the right way.

First of all, I would like to thank my supervisor, professor Oscar Dias, and my co-supervisor Alfons Stams, for all the help and guidance through this work. I would like to thank professor Oscar for providing the opportunity to embrace this challenge.

I would also like to thank Sophia Santos and Fernando Cruz for all the help that they provided and for all the doubts they help me clarify. A big thanks to the two of you.

A special thanks to all my friends and especially to my housemates that without them this journey would have been really boring. An enormous thank you to my three partners that were always there for everything, but they still owe me a reward for winning our hearts competition. Thanks to all of you!

A particular word to my good old friend, my non-blood brother, Nuno Alves for being the greatest friend a person can have and for being there when I needed most. My sincere thanks for all the countless talks and advices.

To Inês, the person that was always there for me when I needed. Thanks for all the help, support, and for making me the happiest person every single day. Thank you *carola* for being everything that I need.

Last but not least, a huge thank to all my family: my mother, my father, my sister, my brother, and my grandparents. A huge thank to my parents that raised me in the right way and always guided me to make the right decisions in life. If I am what I am today, I owe that to all of you.

To all, I thank you deeply!

ABSTRACT

Microbial communities have gained particular interest and have been used for practical applications such as biorefineries, and bioremediation. However, studying these communities has proven to be difficult due to the absence of experimental protocols and computational tools like the ones available for single organisms.

In this work, we present Genome-Scale Metabolic models both for *Methanospirillum hungatei* strain JF1 and *Syntrophobacter fumaroxidans* strain MPOBT, together with a model that combines both into one community model. The genome-scale metabolic model reconstruction of *S. fumaroxidans* was performed in *merlin* whereas, the methane-producing archaeon *M. hungatei* was reconstructed in KBase's environment and the model curation was performed in *merlin*. *OptFlux* and BioCoISO, a tool implemented over COBRApy developed specifically for debugging model pathways, were used for curating and validating both models.

The metabolism of each individual organism was assessed through its model reconstruction. *In silico* simulations demonstrated the production of various compounds of interest such as formate in *M. hungatei* and acetate in *S. fumaroxidans*. The meta-model representing the community composed by both organisms was assembled using FRAMED, and it was able to describe the metabolic exchanges between the formate scavenger *M. hungatei* and the syntrophic partner *S. fumaroxidans*.

The reconstructed models can be used to study further the metabolic interactions between these bacteria.

Keywords: Systems Biology, Genome-Scale Metabolic Models, Metabolic Networks, Constraint-Based Modelling, *merlin*, *Syntrophobacter fumaroxidans*, *Methanospirillum hungatei*, Syntrophic Community, KBase

RESUMO

As comunidades microbianas são de especial interesse e têm sido usadas para aplicações práticas como em biorrefinarias e biorremediação. No entanto, o estudo destas comunidades tem sido difícil devido há falta de protocolos experimentais e ferramentas computacionais, como os que existem para cada organismo individualmente.

Neste trabalho são apresentados os modelos metabólicos à escala genómica para estirpe JF1 de *Methanospirillum hungatei* e a estirpe MPOBT de *Syntrophobacter fumaroxidans*, juntamente com um modelo que combina ambos os modelos criados num modelo de comunidade. A reconstrução do modelo metabólico à escala genómica de *S. fumaroxidans* foi realizada no *merlin*, enquanto que o modelo da bactéria produtora de metano *M. hungatei* foi reconstruído na KBase e a curação manual efetuada no *merlin*. *OptFlux* e BioColSO, uma ferramenta implementada sobre o COBRApy, desenvolvida especificamente para a correção de vias do modelo, foram usadas para a curação e validação de ambos os modelos.

O metabolismo de cada organismo foi acedido através das respetivas reconstruções realizadas para cada um. Simulações *in silico* demonstraram a produção de vários compostos de interesse como o formato no caso de *M. hungatei* e acetato no caso de *S. fumaroxidans*. O meta-modelo criado que representa a comunidade formada por ambos os organismos foi criado a partir de uma ferramenta presente no FRAMED, e este é capaz de descrever as trocas metabólicas entre *M. hungatei* e *S. fumaroxidans*.

Os modelos reconstruídos podem ser usados para estudar no futuro as interações metabólicas entre estas duas bactérias.

Contents

ACKNOWLEDGEMENTS	i
ABSTRACT	ii
RESUMO	iii
INTRODUCTION	1
1.1 CONTEXT AND MOTIVATION	1
1.2 GOALS	2
1.3 STRUCTURE OF THE DOCUMENT	3
STATE-OF-THE-ART	5
2.1 GENOME-SCALE METABOLIC MODELS	5
2.1.1 Background	5
2.1.3 Genome annotation	9
2.1.4 Metabolic Network Assembly	10
2.1.4.1 <i>Genes, Proteins and Reactions Associations</i>	11
2.1.4.2 <i>Spontaneous reactions</i>	12
2.1.4.3 <i>Reaction stoichiometry</i>	12
2.1.4.4 <i>Localization and Compartmentalization</i>	13
2.1.4.5 <i>Manual Curation</i>	13
2.1.5 Stoichiometric Model Assembly	15
2.1.5.1 <i>From Network Reconstruction to Mathematical Model</i>	16
2.1.6 Stoichiometric Model Validation	19
2.2 RELEVANT BIOINFORMATICS TOOLS	20
2.2.1 Cobra ToolBox	21
2.2.2 Pathway tools	22
2.2.3 RAVEN	22
2.2.4 GEMSiRV	23
2.2.5 SuBliMinalL Toolbox	24
2.2.6 <i>merlin</i>	25
2.2.6.1 <i>Enzymes Annotation</i>	26
2.2.6.2 <i>Transporters Annotation</i>	27
2.2.6.3 <i>GSM Model Assembly</i>	28
2.2.7 <i>OptFlux</i>	29
2.2.8 KBase	30
2.2.8.1 <i>KBase Apps for reconstructing a GSM model</i>	32

2.3	SYNTROPHIC BACTERIA COMMUNITY	36
2.3.1	Background	36
2.3.2	<i>Syntrophobacter fumaroxidans</i> strain MPOBT	38
2.3.3	<i>Methanospirillum hungatei</i> strain JF1	39
2.3.5	Syntrophic Relationship of <i>S. fumaroxidans</i> strain MPOBT and <i>M. hungatei</i> strain JF1	40
	MATERIALS AND METHODS	42
3.1	GENOME ANNOTATION	42
3.1.1	Enzymes and transporters annotations for <i>M. hungatei</i> using KBase	42
3.1.2	Building <i>GSM</i> draft model of <i>M. hungatei</i> in KBase	43
3.1.4	Enzymes annotation in <i>merlin</i> for <i>S. fumaroxidans</i>	45
3.1.3	Transporter Proteins	46
3.2	<i>Syntrophobacter fumaroxidans</i> MPOB ^r DRAFT NETWORK ASSEMBLY	47
3.2.1	Metabolic Data Integration	47
3.2.2	Transporter Proteins Data Integration	47
3.2.3	Exchange reactions Integration and Compartmentalization	47
3.2.4	Manual Curation of the Draft Network of <i>S. fumaroxidans</i>	48
3.2.4.1	<i>Pathway-by-pathway Analysis</i>	48
3.2.4.2	<i>Gap Filling</i>	49
3.2.4.3	<i>Mass Balance</i>	50
3.3	<i>M. hunagtei</i> DRAFT METABOLIC NETWORK CURATION IN <i>merlin</i>	50
3.3.1	<i>M. hungatei</i> draft <i>GSM</i> model integration in <i>merlin</i>	50
3.3.2	Manual curation of the Draft Network of <i>M. hunagtei</i>	50
3.4	BIOMASS AND ENERGY REQUIREMENTS FORMULATION FOR <i>M. hungatei</i> and <i>S. fumaroxidans</i>	52
	RESULTS AND DISCUSSION	54
4.1	GENOME-SCALE METABOLIC MODEL OF <i>S. fumaroxidans</i> strain MPOBT	54
4.1.2	Manual curation of the draft metabolic network of <i>S. fumaroxidans</i>	54
4.1.2.1	<i>Pathway-by-pathway analysis</i>	55
4.1.2.2	<i>Transport reactions</i>	57
4.1.2.3	Gap Filling	57
4.1.2.4	<i>Mass Balance</i>	57
4.1.3	Biomass and energy requirements formulation	58
4.2	GENOME-SCALE METABOLIC MODEL OF <i>M. hungatei</i> strain JF1	62
4.2.1	Manual curation of the draft metabolic network of <i>M. hungatei</i> JF1	62

4.2.1.1	<i>Pathway-by-Pathway analysis</i>	62
4.2.1.2	<i>Gap Filling</i>	63
4.2.1.3	Mass Balance	68
4.2.2	Biomass and energy requirements formulation	69
4.3	<i>GSM</i> MODELS COMPARISON	73
4.3.1	KBase <i>GSM</i> models	73
4.3.2.	KBase <i>GSM</i> model's vs <i>merlin</i> <i>GSM</i> model's	74
4.3.3	<i>GSM</i> models comparison with literature	75
4.4	META-MODEL ASSEMBLY	76
4.4.1	Single model troubleshooting	76
4.4.2	Meta-Model Assembly troubleshooting	77
CONCLUSIONS AND FUTURE WORK		80
5.1	General conclusions	80
5.2	Future work	81
BIBLIOGRAPHY		83

LIST OF FIGURES

Figure 1. Illustration representing the iterative processes during metabolic reconstruction. Firstly, it starts with a complex compilation of the information available for the microorganism metabolism from different information sources. A reaction set is built, and debugging is performed to build a steady-state metabolic model. Next, the comparison of in silico simulation results with experimental data is made and when the last is in accordance with in silico predictions, the model is ready to be use in biotechnology applications. If the in silico predictions don't match the experimental results it is necessary to make a revision on the information sources. Adapted from [19]

7

Figure 2. Example of a metabolic network with 6 metabolites (A to F) and 10 fluxes (V1 to V10). A reaction scheme is presented in (1) with outlined system boundaries. The fluxes V1 to V4 represent the exchange fluxes of metabolite substrate (A) and products (B, F and E). The reversible reactions are shown with double arrows, and the irreversible reactions are identified with a forward arrow. In section (2) is presented the stoichiometric of the network. Section (3) represents the steady-state mass balances, and section (4) shows the constraints around the flux values. The final section (5) illustrates the representation of the mass balances in matrix format.

18

Figure 3. COnstraints-Based Reconstruction and Analysis of biological networks: (A) -Data sources for network reconstruction; (B) -Network reconstruction; (C) - Application of constraints to the network; (D)- Analysis of the network model in other to achieve an optimal solution. Adapted from [2].

21

Figure 4. Flow diagram of the chained modules present in SuBliMinaL Toolbox working together in order to form the draft model. The names of the boxes refer to individual modules of the tool. The grey right-hand branch signifies that existing reconstructions or individual pathways can be added to the pipeline of work allowing for the generation high-quality drafts. Adapted from [75]

Erro! Marcador não definido.

Figure 5. Enzyme annotation data collected by merlin after BLAST. The most relevant data for each gene present in the genome fasta file is stored. Also, data for every homologue identified for a certain gene is saved [4].

26

Figure 6. Illustration of a Narrative in KBase where (1) refers to the data field where the user stores all the imputed/generated data along the narrative. (2) is the apps field, the user has access for over than 160 applications. (3) Represent the analysis steps of each app. The (4) field is pointing towards the share symbol, meaning that the users can share their narratives with other users. In field (5), Markdown cell, it is possible to add commentaries. The last field, (6), points out the custom scripts with which the user can a python script. Adapted from [88].

31

- Figure 7.** Schematic drawing of “classical” syntrophy depending on the environmental conditions (methanogenic environment). The thicker dashed line represents the border of the system. 37
- Figure 8.** Microscopy photograph of cells of *S. fumaroxidans*. Adapted from [108]. 39
- Figure 9.** Cells of *M. hungatei* showing tufts of polar flagella. Black arrow pointing towards tufts. Adapted from [110] 40
- Figure 10.** Simplified illustration of the syntrophic relationship between *S. fumaroxidans* and *M. hungatei*. *S. fumaroxidans* uses propionate from the environment to produce acetate and in the process, it can produce hydrogen (H_2). The hydrogen can later be very important in the metabolism of *M. hungatei* more exactly in stage (A) where it will participate in the reduction of the ferredoxins that assist the reaction that converts intermediate 3 into intermediate 4. Adapted from [116] 41
- Figure 11.** TRIAGE standards for retrieving genes encoding transport systems in *S. fumaroxidans* MPOBT 46
- Figure 12.** Example of a partial metabolic KEGG’s pathway map coloured by *merlin* tool. 49
- Figure 13.** *merlin*’s e-biomass equation tool used to formulate the biomass composition of *Methanospirillum hungatei* 51
- Figure 14.** *merlin*’s e-biomass equation tool used to formulate the biomass composition of *Syntrophobacter fumaroxidans* 52
- Figure 15.** Schematic representation of the peptidoglycan structure in *S. fumaroxidans*. 61
- Figure 16.** Methanogenic pathway in *Methanospirillum hungatei* JF1. The metabolites coloured in red were the ones identified as gaps in this pathway. CoM – Coenzyme M; H4MPT – tetrahydroneopterin; HTP or CoB – Coenzyme B; Adapted from [146] 68
- Figure 17.** Schematic representation of the metabolic interactions between *S. fumaroxidans* and *M. hungatei*; - Hydrogenase -
Formate dehydrogenase 76
- Figure 18.** Scheme representing the pipeline of the tool developed to modify both models so they could be merged by the COMMUNITY tool. Functions are highlighted by grey boxes. 79

LIST OF TABLES

Table 1. Online Bioinformatic resources	8
Table 2. The table below lists the sources and the respective information they provide while creating the GPR associations.	12
Table 3. List of all available tools in <i>merlin</i> 's repertoire for curation of draft metabolic networks.	28
Table 4. KBase supported data objects	30
Table 5. Format of TSV or Excel file to be uploaded into KBase's Narrative.	34
Table 6. Comparison of the features of software tools developed for aiding the reconstruction of genome-scale metabolic models, mentioned above.	35
Table 7. Complete basal bicarbonate-buffered medium used for gapfilling. Adapated from [120].	43
Table 8. Standards used by <i>merlin</i> to colour the reactions and enzymes in the network	48
Table 9. List of metabolic pathways available in the <i>GSM</i> model of <i>Syntrophobacter fumaroxidans</i> MPOBT.	55
Table 10. Reactions added to the fatty acid biosynthesis pathway.	56
Table 11. Summary of reactions corrected according to the mass balance curation.	58
Table 12. Biomass composition of <i>Syntrophobacter fumaroxidans</i> MPOBT.	58
Table 13. Protein composition of <i>Syntrophobacter fumaroxidans</i> MPOBT. The R present in every chemical formula represents the R group abbreviation, meaning it represents any group or any formula linked to a carbon or hydrogen atom on the rest of the molecule.	58
Table 14. DNA composition of <i>Syntrophobacter fumaroxidans</i> MPOBT.	59
Table 15. RNA composition of <i>Syntrophobacter fumaroxidans</i> MPOBT.	59
Table 16. Cofactors composition of <i>Syntrophobacter fumaroxidans</i> MPOBT.	60
Table 17. <i>S. fumaroxidans</i> fatty acid composition	60
Table 18. <i>S. fumaroxidans</i> peptidoglycan composition	61
Table 19. Energy requiremnts of <i>Syntrophobacter fumaroxidans</i> MPOBT.	61
Table 20. List of metabolic pathways available in the <i>GSM</i> model of <i>Methanospirillum hungatei</i> JF1.	62
Table 21. Reactions added to the metabolic network of <i>M. hungatei</i> JF1 in order to create a pathway for methanofuran biosynthesis. The following letters are unique for this table: A - 4-[N-gamma-L-glutamyl-)-p-(beta-aminoethyl)phenoxy-methyl]-2-(aminomethyl)furan_c0; B - 4-[N-gamma-L-glutamyl-gamma-L-glutamyl-)-p-(beta-aminoethyl)phenoxy-methyl]-2-(aminomethyl)furan_c0.	64
Table 22. Reactions added to the metabolic network of <i>M. hungatei</i> JF1 in order to create a pathway for H4MPT biosynthesis. The following letters are unique for this table: A – 4-(?-D-ribofuranosyl)hydroxybenzene 5'-phosphate; B – 4-(?-D-ribofuranosyl)-N-succinylaminobenzene 5'-	

phosphate; C – [1-(2-amino-7-methyl-4-oxo-7,8-dihydro-3H-pteridin-6-yl)]ethyl diphosphate; D – [1-(2-amino-7-methyl-4-oxo-7,8-dihydro-3H-pteridin-6-yl)]ethyl-4-(?-D-ribofuranosyl)aminobenzene 5'-phosphate; E – [1-(2-amino-7-methyl-4-oxo-7,8-dihydro-3H-pteridin-6-yl)]ethyl-(4-aminophenyl)-1-deoxy-D-ribitol 5'-phosphate; F – [1-(2-amino-7-methyl-4-oxo-7,8-dihydro-3H-pteridin-6-yl)]ethyl-(4-aminophenyl)-1-deoxy-D-ribitol; G – [1-(2-amino-7-methyl-4-oxo-7,8-dihydro-3H-pteridin-6-yl)]ethyl-(4-aminophenyl)-1-deoxy-5-[1-?-D-ribofuranosyl 5-phosphate]-D-ribitol; H – [1-(2-amino-7-methyl-4-oxo-7,8-dihydro-3H-pteridin-6-yl)]ethyl-(4-aminophenyl)-1-deoxy-5-[1-?-D-ribofuranosyl triphosphate]-D-ribitol	65
Table 23. Reactions present in the coenzyme M biosynthesis.	65
Table 24. Reactions present in the coenzyme B biosynthesis pathway.	66
Table 25. Reactions present in the coenzyme F420 biosynthesis pathway.	66
Table 26. Biomass composition of <i>Methanospirillum hungatei</i> JF1	69
Table 27. Protein composition of <i>Methanospirillum hungatei</i> JF1.	69
Table 28. DNA composition of <i>Methanospirillum hungatei</i> JF1.	70
Table 29. RNA composition of <i>Methanospirillum hungatei</i> JF1.	70
Table 30. Cofactors composition in <i>Methanospirillum hungatei</i> JF1.	71
Table 31. Lipid composition in <i>Methanospirillum hungatei</i> JF1. Note that the sum of all stoichiometric coefficients is accounted to one.	72
Table 32. Peptidoglycan composition in <i>Methanospirillum hungatei</i> JF1.	72
Table 33. Energy requirements of <i>Methanospirillum hungatei</i> JF1.	72
Table 34. [c0] – cytosol compartment; (*) – Stoichiometric coefficient relative to biomass equation	73
Table 35. Summary of the principle characteristic of KBase <i>GSM</i> models and <i>merlin GSM</i> models. (*) This <i>GSM</i> model was first reconstructed in KBase and later integrated into <i>merlin</i> .	74
Table 36. Summary of the comparison of the four <i>GSM</i> models built using different approaches. *(a)- GPR association only for some reactions; *(b)- The GPR associations are not yet present in the model; *(c)- The validation was performed upon a draft model	75

ACRONYMS

ADP : Adenosine Pyrophosphate	16
ATP : Adenosine triphosphate	13, 16, 44, 72
BIGG : Biochemical, Genetics and Genomic Models	8, 21
BLAST : Basic Local Alignment Search Tool	6, 25, 32, 45
BRENDA : Braunschweig Enzyme Database	8, 9, 11, 12
CDW : cellular dry weight	58, 69
CoA : coenzyme A	38, 56
COBRA : COntstraint-Based Reconstruction and Analysis	21, 35
DNA : Deoxyribonucleic Acid	10, 50, 59, 70
DOE : Department of Energy	30
EC : Enzyme Classification	10, 26, 45, 47
EEGC : Enzyme Encoding Gene Candidate	45
FBA : Flux Balance Analysis	19, 21, 23
FVA : Fluxes Variability Analysis	33
GEMSiRV : Genome scale Metabolic model Simulation, Reconstruction and Visualization	23
GLPK : GNU Linear Programming Kit	35
GO : gene ontology	10
GPR : gene-protein-reaction	11, 25, 28, 33, 49, 75
GSM : Genome-Scale Metabolic	1, 2, 6, 10, 13, 19, 20, 21, 23, 25, 27, 35, 50, 54, 56, 68, 78, 80
HTP : coenzyme B	65, 67
JGI : Joint Genome Institute	30, 38
JSON : JavaScript Object Notation	51, 74
KEGG : Kyoto Encyclopedia of Genes and Genomes	8, 10, 12, 14, 20, 23, 25, 38, 44, 47, 48
KO : KEGG Orthology	28
MCA : Metabolic Control Analysis	29
ME : Metabolic engineering	1
MIRIAM : Minimal Information Required In the Annotation of Models	24
MOMA : Minimization of Metabolic Adjustment	29
NAD : Nicotinamide Adenine Dinucleotide	41
NADP : Nicotinamide Adenine Dinucleotide Phosphate	41
NCBI : National Center for Biotechnology Information	8, 42, 45
ORF : Open Reading Frames	10
RAVEN : Reconstruction, Analysis, and Visualization of Metabolic Networks	22, 35

RNA : Ribonucleic Acid	14, 50, 59, 70
ROOM : Regulatory on/off Minimization of Metabolic flux changes	29
rRNA : Ribosomal Ribonucleic Acid	52
SB : System Biology	1, 5
SBML : Systems Biology Markup Language	18, 22, 24, 50, 73, 74
SCIP : Solving Constraint Integer Programs	35
system biology	5
TC : Transport Classification	11, 27, 46
TCDB : Transport Classification Database	8, 9, 12, 27
TMHMM : TransMembrane prediction using Hidden Markov Models	27
TRIAGE : Transport Proteins Annotation and Reactions Generation	27, 45, 47, 57
tRNA : Transfer Ribonucleic Acid	52
TSV : Tab-Separated Values	34, 74
UniProt : Universal Protein Resource	8, 9, 13, 45
URL : Uniform Resource Locator	42, 54, 62, 77

CHAPTER 1

INTRODUCTION

The current chapter aims to present the context, motivation and goals of this thesis.

1.1 CONTEXT AND MOTIVATION

Metabolic engineering (ME) is a branch of the engineering that, through the modification of biochemical reactions, can improve cellular properties, such as increasing the production of a particular metabolite in the organism. This field plays a vital role in the manipulation of metabolic fluxes and is also focused on metabolic pathways that are all biochemical reactions steps that connect a specific set of input and output metabolites [1]. At present, many tools can be used to perform ME in different ways by using new advanced methodologies and Systems Biology (SB) tools, such as COBRA Toolbox [2], to improve strain optimization and predict cellular behavior.

Nowadays, the possibility of sequencing and performing automatic genome annotations, using SB tools, such as Rapid Annotations using Subsystems Technology (RAST) [3], allow creating network reconstructions at the genome scale, enabling the development of genome-scale metabolic (*GSM*) models for all organisms that have their genome sequenced [4]. *GSM* models represent the organism at the mathematical level and are used to predict the phenotypical role.

Computational tools are essential in the automation of specific tasks, such as genome annotation and metabolic network reconstructions. Data obtained from these tasks can be studied along with other sets of data from a different organism, for instance when the goal is to understand

the metabolic interaction between such organisms. Interactions between dependent microbial partners, which are called syntrophic relationships, are common in microbial communities [5].

Microbial communities have gained particular interest and are used for practical applications such as biorefineries, bioelectricity generation and bioremediation [6]. However, studying these communities has proven to be difficult due to the absence of experimental protocols and computational tools like the ones available for single organisms [7].

1.2 GOALS

The primary goal of this work will be the development of a *GSM* model of a syntrophic community of *Syntrophobacter fumaroxidans* (*S. fumaroxidans*) and *Methanospirillum hungatei* (*M. hungatei*). For this purpose, both *S. fumaroxidans* and *M. hungatei* *GSM* models will be reconstructed and later integrated into a single consortium model.

The main scientific/technological objectives were:

- Reconstructing a *GSM* model for *S. fumaroxidans*, using *merlin* to carry out the following tasks:
 - Generate an up-to-date, high quality functional annotation of the *S. fumaroxidans* genome;
 - Obtain a *GSM* draft network utilizing genome annotation;
 - Execute manual curation and refinement of the draft network using information obtained from literature or even by experimental data;
 - Build a stoichiometric model of the draft metabolic network;
 - Collect experimental data for refining and validation of the *GSM* model;
 - Test and validate the *GSM* model;
- Reconstructing a *GSM* model for *M. hungatei* using KBase and *merlin* to perform the following steps:
 - Obtain an up-to-date automatic functional annotation of the *M. hungatei* genome;
 - Automatically generate a *GSM* draft network using genome annotation;
 - Perform manual curation and refinement of the draft metabolic network utilizing information retrieved from literature and experimental data.
 - Develop the stoichiometric model of the draft metabolic network;
 - Collect experimental data for refining and validation of the *GSM* model;

- Test and validate the *GSM* model;
- Integrate both models into a metamodel, build a common objective function, and create compartment exchanges.
- Perform *in silico* simulations for metamodel validation.
- Analyze the genome and search for valuable information.

1.3 STRUCTURE OF THE DOCUMENT

This document is organized by the following structure:

- **Chapter 2 – State-of-the-art**
 - A revision of the tasks and methodologies associated with the reconstruction of *GSM* models.
 - An introduction about the computational tools currently available for reconstructing *GSM* models.
 - A brief explanation concerning syntrophic communities, including both *S. fumaroxidans* and *M. hungatei* metabolisms.
- **Chapter 3 – Material and Methods**
 - A detailed description of the tasks related to genome annotation.
 - Methods used to assemble both draft metabolic models including manual curation.
 - Biomass formulation.
 - Report on the tools and approaches used to troubleshoot the constructed model.
 - Description of the tools and strategies used to validate the assembled model.
- **Chapter 4 – Results**
 - Results obtained from the genome annotation tasks.
 - Results of both manual curations.
 - Models reconstructions issues.
 - Models imprecisions.
 - Description of the metabolic models for both *S. fumaroxidans* and *M. hungatei*.
 - Description of the reconstructed metabolic community model.
 - Succinct description for both organisms' networks characteristics.

- Chapter 5 – Conclusion and Future Work
 - Assessment for the single-species and community models.
 - Applications of the GSM

CHAPTER 2

STATE-OF-THE-ART

*This chapter presents the introduction to state-of-the-art methodologies and relevant computational tools. Furthermore, syntrophic bacteria metabolic landscape and studies of *Syntrophobacter fumaroxidans* and *Methanospirillum hungatei* are presented as well.*

2.1 GENOME-SCALE METABOLIC MODELS

2.1.1 Background

The human body is made up of many different systems, such as molecules, cells, genes, proteins, and regulators of proteins and they all interact in innumerable ways. The volume of information is overwhelming, and the language of communication between the interacting parts are unknown or only somewhere understood. To better treat diseases, the principles that govern the design, function, and interaction of these systems have to be well understood. In case of a disruption of the system caused by a virus, one of the main goals for scientists is to track the network of genes and proteins responsible for processing information in disease cells which involves monitoring hundreds of thousand genes and cells, and communication channels simultaneously [8]. Genes and proteins are always interacting with one another, so the first step in system biology (SB) research is to collect the data and then analyze it. Systems biologists need a background in math, a good understanding of physics and in-depth know-how in bioinformatics

to be able to analyze massive datasets. The synthesis of insights and genomics like the sequencing of the human genome has catapulted biology to a new level, enabling scientists to explore health and sickness in a brand-new way. Classic biology looks at individual parts that comprise the human body. SB, by contrast, takes a holistic approach a “bird’s eye” view and has shifted the lens through which biology is studied and understood. Therefore, systems biology quantifies the components and analyses the interactions between organisms to understand their organizations and to predict their behavior [9].

