University of Minho
School of Engineering
Department of Informatics

Renato Alexandre Azevedo Cruz

Nitrobacter vulgaris: genome-scale model reconstruction and interactions with *Nitrosomonas europaea*

October 2018

University of Minho
School of Engineering
Department of Informatics

Renato Alexandre Azevedo Cruz

Nitrobacter vulgaris*: genome-scale model reconstruction and interactions with *Nitrosomonas europaea

Master dissertation

Master Degree in Bioinformatics – Technologies of Information

Dissertation supervised by

Oscar Manuel Lima Dias

Jorge Manuel Padrão Ribeiro

October 2018

DECLARAÇÃO

Nome: Renato Alexandre Azevedo Cruz

E-mail: renato_cruz95@hotmail.com

Contacto: +351 914 033 685

Cartão de Cidadão: 14762527

Título de dissertação: *Nitrobacter vulgaris*: genome-scale model reconstruction and interactions with *Nitrosomonas europaea*

Orientadores:

Prof. Doutor Oscar Manuel Lima Dias

Doutor Jorge Manuel Padrão Ribeiro

Ano de conclusão: 2018

Mestrado em Bioinformática

É AUTORIZADA A REPRODUÇÃO PARCIAL DESTA DISSERTAÇÃO (indicar, caso tal seja necessário, nº máximo de páginas, ilustrações, gráficos, etc.), APENAS PARA EFEITOS DE INVESTIGAÇÃO, MEDIANTE DECLARAÇÃO ESCRITA DO INTERESSADO, QUE A TAL SE COMPROMETE.

Universidade do Minho, 31 / 10 / 2018

Assinatura: Renato Alexandre Azevedo Cruz

Acknowledgments

I would like to thank Oscar Dias and Jorge Padrão for all the support and motivation given throughout the development of this project. This project is as much fruit of my work as of their support, time and patience. I would also like to express my thanks for the opportunity to work alongside them and for integrating me in their fields of work to which I have grown a great interest.

I would also like to thank Ana Nicolau for allowing me to work in the LMAMB laboratory.

A special thank Sophia Santos for all the time she spent guiding me in the laboratory and for sharing her experience in the reconstruction of *GSM* models. Without her help the work would not be as complete and accurate.

I thank Pedro Raposo for letting me participate in the reconstruction of the *GSM* model of *N. europaea* and for the guidance and experience he offered in the development of this work.

I thank to all the people in CEB that always showed availability and support in my work, particularly to the technicians Maura Guimarães, Diana Vilas Boas and Nicole Dias and to Sónia Barbosa. Also, I thank to the LMAMB researchers, for their support, patience and good mood.

I thank to all the people in the BioSystems group for all the help, support and good mood in the office. A particular thank for the *merlin* staff team, Amaro Morais and Davide Lagoa for the help with *merlin*.

Finally, the biggest thanks goes to my family, specially my mother, step-father and brothers, for all the support in the best and worst times, without them this project could never be finished.

This study was supported by the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UID/BIO/04469 unit and COMPETE 2020 (POCI-01-0145-FEDER-006684) and BioTecNorte operation (NORTE-01-0145-FEDER-000004) funded by the European Regional Development Fund under the scope of Norte2020 - Programa Operacional Regional do Norte.

Abstract

The fast growing pace of human industrialization and agriculture has led to an increasing contamination of nitrogen reactive species both in soil and water. This contamination is a recalcitrant problem, damaging the biotic soil communities and causing eutrophication of aquatic systems. The main focus of this work is not only to solve the nitrogen contamination problem, but to take advantage of it. In this work we demonstrate how Metabolic Engineering and two bacteria, *Nitrosomonas europaea* and *Nitrobacter vulgaris*, could offer a biotechnological solution to the nitrogen problem.

N. europaea is a well-studied species of ammonia-oxidizing bacteria, capable of consuming ammonia and producing nitrite. On the other hand, *N. vulgaris*, was subjected to few studies, and is a nitrite-oxidizing bacteria, capable of oxidize nitrite into nitrate. In this work, a *Genome-Scale Metabolic* model of *N. vulgaris* was reconstructed using a specialized software, *merlin,* and it was combined with a previously reconstructed model of *N. europaea,* resulting in a *community* model. This allows *in silico* simulation of these bacteria, providing crucial information about both species, their interactions and uses.

A semi-automatic annotation of the *N. vulgaris* genome was performed, as well as curation of its metabolic pathways and reactions. A steady-state culture of *N. vulgaris* was established *in vivo*, suppling data that was used to validate and shape the model.