In the field of SB the study of the metabolism can be considered as an integrated study because of the data availability and the importance of applications. The volume of available data is so large that it allows the generation of high-quality models, which allow simulating the organism behavior when it is subject to different conditions [10]. Nowadays, it is easy to apply the knowledge of the metabolism, learnt for certain organism, to other organisms by performing phylogenetic similarity using bioinformatics tools such as the Basic Local Alignment Search Tool (BLAST) [11]. On the other hand, it is not easy to do the same for other functions such as transcription regulation and signaling, so there is a bigger variety of metabolic models constructed when compared with regulatory or signaling models.

The major challenge in SB is not only to generate high-quality models capable of mimicking the cells compartments but also to predict their behavior[12]. SB is closely related with industrial biotechnology processes, because the models generated in this field, together with a large number of bioinformatics tools, allow identifying genetic targets for increasing technical factors such as yields and productiveness [13].

Nowadays, the availability of whole-genome sequence for many organisms and high-quality knowledge of biochemical reactions present in several biological databases [13], allow the creation of metabolic networks at the genome scale. The *GSM* networks are a set of biological reactions from the enzymes encoded in the target organism’s genome. These networks allow determining the physiological and biochemical properties of the cells. Despite the contributions of the *GSM* networks only the *GSM* models can be used for predicting the capabilities of the metabolic system. For instance, *GSM* models are being used to predict, *in silico*, to identify potential candidate drug targets and to study the response of microorganisms to perturbations[14].

GSM models encompass the total metabolic potential encoded in the genome of an organism. The main steps for creating a *GSM* reconstruction include the creation of a draft

reconstruction based on the genome sequence, the identification (and amendment) of errors and gaps in the network, characterization of drains and the biomass equation and finally to validate the model with additional experiments [15]. The reconstruction of GSM models is supported by information available in several online databases [16, 17]. These online platforms provide genome sequence information, annotation data, and for some cases the functional capabilities of the proteins [18].

The reconstruction of a *GSM* model is a repetitive process in which the information retrieved from various data sources is collected and used for assembling a draft *GSM* network. The results from genome annotation are compiled and used to obtain the initial metabolic reconstruction. Then there is a search for errors in the system, and finally, the network is converted into a *GSM* model by adding an equation representing the biomass formation and other constraints. In the end, the model obtained must be validated through the comparison of experimental data with the *GSM* model (Figure 1).

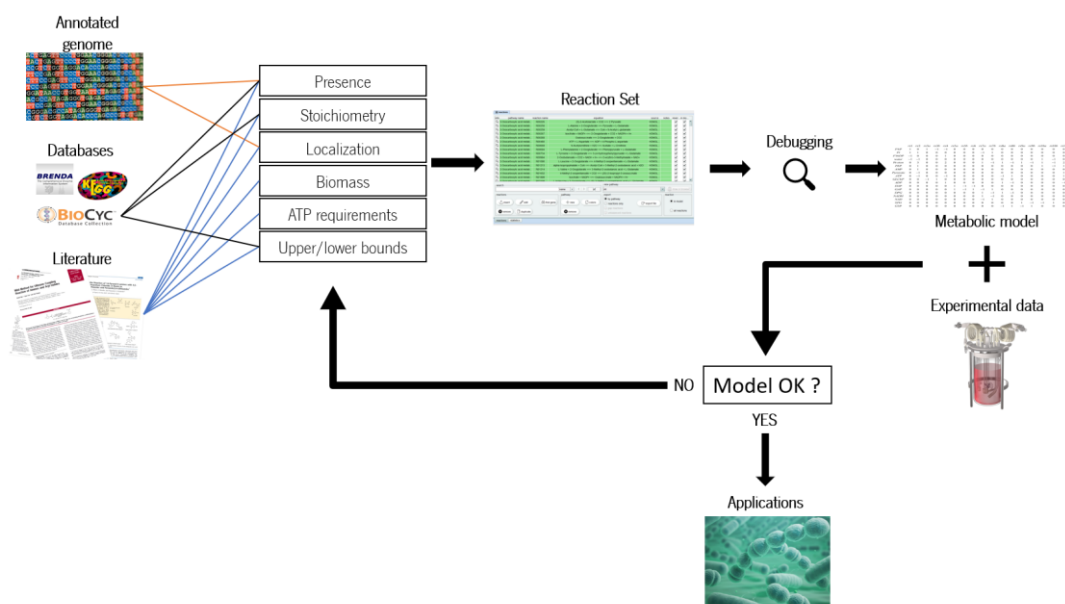


Figure 1. Illustration representing the iterative processes during metabolic reconstruction. Firstly, it starts with a complex compilation of the information available for the microorganism metabolism from different information sources. A reaction set is built, and debugging is performed to build a steady-state metabolic model. Next, the comparison of *in silico* simulation results with experimental data is made and when the last is in accordance with *in silico* predictions, the model is ready to be use in biotechnology applications. If the *in silico* predictions don't match the experimental results it is necessary to make a revision on the information sources. Adapted from [19]

2.1.2 Online Bioinformatic Resources

It is essential to have the most up-to-date available information about the organism being studied to have a reasonable reconstruction of *GSM* models. Several data are needed, such as well-annotated genome sequences, functional and molecular knowledge of enzymes or functional information of proteins present in membrane transport processes, biochemical and stoichiometric information of metabolic reactions. Several online databases are listed in Table 1. These databases are sources of the information mentioned above.

Table 1. Online Bioinformatic resources

Database	Acronym	WebAddress	Reference
Kyoto Encyclopedia of Genes and Genomes	KEGG	http://www.kegg.jp/	[20]
MetaCyc		http://www.metacyc.org/	[21]
BioCyc		http://biocyc.org/	[21]
Biochemical, Genetics and Genomic Models	BIGG	http://bigg.ucsd.edu/	[22]
Universal Protein Resource	UniProt	http://www.uniprot.org/	[18]
National Center for Biotechnology Information	NCBI	http://ncbi.nlm.nih.gov/	[23]
Transport Classification Database	TCDB	http://www.tcdb.org/	[24]
ModelSEED		http://modelseed.org/	[25]
Braunschweig Enzyme Database	BRENDA	http://www.brenda-enzymes.org/	[26]

One of the sources that can be used for the high-level understanding of a biological system is **KEGG**[20]. It is an online database resource of metabolic data, where the user can find an extensive collection of information about genes, metabolites, enzymes, reactions, and pathways. All the information is gathered from molecular-level data. First, the data is retrieved from genome sequencing and other high-throughput experimental technologies and then is coupled into large-scale molecular datasets [16].

Another database is **MetaCyc** which collects metabolic pathways from different organisms [21]. It is a curated database and has good quality information about every single organism.

Similar to MetaCyc, **ByoCyc** is a database that details each genome and metabolic pathway of an organism [17].

BIGG models is an online database with more than 75 well curated genome-scale metabolic network reconstructions and create patterned identifiers for metabolites named BIGG IDs [22].

The **UniProt** is a knowledge-base of functional information about proteins, that uses an accurate, logical and productive method to gather all the most critical data of proteomics [18]. This database is split into two branches: UniProt/Swiss-Prot and UniProt/Translated EMBL Database (TrEMBL). The first one relies on manually curated annotations extracted from literature or computational analysis. The second one provides unreviewed data waiting for manual curation.

One of the most useful databases in the reconstruction of genome-scale metabolic models is **TCDB** which gives transport proteins information [27]. In it, there is available functional information about transport proteins for a wide variety of organisms. The TCDB database establishes a classification system called the Transport Classification (TC) that has a vast range of information, such as structural, functional, mechanical, evolutionary and disease/medical data.

The **ModelSEED** online resource can be classified as an open platform with tools that allows the reconstruction, comparison, and analysis of metabolic models [25]. It is a curated database which contains mass and charged balanced reactions, standardized to aqueous conditions at neutral pH. This database integrates biochemistry included in KEGG, MetaCyc, EcoCyc, Plant Metabolic Networks [28], and Gramene [29]. It also presents access to biochemical reactions and genome annotations by having *GSM* models integrated.

BRENDA is a source of professional curated data for every enzyme [26]. Most enzymatic data can be manually retrieved in this online database. The literature search is the method used to classify each enzyme function. BRENDA database uses a classification system named Enzyme Commission (EC) to organize the available information.

2.1.3 Genome annotation

The first stage of reconstructing a *GSM* model is to generate a draft reconstruction based on the genome annotation of the target organism and biochemical databases. This automated reconstruction is a collection of metabolic functions encoded in the genome, and some of them are missing due to wrong or incomplete annotations.

The genomic information provided by the organism's genome is important to define the gene properties unambiguously. The draft reconstruction profoundly relies on the genome annotation, thus it is important to work with the most recent genome version available because this genome will contain all updates and corrections since the genome's original publication. Hence, the genome annotation stage is crucial to the reconstruction quality [30].

The metabolic genes identified in the genome annotation need to be retrieved by using keywords or gene ontology (GO) categories [31] during the generation of the draft reconstruction. The connection between metabolic reactions catalysed by the identified gene products and the draft reconstruction is achieved by using the EC numbers [32] and biochemical reaction databases such as KEGG and the Brenda database [33]. The list of candidates generated by genome annotation may contain many false-positives, for example, proteins involved in Deoxyribonucleic Acid (DNA) methylation also have EC numbers, but their functions are rarely considered in metabolic reconstructions[34]. Biochemical databases like BRENDA can be used to reduce the number of false-positives, but this strategy does not replace manual curation [35].

As mentioned before, the quality of the curated genome annotation is directly correlated with the quality of the reconstructed model, and for some instances, the reannotation of the previously annotated genome may be required [30]. The genome reannotation consists in the process of looking for specific data such as Open Reading Frames (ORF) names, product names, and EC numbers[36]. The metabolic genes, the ones encoding for enzymes and transport systems, are the only set of genes required for the development of *GSM* models [37].

2.1.4 Metabolic Network Assembly

In this stage, the draft reconstruction will be re-evaluated and refined. It is the second stage of the reconstruction process, and it is responsible for the curation and refinement of the network content [30]. The metabolic functions and reactions retrieved from the draft reconstruction need special attention, so they are individually evaluated against organism-specific literature. This process is called manual curation, and it is important because not all annotations have a high-confidence score, and biochemical databases are mostly organism-unspecific, thus listing enzymes activities found in various organisms, and maybe not all of them are present in the target organism. The inclusion of organism-unspecific reactions can deeply affect the predictive behavior of *GSM* models [34].

2.1.4.1 *Genes, Proteins and Reactions Associations*

The reconstruction of a metabolic model consists of performing a number of tasks. One of the tasks is to “build” gene-protein-reaction (GPR) association where all annotated metabolic genes are linked to proteins and reactions. The TC and EC numbers are assigned during genome annotation give information about enzymatic and transport reactions, respectively.

The classification system used by BRENDA (EC number) as previously mentioned is responsible for classifying the enzymes by their functions [38]. This system uses a four-digit code to order the enzymes by the chemical reactions they catalyze. By using this classification, the enzymes are categorized into seven categories, transferases, oxidoreductases, hydrolases, ligases, lyases, isomerases, and translocases. The first digit of the four-digit code represents the category of the enzyme. The following numbers will progressively restrict and specify the enzyme classification.

The classification system used for membrane transport reactions (TC number) is similar to the EC system, and it also specifies the protein regarding phylogenetic information [24]. Instead of the four-digit code used in the EC system, the TC classification system uses a five-element code: four digits and one letter. There are seven main classes of membrane transport proteins, namely, primary active transporters, accessory factors present in transport, channels, incompletely characterized transport systems, group translocators, and electrochemical potential-driven transporters. The first digit in the TC classification system like in the EC digit code also represents the transporter class they belong. Next, to the first digit, there is a letter and after it come the remaining three digits.

The GPR association information often comes for the genome annotation, and these associations allow to associate genes to reactions [35]. This step includes determining (i) if the functional protein is a heteromeric enzyme complex, (ii) if the enzyme complex can promote more than one reaction and (iii) if distinct proteins have the same function (i.e., isoenzymes).

Regarding the first case, the genome annotation has refined information, which suggests that at least one more subunit is required for the formation of the protein complex. Databases, such as KEGG, list the subunits for these protein complexes in some cases. Most of the times, a more comprehensive database like TCDB [24] and literature is required. For the second case and third case, information can also be retrieved from biochemical databases and literature [16]. The correct assignments in the GPR associations will be directly related to results of *in silico* gene deletion studies.

The information retrieved from databases such as KEGG, BRENDA, and TCDB when crossed with genome annotation information generates a reaction set with all the data compiled (Table 2).

Table 2. The table below lists the sources and the respective information they provide while creating the *GPR* associations.

Source	Information
<i>KEGG</i>	Names or identifiers of the reactions; Reactants and products of the reactions; Equation of the reaction
<i>BRENDA</i>	EC number; names of enzymes
<i>TCDB</i>	TC number; names of membrane transport proteins
<i>Genome annotation</i>	Names or identifiers of the genes

The next step is to complete the reaction set with exchange, non-enzymatic, spontaneous and reactions that occur in the organism. In the case of missing genes for some of these reactions, a literature search must be performed [39].

The draft *GSM* network is complete when the first set of reactions is put together. As the name indicates it is just a draft and for that reason has many imprecisions. For this reason, the *GSM* network may have some false positives like, having enzymes involved in the nucleic acid metabolism and signal transfer that generally are not used in the *GSM* network assembly [40].

2.1.4.2 *Spontaneous reactions*

The spontaneous reactions to be added to the reconstruction must have all metabolites connecting them to the network. This approach will avoid dead-end metabolites caused by spontaneous reactions [30]. These reactions can be found in literature or online web sources, such as KEGG.

2.1.4.3 *Reaction stoichiometry*

Metabolites present in databases are usually listed with their uncharged formula, but in medium and in cells, many of them are protonated or deprotonated. The pH of interest will influence the charged formula by modifying the protonation state [41]. Depending on the environmental conditions and target organism the pH value may vary so that the same metabolites can have different charges in different organisms.

After determining the charge for each metabolite, the reaction stoichiometry can be determined by counting the different elements on both sides of the reaction. In some cases, water and protons may be required to balance the total charge of the reaction. Notice that unbalanced reactions may lead to the synthesis of protons or adenosine triphosphate (ATP) out of nothing [42].

2.1.4.4 *Localization and Compartmentalization*

The localization of enzymes inside compartments or outside the cell is vital for the development of *GSM* models, as it identifies the organelles in which the enzymes operate. The cellular localization of proteins based on nucleotide or amino acid sequences can be determined using algorithms such as *PSort* [43] and *TargetP* [44]. The information retrieved from these databases is then used to determine the compartments of the metabolites. Other databases like UniProt can also contain information about the localization of enzymes and reactions, which can be useful to constrain the reactions to a compartment. By default, enzymes are assumed to be in the cytosol. Additional gaps in the network can be formed as a result of incorrect compartmentalization and consequently lead to misrepresentation of the network properties.

The distribution of reactions among different compartments must be performed when assembling the metabolic network reconstruction. The same metabolite may be present in similar reactions, and these may be located in different places inside the cell. For that reason, the name and identifier must identify the metabolite localization with the respective compartment.

While identifying the localization of the metabolites, the extracellular location should be included. This location will resemble the extracellular space and exchanges reactions that need to be integrated for reactants and products that are located outside the cell [4].

2.1.4.5 *Manual Curation*

Although automatic methods used in automatic draft reconstruction are very useful, these methods are fallible [45]. It will lead to an incomplete draft reconstruction because it will have missing reactions and it may contain reactions irrelevant to the *GSM* model [37]. The revision of literature, like publications and textbooks, organism-specific databases, for the validation of the reactions, is the last step of the *GSM* model reconstruction. It is a slower method but very important to the final quality of the model. In manual curation step, each reaction of the model should be confirmed.

One strategy used by researchers for model refinement is to analyze every pathway by a particular order [40]. The analysis should start in the central pathways and end in the secondary pathways. All the *GSM* network characteristics and reactions properties should be deeply examined. For instance, routes for known carbon, nitrogen, sulfur and phosphorous sources to the biomass' precursors must be present and have to be cleared of gaps.

Errors can occur while building the reaction set. As an example, some reactions may not be included due to ambiguous identifiers, or they do not exist in that specific organism [37].

The removal of generic terms is important since these metabolites are often created due to the lack of information about them. A few examples of these terms are DNA, Ribonucleic Acid (RNA) and protein. By removing these from the dataset, the quality of the reconstruction will improve.

Literature research is fundamental when performing a manual curation of the network model. For instance, each organism uses specific substrates or cofactors besides organism-unspecific databases indicating they have a wide range of possibilities. Information about substrate and cofactor usage must be conferred if available.

To achieve a high-quality network reconstruction the model must be stoichiometrically and massed balanced. Each reaction in the dataset must be balanced, which means that the total number of elements (i.e., atoms such as, carbon and hydrogen) should be equal in each side of the reaction. If not, the flux distribution of the *GSM* network will become impaired, thus reducing the quality of the model because these unbalanced reactions will be blocked [4].

Regarding the network model, the last step is to identify missing metabolic functions in the reconstruction, the so-called network gaps [46]. These gaps are a consequence of the iterative process of the *GSM* model reconstruction, specifically, by repeating the second and third stage partially. After the manual curation, it is recommended to perform a first gap analysis because it will reduce the number of "bugs" in the model [30]. To fulfil these gaps, a manual gap-filling analysis should be performed. Web services such as KEGG are used in the process of manual gap-filling because they provide an extensive collection of manually drawn pathway maps which represent molecular interactions and reaction networks. On the other hand, some algorithms can be used to turn gap-filling in an automatic process [47].

2.1.5 Stoichiometric Model Assembly

In this stage, the network is converted into a mathematical format, and most of the process can be easily automated. At this time, systems boundaries will be defined, thus turning the *GSM* network into a condition-specific model by providing constraints. Errors in a simulation are, most of the times, a direct result of the constraints not being correctly set [37]. The reversibility of the reactions are the major factors responsible for the main restrictions that should be added to the model. An irreversible reaction is the one that is not constrained, whereas irreversible reactions are limited between the minimum/maximum (depending on the direction of the reaction) flux and zero.

The first thing that should be done before converting the *GSM* network into a model is to add an equation representing the biomass formation to the reactions set. That equation represents the macromolecular composition of the cell, and the reactants are macromolecules and other smaller biomass units named building blocks [48]. For the organism to produce the larger precursors (macromolecules) reactions that represent the assembling of the building blocks must be present in the *GSM* network.

The biomass equation can be expressed as:

$$\sum_{k=1}^P C_k \cdot X_k \rightarrow \text{biomass} \quad (1)$$

P – number of biomass constituents

C_k – Coefficient of a metabolite

X_k – Metabolite

The growth or specific growth of an organism (h^{-1}) is represented by the flux that is associated with the biomass reaction, and it is normalized to 1 gram of biomass. The unit representing the growth express how many grams of biomass are produced per gram of biomass, already in the medium, per hour [19].

Using organism-specific literature to retrieve information about the biomass composition is one of the strategies used by researchers, while reconstructing the *GSM* network. Another method is to experimentally study the organism to achieve a detailed biomass composition [49]. If it is not possible to perform one of the above strategies the biomass composition known for a strictly phylogenetic relative must be used.

All organisms require energy to survive. Thus, the biomass equation should include growth-associated energy requirements in terms of ATP molecules per mass of biomass synthesized. In general, the growth-associated energy is reflected in the biomass reaction as the hydrolysis of ATP into Adenosine Pyrophosphate (ADP) and orthophosphate.

Usually, chemostat growth experiments are used to retrieve information about energy requirements [40]. Otherwise, if such data is not available, organism-specific literature or closely phylogenetic relative organism's data should be used. Additionally, a manual estimation of the growth-associated energy can be done by determining the energy required for each macromolecular synthesis [19].

As mentioned before the flux associated with the biomass reaction represents the growth rate of an organism. The *GSM* network should also contain non-growth energy requirements information. The same strategies used to determine the energy requirements for growth are used to calculate the non-growth conditions. This energy is in charge of cell maintenance (cell maintenance energy), more precisely in controlling the membrane potential [50].

2.1.5.1 *From Network Reconstruction to Mathematical Model*

The conversion of the reaction set into a stoichiometric matrix is one of the last steps in the *GSM* network reconstruction. At this stage, the classical principles of chemical engineering are used to define the dynamic balances of a metabolite. For this purpose, differential equations are created for every single metabolite present in the metabolic network [48].

Equation 2, represents the behavior of the concentration of a metabolite throughout time:

$$\frac{dX_i}{dt} = \sum_{j=1}^N S_{ij} v_j + \mu X_i, \quad i=1,\dots,M \quad (2)$$

X_i - Metabolite i concentration

V_j - Rate of reaction

S_{ij} - Stoichiometric coefficient of the metabolite i in the reaction j

μX_i - Growth rate

As mentioned above, the method of associating each metabolite in the network with a dynamic mass balance will generate a set of differential equations. However, the kinetic expressions and parameters data is insufficient, and as a consequence, it is only possible to simulate dynamic conditions to specific pathways of the organism metabolism [19]. To solve this

data gap, usually, a steady-state approximation is applied to the network, reducing the mass balances to a set of linear equations. If a steady-state paradigm is confined to the model, it means that the concentration of a metabolite throughout time will remain constant. In the end, these set of linear equations [51] are converted into a stoichiometric matrix. The matrix should also contain the exchange fluxes.

Equation 3, steady-state representation of equation 2

$$S * v = 0 \quad (3)$$

v – Flux vector

S – Stoichiometric matrix; reactions are represented in columns and metabolites in rows.

One of the problems of a stoichiometric model is that the number of fluxes is greater than the number of mass balance constraints. For this reason, a stoichiometric model is considered to be an underdetermined system, because there is an infinite number of flux distributions that accomplish the mass balance constraints and it is impossible to achieve a single solution out of it [52]. This is generally classified as the null space of the matrix S [53].

To reduce the space of potential solutions for the system, constraints are used. Using this methodology, it is possible to retrieve a set of feasible solutions and giving each one a set of conditions.

The constraints are defined by imposing limit values to the flux of each reaction. As a result, both upper and lower bounds are added as model constraints of the reactions fluxes. Most of these constraints can be added to the model as inequalities' as shown in equation 4 below.

$$\alpha_j \leq v_j \leq \beta_j, \quad j=1,\dots,N. \quad (4)$$

α_j - Lower bound

v_j - Flux vector

β_j - Upper bound

The definition of reversible reaction in a stoichiometric model implies that the bounds are set from minus to plus infinity. Differently, irreversible reactions bounds must be configured with minus or plus infinity to zero, counting on the directionality [19].

After the addition of the biomass equation and the non-growth ATP requirements to the network model, all reactions can be represented in the form of a stoichiometric matrix. In Figure

2, it is described the conversion of a metabolic network with six metabolites and ten fluxes into a stoichiometric matrix. This mathematical model can be used to predict the dynamic behavior of the metabolite concentration, by performing dynamic balances with ordinary differential equations [54]. In the end, the mathematical model created is advised to be saved in the Systems Biology Markup Language (SBML) standard format [55].

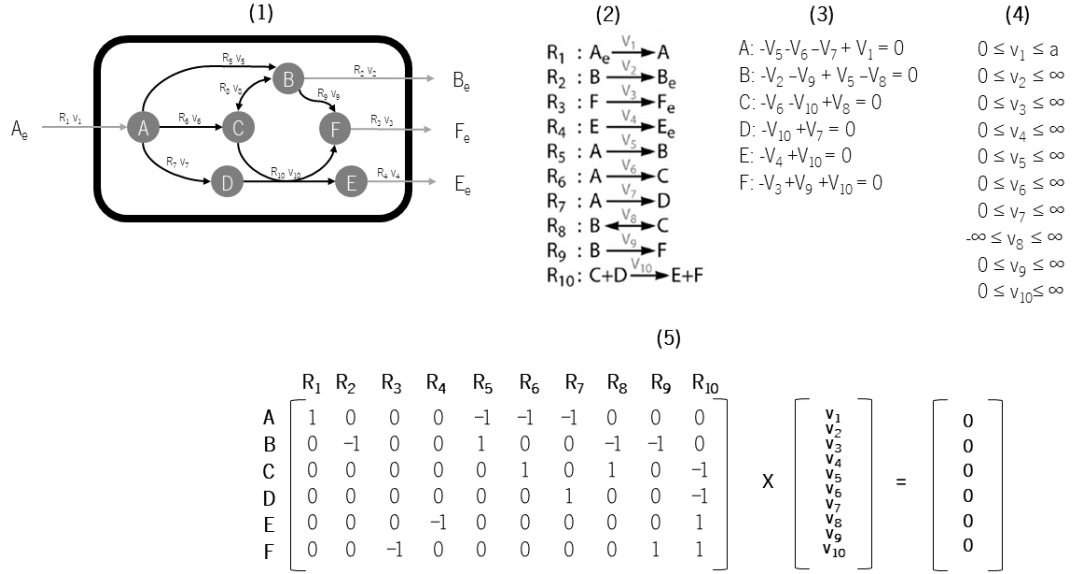


Figure 2. Example of a metabolic network with 6 metabolites (A to F) and 10 fluxes (V1 to V10). A reaction scheme is presented in (1) with outlined system boundaries. The fluxes V1 to V4 represent the exchange fluxes of metabolite substrate (A) and products (B, F and E). The reversible reactions are shown with double arrows, and the irreversible reactions are identified with a forward arrow. In section (2) is presented the stoichiometric of the network. Section (3) represents the steady-state mass balances, and section (4) shows the constraints around the flux values. The final section (5) illustrates the representation of the mass balances in matrix format.

2.1.6 Stoichiometric Model Validation

The final stage of the *GSM* model reconstruction process is to validate the model generated by verifying the network, evaluate it and finally validate it. One of the main tests consists in studying the ability of the model to synthesize biomass precursors, such as amino acids and lipids.

The validation step consists in predicting the behavior of the target organism and compare it to experimental data. If the predictions are not in agreement with experimental data, the previous stoichiometric model should be further reviewed. At this stage, all the errors made while performing the reconstructing process will be reflected in the model predictions. Moreover, the *GSM* model will only be ready when the predictions match the experimental data.

Using high-throughput growth phenotyping data, retrieved from the use of bioinformatics tools own by companies, such as Biolog Inc. [56], will allow the comparison between the simulated results. This method enables the model to be tested for growth in many limiting substrates and the results compared with the high- throughput data.

One of the methods used in the *GSM* model validation is the quantitative evaluation of growth rate and by-product formation of the organism. If the target organism has slow growth, it could mean that at least one of the precursors of the biomass function cannot be synthesized sufficiently. Depending on the precursor, the model's biomass production is limited by either carbon, nitrogen, oxygen, sulphur, or phosphate.

A different approach for the *GSM* model validation is to perform an analysis of false positive and false negative predictions, helping to refine the network content further if the information is available or instigate new experimental studies. Single gene deletion phenotypes accomplish this approach, and this deletion simulated in the *GSM* model may replicate the experimental data, if not, an examination of the genome annotation should be performed. For instance, if there are inconsistencies between the results obtained from *GSM* models' predictions and experimental results, the model should reformulate from the second stage ahead.

One technique used by researchers to compute a single solution for the *GSM* model is to perform a Flux Balance Analysis (FBA) [57], which relies on linear optimization to determine the steady state reaction flux distribution. This method seeks to minimize or maximize an objective function represented in equation 5, where C represents a vector of weights indicating how much each reaction contributes to the objective function.

$$Z = C^T * v \quad (5)$$

Z - Linear objective function

C^T - Vector of weights

v - Fluxes

The FBA method was found to be very useful in many reconstruction stages, such as gap filling, model validation and model refinement [58].

Usually, the objective function used is the maximization of biomass formation. For the majority of the organisms, the maximization of growth rate is what most researchers seek [1]. It has been proved that an organism tends to maximize growth rate and biomass formation when restricted to limitations of the carbon source or gene deletions [59]. Moreover, other objective functions can be used, and a few examples are the maximization or minimization of ATP production or maximization/minimization of a specific metabolite production [60].

To conclude, a final validation of the *GSM* model should consider all the decisions that were taken in the manual curation step, and the model will only be ready when the prediction behavior match the experimental data.

2.2 RELEVANT BIOINFORMATICS TOOLS

It can be hard for biologists to handle large amounts of data collected during research, as these have to be analyzed with advanced mathematics and computation methods [61]. Bioinformatics is extremely useful and offers several tools to perform analysis of genome-scale datasets. The genome annotation step could take a long time to be completed, but it may be automated using bioinformatics tools that use biological databases such as KEGG the process may be hastened. In this line of view, this section presents some bioinformatics tools that can be useful in the generation of *GSM* models or tools that have methods which use these models to model the organism behavior.

2.2.1 Cobra ToolBox

The Cobra ToolBox 2.0 [2] is a bioinformatics tool that researchers can use to predict a variety of metabolic phenotypes using *GSM* models, and it has been used over the past decade in several fields [62–64]. The Cobra ToolBox is released as a MATLAB® package for implementing COBRA methods. The name COBRA stands for CONstraint-Based Reconstruction and Analysis, and this software has been applied to the field of microbial metabolic reconstructions with success [65]. This software is used as a guide in metabolic pathway engineering to model pathogens [66] and host-pathogen interactions [67] and is also used to study the effect of diseases in the human metabolism [68]. Often, it is used for modelling, analyzing and predicting a variety of metabolic phenotypes using genome-scale biochemical networks.

The COBRAtoolbox 2.0 core pipeline is to employ physicochemical, data-driven, and biological constraints to reach the possible phenotypic states of a reconstructed biological network (Figure 3). The methods available in COBRA may not provide an optimal solution, but they can provide a reduced set of solutions that can be used by researchers to formulate a biological hypothesis.

The *GSM* networks in COBRA are created using knowledgebase such as BIGG [69]. This database uses manually curated genomes' annotations that relate biological functions to the genome, by using the information of gene-protein-reactions [70]. The BIGG strategy has shown good results when applied to metabolism, and several *GSM* models' reconstructions are available, for various organisms [71, 72].

The COBRA toolbox 2.0 provides the user with access to different methods such as FBA, gene essentiality analysis, gap filling[73], metabolic engineering[74], and visualization of computational models of metabolism.

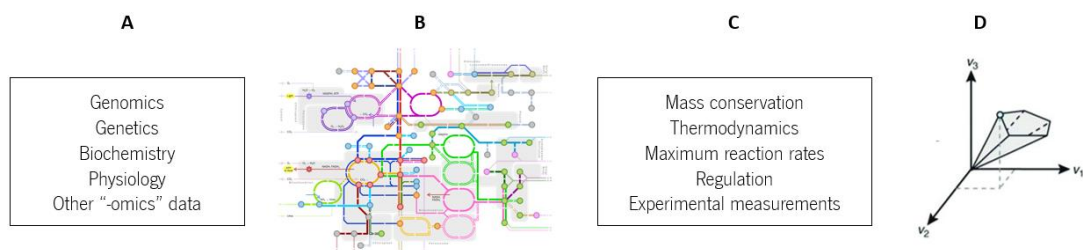


Figure 3. CONstraints-Based Reconstruction and Analysis of biological networks: (A) -Data sources for network reconstruction; (B) -Network reconstruction; (C) - Application of constraints to the network; (D)- Analysis of the network model in order to achieve an optimal solution. Adapted from [2].

2.2.2 Pathway tools

Pathway Tools software provides query, editing and visualization operations for pathways and genome databases, for EcoCyc and MetaCyc databases.

MetaCyc is a vast metabolic-pathway database for many organisms. This database defines pathways and enzymes for different organisms. EcoCyc is a Pathway/Genome database focused on a specific organism named *Escherichia coli* (*E. coli*). Pathway Tools is a bioinformatics tool that enables the user to curate and computes a genome annotation. This tool can organize objects that share similar properties and attributes within classes.

MetaCyc database contains information about metabolic pathways, reactions, enzymes, and substrate compounds. This tool serves as a reference for computational prediction of metabolic pathways for organisms that have their genome sequenced. The pipeline of MetaCyc is to gather information about pathways that had been reported in the literature, and then label the pathway with the organism(s) wherever that pathway is known to occur. One disadvantage of MetaCyc, when compared with EcoCyc, is that the metabolic maps are not so well studied as the metabolic model for *E. coli* [17].

2.2.3 RAVEN

The software name RAVEN[75] stands for Reconstruction, Analysis, and Visualization of Metabolic Networks. Like the COBRA Toolbox mentioned above, this tool also runs in MATLAB®. It can perform reconstruction, analysis, simulation, and visualization of *GSM* models. The software imports and exports the information in two formats, which are: SBML format and a Microsoft® Excel model representation. These formats allow a bigger annotation of model components, such as databases for reactions and genes. Models from COBRA Toolbox format can also be imported.

RAVEN Toolbox is capable of creating a general network, generate functional models, assign sub-cellular localization, use user-defined models, integrate gap filling, running offline, and *GSM* models visualization [75]. The mains focus of the software are:

- Perform the automatic reconstruction of *GSM* models, based on protein homology
- Network analysis, modelling, and interpretation of simulation results
- Using pre-drawn metabolic network maps, visualize the *GSM* models

This software uses both manual and automatic curation for generating draft models. The automatic curation relies on the KEGG database for the automatic identification of new metabolic functions not included in the manually curated model. Hence, RAVEN initially analyses the manually curated model and then uses online databases to find further information [75].

2.2.4 GEMSiRV

The Genome scale Metabolic model Simulation, Reconstruction and Visualization (GEMSiRV) [76] is another bioinformatics tool that can be used to reconstruct metabolic networks and provide easy editing, visualization and perform flux balance analysis of the generated models. This tool can be used for the *GSM* model's reconstruction, computational studies, display, and manual curation.

It is free software, able to run on the user local computer/server, decreasing the restrictions related to data size or internet speed. Another advantage of using GEMSiRV is that it enables import and manual curation of existing models, thus making it very interesting for *GSM* network reconstructions. This tool also has an interactive interface which allows the user to perform gap filling and to visualize the changes in the network further.

GEMSiRV has three main modules:

- The metabolic network reconstruction model – involves the tasks of model importing data and editing, construction of the references databases, draft reconstruction and model refinement.
- Simulation model – relies on dead-end metabolite identification, objective optimization, FBA, robustness analysis, gene/reaction essentiality analysis, and gene deletion analysis.
- Visualization model – GEMSiRV provides an interface for editing and visualizing the metabolic networks of interest.

GEMSiRV software can accelerate the development of biomedical applications of metabolic reconstructions [76]. The projects-in-progress using this tool can be easily shared between researchers and therefore aids the share of information exchanges in the researchers' community to get high-quality *GSM* models

2.2.5 SuBliMinaL Toolbox

The SuBliMinaL Toolbox[77] consists of software capable of automating the steps of *GSM* models reconstruction and analysis. This tool has several independent modules that can be used independently or chained together to plan a reconstruction workflow allowing the generation of an initial draft of a metabolic reconstruction. The reconstructions are generated in the SBML format. The COBRA Toolbox can use the models created by SuBliMinaL Toolbox because of the software creates a generic biomass function [77]. Thus, allowing to perform constraint-based analyses of the model by using techniques such as FBA.

This JAVA™ software uses web-services such as KEGG and MetaCyc to retrieve the required biochemical data. The models used to extract the information from the online databases frequently generate formatted models representing the union of all metabolic pathways detailed in each resource. These models can be merged, and their annotations can be used in other modules.

The models used for annotation are dependent on the initial draft. The draft model needs to have unambiguous identifiers according to the Minimal Information Required In the Annotation of Models (MIRIAM) standard [78].

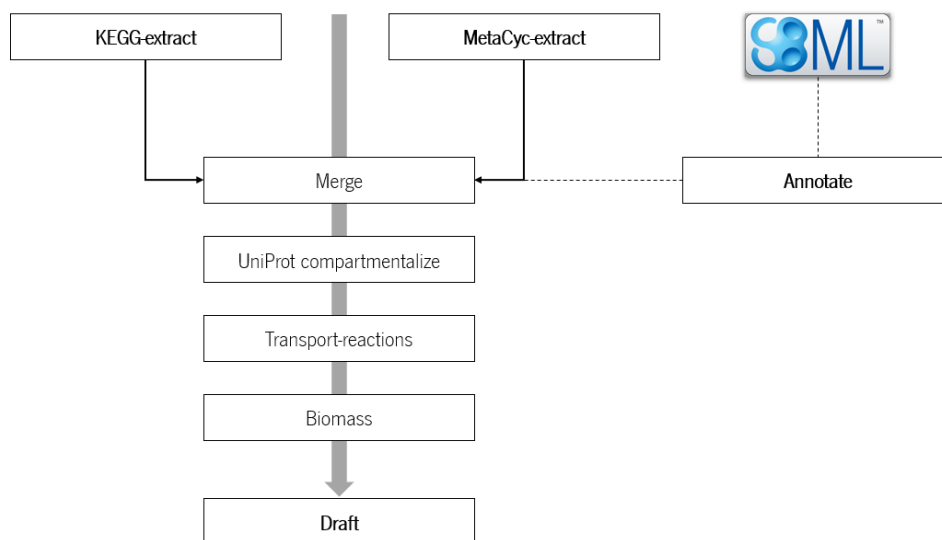


Figure 4. Flow diagram of the chained modules present in SuBliMinaL Toolbox working together in order to form the draft model. The names of the boxes refer to individual modules of the tool. The grey right-hand branch signifies that existing reconstructions or individual pathways can be added to the pipeline of work allowing for the generation high-quality drafts. Adapted from [77]

2.2.6 *merlin*

merlin is an application fully implemented in Java™ that uses genome-scale information to reconstruct high-quality, large-scale metabolic models, for any organism that has its genome sequenced. *merlin* can identify and annotate transport protein-encoding genes and assemble transport reactions for those carriers. Also, GPR rules are automatically created and included in the *GSM* model. *merlin* allows the users to visualize the *via* of the preliminary biochemical network, using KEGG pathways, and to curate the network manually.

This software can accelerate the reconstruction process by performing an optimized genome re-annotation while allowing the local manual curation without the need for commercial software. *merlin* uses MySQL® or H2 relational databases for the local data repository and uses Java libraries to access web services.

merlin has two main independent modules [4], namely:

- Internal model database –used to retrieve and load an initial set of metabolic data such as metabolites, enzymes, and reactions into *merlin*'s internal database, which enables the assembly of the *GSM* model draft.
- Annotation module, subdivided into three sub-modules:
 - Enzymes annotation – is the submodule responsible for the annotation of enzymatic functions to proteins encoded in the genome using homology search tools such as BLAST and HMMER [79]. Data retrieved for each homologous gene identified in the similarity search is processed individually, and the following data are retrieved: locus identifier, expected value, score, and organism.
 - Transporters annotation – these reactions are often only included in models if there are pieces of evidences supported in experimental data or literature. However, this approach usually originates a minimal number of transporters and does not allow performing GPR associations, as often the associated gene is unknown. The technology used to fix the problem automatically annotates carriers with TC family numbers and generates transport reactions for all metabolites transported by these carriers.
 - Compartments prediction module: LocTree performs the determination of the proteins' localization in eukaryotic organisms whereas prokaryotic organisms may use LocTree 3 [80] as well as PSORTb 3.0[81].

2.2.6.1 Enzymes Annotation

One of the advantages of using *merlin* for *GSM* model reconstruction is to have a complete framework capable of retrieving enzymatic, transport, and localization information, in a semi-automated way. *merlin* framework can be used for both prokaryotes and eukaryotes. The data output from the tools used for enzymes annotations, namely, BLAST and HMMER can be worked by *merlin* in order to retrieve relevant information (Figure 5) [4].

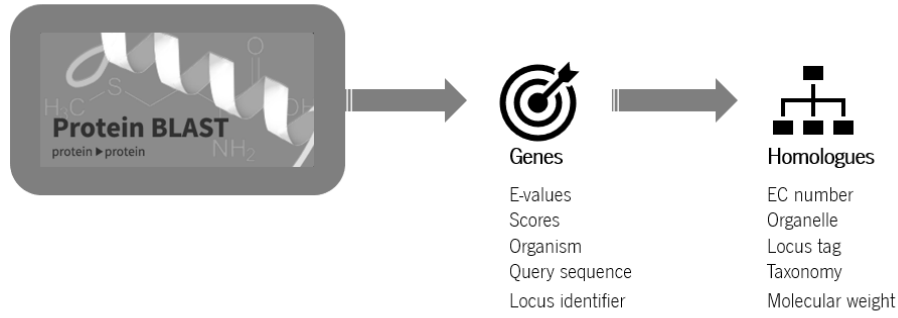


Figure 5. Enzyme annotation data collected by *merlin* after BLAST. The most relevant data for each gene present in the genome *fasta* file is stored. Also, data for every homologue identified for a certain gene is saved [4].

For each gene resulting from enzymatic annotation a numeric confidence score is calculated. The score will help to choose the product name and the EC number that should be given to every gene. This score, also referred to as annotation score, varies from 0 to 1 and it is calculated for each EC number and product name, as demonstrated in equation 6. The criteria to select the EC number and product name for each gene rely on choosing the one with the highest score. A parameter α controls both the weights of frequency and taxonomy.

$$\text{score}_{\text{annotation}} = \alpha * \text{score}_{\text{frequency}} + (1-\alpha) * \text{score}_{\text{taxonomy}} \quad (6)$$

For a given gene, the frequency score ($\text{score}_{\text{frequency}}$), calculates the number of occurrences of certain EC number among all homologues of that gene. The taxonomy score ($\text{score}_{\text{taxonomy}}$) is defined by most common taxonomy data between the first n homologues identified by the blast and the target organism. Additionally, the alpha value is the variable that allows *merlin*'s users to choose which parameter has more influence in genome annotation.

2.2.6.2 *Transporters Annotation*

Usually, most of the transport reactions present in a *GSM* model are identified using experimental data and literature. The Transport Proteins Annotation and Reactions Generation (TRIAGE) [82] is a tool present in *merlin*'s framework that allows the user to identify and classify all potential membrane transporter proteins of a target genome. The tool will automatically generate transport reactions for detected metabolites. Moreover, the data collected can be integrated into the *GSM* model.

TRIAGE has a database of transport reactions and the information needed to create them was retrieved from TCDB. As mentioned before, the TCDB database associates each transporter protein with a unique TC number. The transport reactions associated with genes are usually catalyzed by proteins present in membranes [83]. For this reason, all proteins identified with transmembrane domains are plausible candidates to potential transport systems, including transport reactions of genes. The tool used by TRIAGE for the identification of transmembrane proteins is the TransMembrane prediction using Hidden Markov Models (TMHMM) [84] tool. The methodology used by TMHMM is to search in the target organism genome for genes encoding proteins with transmembrane helices [82]. After, the genes that could encode transmembrane proteins are identified and aligned using the Smith-Waterman (SW) algorithm for local alignments, thus improving the sensitivity when looking for homologous sequences. Furthermore, all the genes are associated with the TC family numbers that they match and a wide range of metabolites. TRIAGE selects the TC family number and all metabolites associated with each gene using the same method mentioned in the enzymatic annotation (Equation 6).

The final candidates' annotations for each gene provided by TRIAGE contain a TC family number and all metabolites that may be transported. After TRIAGE have finished building the database, it can create transport reactions by getting the information needed from its internal database [82].

2.2.6.3 *GSM Model Assembly*

Later, all the information collected in these modules is uploaded into *merlin's* database. After the compartmentalization module, the data collected during genome annotation is integrated into a draft *GSM* model. *merlin* also allows the user to perform manual and automatic curations using several GUIs.

After the enzyme's annotation *merlin* can combine this information with more information retrieved from KEGG metabolic data and create a draft metabolic model containing all the essential reactions to assemble a stoichiometric model. The reactions from the draft network can be visualized and edited in a reactions view. Additionally, the user can curate in a "freeway", meaning that the model can be manually modified by adding or removing reactions to the network or from the network. Users can also remove entire KEGG pathways from the draft metabolic network. A KEGG pathway is a set of reactions linked by matching metabolites. One particularity of these group of reactions is that they are associated with genes and EC numbers by KEGG Orthology (KO).

All the information gathered by *merlin* tools such as transport protein annotations, transport reactions, and GPR associations is going to be integrated into a draft model. The user, after the integration, can use *merlin's* tools presented in Table 3, to edit the draft model generated manually.

Table 3. List of all available tools in *merlin's* repertoire for curation of draft metabolic networks.

Tool	Function
blocked reactions	Identifies unconnected metabolites
unbalanced reactions	Identifies possible unbalanced reactions
drains (create)	Automatically creates drains/exchange reactions
Gene-protein-reaction rules	Generates GPR rules using KO's information
e-biomass equation	Generates biomass reaction using genome sequencing data
correct reversibility	Corrects the reversibility of reactions in the network
Export	Allow to export at any time the stoichiometric model in the SBML format, export genome files and reactions.

2.2.7 *OptFlux*

OptFlux[85] is another bioinformatics tool that can aid in metabolic engineering tasks, such as Metabolic Control Analysis (MCA), by using available models of metabolisms together with mathematical tools or experimental data to identify, for example, targets for genetic engineering [86]. This tool is implemented in Java™ language. *OptFlux* is divided into four main modules and they are (i)-Model Handling, (ii)-Simulation, (iii)-Optimization and (iv)-Pathway Analysis.

Regarding the first module, *OptFlux* provides the user with several operations to visualize, import and export stoichiometric metabolic models, containing reactions, metabolites, equations and, if available, gene-reaction associations. The user can also upload models either from text files containing the lists of reactions, metabolites, the stoichiometric matrix, from files in SBML format or text files following the Metatool [87]. The model should identify external metabolites and biomass formation reactions from the input files based, for example, on explicit information. The user can later validate or edit this information.

The second module corresponds to the metabolic phenotype simulation methods implemented in *OptFlux*. One method is to use algorithms to calculate the values for the fluxes over the whole set of reactions in the model. The results obtained for this model include flux values and net conversions. In the simulation module, the user can define specific environmental conditions. *OptFlux* provides the user to perform simulations using three different methods: FBA, Minimization of Metabolic Adjustment (MOMA) [88] or Regulatory on/off Minimization of Metabolic flux changes (ROOM) [57].

In the third module, the user can identify sets of reactions deletions that maximize a given objective function connected with an industrial objective. All the algorithms implemented in this module are used so the user can identify genetic modifications that force the microorganism to produce a specific metabolite, but at the same time continue to maximize the biomass production.

The fourth module, Pathways visualization, includes visualization which allows the user to have a graphic visualization of the pathways of the model. This tool associate's numerical values to the different types of nodes (i.g., metabolites, enzymes, reactions) and edges (connections). Thus, allowing the visualization of the metabolic network overlapped by the values of the fluxes obtained in a simulation.

2.2.8 KBase

The United States Department of Energy (DOE) Systems Biology Knowledgebase (KBase) is a software and an online data platform that provides the user with many resources for analyzing public data together with their experiments, thus enabling a better understanding of the results they obtained. Using the KBase platform is possible to get biochemical information about the genome being studied. Information's, such as biochemical species (compounds), reactions, roles, media, metabolic maps, and metabolic pathways, are given to the user if he uses the annotation applications provided by KBase. One of the functionalities of this software is to predict and design biological functions. It provides the user with data and tools in a simple GUI, which is very similar among all the applications that KBase provides.

KBase has over than 160 apps that can offer the user different approaches for (meta)genome assembly, contig binning, sequence homology analysis, tree building, comparative genomics, metabolic modelling, community modelling, gap-filling, genome annotation, RNA-seq processing, and expression analysis. Additionally, in Table 4 are listed the KBase supported data objects. It also provides data integration along with easy access to shared user analyses of public microbial reference data from external resources like the NCBI and the DOE Joint Genome Institute (JGI). They allow the user to complete the task and get to the results that they will later study. KBase purpose is to make it easier for scientists to create, execute and share analyses of their biological data.

Table 4. KBase supported data objects

Supported Data Objects
reads
contigs
genomes
metabolic models
growth media
RNA-seq
expression
growth phenotype data
flux balance analysis conclusions

This platform is expected to grow wider, and as it does, the data, analysis tools, and computational experiments contributed by users should also increase, aiming to broader biological applications with enhanced support for functional prediction and comparison.

KBase's GUI is called Narrative, and it enables researchers to efficiently work together within the same platform (**Figure 6**). This GUI is built on the Jupyter® Notebooks, allowing researchers to design, carry out, record and share computational experiments in the form of Narratives. These are interactive documents that consist of all the data, analysis steps, parameters, visualizations, scripts, commentary, results, and conclusions of an experiment. KBase also has public Narratives that can be seen as tutorials for users. The main advantage of Narratives being built upon the Jupyter® Notebooks framework is that users can create and run scripts within a narrative using a “code cell”. The user can also use the flexibility of the code cells to customize analysis steps into their Narratives.



Figure 6. Illustration of a Narrative in KBase where (1) refers to the data field where the user stores all the imputed/generated data along the narrative. (2) is the apps field, the user has access for over than 160 applications. (3) Represent the analysis steps of each app. The (4) field is pointing towards the share symbol, meaning that the users can share their narratives with other users. In field (5), Markdown cell, it is possible to add commentaries. The last field, (6), points out the custom scripts with which the user can a python script. Adapted from [89].

KBase has a reference database which includes all public genome sequences from RefSeq [90] and Phytozome [91]. The software keeps the genomes with their original IDs and annotations, and the pipeline of KBase can maintain gene calls and annotations updated. This platform stores lots of data sets shared by users.

2.2.8.1 *KBase Apps for reconstructing a GSM model*

When the genome from an organism is uploaded to KBase interface, it should be re-annotated using the application named “*Annotate Microbial Genome*”. This annotation app uses components from the Rapid Annotation using Subsystem Technology (RAST) toolkit, to update the annotations of a genome, or to perform computations on multiple of genomes. The tool receives as input an annotated genome and allows the user to re-annotate it, so the annotations are in accordance with other KBase genomes. As a result, the re-annotated genome is ready to be further analyzed by other KBase apps. The identifiers generated by this tool are consistent with SEED subsystem naming conventions.

This application has the following pipeline:

1. *Annotate protein-encoding genes with k-mers.*

It consists in defining a set of signature k-mers (all the possible substrings of length k that are contained in a string, for this case amino acid 8-mers) built from information regarding annotations in the CoreSEED. The CoreSEED is a microbial genome database and is most used by RAST for manual annotations.

This annotation strategy promotes a better estimative of the core gene functions.

2. *Annotate remaining hypothetical proteins with k-mers*

In this stage will be used a set of k-mers that was built using the public annotation version of the SEED database named PubSEED.

3. *Find close neighbours and Annotate proteins similarity*

All the hypothetical proteins that were possibly missed in steps 1 and 2 are going to be annotated by searching against close relative genomes. BLAST is used in the search.

The next stage in automated *GSM* model reconstruction in KBase is to use the app “*Build Metabolic Model*”. The pipeline consists of steps, namely, step 1) Re-annotating Imported Genomes, step 2) Preliminary Reconstruction, step 3) Initial Gapfilling, step 4) Flux Balance Analysis.

Step 1 - Re-annotating Imported Genomes

This method was already described above using the application “*Annotate Microbial Genome*”.

Step 2 – Preliminary Reconstruction

After the genome reannotation using RAST at step 1, the genome can be inserted in the pipeline for preliminary reconstruction. At this step, the annotations retrieved by RAST at step 1 are used to generate draft metabolic models. The draft model will have a biomass reaction that is organism-specific, which was created using a template biomass reaction. When the template biomass reactions, such as, for, gram-negative, gram-positive, and plant were generated they were based on the use of RAST functional annotations to commit non-universal biomass components (i.e., cofactors) that serve as unique biological functions within all a wide range of organisms or a smaller set of organisms. The draft model uses GPR associations to represent the mapping between the biochemical reactions and the functional gene rules assigned when the RAST annotation was performed. The primary goal of creating GPR associations is to allow the pipeline to comprehend between different scenarios such as cases where protein products from different genes form an enzymatic complex to catalyze a reaction and cases where proteins products from different genes can individually catalyze the same reaction. In this step, spontaneous reactions are added. At the end of this step, the draft metabolic model will contain all reactions associated with one or more enzymes encoded in the genome that are identified in the annotations.

Step 3 – Initial Gapfilling

It is an optional step, but it runs by default thus meaning it is recommended to be done. This step will enhance the quality of the draft metabolic model because most of the genomes are not thoroughly annotated and therefore usually draft metabolic models have gaps that inhibit the production of some biomass precursors. An optimization algorithm runs at this step, to identify the minimal set of reactions that have to add to each model to fill the gaps. This step will ensure that every model can simulate growth. The reactions inserted or modified during gapfilling are retrieved from the ModelSEED biochemistry database.

Step 4 – Flux Balance Analysis

The FBA method can be performed when the model reconstruction is complete in order to get knowledge of the capacity of reactions to carry flux and reaction essentiality. Additionally, KBase uses Fluxes Variability Analysis (FVA) to classify the reactions in essential, active or blocked. An essential reaction is defined as a reaction that must carry flux for growth. Active reactions are the ones that only optionally carry flux, and blocked reactions are not able to carry flux.

The “Gapfilling” application can be used inside the “Build Metabolic Model” app, or it can be done separately. With this application the user can choose the minimal set of reactions to add to a draft metabolic model, thus enabling it to produce biomass in a specific media. Using this app, it is also possible to set the media condition; for example, it enables the user to study the metabolites involved in the growth of the organism being studied. By default, KBase’s gapfilling app uses a “complete” media (media containing all metabolites present in KBase’s biochemistry database). Additionally, KBase provides more than 500 media conditions that can be used for gapfilling. Besides this wide media range alternatives for default media, the user can also upload their media or create it by using the app “Create or Edit Media”.

Regarding the upload method, the media as to be in a personalized Tab-Separated Values (TSV) or Excel file with four columns like in Table 5.

Table 5. Format of TSV or Excel file to be uploaded into KBase’s Narrative.

Column	Name	Data
1st	Compound identifier	ModelSEED identifier, KEGG identifier, PubChem identifier, or compound name such as “lactose”
2sd	Concentration	Concentration of compound in mol/L
3rd	Min flux	Minimum allowed uptake/excretion of compound
4th	Max flux	Maximum allowed uptake/excretion of compound

The “minflux” and “maxflux” columns are the connection between the media to an FBA model. Absolute units must be used in these two columns, and they represent the range of possible fluxes for exchange reactions (i.e., reactions that transport the media compounds into an out of the cell. Negative values represent the excretion, transport out of the cell, and positive values

correspond to uptake, transport into the cell. Both minflux and maxflux units are mmol per gram cell dry weight per hour.

If the user runs more than one time this app, all the multiple gapfilling solutions are gathered into the same model. For example, if in the “Build Metabolic Model” app the user chooses complete media the gapfill will only add all the reactions needed for simulating growth. After, if the user runs again this app for an already gapfilled model but this time it uses a minimal media, the algorithm will only add additional reactions needed to grow on the minimal media.

KBase uses two solvers for Gapfill optimization, namely, GNU Linear Programming Kit (GLPK) and Solving Constraint Integer Programs (SCIP). The first one is used for most pure-linear optimizations and the second one is used for more complex problems.

Most of the apps built on KBase software were based on ideas from previous systems designed to hold large-scale bioinformatics analysis and model building. KBase is very different from other systems such as COBRA toolbox, Pathway Tools, and RAVEN Toolbox, which are capable of predictive modelling of metabolism because of these lack support for genome assembly, annotation, and comparison. In contrast, systems such as Galaxy[92], Taverna [93] and GenePattern [94] allow the user to develop and run bioinformatics workflows, but lack tools for genome annotation or predictive modelling.

As shown in **Table 6**, KBase and *merlin* are currently the two frameworks more suitable for reconstructing *GSM* models, according to these criteria. KBase is a broader platform because it enables the user to perform not only *GSM* reconstructions but also genome assembly and comparative genomics.

Table 6. Comparison of the features of software tools developed for aiding the reconstruction of genome-scale metabolic models, mentioned above.

Software	KBase	<i>merlin</i>	RAVEN	Pathway tools
Includes general network	x	x	x	x
Generates functional models	x	x	x	x
Assigns sub-cellular localization	x	x	x	-
Can use user defined models	x	x	x	x
Integrates gap filling	x	x	x	
Offline software		x	x	x
Graphical interface for manual curation	x	x	-	x
Reactions stoichiometry validation	x	x	-	x
Includes visualization	x	x	x	x
Enzymes annotation	x	x	x	
Transporters annotation	x	x	-	x

2.3 SYNTROPHIC BACTERIA COMMUNITY

2.3.1 Background

First, it is essential to understand the meaning of the term “Syntrophy”. This definition is used to characterize microbial cross-feeding. The term syntrophy has several classical definitions, such as:

- The cooperation between two microorganisms that depend on each other to perform their metabolic activity and in which the mutual dependence cannot be overcome by the addition of a co-substrate or any nutrient [95].
- A “thermodynamically independent lifestyle where the degradation of a compound, such as a fatty acid occurs only when degradation end products, usually hydrogen, formate, and acetate, are maintained at deficient concentrations” [96].
- A “nutritional situation in which two or more organisms combine their metabolic capabilities to catabolise a substrate that cannot be catabolised by either one of them alone” [97].
- “Relationships in which both partners depend on each other for energetic reasons and perform together a fermentation process that neither one or both could run on its own” [98].
- An “obligated mutualistic metabolism” [5].

This relationship resembles a beneficial symbiosis (i.e., mutualism [99]). These two relationships differ on one another because symbiotic relationships are not necessarily based on metabolism but, for example, on protection against chemical or mechanical stress [100]. The type of metabolism in a syntrophic relationship between two organisms, most of the times, can be defined as one partner providing a chemical compound that is consumed by the other in exchange for something. The chemical outcome of syntrophic activities differs from the outcome of each organism when they act separately. Usually, the benefits of the metabolic interaction between two microorganisms come at the cost of low energetic yields and slower growth rates [5].

Syntrophic microorganisms play an important role in carbon cycling, for example under anoxic conditions. The metabolic process of a syntrophic organism can change the environment around them. Studying the mechanisms behind metabolic interactions of syntrophic organisms will allow the microbial process to be better engineered to, for instance, treat wastewater or recover methane gas from proliferous resources.

The general process of such syntrophic relationship consists of anaerobic degradation of complex chemical compounds. It is often two- or three-step process, in which polysaccharides, proteins, nucleic acids, and lipids are primarily fermented to intermediates, such as hydrogen, formate, carbon dioxide (CO₂), and acetate, for methanogenesis and to smaller organic compounds (i.e., lactate and butyrate) (Figure 7). These intermediate products are further degraded by secondary fermentation process if the microorganisms live in an environment that lacks external electron acceptor. Thus, allowing the production of substrates that can enter directly into methanogenic pathways [98].

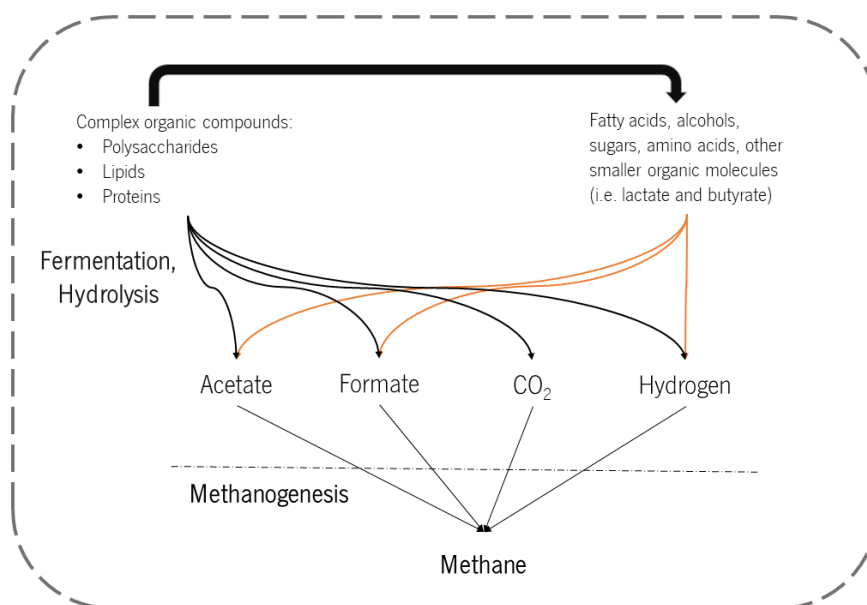


Figure 7. Schematic drawing of “classical” syntrophy depending on the environmental conditions (methanogenic environment). The thicker dashed line represents the border of the system.

Regarding methanogenesis, it plays a vital role in carbon cycling, leading to the formation of methane from small carbon sources. Methanogenic archaea directly influence the metabolism of syntrophic microorganisms, because they are responsible for the removal of significant electron carriers, such as hydrogen and formate, in the lack of other terminal electron acceptors. Also, methanogens need fermenting microorganisms to produce their substantial metabolic intermediate products.

2.3.2 *Syntrophobacter fumaroxidans* strain MPOBT

In nature, there are four species of the genus *Syntrophobacter*. The name of the genus refers to a rod-shaped bacterium that establishes a syntrophic association with hydrogen and formate using microorganisms [101]. Species from this genus can grow on propionate. This substrate is an essential intermediate for microorganisms that live in methanogenic environments. To the moment, only seven syntrophic propionate-oxidation bacteria species have been discovered. Between them, a group of four mesophilic (warm temperatures) species can be identified, they are related to the mesophilic sulphate reducers. They are *Syntrophobacter wolinii* [102], *S. fumaroxidans* [101], *Syntrophobacter pfennigii* and *Smithella propionica* [103], and they are all Gram-negative bacteria. All members of the genus *Syntrophobacter* are able to use sulphate as an electron acceptor and oxidize propionate via the methylmalonyl coenzyme A (CoA) pathway [104].

Regarding the organism of interest, *S. fumaroxidans*, the species name derives from the Latin word “fumaricum” referring to fumaric acid and “oxidans” that refers to oxidizing, pointing out to fumarate fermentation. The most studied species of the genus *Syntrophobacter* is the strain MPOB^T. The cells from *S. fumaroxidans* strain MPOB^T have a rod shape (Figure 8) and do not form endospores. The metabolism of these organisms can be respiratory or fermentative, but it is essentially strictly anaerobic [104].

In pure culture *S. fumaroxidans* uses sulphate or fumarate as an electron acceptor, but when in co-culture with a hydrogen and formate-scavenger methanogen organism it uses propionate syntrophically via the methylmalonyl-CoA pathway [105].

The genome of *S. fumaroxidans* was sequenced at JGI [106]. Genome annotation was performed using Critica [107] supplemented with the output of the Generation and Glimmer models [108]. All the predicted coding DNA sequences (CDSs) were translated and searched in databases, such as NCBI's non-redundant database, KEGG, and UniProt.

S. fumaroxidans genome has a length of 4,990,251 bp and holds a circular chromosome with a 59.95% GC content. The genome annotation also showed that *S. fumaroxidans* has 4,179 predicted genes, of which 4,098 were protein coding genes, 81 RNAs and 34 pseudogenes. The percentage of genes that have been given a supposed function is around 67%, and the remaining genes are considered to be hypothetical proteins [109].

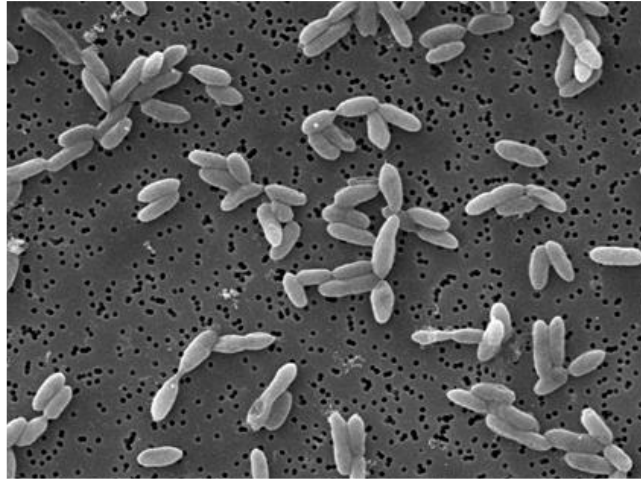


Figure 8. Microscopy photograph of cells of *S. fumaroxidans*. Adapted from [109].

2.3.3 *Methanospirillum hungatei* strain JF1

In methanogenic environments, it is common to find many different species able to produce biogenic methane. Between them there is one of particular interest, that is *M. hungatei* which belongs to the family of *Methanospirillaceae*, and the only genus is *Methanospirillum*. The species epithet comes from the Latin in honour of Dr. R. E. Hungate. He was the man responsible for the invention of methodologies for modern isolation and cultivation of strictly anaerobic bacteria and archaea [110]. The strain JF1 was the first isolated type species for *M. hungatei*, and it was first isolated from sewage sludge [111]. *M. hungatei* strain JF1 is Archaeal hydrogen- and formate-utilizing, that produces methane. The cells of this microorganism are narrow and curved rods (**Figure 9**). The metabolism of this organism is strictly anaerobic, and it uses hydrogen plus carbon dioxide and/or formate as the methanogenic substrate. *M. hungatei* uses acetate as the main supply for cell carbon [112]. The ideal growth temperature for this organism is around 37 °C.

Microorganisms like *M. hungatei* that can produce methane have a crucial role in the global carbon cycle, and they are also used in the treatment of organic waste, and they can also be used to produce biofuel from biomass [95].

The genome of *M. hungatei* strain JF1, like the genome of *S. fumaroxidans*, was also sequenced at the JGI. *M. hungatei* had the genome annotated using Prodigal [113] and pursued by a manual curation using the JGI GenePRIMP pipeline [114]. The procedure used in the characterization of CDSs is the same used for *S. fumaroxidans*, with some differences in the databases that were used for the identification of transporter proteins. For this microorganism, membrane transport protein analysis was performed with IMG [115].

The genome of the formate-utilizing organism *M. hungatei* is one circular chromosome of 3,544,738 bp with 3,307 predicted genes of which 3,239 are protein-coding genes and a GC content of 45.15%. About 61%, of the protein-coding genes, have a putative function and the remaining genes do not have a function assigned [116]. This genome is among the largest within the *Archaea* domain, and it suggests the presence of unrecognized biochemical/physiological properties that probably extend to close organisms belonging to the family *Methanospirillaceae* and the ability to interact with syntrophic bacteria, like *S. fumaroxidans*. *M. hungatei* strain JF1 has multiple archaeal-type flagella filaments at the cell ends [111]. These filaments are very important to this microorganism because they may function in cell-cell adhesion or cell-cell communication and genes for multiple hydrogenases and formate dehydrogenases to metabolize certain compounds, such as formate.

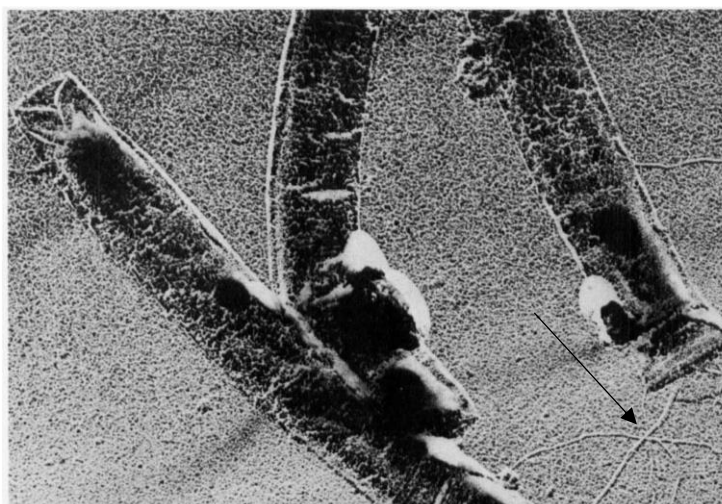


Figure 9. Cells of *M. hungatei* showing tufts of polar flagella. Black arrow pointing towards tufts. Adapted from [111]

2.3.5 Syntrophic Relationship of *S. fumaroxidans* strain MPOBT and *M. hungatei* strain JF1

S. fumaroxidans is a bacterium that degrades propionate in a syntrophic association with formate scavengers, such as *M. hungatei*. Studies revealed that the transference of both hydrogen and formate are essential mechanisms of electron transference between species [117]. The degradation of propionate in the absence of an external electron acceptor requires low hydrogen and formate concentrations, and *M. hungatei* is the ideal partner for *S. fumaroxidans* because this organism can maintain these low concentrations. *M. hungatei* uses both hydrogen and carbon dioxide in order to produce methane (Figure 10).

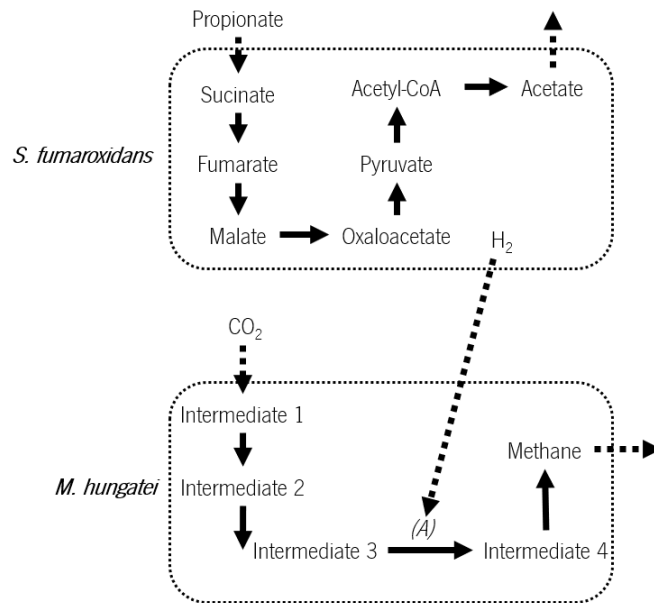


Figure 10. Simplified illustration of the syntrophic relationship between *S. fumaroxidans* and *M. hungatei*. *S. fumaroxidans* uses propionate from the environment to produce acetate and in the process, it can produce hydrogen (H_2). The hydrogen can later be very important in the metabolism of *M. hungatei* more exactly in stage (A) where it will participate in the reduction of the ferredoxins that assist the reaction that converts intermediate 3 into intermediate 4. Adapted from [117]

The metabolism of both organisms is dependent on the electron flux, allowing for all reduction and oxidation reactions to happen. This flux is generated by the action of several hydrogenases and dehydrogenases present in each organism [118]. Hydrogenases are enzymes that catalyse the reversible oxidation of molecular hydrogen, and they have an important role in anaerobic metabolism [119]. Dehydrogenases are another kind of enzymes that promote the oxidation of substrates through the reduction of an electron acceptor, usually Nicotinamide Adenine Dinucleotide (NAD^+)/ Nicotinamide Adenine Dinucleotide Phosphate (NADP).

CHAPTER 3

MATERIALS AND METHODS

*This chapter is focused on the materials and methods used in the reconstruction of the GSM models for *M. hungatei* and *S. fumaroxidans*.*

3.1 GENOME ANNOTATION

In this section, the strategies that were used to annotate the genome of *Syntrophobacter fumaroxidans* and *Methanospirillum hungatei* will be presented. Since the GSM's model' reconstructions for both *M. hungatei* and *S. fumaroxidans* were performed in different ways the methods used must be presented separately.

3.1.1 Enzymes and transporters annotations for *M. hungatei* using KBase

The enzymatic annotation for *M. hungatei* was performed in KBase. First, a *fasta* file containing *M. hungatei*'s amino acid sequences was uploaded into the Narrative. This file was retrieved from the NCBI database using the free Uniform Resource Locator (URL) access: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/013/445/GCF_000013445.1_ASM1344v1 (accessed at 19 of February 2018). There are two methods for importing the files into the KBase interface, one to download the documents from NCBI manually, and the other is, to use a tool named “Add Data” that allows automatic access to public databases such as NCBI. The last-mentioned method was the one it was used.

After uploading the genome to the Narrative, it was reannotated using the app “Annotate Microbial Genome”. This tool used the RAST algorithm to annotate the genome, and as a result, it created a file containing 2085 annotated genes with 1390 distinct functions.

3.1.2 Building *GSM* draft model of *M. hungatei* in KBase

In KBase the models are generated based on template models, such as Gram-negative microbe model, Gram-positive microbe model, Core pathways microbe model, and Plant model. The main difference in these models is the biomass composition and biochemical reaction functional associations. The “Build Metabolic Model” app present in KBase was used to generate a draft model for *M. hungatei*. In order to run this app, the genome previously uploaded to the Narrative was used as the input genome together with a complete media (Table 7) for gapfilling. The template model chosen was the Gram-negative microbe model to match the fact that the organism being studied is also of the same Gram stain.

Table 7. Complete basal bicarbonate-buffered medium used for gapfilling. Adapted from [120].

Category	Formula	Concentration	Cont. (...)	
	Na ₂ HPO ₄	0.53 g/L	Alkaline trace solution	Na ₂ SeO ₃ 0.10 mmol/L
	KH ₂ PO ₄	0.41 g/L		Na ₂ WO ₄ 0.10 mmol/L
	NH ₄ Cl	0.30 g/L		Na ₂ MoO ₄ 0.10 mmol/L
	CaCl ₂	0.11 g/L		NaOH 10.00 mmol/L
	MgCl ₂	0.10 g/L	Vitamins	Biotin 0.02 g/L
	NaCl	0.30 g/L		Niacin 0.20 g/L
	NaHCO ₃	4.00 g/L		Pyridoxine 0.50 g/L
Acid Trace Solution	FeCl ₂	7.50 mmol/L		Riboflavin 0.10 g/L
	H ₃ BO ₄	1.00 mmol/L		Thiamine 0.20 g/L
	ZnCl ₂	0.50 mmol/L		Cyanocobalamin 0.10 g/L
	CuCl ₂	0.10 mmol/L		<i>p</i> -aminobenzoic acid 0.10 g/L
	MnCl ₂	0.50 mmol/L		Pantothenic acid 0.10 g/L
	CoCl ₂	0.50 mmol/L		
	NiCl ₂	0.10 mmol/L		
	HCl	50.00 mmol/L		

The output of the app generated a draft model as an “FBAmol” object in the Narrative. The file type only reflects the type of model. The output of gapfilling is also added to the Narrative

panel. The draft created has eight tabs for browsing the data, namely, *Overview*, *Reactions*, *Compounds*, *Genes*, *Compartments*, *Biomass*, *Gapfilling*, and *Pathways*.

- *Overview* : The key information about the model, such as the associated genome, the number of reactions, and the number of compounds are summarized.
- *Reactions* : All reactions are identified by a reaction ID, name, biochemical equation, associated gene IDs, and an indication informing if the reaction was added by the gapfilling stage or not.
- *Compounds* : Presents the information about the compounds in the model such as chemical formula and charge.
- *Genes* : Represent the genes by gene IDs and associated reaction IDs.
- *Compartments* : Detailed information about subcellular localization of the compounds and enzymes. Usually, there are different types of compartments in microbes, namely, Cytosol (c), Periplasm (p), and Extracellular (e). For each compartment all the compounds and reactions are identified using compartment notation (i.e., rxn00001[c0], cpd00001[c0]).
- *Biomass* : Indicates the biomass composition of the model. In the majority of models, the biomass is defined as an equation where biomass compounds and ATP would make 1 gram of biomass. In this tab, it is also possible to check the coefficient of each biomass compound in the Coefficient column. The positive and negative coefficients indicate if the compounds are on the right or the left side of the biomass equation, respectively.
- *Gapfilling* : In this section, the reactions that were added in order to fill metabolic gaps as a direct result of missing genes or inconsistencies annotations. The objective of the gapfilling process is to add a minimal number of reactions and compounds to make the *GSM* network generate biomass.
- *Pathways* : This tab has KEGG maps representing the metabolic network of the model.

Using the draft model of *M. hungatei* present in the Narrative, it was possible to browse the linkage between reactions and protein-encoding genes, search for all compounds used in a model, and identify possible gaps in the model that were not found by the automatic annotation tools.

3.1.4 Enzymes annotation in *merlin* for *S. fumaroxidans*

Different from *M. hungatei* the enzymatic annotation for *S. fumaroxidans* was performed using *merlin*'s framework. First, a workspace was created in *merlin* with the name "sfumaroxidansMPOBT". A plugin present in *merlin*'s interface named "open" was used to open the workspace. In this plugin the taxonomy number of *S. fumaroxidans* (taxonomy identifier: 335543) was added and a button named "search" was pressed in order to retrieve information from different online databases such as NCBI automatically. Then, inside *merlin*'s interface, the NCBI locus tag identifier (i.g., "Sfum_0001") was nominated to each enzyme encoding gene.

Afterward, the BLAST algorithm was used to make a homology search for every gene. The databases used as sources of homologous amino acid sequences in BLAST were from UniProt, namely, UniProt/Swiss-Prot (curated database) and UniProt/TrEMBL (non-curated database). The maximum e-value was inferred as 1E-30 to get a high-quality set of homologies. The e-value measures the statistical difference of a BLAST hit and for that reason can display the quality of a homology. Values of e-value near to zero (low e-values) represent good homologies results and indicate a high similarity between sequences. The results field was limited to a maximum of 100 hits per gene.

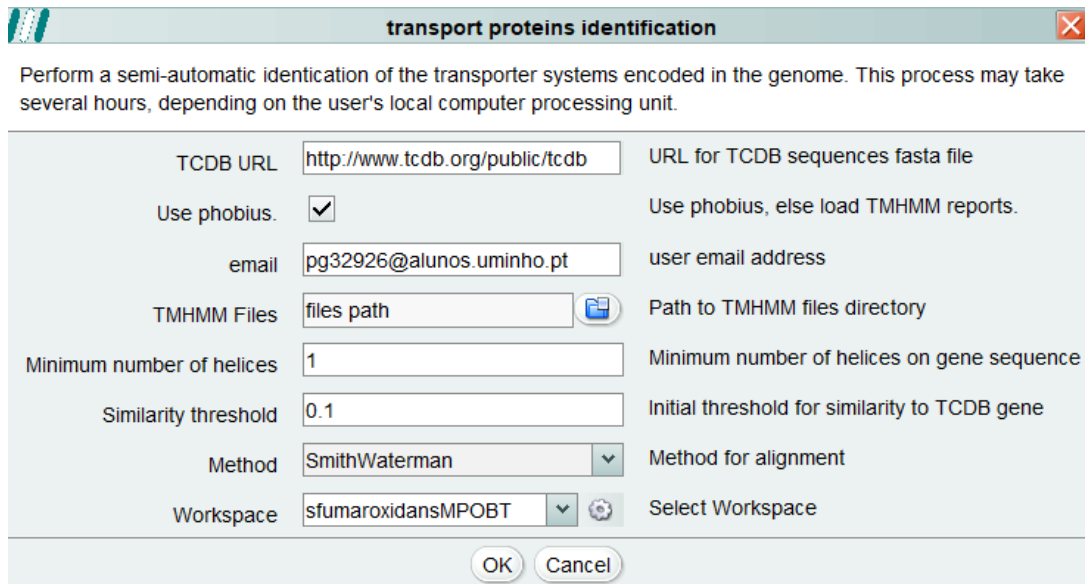
merlin automatically processed the information provided by BLAST and the annotations of the metabolic genes, uniformly labelled as Enzyme Encoding Gene Candidate (EEGC), were analysed. For a gene to be considered an EEGC, at least, one of its homologous' genes has an enzymatic function associated with an EC number.

The next stage was to automatically annotate the genome using a new plugin in *merlin* named "automatic workflow". Note that this plugin is not yet available in the current version of *merlin*. It consists of selecting seven close phylogenetic relatives to *S. fumaroxidans* and select different e-values for each of them. The lowest e-values (1E-10/1E-20) are assigned for the three organisms that have information curated in UniProt's databases, more precisely in UniProt/SWISS-Prot. The remaining five organisms are assigned with an e-value higher than 1E-20 (1E-30/1E-40) and are the ones that do not have curated information. The annotations of these organisms are retrieved from the non-curated UniProt's database, the UniProt/TrEMBL.

3.1.3 Transporter Proteins

TRIAGE method is implemented in merlin as previously mentioned and was the one chosen to perform the transporters annotation. This *merlin*'s tool was able to identify genes encoding transport systems based on *S. fumaroxidans* strain MPOBT genome information. Using the following URL <http://www.tcdb.org/public/tcdb>, TRIAGE was able to access the TCDB sequences *fasta* files. Next, the algorithm used was SW in order to align the sequences and the similarity threshold value was defined as 10%. Additionally, the minimum number of helices was set to 1 (Figure 11).

The tool generated records with TC family number and UniProt accession number for each gene homology. The TC number is used to access transport protein description available on TCDB database.





transport proteins identification	
Perform a semi-automatic identification of the transporter systems encoded in the genome. This process may take several hours, depending on the user's local computer processing unit.	
TCDB URL	<input type="text" value="http://www.tcdb.org/public/tcdb"/> URL for TCDB sequences fasta file
Use phobius.	<input checked="" type="checkbox"/> Use phobius, else load TMHMM reports.
email	<input type="text" value="pg32926@alunos.uminho.pt"/> user email address
TMHMM Files	<input type="text" value="files path"/>  Path to TMHMM files directory
Minimum number of helices	<input type="text" value="1"/> Minimum number of helices on gene sequence
Similarity threshold	<input type="text" value="0.1"/> Initial threshold for similarity to TCDB gene
Method	<input type="text" value="SmithWaterman"/> Method for alignment
Workspace	<input type="text" value="sfumaroxidansMPOBT"/>  Select Workspace
<input type="button" value="OK"/> <input type="button" value="Cancel"/>	

Figure 11. TRIAGE standards for retrieving genes encoding transport systems in *S. fumaroxidans* MPOBT

3.2 *Syntrophobacter fumaroxidans* MPOB^r DRAFT NETWORK ASSEMBLY

This section describes all the strategies used during the assembly of the draft network of *Syntrophobacter fumaroxidans* and the manual curation methods used to curate the draft network.

3.2.1 Metabolic Data Integration

merlin can automatically retrieve KEGG metabolic data, such as metabolites, proteins/enzymes, and biochemical reactions. Next, the enzymatic annotation was fully automatically integrated, thus generating a draft network. *merlin* automatically gathered all metabolic reactions and associated them with complete EC numbers creating KEGG metabolic pathways. Moreover, all spontaneous and non-enzymatic reactions were integrated into the network.

3.2.2 Transporter Proteins Data Integration

The TRIAGE plugin generated a list of metabolites that can be transported for each transported protein that was previously annotated, but only the ones with a score above a specific threshold were integrated into the network. The strategy adopted allowed to select a set of reactions associated with the assigned transporter protein encoding genes and integrate them into the network.

3.2.3 Exchange reactions Integration and Compartmentalization

As a result, of using TRIAGE to build the transporter reactions, an outside compartment for the extracellular space was created. *merlin* automatically generated the exchange reactions, that are reactions slightly different from the usual reactions. These only have a single reactant metabolite and are only created for every metabolite present in the outside compartment.

Merlin also created a compartment for the intracellular space (inside compartment) that is separated from the extracellular space (outside compartment) by the cellular membrane. Additionally, the core of the metabolic network is allocated in the inside compartment.

3.2.4 Manual Curation of the Draft Network of *S. fumaroxidans*

In this stage will be presented all the strategies used to revise the draft network of *Syntrophobacter fumaroxidans*.

3.2.4.1 *Pathway-by-pathway Analysis*

In *merlin*, it is possible to visualize all the KEGG pathways present in the draft network, previously built, on a web browser (**Figure 12**). A KEGG pathway is a set of reactions that belong to the network, paired with the corresponding enzymes (associated with an EC number) connected by metabolites. *merlin* can highlight the reactions and enzymes present in each KEGG pathway map and allows the user to have a better visualization of that section of the metabolic network.

Both reactions and EC numbers available in the network are colored by *merlin* using the standards present in **Table 8**.

Table 8. Standards used by *merlin* to colour the reactions and enzymes in the network

Colour	Meaning
Green	Presence of the enzyme and reaction in the pathway
Blue	Presence of the reaction in the pathway but absence of the enzyme
Cyan blue	Presence of the enzyme in the pathway but the reaction is a connection to a dead-end of the draft network
Red	Presence of the enzyme but the reaction is a dead end of the draft network
Colorless	Absence of both reaction and enzyme in the pathway

A pathway-by-pathway analysis was performed in order to proceed with the manual curation. Both KEGG's and MetaCyc's *S. fumaroxidans* reference pathway were used to possible undercover issues with the network. One by one, the core metabolic pathways (pathways for the synthesis of essential metabolites) of the draft network were compared against the reference pathways from KEGG and MetaCyc, to find potential missing EC numbers and reactions.

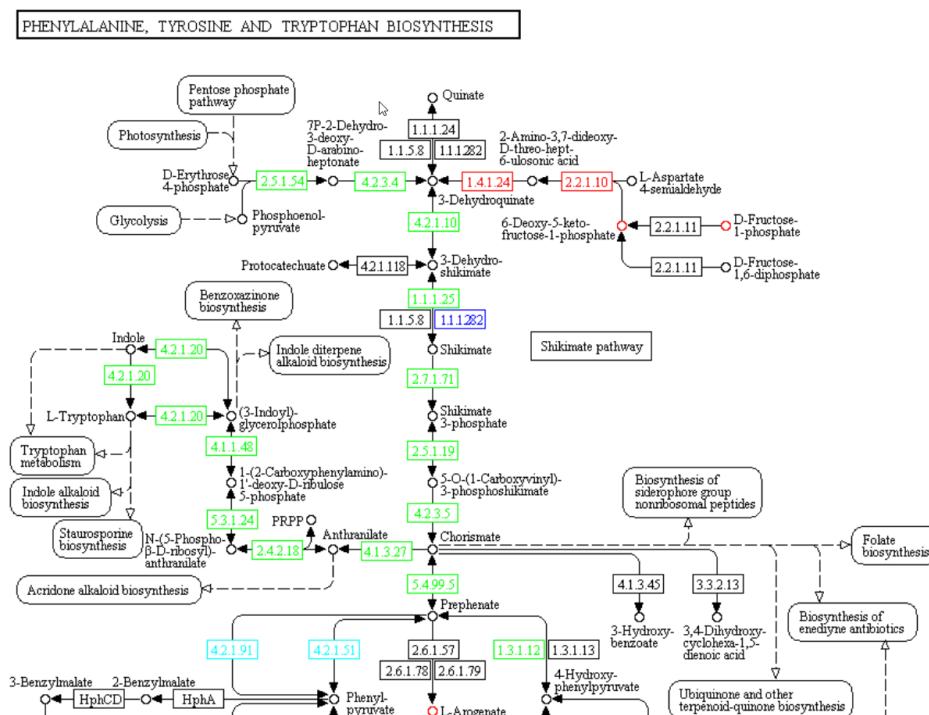


Figure 12. Example of a partial metabolic KEGG's pathway map coloured by *merlin* tool.

3.2.4.2 Gap Filling

For gap filling information about *S. fumaroxidans* genome and metabolism was retrieved from literature. The main topics about the metabolism, such as the central carbon metabolism, the carbohydrate metabolism, and biosynthesis of macromolecules were investigated. Relevant literature about the metabolism of strictly phylogenetic organisms of *S. fumaroxidans* was also investigated to differentiate the gaps that had to be filled from the gaps that were originated by auxotrophies, an absence of functional genes (i.e., pseudogenes), and metabolic incapacities.

While investigating for potential errors when performing the pathway-by-pathway analysis, sometimes some reactions or even GPR associations were added to the network to fill gaps, but only if supporting evidence was available in the literature or genome annotation. For a few cases, the literature was the only source available to corroborate the veracity of the reactions due to the lack of information at the genome level. These set of reactions were, therefore, added to the network but they have missing enzymes or enzyme encoding genes associated.

3.2.4.3 *Mass Balance*

Inside *merlin*'s framework, there is a semi-automatic tool to identify unbalanced reactions. This tool will highlight as bold and italicized all the reactions that have an unequal number of atoms between reactants and products.

Regarding the charge balance, only neutral and unprotonated formulas of the compounds are used by KEGG. *merlin*'s tool allowed to define the proton name and for that purpose was established as "H⁺". To all the highlighted unbalanced reactions was added a tab corresponding to the balance mass, and it indicated the numerical difference of chemical elements present on both sides of the reaction. After investigating the reactions in terms of stoichiometry and mass balance, the reactions were modified and updated in *merlin* by the addition or removal of chemical elements.

3.3 *M. hunagtei* DRAFT METABOLIC NETWORK CURATION IN *merlin*

This section describes the manual curation methods used to curate the draft network of *Methanospirillum hungatei*.

3.3.1 *M. hungatei* draft *GSM* model integration in *merlin*

Besides the easy and fast generation of a *GSM* model in KBase, performing manual annotation in the KBase's narrative was very hard, and it would take too long to be finished. In this line of view, the draft *GSM* model was exported from KBase in the SBML format and integrated into *merlin*.

For this work, a tool named BioCOISO, previous developed for *merlin* (not yet published) was used and it is able to import models in the SBML format. This tool is available at the following URL: https://drive.google.com/open?id=0B8m5ZT09b_cCbUhLb2lVeHNZWWM. To that date, *merlin* could not import *GSM* models from other genome-scale metabolic model reconstruction tools. The tool was developed based on the SMBL files generated by KBase, and for this reason, the tool is better suited to accept SBML files with modelSEED identifiers.

3.3.2 Manual curation of the Draft Network of *M. hunagtei*

The first task performed in the manual curation of the draft network occurred inside KBase though it involved the use of the *merlin* tool named "e-biomass equation". The biomass reaction generated by KBase includes three macromolecules, namely, DNA, RNA, and Protein without

reactions to biosynthesize these compounds. *merlin*'s tool was used to identify the precursors and assign a stoichiometric value to each one. Note that this method is identical to the one used in the next section since both organisms are of the type Gram-negative, which means that the tool will be performed using the same standards, only differing in the information retained in the genome files (Figure 13). For more information about the tool, read the next section.

Figure 13. *merlin*'s e-biomass equation tool used to formulate the biomass composition of *Methanospirillum hungatei*

The output of the tool allowed the addition of precursors to the three macromolecules together with the respective content values. The three synthesis reactions for these three macromolecules were modified inside KBase using an application named “Edit Metabolic Model”, to edit those reactions manually. After modifying these reactions, the model was exported format KBase using the export tool, which enables it to be exported in different formats such as JavaScript Object Notation (JSON), EXCEL and SBML.

The pipeline used in the manual curation in *merlin* of *S. fumaroxidans* was adapted to perform the manual curation of *M. hungatei*. The same tasks were performed, starting in a pathway-by-pathway analysis following the same standards used before. The next step, gap filling, was conducted in the same way although the network has been previously gap-filled using KBase's tool for gap filling. The last task, mass balance, was also carried out in the same standards.

3.4 BIOMASS AND ENERGY REQUIREMENTS FORMULATION FOR *M. hungatei* and *S. fumaroxidans*

The biomass reaction, as mentioned before, is the reaction responsible for the formation of biomass from precursors or macromolecules. In this work, the strategy used was to use complex macromolecules has the precursors in the biomass reaction. For this purpose, eight entities were created. Each one of these represented a macromolecule, and when combined, they accounted for the overall composition of *Syntrophobacter fumaroxidans* MPOBT. The stoichiometry of each reactant (macromolecule) in the biomass reaction was defined as the value of a gram of macromolecule per gram of biomass.

merlin's “e-biomass equation” was used to formulate the biomass reaction [121]. This tool generated a draft biomass reaction using the genome information (Figure 12). It uses the *fasta* files of assembled whole-genome sequences, such as genome nucleotide sequence, genome amino acid, genome Transfer Ribonucleic Acid (tRNA) sequence, genome mRNA sequence and genome Ribosomal Ribonucleic Acid (rRNA). These files were automatically downloaded from the NCBI database when the workspace was created. The contents of the macromolecules for Gram-negative bacteria were previously estimated in the study mentioned above [121] and integrated into *merlin's* tool. The draft biomass reaction for *S. fumaroxidans* MPOBT was then retrieved by using the standards illustrated in Figure 14.

Figure 14. *merlin's* e-biomass equation tool used to formulate the biomass composition of *Syntrophobacter fumaroxidans*

Another method was used to complement the biomass equation. Firstly, research was performed in order to identify the phylogenetic closest organisms. Among the organisms identified the *Geobacter sulfureducens* (*G. sulfureducens*) [122] was the only one to have a *GSM* model available. This *GSM* was then used to help in the formulation of the biomass equation of the *Syntrophobacter fumaroxidans* *GSM* model. Two new macromolecular entities were assigned to the *S. fumaroxidans* model, the “e-Fatty Acid” and “e-Peptidoglycan” based on *G. sulfureducens* model. In *merlin*, for each macromolecule, a reaction was created representing the bio assembly of the monomers into the macromolecules

Two reactions associated with lipid metabolic biosynthesis and metabolism were added to the draft model in the biomass pathway. These reactions were assigned with the following IDs:

- “R_e-Acyl-CoA”
- “R_e-Fatty_Acid”

The first reaction, “R_e-Acyl-CoA”, represents the link between the “Fatty Acid” and the Acyl-CoA. The term “Fatty Acid” was chosen to represent the fatty acid profile in *S. fumaroxidans* MPOBT, and it is an average long-chain fatty acid. Note that Acyl-CoA metabolite makes the connection between CoA and a long-chain fatty acid and therefore it is a lipid precursor.

In the *GSM* model of *G. sulfureducens*, the lipid fraction belonging to the biomass is represented by the monomeric units of lipids, the fatty acids. The same approach was used to create the “e-Fatty Acid” present in *S. fumaroxidans*. Therefore, the lipid content in the biomass equation of *S. fumaroxidans* corresponds to the fatty acids.

CHAPTER 4

RESULTS AND DISCUSSION

*In this chapter, the results obtained from the reconstruction
Syntrophobacter fumaroxidans and Methanospirillum hungatei
GSM models will be discussed.*

4.1 GENOME-SCALE METABOLIC MODEL OF *S. fumaroxidans* strain MPOBT

The *GSM* model reconstruction of *S. fumaroxidans* strain MPOBT 1 is presented in this section. The model can be accessed by the following URL:
<https://nextcloud.bio.di.uminho.pt/s/FMpWBpYeySykeJA>

GSM models of *Geobacter sulfurreducens* ATCC 51573 [122], *Geobacter metalreducens* ATCC:53774 [123] were used to guide in the reconstruction of *S. fumaroxidans* strain MPOBT *GSM* model.

4.1.1 Manual curation of the draft metabolic network of *S. fumaroxidans*

In this section will be presented all the modifications made to the *S. fumaroxidans GSM* model while performing the manual curation.

4.1.2.1 *Pathway-by-pathway analysis*

Several KEGG metabolic pathways during manual curation were modified, and others were ignored. The ignored KEGG pathways were removed since they did not have reactions connected to the remaining network. Additionally, in Table 9 lists all the pathways, with the respective number of reactions, which were studied when performing the manual curation phase.

Table 9. List of metabolic pathways available in the *GSM* model of *Syntrophobacter fumaroxidans* MPOBT.

Pathway	Number of reactions
Glycolysis / Gluconeogenesis	26
Citrate cycle (TCA cycle)	18
Pentose Phosphate Pathway	14
Pentose and glucuronate interconversions	4
Fructose and mannose metabolism	13
Galactose metabolism	7
Starch and sucrose metabolism	17
Amino sugar and nucleotide sugar metabolism	32
Pyruvate metabolism	30
Propanoate metabolism	18
Butanoate metabolism	12
Nitrogen metabolism	15
Sulphur metabolism	14
Fatty acid biosynthesis	43
Glycerolipid metabolism	6
Glycerophospholipid metabolism	12
Biosynthesis of unsaturated fatty acids	5
Purine metabolism	75
Pyrimidine metabolism	32
Alanine, aspartate and glutamate metabolism	20
Glycine, serine and glutamate metabolism	28
Cysteine and methionine metabolism	21
Valine, leucine and isoleucine metabolism	41
Lysine biosynthesis	14
Arginine and proline metabolism	27
Histidine metabolism	15
Tyrosine metabolism	10
Phenylalanine metabolism	5
Phenylalanine, tyrosine and tryptophan biosynthesis	23
Selenocompound metabolism	12
D-Glutamine and D-glutamate metabolism	6
D-Alanine metabolism	2
Glutathione	12
Peptidoglycan biosynthesis	19
Thiamine metabolism	13
Riboflavin metabolism	10
Vitamin B6 metabolism	8

Nicotinate and nicotinamide metabolism	22
Pantothenate and CoA biosynthesis	16
Biotin metabolism	13
Lipoic acid metabolism	4
Folate biosynthesis	25
One carbon pool by folate	11
Porphyrin and chlorophyll metabolism	35
Terpenoid backbone biosynthesis	11
Polyketide sugar unit biosynthesis	6
Aminoacyl-tRNA biosynthesis	25
2-Oxocarboxylic acid metabolism	31
Biomass Pathway	9
Non-enzymatic	2
Spontaneous	122
Transporters Pathway	97
Drains Pathway	64
Non-associated to Pathways	238

After the pathway-by-pathway analysis 52 metabolic reactions were added to the draft metabolic network. As mentioned before in this chapter, the *G. sulfurreducens* GSM model was used to infer some metabolites for the biomass composition. Metabolites identified in the fatty acid composition were assumed to be the same composing the fatty acid group in *S. fumaroxidans*. Likewise, reactions present in Table 10 were added to the model to enable the production of the target metabolites.

Table 10. Reactions added to the fatty acid biosynthesis pathway.

Reaction ID	Reaction
R00742	ATP + Acetyl-CoA + HCO ₃ ⁻ + H ⁺ <=> ADP + Orthophosphate + Malonyl-CoA
R01624	Acetyl-CoA + Acyl-carrier protein <=> CoA + Acetyl-[acyl-carrier protein]
R01706	Hexadecanoyl-[acp] + H ₂ O <=> Acyl-carrier protein + Hexadecanoic acid
R02814	Oleoyl-[acyl-carrier protein] + H ₂ O <=> Acyl-carrier protein + (9Z)-Octadecenoic acid
R03370	Octadecanoyl-[acyl-carrier protein] + 2 Reduced ferredoxin + Oxygen + 2 H ⁺ <=> Oleoyl-[acyl-carrier protein] + 2 Oxidized ferredoxin + 2 H ₂ O
R04430	Butyryl-[acp] + NADP ⁺ <=> But-2-enoyl-[acyl-carrier protein] + NADPH + H ⁺
R04725	Dodecanoyl-[acyl-carrier protein] + NADP ⁺ <=> trans-Dodec-2-enoyl-[acp] + NADPH + H ⁺
R04956	Hexanoyl-[acp] + NADP ⁺ <=> trans-Hex-2-enoyl-[acp] + NADPH + H ⁺
R04959	Octanoyl-[acp] + NADP ⁺ <=> trans-Oct-2-enoyl-[acp] + NADPH + H ⁺
R04962	Decanoyl-[acp] + NADP ⁺ <=> trans-Dec-2-enoyl-[acp] + NADPH + H ⁺
R04967	Tetradecanoyl-[acp] + NADP ⁺ <=> trans-Tetradec-2-enoyl-[acp] + NADPH + H ⁺
R04970	Hexadecanoyl-[acp] + NADP ⁺ <=> trans-Hexadec-2-enoyl-[acp] + NADPH + H ⁺

R07764	(R)-3-Hydroxyoctadecanoyl-[acp] \rightleftharpoons (2E)-Octadecenoyl-[acp] + H ₂ O
R07765	(2E)-Octadecenoyl-[acp] + NADH + H ⁺ \rightleftharpoons Octadecanoyl-[acyl-carrier protein] + NAD ⁺
R08159	Tetradecanoyl-[acp] + H ₂ O \rightleftharpoons Acyl-carrier protein + Tetradecanoic acid
R08161	Hexadecanoyl-[acp] + 2 Reduced ferredoxin + Oxygen + 2 H ⁺ \rightleftharpoons Hexadecenoyl-[acyl-carrier protein] + 2 Oxidized ferredoxin + 2 H ₂ O
R08162	Hexadecenoyl-[acyl-carrier protein] + H ₂ O \rightleftharpoons Acyl-carrier protein + (9Z)-Hexadecenoic acid
R08163	Octadecanoyl-[acyl-carrier protein] + H ₂ O \rightleftharpoons Acyl-carrier protein + Octadecanoic acid

4.1.2.2 *Transport reactions*

TRIAGE automatically identified and integrated 87 transport reactions into the *GSM* model. All these reactions were manually verified before they were integrated into the model. Additionally, transport reactions were created and included in the model in order to fulfil the absence of end-products transportation. These transport reactions described the metabolites facilitated diffusion, meaning that only one metabolite was present in those reactions. Moreover, transport reactions for several ions, cofactors, and vitamins were also assembled into the model.

4.1.2.3 *Gap Filling*

In this phase of the manual curation different reactions were added or corrected to fill metabolic gaps. A total of 52 reactions were assembled into the model in order to fulfil the metabolic gaps found during manual curation of metabolic pathways. Different gaps were found in different metabolic pathways of the gram-negative bacterium *Syntrophobacter fumaroxidans* MPOBT.

4.1.2.4 *Mass Balance*

merlin's tool “unbalanced reactions” unveiled 51 reactions, which were manually corrected. The most frequent cases of unbalanced reactions were the ones where protons were either missing or in excess (46 reactions). *merlin* also flagged as unbalanced reactions the ones containing KEGG metabolites without chemical formula associated. Special attention was given to these reactions to ensure that they would not impair the flux through the network if one these reactions were able to make the flux unstable it was immediately removed (35 reactions removed). Additionally, 70 drains were ignored since they were drains or reactions associated with the biomass pathway. In Table 11 are summarized the number of occurrences for the primary cases during the mass balance curation procedure.

Table 11. Summary of reactions corrected according to the mass balance curation.

Cases	Number of Reactions
Protons added to products	16
Products added to reactants	24
Protons removed from products	2
Protons removed from reactants	4
Removed from the model	35
Others	7
Ignored	70

4.1.2 Biomass and energy requirements formulation

In this section, the biomass was “break down” into six entities, classified as “e-Metabolites” (electronic metabolites). The purpose of this classification is enabling the model to represent biological macromolecules and cell structures. In the six entities are summarized, the fraction of each one in the overall biomass composition, grams of cellular dry weight (CDW) normalized to 1 gram of biomass, and the reference study.

Table 12. Biomass composition of *Syntrophobacter fumaroxidans* MPOBT.

e-Metabolite	Stoichiometry (wt/wt)	Reference
e-DNA	0.025	[48]
e-RNA	0.152	
e-Protein	0.591	
e-Lipid	0.094	
e-Cofactor	0.054	
e-Peptidoglycan	0.084	

The “e-Protein” represents the average cellular protein composition. The merlin’s tool “e-biomass” was able to retrieve the stoichiometric coefficients of every amino acid, listed in Table 13, by performing the codon usage method [40]. The “e-Protein” metabolite content was defined as 0.591 grams of protein per 1 gram of biomass according to [48].

Table 13. Protein composition of *Syntrophobacter fumaroxidans* MPOBT. The R present in every chemical formula represents the R group abbreviation, meaning it represents any group or any formula linked to a carbon or hydrogen atom on the rest of the molecule.

e-Protein precursor	Chemical Formula	Stoichiometric coefficient
L-Cysteinyl-tRNA(Cys)	C18H26N6O11PSR(C5H8O6PR)n	0.1060
L-Histidyl-tRNA(His)	C16H24N3O11PR2(C5H8O6PR)n	0.1706
L-Tryptophanyl-tRNA(Trp)	C26H31N7O11PR(C5H8O6PR)n	0.0995
L-Methionyl-tRNA	C20H30N6O11PSR(C5H8O6PR)n	0.1982
L-Prolyl-tRNA(Pro)	C15H24N3O11PR2(C5H8O6PR)n	0.3867

Glutaminyl-tRNA	C20H29N7O12PR(C5H8O6PR)n	0.5250
L-Tyrosyl-tRNA(Tyr)	C24H30N6O12PR(C5H8O6PR)n	0.2198
L-Arginyl-tRNA(Arg)	C21H33N9O11PR(C5H8O6PR)n	0.6083
L-Asparaginyt-tRNA(Asn)	C14H23N2O12PR2(C5H8O6PR)n	0.2251
L-Phenylalanyl-tRNA(Phe)	C19H26N11PR2(C5H8O6PR)n	0.3336
L-Threonyl-tRNA(Thr)	C14H24N12PR2(C5H8O6PR)n	0.3734
Glycyl-tRNA(Gly)	C12H20N11PR2(C5H8O6PR)n	0.5682
L-Aspartyl-tRNA(Asp)	C14H22N13PR2(C5H8O6PR)n	0.4057
L-Seryl-tRNA(Ser)	C13H22N12PR2(C5H8O6PR)n	0.4286
L-Alanyl-tRNA	C13H22N11PR2(C5H8O6PR)n	0.6731
L-Glutamyl-tRNA(Glu)	C20H28N6O13PR(C5H8O6PR)n	0.2324
L-Valyl-tRNA(Val)	C20H30N6O11PR(C5H8O6PR)n	0.5826
L-Lysyl-tRNA	C16H29N2O11PR2(C5H8O6PR)n	0.3394
L-Isoleucyl-tRNA(Ile)	C21H32N6O11PR(C5H8O6PR)n	0.4297
L-Leucyl-tRNA	C21H32N6O11PR(C5H8O6PR)n	0.7934

The fractions 0.025 and 0.152 grams of DNA and RNA, respectively, per 1 gram of biomass were selected from [48]. *merlin's* tool “e-biomass” assembled the DNA and RNA reactions by using the genome sequence data to retrieve information about the general deoxyribonucleotide and ribonucleotide composition. In Table 14 and Table 15 are presented both DNA and RNA structural units, respectively, along with the corresponding stoichiometric coefficient.

Table 14. DNA composition of *Syntrophobacter fumaroxidans* MPOBT.

e-DNA	Chemical Formula	Stoichiometric coefficient
dATP	C10H16N5O12P3	0.4156
dCTP	C9H16N3O13P3	0.5999
dTTP	C10H17N2O14P3	0.4065
dGTP	C10H16N5O13P3	0.6302

Table 15. RNA composition of *Syntrophobacter fumaroxidans* MPOBT.

e-RNA	Chemical Formula	Stoichiometric coefficient
ATP	C10H16N5O13P3	0.4804
CTP	C9H16N3O14P3	0.4672
UTP	C9H15N2O15P3	0.3718
GTP	C10H16N5O14P3	0.6702

In Table 16 are listed the metabolites belonging to the “e-Cofactor” metabolite. These metabolites are grouped within the cofactors group and are neither building blocks nor energy metabolism associated compounds, but in some cases, they can be associated with the catalysis of enzymes. The stoichiometry coefficient given to these metabolites varies from one another depending on their molecular weight.

Table 16. Cofactors composition of *Syntrophobacter fumaroxidans* MPOBT.

e-cofactor	Chemical formula	Stoichiometric coefficient
Thiamine	C ₁₂ H ₁₇ N ₄ O ₅ S	0.3141
Tetrahydrofolate	C ₁₉ H ₂₃ N ₇ O ₆	0.1871
S-Adenosyl-L-methionine	C ₁₅ H ₂₂ N ₆ O ₅ S	0.2091
Riboflavin	C ₁₇ H ₂₀ N ₄ O ₆	0.2214
Pyridoxal phosphate	C ₈ H ₉ N ₃ O ₃	0.4985
Pantothenate	C ₉ H ₁₇ N ₃ O ₅	0.3801
NADPH	C ₂₁ H ₃₀ N ₇ O ₁₇ P ₃	0.1118
NAD ⁺	C ₂₁ H ₂₈ N ₇ O ₁₄ P ₂	0.1254
Glutathione	C ₁₀ H ₁₇ N ₃ O ₆ S	0.2712
FMN	C ₁₇ H ₂₁ N ₄ O ₉ P	0.1826
FAD	C ₂₇ H ₃₃ N ₉ O ₁₅ P ₂	0.1061
CoA	C ₂₁ H ₃₆ N ₇ O ₁₆ P ₃ S	0.1086

The lipidic composition of *Syntrophobacter fumaroxidans* MPOBT was assumed to be the same present in the *Geobacter sulfurreducens* GSM model. Therefore, the lipidic fraction was assumed to be constituted by the monomeric units the fatty acids. In Table 17 are presented the fatty acids that were retrieved from the *G. sulfurreducens* and are present in the metabolic network of *S. fumaroxidans*. Note that some of the fatty acids present *G. sulfurreducens* metabolic network were not present in *S. fumaroxidans* GSM model. Therefore, the molar fraction of the missing fatty acids was distributed along the ones present in both models.

Table 17. *S. fumaroxidans* fatty acid composition

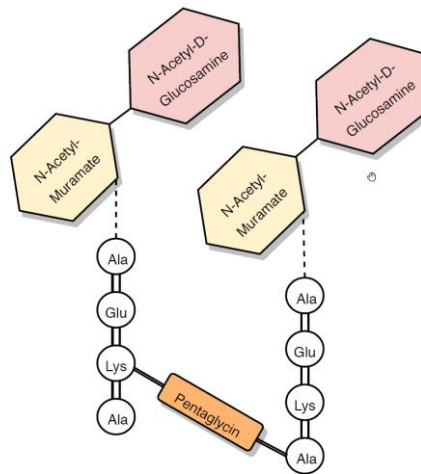
Fatty Acid	Chemical Formula	Stoichiometric coefficient
Tetradecanoic acid	C ₁₄ H ₂₈ O ₂	0.1992
Hexadecanoic acid	C ₁₆ H ₃₂ O ₂	1.5312
(9Z)-Hexadecenoic acid	C ₁₆ H ₃₀ O ₂	1.8695
Octadecanoic acid	C ₁₈ H ₃₆ O ₂	0.0958
(9Z)-Octadecenoic acid	C ₁₈ H ₃₄ O ₂	0.2208

The peptidoglycan composition was copied from the *G. sulfurreducens* model. Usually, the peptidoglycan structure unit has a glycan filament composed by one N-Acetyl-Muramate connected to one N-Acetyl-D-Glucosamine. Next, a peptide subunit composed of four amino acids is covalently linked to the glycan filament. The four amino acids are covalently linked to each other in the following order alanine, glutamate, lysine, and alanine.

Table 18. *S. fumaroxidans* peptidoglycan composition

e-Peptidoglycan	Chemical Formula	Stoichiometric coefficient
UDP-N-acetylmuramate	C ₂₀ H ₃₁ N ₃ O ₁₉ P ₂	0.7893875
UDP-N-acetyl- α -D-glucosamine	C ₁₇ H ₂₇ N ₃ O ₁₇ P ₂	0.5160856
UDP-glucose	C ₁₅ H ₂₄ N ₂ O ₁₇ P ₂	0.2733019

In Figure 15 is presented a schematic representation of the peptidoglycan structure where alanine is connected to the N-Acetyl-Muramate monomer, while lysine is connected to an interpeptide bridge composed of monomers of glycine. Due to the lack of data regarding the amino acid percentage in peptidoglycan composition, the peptidoglycan fraction in *S. fumaroxidans* did not comprise stoichiometric coefficients for the monomeric units of proteins.

**Figure 15.** Schematic representation of the peptidoglycan structure in *S. fumaroxidans*.

In terms of energy requirements, the *GSM* model of *G. sulfureducens* was used once again and the standards for this equation are presented in Table 19. The “R_ATP maintenance” reaction represents the non-growth-associated.

Table 19. Energy requirements of *Syntrophobacter fumaroxidans* MPOBT.

Energy requirement	Reaction	Stoichiometry coefficient
Growth-Associated	e-Biomass	41
Non-Growth-Associated	R_ATP_maintenance	0.45

4.2 GENOME-SCALE METABOLIC MODEL OF *M. hungatei* strain JF1

The *GSM* model reconstruction of *M. hungatei* is presented in this section and it can be accessed by the following URL: <https://nextcloud.bio.di.uminho.pt/s/zfyfRg3BWSzXEES>. The *GSM* model of *Methanosarcina acetivorans* strain C2A [124] was used to guide in the reconstruction of *M. hungatei GSM* model.

4.2.1 Manual curation of the draft metabolic network of *M. hungatei* JF1

At this stage will be presented all the modifications made to the M. hungatei JF1 GSM model during the manual curation.

4.2.1.1 Pathway-by-Pathway analysis

In Table 20 are listed all the pathways, with the respective number of reactions, that were studied during the manual curation phase of the *M. hungatei* draft metabolic network.

Table 20. List of metabolic pathways available in the *GSM* model of *Methanospirillum hungatei* JF1.

Pathway	Number of reactions
2-Oxocarboxylic acid metabolism	27
Alanine, aspartate and glutamate metabolism	15
Amino sugar and nucleotide sugar metabolism	23
Anthocyanin biosynthesis	6
Arginine and proline metabolism	23
Ascorbate and aldarate metabolism	4
Benzoate degradation	7
beta-Alanine metabolism	7
Biosynthesis of amino acids	85
Biotin metabolism	3
Butanoate metabolism	6
Carbon metabolism	52
Chlorocyclohexane and chlorobenzene degradation	4
Citrate cycle (TCA cycle)	8
Cysteine and methionine metabolism	19
D-Glutamine and D-glutamate metabolism	4
Drains pathway	132
Fatty acid biosynthesis	16
Fatty acid metabolism	18
Folate biosynthesis	14
Galactose metabolism	5
Fructose and mannose metabolism	4
Glutathione metabolism	6
Glycerophospholipid metabolism	7

Glycine, serine and threonine metabolism	17
Glycolysis / Gluconeogenesis	19
Glyoxylate and dicarboxylate metabolism	14
Histidine metabolism	15
Inositol phosphate metabolism	7
Lipopolysaccharide biosynthesis	16
Lysine biosynthesis	11
Methane metabolism	91
Naphthalene degradation	8
Nicotinate and nicotinamide metabolism	15
One carbon pool by folate	14
Pantothenate and CoA biosynthesis	17
Pentose and glucuronate interconversions	6
Pentose phosphate pathway	8
Peptidoglycan biosynthesis	12
Phenylalanine metabolism	5
Phenylalanine, tyrosine and tryptophan biosynthesis	21
Porphyrin and chlorophyll metabolism	34
Purine metabolism	45
Pyrimidine metabolism	46
Pyruvate metabolism	19
Starch and sucrose metabolism	5
Streptomycin biosynthesis	6
Terpenoid backbone biosynthesis	13
Thiamine metabolism	8
Transporters pathway	139
Tryptophan metabolism	7
Ubiquinone and other terpenoid-quinone biosynthesis	8
Valine, leucine and isoleucine biosynthesis	15
Valine, leucine and isoleucine degradation	13

During the analysis of the listed pathways above, 66 non-gene associated reactions were manually added to the model gap-fill the model. In the next section, those reactions will be presented separately by metabolite category, thus ordering these by pathways.

4.2.1.2 *Gap Filling*

The initial simulations using the draft metabolic model revealed some gaps in the network. Some of these gaps were found in the methane biosynthesis pathway and were further studied in detail. Regarding the gapfilling of *Methanospirillum hungatei* JF1 draft *GSM* model, it was mainly performed on the pathways of methane production. The original draft model of this organism was not able to produce methane since the metabolite **methanofuran**, a crucial precursor in the methane pathway, was identified as a dead-end metabolite.

The first stage to fill the gap created by the dead-end metabolite was to introduce a complete synthesis pathway for methanofuran. A total of 16 non-gene associated reactions were added to the model based on literature found for methanofuran biosynthesis [125–128]. All the reactions, with the respective reaction id (“_reac”), added to the model to build the methanofuran biosynthesis pathway are listed in Table 21.

Table 21. Reactions added to the metabolic network of *M. hungatei* JF1 in order to create a pathway for methanofuran biosynthesis. The following letters are unique for this table: A - 4-[N-gamma-L-glutamyl)-p-(beta-aminoethyl)phenoxy-methyl]-2-(aminomethyl)furan_c0; B - 4-[N-gamma-L-glutamyl-gamma-L-glutamyl)-p-(beta-aminoethyl)phenoxy-methyl]-2-(aminomethyl)furan_c0.

Reaction ID	Reaction
R_reac002	H_plus__c0 + Phosphoenolpyruvate_c0 + Glycerone_phosphate_c0 => methanofuran biosynththesis intermediate MF1_c0
R_reac003	Methanofuran biosynththesis intermediate MF1_c0 <=> methanofuran biosynththesis intermediate MF2_c0 + Phosphate_c0
R_reac004	methanofuran biosynththesis intermediate MF2_c0 <=> methanofuran biosynththesis intermediate MF3_c0 + H2O_c0
R_reac005	methanofuran biosynththesis intermediate MF3_c0 <=> phosphate-ester-of-dihydrofuran_c0
R_reac006	phosphate-ester-of-dihydrofuran_c0 <=> 2,4-substituted-furan phosphate_c0 + H2O_c0
R_reac007	NADPH_c0 + H_plus__c0 + ATP_c0 + 2,4-substituted-furan phosphate_c0 <=> NADP_c0 + ADP_c0 + Phosphate_c0 + 2-furaldehyde phosphate_c0
R_reac008	L-Alanine_c0 + 2-furaldehyde phosphate_c0 <=> pyruvate + 2-methylamine-furan phosphate_c0
R_reac009	2-methylamine-furan phosphate_c0 + Tyramine_c0 <=> Phosphate_c0 + p-(beta-aminoethyl)phenoxy-methyl-2-(aminomethyl)furan_c0
R_reac010	L_Glutamate_c0 + p-(beta-aminoethyl)phenoxy-methyl-2-(aminomethyl)furan_c0 <=> A_c0 + H2O_c0
R_reac011	A_c0 + L_Glutamate_c0 <=> B_c0 + H2O_c0
R_reac012	2_Oxoglutarate_c0 + Acetyl_CoA_c0 <=> H_plus__c0 + trans-homoaconitate_c0 + CoA_c0
R_reac013	Acetyl_CoA_c0 + trans-homoaconitate_c0 + H2O_c0 <=> H_plus__c0 + pentane-1,3,4,5-tetracarboxylate_c0 + CoA_c0
R_reac014	H_plus__c0 + pentane-1,3,4,5-tetracarboxylate_c0 + CO2_c0 => hexane-6-keto-1,3,4,6-tetracarboxylate_c0 + H2O_c0
R_reac015	H_plus__c0 + hexane-6-keto-1,3,4,6-tetracarboxylate_c0 => hexane-6-ol-1,3,4,6-tetracarboxylate_c0
R_reac016	H_plus__c0 + hexane-6-ol-1,3,4,6-tetracarboxylate_c0 => HTCA_c0 + H2O_c0
R_reac017	B_c0 + HTCA_c0 <=> H2O_c0 + Methanofuran_c0

The gap filling curation method identified another four metabolites without a biosynthesis pathway. Therefore, they were only being consumed and never produced. One of them is named **tetrahydromethanopterin** (H4MPT), and as for the methanofuran metabolite, there was a need to create a biosynthesis pathway for this metabolite too. The use of literature as the information retrieved from the MetaCyc database helped to assemble this pathway [129–134]. In

Table 22 are listed the non-gene associated reactions that compose the H4MPT biosynthesis pathway.

Table 22. Reactions added to the metabolic network of *M. hungatei* JF1 in order to create a pathway for H4MPT biosynthesis. The following letters are unique for this table: A – 4-(?D-ribofuranosyl)hydroxybenzene 5'-phosphate; B – 4-(?D-ribofuranosyl)-N-succinylaminobenzene 5'-phosphate; C – [1-(2-amino-7-methyl-4-oxo-7,8-dihydro-3H-pteridin-6-yl)]ethyl diphosphate; D – [1-(2-amino-7-methyl-4-oxo-7,8-dihydro-3H-pteridin-6-yl)]ethyl-4-(?D-ribofuranosyl)aminobenzene 5'-phosphate; E – [1-(2-amino-7-methyl-4-oxo-7,8-dihydro-3H-pteridin-6-yl)]ethyl-(4-aminophenyl)-1-deoxy-D-ribitol 5'-phosphate; F – [1-(2-amino-7-methyl-4-oxo-7,8-dihydro-3H-pteridin-6-yl)]ethyl-(4-aminophenyl)-1-deoxy-D-ribitol; G – [1-(2-amino-7-methyl-4-oxo-7,8-dihydro-3H-pteridin-6-yl)]ethyl-(4-aminophenyl)-1-deoxy-5-[1-? -D- ribofuranosyl 5-phosphate]-D-ribitol; H – [1-(2-amino-7-methyl-4-oxo-7,8-dihydro-3H-pteridin-6-yl)]ethyl-(4-aminophenyl)-1-deoxy-5-[1-? -D- ribofuranosyl triphosphate]-D-ribitol

Reaction ID	Reaction
R_reac018	chorismate_c0 => 4-hydroxybenzoate_c0 + pyruvate_c0
R_reac019	4-hydroxybenzoate_c0 + PRPP_c0 => A + CO2_c0 + PPi
R_reac020	A + L-aspartate_c0 + ATP_c0 => B + ADP_c0 + Phosphate_c0 + 2H_plus_c0
R_reac021	B => 4-(?D-ribofuranosyl)aminobenzene-5'-phosphate_c0 + fumarate
R_reac022	2 S-adenosyl-L-methionine + 6-hydroxymethyl-dihydropterin + 2 5,10-methylene-tetrahydromethanopterin + NADH => 2-amino-6-[1-hydroxyethyl]-7-methyl-7,8-dihydropterin + 2 5'-deoxyadenosine + 2 L-methionine + 2 7,8-dihydromethanopterin + NAD+
R_reac023	2-amino-6-[1-hydroxyethyl]-7-methyl-7,8-dihydropterin + ATP => C + AMP + H_plus
R_reac024	4-(?D-ribofuranosyl)aminobenzene-5'-phosphate + C => D + PPi
R_reac025	D + NADH + H+ => E + NAD+
R_reac026	E + H2O => F + Phosphate
R_reac027	F + 5-phospho-?D-ribose 1-diphosphate => G + PPi
R_reac028	G + ATP + H+ => H + AMP
R_reac029	H + (S)-2-hydroxyglutarate => 7,8-dihydromethanopterin + PPi
R_reac031	7,8-dihydromethanopterin + NADH + H+ => H4MPT + NAD+

The other metabolite that had to be analyzed was **coenzyme M**, also involved in the methane biosynthesis pathway. Based on the literature and the MetaCyc database a pathway for the biosynthesis of coenzyme M (CoM) was created [135–137]. The non-gene associated reactions assembled into the model for the CoM biosynthesis are listed in

Table 23, and each one has a specific ID starting with “R_M” followed by a number.

Table 23. Reactions present in the coenzyme M biosynthesis.

Reaction ID	Reaction
-------------	----------

R_M001	Sulfite + Phosphoenolpyruvate => (2R)-O-Phospho-3-sulfolactate
R_M002	(2R)-O-Phospho-3-sulfolactate + H ₂ O => (2R)-3-Sulfolactate + phosphate
R_M003	(2R)-3-Sulfolactate + NAD ⁺ <=> 3-Sulfoypyruvate + NADH + H ⁺
R_M004	3-Sulfoypyruvate + H ⁺ => Sulfoacetaldehyde + CO ₂
R_M005	Sulfoacetaldehyde + H ⁺ + NADPH + L-cysteine => NADP + CoM + Pyruvate + NH ₃

The last metabolite identified as a dead-end was **coenzyme B** (HTP) that like the previous two gap metabolites was only consumed in the metabolic network and never produced. By using information retrieved from the literature and MetaCyc database, a coenzyme M biosynthesis pathway was created [138–143]. The non-gene associated reactions added to this pathway are presented in Table 24 with the assigned identifiers.

Table 24. Reactions present in the coenzyme B biosynthesis pathway.

Reaction ID	Reaction
R_M007	Homocitrate => Homoaconitate + H ₂ O
R_M008	homocitrate + NAD ⁺ <=> 2-oxodipate + CO ₂ + NADH
R_M009	2-oxodipate + Acetyl-CoA + H ₂ O <=> dihomocitrate + CoA + H ⁺
R_M010	dihomocitrate => homo2aconitate + H ₂ O
R_M011	homo2aconitate + H ₂ O <=> homo2citrate
R_M012	homo2citrate + NAD ⁺ <=> 2-oxopimelate + CO ₂ + NADH
R_M013	trihomocitrate2-oxopimelate + acetyl-CoA + H ₂ O <=> trihomocitrate + CoA + H ⁺
R_M014	trihomocitrate <=> homo3aconitate + H ₂ O
R_M015	homo3aconitate + H ₂ O => homo3citrate
R_M016	homo3citrate + NAD ⁺ => 2-oxosuberate + CO ₂ + NADH
R_M017	2-oxosuberate + H ⁺ => 7-oxoheptanoate + CO ₂
R_M018	7-oxoheptanoate + H ₂ S + NADPH => 7-mercaptoheptanoate + H ₂ O + NADP
R_M019	7-mercaptoheptanoate + L-threonine + ATP <=> 7-mercaptoheptanoylthreonine + ADP + Phosphate + H ⁺
R_M020	7-mercaptoheptanoylthreonine + ATP <=> HTP + ADP

The last metabolite identified as dead-end in the methanogenic pathway of *M. hungatei* JF1 was the **Coenzyme F420**. The same procedure used in the last three metabolites was used, and therefore, a biosynthesis pathway was created based on the same sources, such as literature and the MetaCyc database [144–146]. All reactions retrieved from those sources, and associated unique identifiers are listed in Table 25.

Table 25. Reactions present in the coenzyme F420 biosynthesis pathway.

Reaction ID	Reaction
R_reac032	2-oxoglutarate + NADH + H ₊ <=> (S)-2-hydroxyglutarate + NAD ⁺
R_reac033	formaldehyde + H ₄ MPT => 5,10-methylene tetrahydromethanopterin + H ₂ O + H ⁺

R_reac034	p-Hydroxyphenylpyruvate + 4-1-D-Ribitylamino-5-aminouracil + 2 S-Adenosyl-L-methionine + H ₂ O => 7,8-Didemethyl-8-hydroxy-5-deazariboflavin + L-Methionine + Deoxyadenosine + Oxalate + NH ₃ + H ⁺
R_reac035	Glycerol-3-phosphate + GTP => lactyl-(2)-diphospho-(5')-guanosine + PPi
R_reac036	7,8-Didemethyl-8-hydroxy-5-deazariboflavin + Lactyl-2-diphospho-5'-guanosine => Coenzyme F420-0 + GMP
R_reac037	Coenzyme F420-0 + GTP + L-Glutamate => Coenzyme F420-1 + GDP + Phosphate
R_reac038	Coenzyme F420-1 + GTP + L-Glutamate => Coenzyme F420 + GDP + Phosphate

As seen above, during manual curation of the methanogenic pathway in *Methanospirillum hungatei* JF1 a total of five metabolites were identified as dead-ends. These metabolites block the flux through this pathway when performing techniques to evaluate the flux distribution. As mentioned before in this section, one of the main pathways in *M. hungatei* is the methanogenic pathway, and therefore, it was studied in detail. This pathway with red colored metabolites representing the gap metabolites blocking the flux distribution is presented in Figure 16.

The strategy used to correct the gaps started by the analysis of the pathway by first correcting the dead-end metabolites at the beginning of the pathway such as methanofuran. The reactions where these metabolites participate as reactants were turned into irreversible to make the flux distribution only to follow one direction to check if the gap was corrected. If in the end, the reaction was able to produce the products the gap was assigned as corrected. The last metabolite to be studied and corrected was HTP which intervenes in the final reaction of the pathway. After all the gaps in this pathway were corrected all the reactions presented flux distribution meaning that the manual curation was successful, and the gram-negative bacterium *Methanospirillum hungatei* JF1 was then able to produce methane.

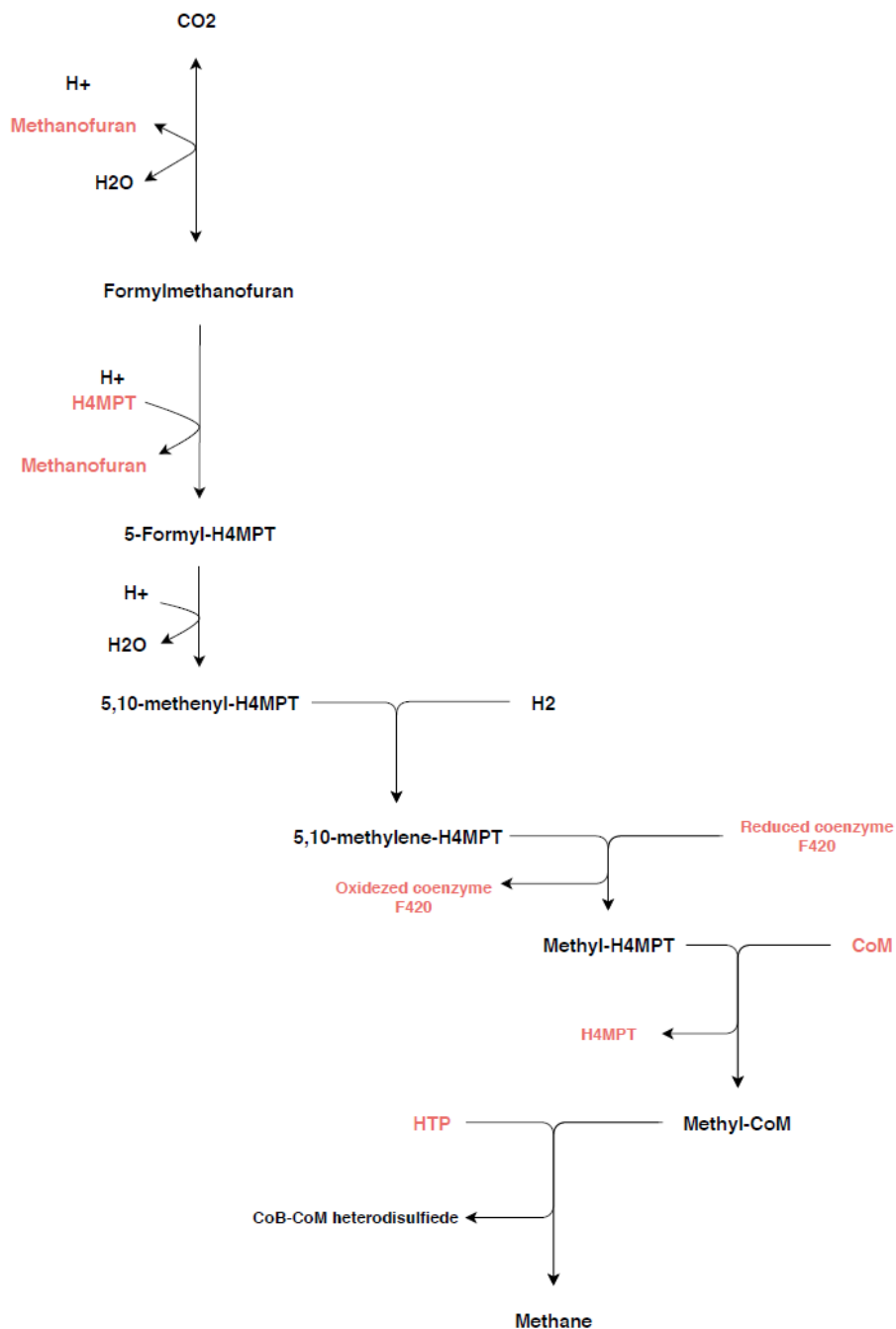


Figure 16. Methanogenic pathway in *Methanospirillum hungatei* JF1. The metabolites coloured in red were the ones identified as gaps in this pathway. CoM – Coenzyme M; H4MPT – tetrahydronopterin; HTP or CoB – Coenzyme B; Adapted from [147]

4.2.1.3 Mass Balance

The mass balance for the *GSM* draft model of *M. hungatei* was performed in the same terms as for *S. fumaroxidans*. The tool “unbalanced reactions” was used and reactions identified as unbalanced were highlighted and later were either removed or balanced.

4.2.2 Biomass and energy requirements formulation

The same methodology used in the formulation of the biomass equation for *S. fumaroxidans* was performed to achieve the biomass equation of *M. hungatei*. As for *S. fumaroxidans*, six entities were created, labelled as “e-Metabolites” that represent bacterial cell components of *M. hungatei* such as biological macromolecules and cell structures. In Table 26 below is listed the macromolecules composing the biomass equation with the respective stoichiometry coefficient in grams per CDW normalized to one gram of biomass.

Initially, the biomass equation comprised all macromolecules and their precursor. In order to obtain a clearer perception of the biomass components, it was divided into the six entities mentioned, namely e-Protein, e-DNA, e-RNA, e-Cofactors, e-Lipid, and e-Peptidoglycan.

Table 26. Biomass composition of *Methanospirillum hungatei* JF1

e-Metabolite	Stoichiometry Coefficient (wt/wt)	Reference
e-DNA	0.025	[48]
e-RNA	0.152	
e-Protein	0.591	
e-Lipid	0.054	KBase
e-Cofactors		
e-Peptidoglycan		

The “Protein biosynthesis” metabolite in Table 27 corresponds to the “e-Protein” metabolite present in Table 26. This metabolite represents the average cellular protein composition. Using *merlin*’s “e-biomass tool”, the stoichiometry coefficients for each amino acid were retrieved.

Table 27. Protein composition of *Methanospirillum hungatei* JF1.

Macromolecule	Metabolite	Formula	Stoichiometry Coefficient (mmol monomer/ g e-Protein)
Protein biosynthesis	L-Phenylalanine	C ₉ H ₁₁ N ₂ O ₂	0.3193
	L-Glutamine	C ₅ H ₁₀ N ₂ O ₃	0.2512
	L-Tryptophan	C ₁₁ H ₁₂ N ₂ O ₂	0.0864
	L-Methionine	C ₅ H ₁₁ N ₂ O ₂ S	0.2080
	L-Glutamate	C ₅ H ₈ N ₂ O ₄	0.2512
	L-Asparagine	C ₄ H ₈ N ₂ O ₃	0.2800
	L-Serine	C ₃ H ₇ N ₂ O ₃	0.4988
	L-Valine	C ₅ H ₁₁ N ₂ O ₂	0.4968
	L-Isoleucine	C ₆ H ₁₃ N ₂ O ₂	0.6612
	L-Proline	C ₅ H ₉ N ₂ O ₂	0.3537

L-Lysine	C ₆ H ₁₅ N ₂ O ₂	0.3968
L-Aspartate	C ₄ H ₆ N ₂ O ₄	0.4473
L-Alanine	C ₃ H ₇ N ₂ O ₂	0.4995
L-Histidine	C ₆ H ₉ N ₃ O ₂	0.1752
L-Arginine	C ₆ H ₁₅ N ₄ O ₂	0.4178
L-Cysteine	C ₃ H ₇ N ₂ O ₂ S	0.1051
L-Leucine	C ₆ H ₁₃ N ₂ O ₂	0.7099
L-Tyrosine	C ₉ H ₁₁ N ₂ O ₃	0.2785
L-Threonine	C ₄ H ₉ N ₂ O ₃	0.4403
Glycine	C ₂ H ₅ N ₂ O ₂	0.4999

The stoichiometric coefficients of both nucleic acids were assumed based on the literature available for the biomass composition of Gran-Negative bacteria [48]. The fractions 0.025 and 0.152 grams of DNA and RNA, respectively, per one 1 gram of biomass were selected. Both DNA and RNA precursors in triphosphate forms are present in Table 28 and Table 29, respectively.

Table 28. DNA composition of *Methanospirillum hungatei* JF1.

Macromolecule	Metabolite	Formula	Stoichiometry Coefficient (mmol monomer/ g e-DNA)
e-DNA	dATP	C ₁₀ H ₁₆ N ₅ O ₁₂ P ₃	0.4156
	dCTP	C ₉ H ₁₆ N ₃ O ₁₃ P ₃	0.5999
	dTTP	C ₁₀ H ₁₇ N ₂ O ₁₄ P ₃	0.4065
	dGTP	C ₁₀ H ₁₆ N ₅ O ₁₃ P ₃	0.6302

Table 29. RNA composition of *Methanospirillum hungatei* JF1.

Macromolecule	Metabolite	Formula	Stoichiometry Coefficient (mmol monomer/ g e-RNA)
e-RNA	ATP	C ₁₀ H ₁₆ N ₅ O ₁₃ P ₃	0.5056
	CTP	C ₉ H ₁₆ N ₃ O ₁₄ P ₃	0.5436
	UTP	C ₉ H ₁₅ N ₂ O ₁₅ P ₃	0.4209
	GTP	C ₁₀ H ₁₆ N ₅ O ₁₄ P ₃	0.5297

The cofactors present in the *GSM* model of *Methanospirillum hungatei* are listed in Table 30. The “e-Cofactor” metabolite represent a group of metabolites constituted by cofactors, enzymes, and ions. All the cofactors were automatically assigned by KBase when building the draft metabolic model.

Table 30. Cofactors composition in *Methanospirillum hungatei* JF1.

e-Metabolite	Metabolite	Formula	Stoichiometry Coefficient (mmol monomer/ g e-Cofactor)
e-Cofactors	Thiamine	C ₁₂ H ₁₇ N ₄ O ₅ S	0.114200
	Tetrahydrofolate	C ₁₉ H ₂₃ N ₇ O ₆	0.068031
	S-Adenosyl-L-methionine	C ₁₅ H ₂₂ N ₆ O ₅ S	0.076054
	Riboflavin	C ₁₇ H ₂₀ N ₄ O ₆	0.080516
	Pyridoxal phosphate	C ₈ H ₉ N ₃ O ₃	0.181282
	Pantothenate	C ₉ H ₁₇ N ₅ O ₅	0.138225
	NADP	C ₂₁ H ₃₀ N ₇ O ₁₇ P ₃	0.040652
	NAD ⁺	C ₂₁ H ₂₈ N ₇ O ₁₄ P ₂	0.045608
	GSH	C ₁₀ H ₁₇ N ₃ O ₆ S	0.098604
	FMN	C ₁₇ H ₂₁ N ₄ O ₉ P	0.066405
	FAD	C ₂₇ H ₃₃ N ₉ O ₁₅ P ₂	0.038576
	CoA	C ₂₁ H ₃₆ N ₇ O ₁₆ P ₃ S	0.039481
	Siroheme	C ₄₂ H ₃₆ FeN ₄ O ₁₆	0.033351
	Spermidine	C ₇ H ₂₂ N ₃	0.204377
	Co ²⁺	Co	0.514221
	Mg	Mg	1.246525
	Mn ²⁺	Mn	0.551566
	Ubiquinone-8	C ₄₉ H ₇₄ O ₄	0.041676
	Fe ²⁺	Fe	0.542628
	Fe ³⁺	Fe	0.542628
	Putrescine	C ₄ H ₁₄ N ₂	0.336066
	Ca ²⁺	Ca	0.756064
	10-Formyltetrahydrofolate	C ₂₀ H ₂₁ N ₇ O ₇	0.064280
	Cu ²⁺	Cu	0.476838
	Cl ⁻	Cl	0.854810
	Zn ²⁺	Zn	0.463420
	TPP	C ₁₂ H ₁₇ N ₄ O ₇ P ₂ S	0.071588
	Menaquinone 8	C ₅₁ H ₇₂ O ₂	0.042257
	K ⁺	K	0.775014
	2-Demethylmenaquinone 8	C ₅₀ H ₇₀ O ₂	0.043100
	5-Methyltetrahydrofolate	C ₂₀ H ₂₃ N ₇ O ₆	0.066245
	Calomide	C ₇₂ H ₁₀₂ CoN ₁₈ O ₁₇ P	0.019160

Different from the approach made for *S. fumaroxidans* the lipid composition of *M. hungatei* was retrieved from KBase. KBase assigned the lipidic composition in *Methanospirillum hungatei* JF1 as for a gram-negative bacterium. The “e-lipid” metabolite represents the lipidic fraction in *M. hungatei*, and at Table 31 are listed all the lipids with the respective stoichiometry coefficient assigned.

Table 31. Lipid composition in *Methanospirillum hungatei* JF1. Note that the sum of all stoichiometric coefficients is accounted to one.

e-Metabolite	Metabolite	Formula	Stoichiometry Coefficient (g monomer/ g _{e-lipid})
e-Lipid	Dianteisoheptadecanoyl phosphatidylglycerol	C40H78O10P	0.08575
	Stearoylcardiolipin	C81H156O17P2	0.16740
	Anteisoheptadecanoylcardiolipin	C77H148O17P2	0.16100
	Diisoheptadecanoyl phosphatidylethanolamine	C39H78N08P	0.08232
	Phosphatidylethanolamine dioctadecanoyl	C41H82N08P	0.08553
	Diisoheptadecanoyl phosphatidylglycerol	C40H78O10P	0.08896
	Isoheptadecanoylcardiolipin	C77H148O17P2	0.08576
	Dianteisoheptadecanoyl phosphatidylethanolamine	C39H78N08P	0.16098

The “e-Peptidoglycan” metabolite represents the peptidoglycan fraction of the gram-negative bacterium *M. hungatei*. KBase tools assigned two peptidoglycans for *M. hungatei*, namely, Core oligosaccharide lipid A and Bactoprenyl diphosphate. Both metabolites have the same stoichiometry coefficient as presented in Table 32.

Table 32. Peptidoglycan composition in *Methanospirillum hungatei* JF1.

e-Metabolite	Metabolite	Formula	Stoichiometry Coefficient (mmol monomer/ g _{e-RNA})
e-Peptidoglycan	Core oligosaccharide lipid A	C176H303N20100P4	0.02501
	Bactoprenyl diphosphate	C55H90O7P2	

The energy requirements for *M. hungatei* in terms of ATP were defined based on the *GSM* model of *Methanosarcina acetivorans* and both growth-associated and non-growth-associated reactions are listed in Table 33.

Table 33. Energy requirements of *Methanospirillum hungatei* JF1.

Energy Requirements	Reaction	Stoichiometry Coefficient
Growth-Associated	e-Biomass	47
Non-Growth-Associated	R_ATP_Maintenance	0.6

4.3 GSM MODELS COMPARISON

In this section is presented a comparison between models reconstructed in KBase and *merlin*.

4.3.1 KBase GSM models

GSM models' reconstructions were performed in KBase for *M. hungatei* JF1 and *S. fumaroxidans* MPOBT, and the results of these reconstructions can be accessed by the following URLs:

- *M. hungatei* narrative - <https://narrative.kbase.us/narrative/ws.35875.obj.1>
- *S. fumaroxidans* narrative - <https://narrative.kbase.us/narrative/ws.36026.obj.1>

In order to access these results, a KBase account must be created. Both GSM draft models reconstructed in KBase were exported in the SBML file format and are available at the following URLs:

- *M. hungatei*: <https://nextcloud.bio.di.uminho.pt/s/caLX9TrqDrDA389>
- *S. fumaroxidans*: <https://nextcloud.bio.di.uminho.pt/s/2GAC5ZFeTRZHKap>

Regarding biomass equations in KBase models, the single biomass equation contains reactants such as DNA replication, Protein biosynthesis and RNA transcription without associated pre-precursor, meaning that the only reaction associated to these metabolites is a drain reaction (Table 34). Hence, reactions to synthesize these macromolecules were added to the model to overcome the problem.

Table 34. [c0] – cytosol compartment; (*) – Stoichiometric coefficient relative to biomass equation

GSM model	DNA equation	Protein equation	RNA equation
<i>M. hungatei</i> GSM model (KBase)	=> DNA replication[c0]	=> Protein biosynthesis[c0]	=> RNA transcription[c0]
<i>M. hungatei</i> GSM model (KBase)			
Stoichiometric coefficient (*)	-1	-1	-1

4.3.2. KBase *GSM* model's vs *merlin* *GSM* model's

As demonstrated in Table 35, using *merlin* for *GSM* models' reconstructions is possible to include more reactions and genes in the model when compared with KBase's model. *merlin*'s models differ from KBase models in most of the parameters, which can be justified by the manual curation performed when reconstructing *merlin*'s *GSM* models instead of using fully automatic tools as occurs in KBase.

Table 35. Summary of the principle characteristic of KBase *GSM* models and *merlin* *GSM* models. (*) This *GSM* model was first reconstructed in KBase and later integrated into *merlin*.

<i>GSM</i> model	N. ° of genes	N. ° of reactions	File format	GPR association
<i>M. hungatei</i> KBase <i>GSM</i> model	727	960	SBML	✓
<i>M. hungatei</i> <i>merlin</i> <i>GSM</i> model (*)	727	1027	SBML	✗
<i>S. fumaroxidans</i> KBase <i>GSM</i> model	891	975	SBML	✓
<i>S. fumaroxidans</i> <i>merlin</i> <i>GSM</i> model	1595	1433	SBML	✓

Additionally, reconstructing models with *merlin* seems to be a more viable solution than reconstructing *GSM* models with KBase because *merlin* is more user-friendly. As explained previously, manual curation of *GSM* models in KBase is an arduous task to be done because every modification in the draft model generates a new draft.

In the table above is showed that the *GSM* model of *M. hungatei*, reconstructed using *merlin* does not have GPR associations due to the lack of capacity of the *merlin* tool "import model" to recognize them when importing into *merlin* database the draft model that was previously reconstructed on KBase.

One of the KBase advantages over *merlin* is the way of exporting models. In *merlin* the only available way to it is by exporting in the SBML file format while in KBase the user can export the draft models in SBML, TSV, EXCEL and JSON file formats. On the other hand, *merlin* *GSM* models showed to be well built due to the comfortable and very user-friendly manual curation an entire model presenting tools not only for identifying biomass macromolecules precursors and their stoichiometric coefficients, but also tools for editing reactions.

4.3.3 *GSM* models comparison with literature

A draft model of *M. hungatei* is available (*iMhu428*) [148]. This *GSM* draft model was built by copying the reactions from a *GSM* model of a close phylogenetic organism *Methanosarcina acetivorans* and only removing certain pathways which lead to the non GPR association for a vast number of reactions in the model. A *GSM* model of *S. fumaroxidans* (*iSfu648*) was also found in literature. Both models found on literature were compared against the ones that were assembled in this work.

Since *iMhu428* is only a draft metabolic model and it was obtained by simply copying the reactions from another model, it was assumed that a new *GSM* model reconstruction for this organism was the logic choice in order to better represent the metabolism behaviour of *M. hungatei*.

The *iSfu648* was built using the semi-automatic reannotation tool RavenToolbox with further manual refinement of certain metabolic pathways. During this work another *GSM* model for *S. fumaroxidans* was also assembled but instead of using RavenToolbox, *merlin* was used.

One of the reasons for developing new *GSM* models for both organisms was because one was only a draft model and the other was built using a semi-automated tool. Another reason was to update both *GSM* models to have more confidence in them and assemble a viable community model.

In Table 36 the four *GSM* models are compared, and the *GSM* model fully reconstructed using *merlin* has more reactions than *iSfu648* showing good prospects for further metabolic behavior studies. Regarding the *M. hungatei*, the model reconstructed in this work has more reactions than *iMhu428*. The reconstruction of all four *GSM* models required using information from close phylogenetic organisms.

Table 36. Summary of the comparison of the four *GSM* models built using different approaches. *(a)- GPR association only for some reactions; *(b)- The GPR associations are not yet present in the model; *(c)- The validation was performed upon a draft model

Model	N.º of reactions	Manual curated	GPR association	Model validation	Close phylogenetic organism <i>GSM</i> model
<i>iSfu648</i>	850	✓	✓	✓	✓
<i>Sfu</i>	1507	✓	✓	✗	✓
<i>iMhu428</i>	721	✗	*(a)	*(c)	✓
Mhu	1027	✓	*(b)	✗	✓

4.4 META-MODEL ASSEMBLY

In this section the process of merging *M. hungatei* *GSM* model with *S. fumaroxidans* *GSM* model is presented.

4.4.1 Single model troubleshooting

The troubleshooting process occurred in the same way for both *GSM* models. The tool used for this stage was BioCOISO.

The tool allowed to identify which metabolites, composing the biomass equation, were not being produced by the target organism. Some functions such as *test_e_precursors* and *test_reaction* were convenient for debugging metabolic pathways in order to allow the production of biomass precursors such as amino acids.

The metabolic syntrophic interaction between *S. fumaroxidans* and *M. hungatei* is represented in Figure 17. The metabolism of *S. fumaroxidans* is able to use propionate to produce acetate, formate, and H₂. These metabolites can be used by *M. hungatei* to produce methane. The metabolic pathways represented in this figure are an example of the pathways studied using BioCOISO. For example, during debugging of the metabolic pathway represented in *S. fumaroxidans*, it was found that acetate was not being produced due to the incorrect direction of the reaction that consumes fumarate to produce malate. The error was only found when using the function *test_reaction* for all the reactions in the pathway. This function uses FBA, and the output indicates if the metabolites are being produced in a particular reaction or not.

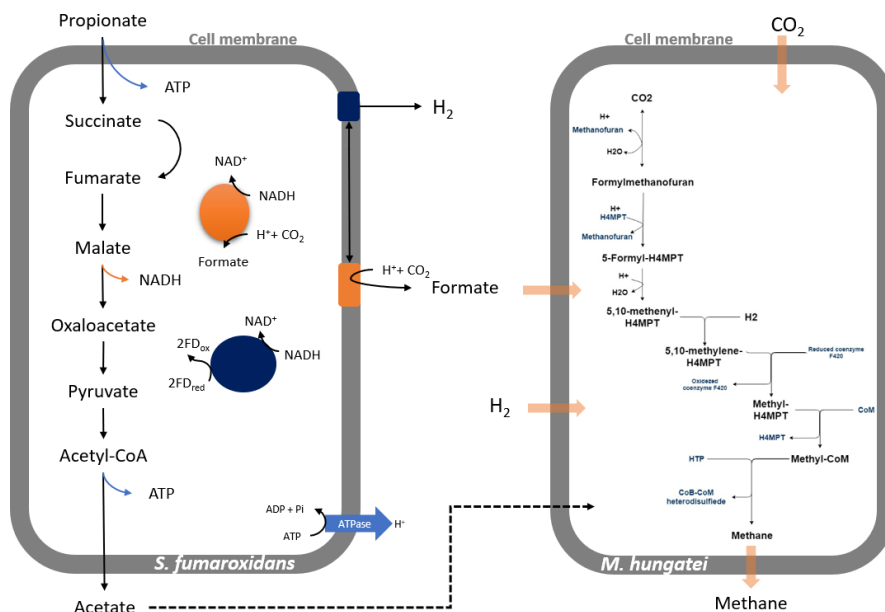


Figure 17. Schematic representation of the metabolic interactions between *S. fumaroxidans* and *M. hungatei*; - Hydrogenase ● - Formate dehydrogenase ●

4.4.2 Meta-Model Assembly troubleshooting

A tool for building a *GSM* metabolic model for a microbial community was used. This tool will merge both *GSM* models that were reconstructed during this work, meaning that a meta-model will be assembled using both *M. hungatei* JF1 and *S. fumaroxidans* MPOBT *GSM* models.

The tool implemented in FRAMED[149] can be accessed in the following URL: <https://github.com/cdanielmachado/framed>. The script allows merging two *GSM* models into a single model with values for biomass reaction and exchange reactions shared by both previous models.

The primary objective of assembling a community model is to study the interactions between the involved organisms. This relationship is represented akin drain reactions in both models. For example, *S. fumaroxidans* excretes acetate, and *M. hungatei* consumes it by using drain reactions.

The *GSM* models are exported from *merlin* in the *SBML* format. The *SBML* file assigns for each metabolite a unique identifier, such as “*M_*****”, M represents “*Metabolite*” and the following five “*” represent the KEGG identifier. Changes were performed in both *GSM* models to merge them, as they were reconstructed in different ways meaning that the internal metabolites identifiers in the *SBML* file will be different for each model. For this reason, in both models, drain reactions with the same metabolite will only differ in the metabolite identifier. This difference would not allow both models to be merged by the selected tool.

The only way to solve the problem was to develop a new tool, capable of overcoming this issue. The developed tool can be accessed in the following URL: <https://nextcloud.bio.di.uminho.pt/s/2eByteTZ47ajLpK>. It has different functions for solving the problems found when trying to merge both models. The first function developed was *get_model1_mets* which retrieves a python dictionary with the names of the metabolites within the keys, and the values are represented by metabolite identifier. These metabolites were retrieved from the exchange (drain) reactions of the first model. An identical function to the previous one, named *get_model2_mets*, was also created to retrieve the exchange metabolites from the second model. Next, a function named *name_matches* was developed only to identify metabolites with the same name in both models, and it retrieves a list with the names of those metabolites.

One of the developed functions named *changeIds_by_name_match* aims to assign the same id in both models for metabolites identified using the *matches* function. The new ids for

these metabolites were assigned as “*NMIC_****” with “*NMIC*” being a diminutive of “*Name Match Identifier Change*” and the following three “*” as numbers.

Another problem found when developing this tool was that in both models, for a few cases, the internal metabolite id in the first model did not match the same metabolite in the second model. The *changeIds_by_id_match* function overcomes this problem by assigning in each model a unique id for those different metabolites. For the first model the ids were changed to the following format “*IMCO_****” and “*IMCO*” stands for “*Identifier Match Change model One*”, and the next three “*” are random numbers. The same classification is applied to the second model but with a slight difference in the format, “*IMCT_****” meaning “*Identifier Match Change model Two*”. The pipeline of the tool is schematized in figure 18.

The names of the compartments in both models were different, and it represented an obstacle for the primary tool to run properly. In this line of view, a straightforward function named *change_compartment_names* was created, and it assigns the names of the first model compartments to the second model compartments names. The function works for this work, but it needs to be developed for cases where the compartment identifier in a model does not match the compartment identifier in the second model.

Some other functions such as *export_model1* and *export_model2* were also added to the tool in order to retrieve the models in the proper way to be used in the primary tool, thus allowing the merge of *S. fumaroxidans* and *M. hungatei* GSM models.

The method developed for implementing the modifications in the model relies on the following principles:

- Metabolites names and ids present in exchange reactions are retrieved for each model and are saved in two different lists;
- If the same metabolite id is found in both lists, each id will be modified to a specific identifier.
- If the metabolite name is found in both lists (match), the id for that metabolite is changed in order to be the same in both models.
- The compartments names from the first model are assigned to the compartment's names of the second model.

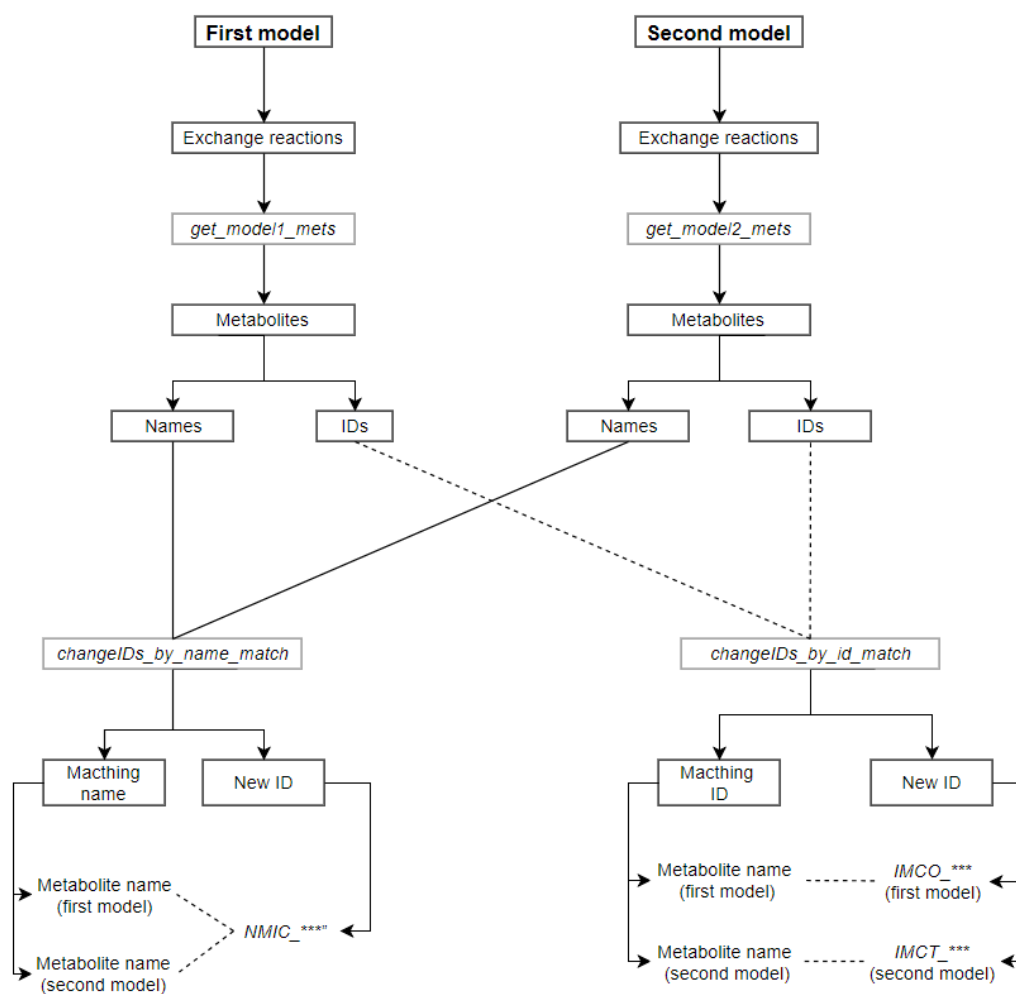


Figure 18. Scheme representing the pipeline of the tool developed to modify both models so they could be merged by the *COMMUNITY* tool. Functions are highlighted by grey boxes.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

In the final chapter of this dissertation, the main conclusions of the present work are drawn, and topics for future work are proposed.

5.1 General conclusions

The main goal of this work was to develop two metabolic models of *Syntrophobacter fumaroxidans* MPOBT and *Methanospirillum hungatei* JF1 based on an up-to-date genome annotation and also to reconstruct a community *GSM* model based on the two reconstructed models. The *GSM* model of *M. hungatei* was first developed in KBase and later imported to *merlin* where manual curation was performed. *S. fumaroxidans* *GSM* model was reconstructed using *merlin*. The reconstructions comprise knowledge retrieved from the organism's genome sequences, biochemical databases and organism-specific literature. These *in silico* assembled models characterize the metabolism behavior of both organisms through complete set of biochemical reactions that occur in each organism to maintain life.

Literature and previous *GSM* models of close phylogenetic organisms to *S. fumaroxidans* and *M. hungatei* were used to determine the overall biomass composition of each organism. An experimental determination of the biomass composition in both organisms would have been a major advantage in this work.

A tool implemented in Python language that uses COBRApy was developed to allow the models to be ready to be merged using one of the FRAMED tools for merging *GSM* models. Other tools not presented on this work were developed in order to retrieve information from *GSM* models of taxonomy close organism such as *Geobacter sulfurreducens*, which speed up some tasks of manual curation of *S. fumaroxidans* and *M. hungatei* *GSM* models.

The comparison of all *GSM* models reconstructed using different tools and different strategies showed that models reconstructed using merlin had more reactions with GPR associated and their reversibility was more consistent with either literature or reliable online biochemical databases.

The *GSM* model of *M. hungatei* JF1 was able to produce methane by consuming formate and acetate as carbon sources when *in silico* simulations were performed.

Regarding the *GSM* model of *S. fumaroxidans* MPOBT, performing *in silico* simulations showed that this model was able to produce acetate and formate by consuming either propanoate or fumarate as carbon sources.

When performing *in silico* simulations on the community *GSM* model it was possible to simulate growth with production and consumption of certain metabolites such as methane and acetate, respectively.

5.2 Future work

The *GSM* models developed in this work can be used separately for studying organism-specific behavior. Additionally, the community *GSM* model reconstructed during this work requires further studying and evaluation. This model should be able to simulate the metabolic behavior of the community formed by *M. hungatei* JF1 and *S. fumaroxidans* MPOBT. In the future, this community model can be used to study the capacity of this community to be used in different applications such as optimizing or minimizing the production of interest metabolites.

In this work two different strategies for reconstructing *GSM* models were used, one consisted of using a semi-automatic tool (KBase), and the other way was to use *merlin* which is not an automatic tool. To date, there is not any *M. hungatei* *GSM* model reconstructed without using a semi-automatic tool, and it would be interesting, in the future, to fully reconstruct a *GSM* model for *M. hungatei* on *merlin* to see if there are significant differences when comparing with the existing *GSM* models.

BIBLIOGRAPHY

1. Stephanopoulos G (1999) Metabolic Fluxes and Metabolic Engineering. *Metab Eng* 1:1–11. <https://doi.org/10.1006/mben.1998.0101>
2. Schellenberger J, Thiele I, Orth JD (2012) Quantitative prediction of cellular metabolism with Constraint based models. *Nat Protoc* 6:1290–1307. <https://doi.org/10.1038/nprot.2011.308>
3. Aziz RK, Bartels D, Best A, et al (2008) The RAST Server: Rapid annotations using subsystems technology. *BMC Genomics* 9:1–15. <https://doi.org/10.1186/1471-2164-9-75>
4. Dias O, Rocha M, Ferreira EC, Rocha I (2015) Reconstructing genome-scale metabolic models with merlin. *Nucleic Acids Res* 43:3899–3910. <https://doi.org/10.1093/nar/gkv294>
5. Morris BEL, Henneberger R, Huber H, Moissl-Eichinger C (2013) Microbial syntrophy: Interaction for the common good. *FEMS Microbiol Rev* 37:384–406. <https://doi.org/10.1111/1574-6976.12019>
6. Thiele I, Heinken A, Fleming RMT (2013) A systems biology approach to studying the role of microbes in human health. *Curr Opin Biotechnol* 24:4–12. <https://doi.org/10.1016/j.copbio.2012.10.001>
7. Mahadevan R, Henson MA (2012) Genome-Based Modeling and Design of Metabolic Interactions in Microbial Communities. *Comput Struct Biotechnol J* 3:e201210008. <https://doi.org/10.5936/csbj.201210008>
8. Kitano H (2002) Systems biology: a brief overview. *Science* 295:1662–4. <https://doi.org/10.1126/science.1069492>
9. Zelezniak A, Pers TH, Soares S, et al (2010) Metabolic network topology reveals transcriptional regulatory signatures of type 2 diabetes. *PLoS Comput Biol* 6:. <https://doi.org/10.1371/journal.pcbi.1000729>
10. Dias O, Rocha I (2015) Systems Biology in Fungi. *Mol Biol Food Water Borne Mycotoxigenic Mycotic Fungi* 69–92
11. Ye J, McGinnis S, Madden TL (2006) BLAST: Improvements for better sequence analysis. *Nucleic Acids Res* 34:6–9. <https://doi.org/10.1093/nar/gkl164>
12. Isalan M (2012) Systems biology: A cell in a computer. *Nature* 488:40–41. <https://doi.org/10.1038/488040a>
13. Otero JM, Nielsen J (2010) Industrial systems biology. *Biotechnol Bioeng* 105:439–460. <https://doi.org/10.1002/bit.22592>
14. Raškevičius V, Mikalayeva V, Antanavičiūtė I, et al (2018) Genome scale metabolic models as tools for drug design and personalized medicine. *PLoS One* 13:1–14. <https://doi.org/10.1371/journal.pone.0190636>
15. Fondi M, Liò P (2015) Bacterial Pangenomics. 1231:233–256. <https://doi.org/10.1007/978-1-4939-1720-4>
16. Kanehisa M (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32:277D–280. <https://doi.org/10.1093/nar/gkh063>
17. Karp PD (2000) The EcoCyc and MetaCyc databases. *Nucleic Acids Res* 28:56–59. <https://doi.org/10.1093/nar/28.1.56>
18. Apweiler R (2009) The universal protein resource (UniProt) in 2010. *Nucleic Acids Res* 38:190–195.

- <https://doi.org/10.1093/nar/gkp846>
19. Rocha I, Forster J, Nielsen J (2008) Rocha, Isabel Jochen Förster, and Jens Nielsen. Microb Gene Essentiality Protoc Bioinforma 416:
 20. Kanehisa M, Sato Y, Furumichi M, et al (2018) New approach for understanding genome variations in KEGG. Nucleic Acids Res. <https://doi.org/10.1093/nar/gky962>
 21. Caspi R, Billington R, Ferrer L, et al (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 44:D471–D480. <https://doi.org/10.1093/nar/gkv1164>
 22. King ZA, Lu J, Dräger A, et al (2016) BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res 44:D515–D522. <https://doi.org/10.1093/nar/gkv1049>
 23. NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 44:D7–D19. <https://doi.org/10.1093/nar/gkv1290>
 24. Saier MH, Reddy VS, Tsu B V., et al (2016) The Transporter Classification Database (TCDB): Recent advances. Nucleic Acids Res 44:D372–D379. <https://doi.org/10.1093/nar/gkv1103>
 25. Henry CS, DeJongh M, Best AA, et al (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol 28:977–982. <https://doi.org/10.1038/nbt.1672>
 26. Jeske L, Placzek S, Schomburg I, et al (2018) BRENDA in 2019: a European ELIXIR core data resource. Nucleic Acids Res 1–8. <https://doi.org/10.1093/nar/gky1048>
 27. Saier MH, Reddy VS, Tsu B V., et al (2016) The Transporter Classification Database (TCDB): recent advances. Nucleic Acids Res 44:D372–D379. <https://doi.org/10.1093/nar/gkv1103>
 28. Chae L, Lee I, Shin J, Rhee SY (2012) Towards understanding how molecular networks evolve in plants. Curr Opin Plant Biol 15:177–184. <https://doi.org/10.1016/j.pbi.2012.01.006>
 29. Tello-ruiz MK, Naithani S, Stein JC, et al (2018) Gramene 2018 : unifying comparative genomics and pathway resources for plant research SEARCH INTERFACE. 46:1181–1189. <https://doi.org/10.1093/nar/gkx1111>
 30. Thiele, Ines; Palsson B (2010) Reconstruction. Nat Protoc 5:93–121. <https://doi.org/10.1038/nprot.2009.203.A>
 31. Consortium TGO (2000) Gene ontologie: Tool for the unification of biology. Nat Genet 25:25–29. <https://doi.org/10.1038/75556.Gene>
 32. Databases E (2010) Data Mining Techniques for the Life Sciences. 609:113–128. <https://doi.org/10.1007/978-1-60327-241-4>
 33. Chang A, Scheer M, Grote A, et al (2009) BRENDA, AMENDA and FRENDA the enzyme information system: New content and tools in 2009. Nucleic Acids Res 37:511–514. <https://doi.org/10.1093/nar/gkn820>
 34. Green ML, Karp PD (2005) Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. Nucleic Acids Res 33:4035–4039. <https://doi.org/10.1093/nar/gki711>
 35. Machado D, Herrgård MJ, Rocha I (2016) Stoichiometric Representation of Gene–Protein–Reaction Associations Leverages Constraint-Based Analysis from Reaction to Gene-Level Phenotype Prediction. PLoS Comput Biol 12:1–24. <https://doi.org/10.1371/journal.pcbi.1005140>

36. Salzberg SL (2007) Genome re-annotation: A wiki solution? *Genome Biol* 8:. <https://doi.org/10.1186/gb-2007-8-1-102>
37. Francke C, Siezen RJ, Teusink B (2005) Reconstructing the metabolic network of a bacterium from its genome. 13:. <https://doi.org/10.1016/j.tim.2005.09.001>
38. NC-IUBMB (2017) Nomenclature Committee of the International Union of Biochemistry and Molecular Biology
39. Essentiality MG (2008) Microbial Gene Essentiality: Protocols and Bioinformatics
40. Thiele, Ines; Palsson B (2010) a Peotocol Generating High Quality Genome-Scale Metabolic Reconstruction. *Nat Protoc* 5:93–121. <https://doi.org/10.1038/nprot.2009.203.A>
41. Petukh M, Steff S, Alexov E (2013) The role of protonation states in ligand-receptor recognition and binding. *Curr Pharm Des* 19:4182–90. <https://doi.org/10.2174/1381612811319230004>
42. Fritzscheier CJ, Hartleb D, Szappanos B, et al (2017) Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. *PLoS Comput Biol* 13:1–14. <https://doi.org/10.1371/journal.pcbi.1005494>
43. Gardy JL, Laird MR, Chen F, et al (2005) PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* 21:617–623. <https://doi.org/10.1093/bioinformatics/bti057>
44. Emanuelsson O, Nielsen H, Brunak S, Von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* 300:1005–1016. <https://doi.org/10.1006/jmbi.2000.3903>
45. Feist AM, Herrg\rd MJ, Thiele I, et al (2009) Reconstruction of Biochemical Networks in Microbial Organisms. *Nat Rev Microbiol* 7:129–143. <https://doi.org/10.1038/nrmicro1949.Reconstruction>
46. Thiele I, Vlassis N, Fleming RMT (2014) FASTGAPFILL: Efficient gap filling in metabolic networks. *Bioinformatics* 30:2529–2531. <https://doi.org/10.1093/bioinformatics/btu321>
47. Thiele I, Vlassis N, Fleming RMT (2014) FASTGAPFILL: Efficient gap filling in metabolic networks. *Bioinformatics* 30:2529–2531. <https://doi.org/10.1093/bioinformatics/btu321>
48. Models RHLM, Ferreira C, Rocha I, et al (2018) Chapter 1 with merlin
49. Benthin S, Nielsen J, Villadsen J (1991) A simple and reliable method for the determination of cellular RNA content. *Biotechnol Tech* 5:39–42. <https://doi.org/10.1007/BF00152753>
50. Feist AM, Henry CS, Reed JL, et al (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 3:1–18. <https://doi.org/10.1038/msb4100155>
51. Song B, Büyüktaktakin IE, Ranka S, Kahveci T (2011) Manipulating the steady state of metabolic pathways. *IEEE/ACM Trans Comput Biol Bioinforma* 8:732–747. <https://doi.org/10.1109/TCBB.2010.41>
52. Palsson B (2000) The challenges of in silico biology. *Nat Biotechnol* 18:1147–1150. <https://doi.org/10.1038/81125>
53. Ma ZM, Wang CS, Chen H De, Tian JC (2005) Study on digital geologic map and its relational data model. *J Chengdu Univ Technol (Science Technol Ed)* 32:200–207. <https://doi.org/10.1038/nbt.1614>
54. MacGillivray M, Ko A, Gruber E, et al (2017) Robust analysis of fluxes in genome-scale metabolic

- pathways. *Sci Rep* 7:1–20. <https://doi.org/10.1038/s41598-017-00170-31>
55. Chaouiya C, Béranguier D, Keating SM, et al (2013) SBML qualitative models: A model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools. *BMC Syst Biol* 7:. <https://doi.org/10.1186/1752-0509-7-135>
 56. Holmes B, Costas M, Ganner M, et al (1994) Evaluation of biolog system for identification of some gram-negative bacteria of clinical importance. *J Clin Microbiol* 32:1970–1975
 57. Shlomi T, Berkman O, Ruppin E (2005) Regulatory on $\overline{\text{off}}$ minimization of metabolic flux. *Pnas* 102:7695–7700. <https://doi.org/10.1073/pnas.0406346102>
 58. Reding D (2011) Predictions of. 125–130. <https://doi.org/10.1038/84379>
 59. Shen CR, Liao JC (2013) Synergy as design principle for metabolic engineering of 1-propanol production in *Escherichia coli*. *Metab Eng* 17:12–22. <https://doi.org/10.1016/j.ymben.2013.01.008>
 60. Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol Syst Biol* 3:. <https://doi.org/10.1038/msb4100162>
 61. Santos F, Boele J, Teusink B (2011) A practical guide to genome-scale metabolic models and their analysis, 1st ed. Elsevier Inc.
 62. Dmem H, Chen J, Covert MW, et al (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429:2–6. <https://doi.org/10.1038/nature02514>.Published
 63. Papin JA, Hunter T, Palsson BO, Subramaniam S (2005) Reconstruction of cellular signalling networks and analysis of their properties. *Nat Rev Mol Cell Biol* 6:99–111. <https://doi.org/10.1038/nrm1570>
 64. Thiele I, Jamshidi N, Fleming RMT, Palsson BO (2009) Genome-scale reconstruction of *Escherichia coli*'s transcriptional and translational machinery: A knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput Biol* 5:. <https://doi.org/10.1371/journal.pcbi.1000312>
 65. Oberhardt MA, Palsson B, Papin JA (2009) Applications of genome-scale metabolic reconstructions. *Mol Syst Biol* 5:1–15. <https://doi.org/10.1038/msb.2009.77>
 66. Becker SA, Palsson B (2005) Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: An initial draft to the two-dimensional annotation. *BMC Microbiol* 5:1–12. <https://doi.org/10.1186/1471-2180-5-8>
 67. Thiele I, Price ND, Vo TD, Palsson B (2005) Candidate metabolic network states in human mitochondria. Impact of diabetes, ischemia, and diet. *J Biol Chem* 280:11683–11695. <https://doi.org/10.1074/jbc.M409072200>
 68. Bordbar A, Lewis NE, Schellenberger J, et al (2010) Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. *Mol Syst Biol* 6:. <https://doi.org/10.1038/msb.2010.68>
 69. Schellenberger J, Park JO, Conrad TM, Palsson BØ (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* 11:213. <https://doi.org/10.1186/1471-2105-11-213>
 70. Reed JL, Famili I, Thiele I, Palsson BO (2006) Towards multidimensional genome annotation. *Nat Rev Genet* 7:130–141. <https://doi.org/10.1038/nrg1769>

71. Raghunathan A, Reed J, Shin S, et al (2009) Constraint-based analysis of metabolic capacity of *Salmonella typhimurium* during host-pathogen interaction. *BMC Syst Biol* 3:1–16. <https://doi.org/10.1186/1752-0509-3-38>
72. Gonzalez O, Gronau S, Falb M, et al (2008) Reconstruction, modeling & analysis of *Halobacterium salinarum* R-1 metabolism. *Mol BioSyst* 4:148–159. <https://doi.org/10.1039/B715203E>
73. Satish Kumar V, Dasika MS, Maranas CD (2007) Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* 8:1–16. <https://doi.org/10.1186/1471-2105-8-212>
74. Burgard AP, Pharkya P, Maranas CD (2003) OptKnock: A Bilevel Programming Framework for Identifying Gene Knockout Strategies for Microbial Strain Optimization. *Biotechnol Bioeng* 84:647–657. <https://doi.org/10.1002/bit.10803>
75. Agren R, Liu L, Shoaie S, et al (2013) The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for *Penicillium chrysogenum*. *PLoS Comput Biol* 9:. <https://doi.org/10.1371/journal.pcbi.1002980>
76. Liao YC, Tsai MH, Chen FC, Hsiung CA (2012) GEMSiRV: A software platform for GEnome-scale metabolic model simulation, reconstruction and visualization. *Bioinformatics* 28:1752–1758. <https://doi.org/10.1093/bioinformatics/bts267>
77. Swainston N, Smallbone K, Mendes P, et al (2011) The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks. *J Integr Bioinform* 8:186. <https://doi.org/10.2390/biecoll-jib-2011-186>
78. Le Novère N, Finney A, Hucka M, et al (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol* 23:1509–1515. <https://doi.org/10.1038/nbt1156>
79. Eddy SR (2011) Accelerated profile HMM searches. *PLoS Comput Biol* 7:. <https://doi.org/10.1371/journal.pcbi.1002195>
80. Goldberg T, Hecht M, Hamp T, et al (2014) LocTree3 prediction of localization. *Nucleic Acids Res* 42:350–355. <https://doi.org/10.1093/nar/gku396>
81. Yu NY, Wagner JR, Laird MR, et al (2010) PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26:1608–1615. <https://doi.org/10.1093/bioinformatics/btq249>
82. Dias O, Gomes D, Vilaca P, et al (2016) Genome-wide Semi-automated Annotation of Transporter Systems. *IEEE/ACM Trans Comput Biol Bioinform* XX:1–14. <https://doi.org/10.1109/TCBB.2016.2527647>
83. Saier, Jr. MH (2000) A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol Mol Biol Rev* 64:354–411. <https://doi.org/10.1128/MMBR.64.2.354-411.2000>
84. Krogh A, Larsson B, Von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol Biol* 305:567–580. <https://doi.org/10.1006/jmbi.2000.4315>
85. Gavrilescu M, Chisti Y (2005) Biotechnology - A sustainable alternative for chemical industry. *Biotechnol Adv* 23:471–499. <https://doi.org/10.1016/j.biotechadv.2005.03.004>
86. Pinto JP, Pereira R, Cardoso J, et al (2013) TNA4OptFlux - A software tool for the analysis of strain optimization strategies. *BMC Res Notes* 6:. <https://doi.org/10.1186/1756-0500-6-175>

87. von Kamp A, Schuster S, Shlomi T, et al (2005) Metatool 5.0: Fast and flexible elementary modes analysis. *Bioinformatics* 22:7695–7700. <https://doi.org/10.1073/pnas.0406346102>
88. Segre D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci* 99:15112–15117. <https://doi.org/10.1073/pnas.232349399>
89. Arkin AP, Cottingham RW, Henry CS, et al (2018) KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat Biotechnol* 36:566–569. <https://doi.org/10.1038/nbt.4163>
90. O'Leary NA, Wright MW, Brister JR, et al (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 44:D733–D745. <https://doi.org/10.1093/nar/gkv1189>
91. Goodstein DM, Shu S, Howson R, et al (2012) Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res* 40:1178–1186. <https://doi.org/10.1093/nar/gkr944>
92. Goecks J, Nekrutenko A, Taylor J, The Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computation research in the life sciences. *Genome Biol* 11:R86
93. Oinn T, Addis M, Ferris J, et al (2004) Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 20:3045–3054. <https://doi.org/10.1093/bioinformatics/bth361>
94. Reich M, Liefeld T, Gould J, et al (2006) GenePattern 2.0 [2]. *Nat Genet* 38:500–501. <https://doi.org/10.1038/ng0506-500>
95. Schink B (1997) Energetics of syntrophic cooperation in methanogenic degradation. *Microbiol Mol Biol Rev* 61:262–280. [https://doi.org/10.1092-2172/97/\\$04.0010](https://doi.org/10.1092-2172/97/$04.0010)
96. Mcinerney MJ, Sieber JR, Gunsalus RP (2010) NIH Public Access. *Curr Opin Biotechnol* 20:623–632. <https://doi.org/10.1016/j.copbio.2009.10.001>. Syntrophy
97. Stams AJM, Plugge CM (2009) Electron transfer in syntrophic communities of anaerobic bacteria and archaea. *Nat Rev Microbiol* 7:568–577. <https://doi.org/10.1038/nrmicro2166>
98. Schink B (2002) Synergistic interactions in the microbial world. *Antonie van Leeuwenhoek, Int J Gen Mol Microbiol* 81:257–261. <https://doi.org/10.1023/A:1020579004534>
99. Manuscript A, Invertebrates B (2010) NIH Public Access. 34:41–58. <https://doi.org/10.1111/j.1574-6976.2009.00193.x>. Common
100. Stewart PS (2002) Mechanisms of antibiotic resistance in bacterial biofilms. *Int J Med Microbiol* 292:107–113. <https://doi.org/10.1078/1438-4221-00196>
101. Harmsen HJM, Kuijk BLM Van, Plugge CM, et al (1998) Reducing Bacterium. *Int J Syst Bacteriol* 1383–1388
102. Boone DR, Bryant MP (1980) Propionate-Degrading Bacterium, *Syntrophobacter wolinii* sp. nov. gen. nov., from Methanogenic Ecosystems. 40:626–632
103. Liu Y, Balkwill DL, Henry CA, et al (1999) Characterization of the anaerobic propionate-degrading syntrophs *Smithella propionica*. *Int J Syst Bacteriol* 49:545–556. <https://doi.org/10.1099/00207713-49-2-545>
104. Manuscript A, Blood W, Count C (2009) NIH Public Access. 49:1841–1850. <https://doi.org/10.1016/j.jacc.2007.01.076>. White

105. Plugge CM, Dijkema C, Stams a. JM (1993) Acetyl-CoA cleavage pathway in a syntrophic propionate oxidizing bacterium growing on fumarate in the absence of methanogens. *FEMS Microbiol Lett* 110:71–76. [https://doi.org/10.1016/0378-1097\(93\)90244-V](https://doi.org/10.1016/0378-1097(93)90244-V)
106. Ewing B, Ewing B, Hillier L, et al (2005) Base-Calling of Automated Sequencer Traces Using. *Genome Res* 175–185. <https://doi.org/10.1101/gr.8.3.175>
107. Badger JH, Olsen GJ (1999) CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol* 16:512–524. <https://doi.org/10.1093/oxfordjournals.molbev.a026133>
108. Delcher A (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 27:4636–4641. <https://doi.org/10.1093/nar/27.23.4636>
109. Plugge CM, Henstra AM, Worm P, et al (2012) Complete genome sequence of *Syntrophobacter fumaroxidans* strain (MPOBT). *Stand Genomic Sci* 7:91–106. <https://doi.org/10.4056/sigs.2996379>
110. Hungate RE (1950) The Anaerobic Mesophilic Cellulolytic Bacteria. *Bacteriol Rev* 14:1–49
111. Ferry JG, Smith PH, Wolfe RS (1974) *Methanospirillum*, a new genus of methanogenic Bacteria , and characterization of *Methanospirillum hungatii* sp . nov . *Int J Syst Bacteriol* 24:465–469. <https://doi.org/10.1099/00207713-24-4-465>
112. Ferry JG, Wolfe RS (1977) Nutritional and biochemical characterization of *Methanospirillum hungatii*. *Appl Environ Microbiol* 34:371–376
113. Hyatt D, Chen GL, LoCascio PF, et al (2010) Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:. <https://doi.org/10.1186/1471-2105-11-119>
114. Pati A, Ivanova NN, Mikhailova N, et al (2010) GenePRIMP: A gene prediction improvement pipeline for prokaryotic genomes. *Nat Methods* 7:455–457. <https://doi.org/10.1038/nmeth.1457>
115. Markowitz VM, Mavromatis K, Ivanova NN, et al (2009) IMG ER: A system for microbial genome annotation expert review and curation. *Bioinformatics* 25:2271–2278. <https://doi.org/10.1093/bioinformatics/btp393>
116. Gunsalus RP, Cook LE, Crable B, et al (2016) Complete genome sequence of *Methanospirillum hungatei* type strain JF1. *Stand Genomic Sci* 11:2. <https://doi.org/10.1186/s40793-015-0124-8>
117. Worm P, Stams AJM, Cheng X, Plugge CM (2011) Growth- and substrate-dependent transcription of formate dehydrogenase and hydrogenase coding genes in *Syntrophobacter fumaroxidans* and *Methanospirillum hungatei*. *Microbiology* 157:280–289. <https://doi.org/10.1099/mic.0.043927-0>
118. Sieber JR, McInerney MJ, Gunsalus RP (2012) Genomic Insights into Syntrophy: The Paradigm for Anaerobic Metabolic Cooperation. *Annu Rev Microbiol* 66:429–452. <https://doi.org/10.1146/annurev-micro-090110-102844>
119. Yagi T, Higuchi Y (2013) Studies on hydrogenase. *Proc Jpn Acad, Ser B* 89:16–33. <https://doi.org/10.2183/pjab.89.16>
120. Stams AJM, Van Dijk JB, Dijkema C, Plugge CM (1993) Growth of syntrophic propionate-oxidizing bacteria with fumarate in the absence of methanogenic bacteria. *Appl Environ Microbiol* 59:1114–1119. <https://doi.org/10.1007/BF01912397>
121. Santos S, Rocha I (2016) Estimation of biomass composition from genomic and transcriptomic information. 13:. <https://doi.org/10.2390/biecoll-jib-2016-285>

122. Mahadevan R, Bond DR, Butler JE, Coppi M V (2006) Characterization of Metabolism in the Fe (III) - Reducing Organism *Geobacter sulfurreducens* by Constraint-Based Modeling †. 72:1558–1568. <https://doi.org/10.1128/AEM.72.2.1558>
123. Sun J, Sayyar B, Butler JE, et al (2009) BMC Systems Biology. 15:1–15. <https://doi.org/10.1186/1752-0509-3-15>
124. Benedict MN, Gonnerman MC, Metcalf WW, Price ND (2012) Genome-scale metabolic reconstruction and hypothesis testing in the methanogenic archaeon *Methanosarcina acetivorans* C2A. J Bacteriol 194:855–865. <https://doi.org/10.1128/JB.06040-11>
125. DiMarco AA, Bobik TA, Wolfe RS (1990) Unusual coenzymes of methanogenesis. Annu Rev Biochem 59:355–94. <https://doi.org/10.1146/annurev.bi.59.070190.002035>
126. Kezmarsky ND, Xu H, Graham DE, White RH (2005) Identification and characterization of a L-tyrosine decarboxylase in *Methanocaldococcus jannaschii*. Biochim Biophys Acta 1722:175–82. <https://doi.org/10.1016/j.bbagen.2004.12.003>
127. White RH (1988) Biosynthesis of the 2-(aminomethyl)-4-(hydroxymethyl)furan subunit of methanofuran. Biochemistry 27:4415–20
128. Cole SH, Carney GE, McClung CA, et al (2005) Two functional but noncomplementing *Drosophila* tyrosine decarboxylase genes: distinct roles for neural tyramine and octopamine in female fertility. J Biol Chem 280:14948–55. <https://doi.org/10.1074/jbc.M414197200>
129. Maden BE (2000) Tetrahydrofolate and tetrahydromethanopterin compared: functionally distinct carriers in C1 metabolism. Biochem J 350 Pt 3:609–29
130. Mashhadi Z, Xu H, White RH (2009) An Fe²⁺-dependent cyclic phosphodiesterase catalyzes the hydrolysis of 7,8-dihydro-D-neopterin 2',3'-cyclic phosphate in methanopterin biosynthesis. Biochemistry 48:9384–92. <https://doi.org/10.1021/bi9010336>
131. Raemakers-Franken PC, Voncken FG, Korteland J, et al (1989) Structural characterization of tatiapterin, a novel pterin isolated from *Methanogenium tationis*. Biofactors 2:117–22
132. Rasche ME, White RH (1998) Mechanism for the enzymatic formation of 4-(beta-D-ribofuranosyl)aminobenzene 5'-phosphate during the biosynthesis of methanopterin. Biochemistry 37:11343–51. <https://doi.org/10.1021/bi973086q>
133. White RH (1996) Biosynthesis of methanopterin. Biochemistry 35:3447–56. <https://doi.org/10.1021/bi952308m>
134. White RH (1998) Methanopterin biosynthesis: methylation of the biosynthetic intermediates. Biochim Biophys Acta 1380:257–67
135. Graham DE, Xu H, White RH (2002) Identification of coenzyme M biosynthetic phosphosulfolactate synthase: a new family of sulfonate-biosynthesizing enzymes. J Biol Chem 277:13421–9. <https://doi.org/10.1074/jbc.M201011200>
136. Graupner M, Xu H, White RH (2000) Identification of the gene encoding sulfopyruvate decarboxylase, an enzyme involved in biosynthesis of coenzyme M. J Bacteriol 182:4862–7
137. Denger K, Mayer J, Buhmann M, et al (2009) Bifurcated degradative pathway of 3-sulfolactate in *Roseovarius nubinhibens* ISM via sulfoacetaldehyde acetyltransferase and (S)-cysteate sulfolyase. J

- Bacteriol 191:5648–56. <https://doi.org/10.1128/JB.00569-09>
138. Deppenmeier U (2002) The unique biochemistry of methanogenesis. *Prog Nucleic Acid Res Mol Biol* 71:223–83
 139. Drevland RM, Jia Y, Palmer DRJ, Graham DE (2008) Methanogen homoacetylase catalyzes both hydrolyase reactions in coenzyme B biosynthesis. *J Biol Chem* 283:28888–96. <https://doi.org/10.1074/jbc.M802159200>
 140. Howell DM, Harich K, Xu H, White RH (1998) Alpha-keto acid chain elongation reactions involved in the biosynthesis of coenzyme B (7-mercaptoheptanoyl threonine phosphate) in methanogenic Archaea. *Biochemistry* 37:10108–17. <https://doi.org/10.1021/bi980662p>
 141. Noll KM, Donnelly MI, Wolfe RS (1987) Synthesis of 7-mercaptoheptanoylthreonine phosphate and its activity in the methylcoenzyme M methylreductase system. *J Biol Chem* 262:513–5
 142. White RH (1994) Biosynthesis of (7-mercaptoheptanoyl)threonine phosphate. *Biochemistry* 33:7077–81
 143. Howell DM, Graupner M, Xu H, White RH (2000) Identification of enzymes homologous to isocitrate dehydrogenase that are involved in coenzyme B and leucine biosynthesis in methanoarchaea. *J Bacteriol* 182:5013–6
 144. Ashton WT, Brown RD, Jacobson F, Walsh C (1979) Synthesis of 7,8-didemethyl-8-hydroxy-5-deazariboflavin and confirmation of its identity with the deazaalloxazine chromophore of *Methanobacterium* redox coenzyme F420. *J Am Chem Soc* 101:4419–4420. <https://doi.org/10.1021/ja00509a083>
 145. Ebert S, Rieger PG, Knackmuss HJ (1999) Function of coenzyme F420 in aerobic catabolism of 2,4, 6-trinitrophenol and 2,4-dinitrophenol by *Nocardioides simplex* FJ2-1A. *J Bacteriol* 181:2669–74
 146. Graupner M, White RH (2003) *Methanococcus jannaschii* coenzyme F420 analogs contain a terminal alpha-linked glutamate. *J Bacteriol* 185:4662–5
 147. Núñez VTS ENERGY CONSERVATION MECHANISMS AND ELECTRON TRANSFER IN SYNTROPHIC PROPIONATE-OXIDIZING MICROBIAL CONSORTIA
 148. Hamilton JJ, Calixto Contreras M, Reed JL (2015) Thermodynamics and H₂ Transfer in a Methanogenic, Syntrophic Community. *PLoS Comput Biol* 11:1–20. <https://doi.org/10.1371/journal.pcbi.1004364>
 149. Machado D, Andrejev S, Tramontano M, Patil KR (2018) Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res* 46:7542–7553. <https://doi.org/10.1093/nar/gky537>