Data obtained *in vivo* revealed that the *N. vulgaris* model accurately represents the organism. The *N. vulgaris* model is fully functional and helped the understanding of the bacterium reaction in different conditions.

The community model representing the *N. europaea* – *N. vulgaris* system shows that both bacteria can cooperate in the nitrogen species oxidation process.

With these models at our disposal, an optimized approach that removes ammonia and nitrite from wastewater or recirculating aquaculture systems (RAS) and nitrate can be collected. Therefore, in addition to providing decontamination from several nitrogen reactive species (ammonia, ammonium and nitrite) producing of nitrate, a valuable compound fertilizers, that is still currently collected in its impure forms from mines.

Keywords: Metabolic Engineering, *Nitrosomonas europaea*, *Nitrobacter vulgaris*, *Genome Scale Metabolic* model, community model.

Resumo

O rápido crescimento da industrialização e agricultura levou a um aumento na contaminação de espécies azotadas ativas em ambos solos e águas. Esta contaminação é um problema recalcitrante e em curso, tornando os solos inférteis e sistemas aquáticos eutrofizados. O objetivo principal deste trabalho é não só resolver o problema de contaminação de azoto, mas também tirar proveito dele. Neste trabalho demonstramos como Engenharia Metabólica e duas bactérias, *Nitrosomonas europaea* e *Nitrobacter vulgaris*, podem oferecer uma solução ao problema do azoto reativo antropogénico. *N. europaea* é uma espécie bem estudada de bactéria oxidante de amoníaco capaz de consumir amoníaco e produzir nitrito. *N. vulgaris*, por outro lado, foi sujeita a poucos estudos e é uma bactéria oxidante de nitrito, capaz de oxidar nitrito a nitrato. Neste trabalho, um modelo *Genómico à Escala Metabólica* de *N. vulgaris* foi reconstruído usando o *merlin* (um software especializado para este processo) e foi combinado com um modelo já existente de *N. europaea,* resultando num modelo de comunidade. Isto permitirá simulações *in silico* destas bactérias, fornecendo informação crucial sobre ambas as espécies, as suas interações e usos.

Foi realizada uma anotação semi-automática ao genoma de *N. vulgaris*, assim como a curação das suas vias e reações metabólicas. Foi estabelecida uma cultura contínua *in vivo*, que permitiu recolher dados que foram usados para validar e moldar o modelo.

Dados obtidos *in vivo* revelaram que o modelo de *N. vulgaris* representa o organismo com precisão. O modelo da *N. vulgaris* é funcional e permitiu na compreensão da resposta da bactéria quando sujeita a diferentes condições

O modelo de comunidade que representa o sistema *N. europaea* – *N. vulgaris* demonstra que ambas as espécies podem cooperar na oxidação de espécies azotadas.

Com estes modelos à nossa disposição, uma abordagem para a remoção de amoníaco e nitrito de águas residuais ou sistemas de recirculação de aquaculturas e nitrato pode ser alcançado. Isto, em conjunto com a descontaminação de espécies reativas de nitrogénio (amoníaco, amónio e nitrito) produzindo nitrato, um composto importante composto para fertilizantes, que é atualmente obtido em minas nas suas formas impuras.

Palavras-chave: Engenharia Metabólica, *Nitrosomonas europaea*, *Nitrobacter vulgaris,* modelo *Genómico à Escala Metabólica*, modelo comunidade.

Index

List of Figures ***x***

List of Tables ***xi***

List of Equations ***xii***

Acronyms ***xiii***

1. Introduction ***1***

1.1 Context and Motivation **1**

1.2 Goals **3**

1.3 Structure of the Document **4**

2. State of Art ***6***

2.1 Environmental and Energetic Problem **6**

2.2 Causes **8**

2.3 Solutions **9**

2.4 *Nitrosomonas europaea* **10**

2.5 *Nitrobacter vulgaris* **11**

2.6 Community of *N. europaea* and *N. vulgaris* **11**

2.7 Metabolic Engineering **13**

2.8 Systems Biology **14**

2.9 Biologic Databases **16**

2.10 Simulation Methods **17**

2.11 Available Software for *GSM Models* Reconstruction and/or Simulation **18**

3. Methods and Materials ***25***

3.1 Wet-lab Materials and Methods **25**
3.1.1 Organisms 25
3.1.2 Medium Preparation 26
3.1.3 Organisms Compatibility 27
3.1.4 Chemostat Setup 27
3.1.5 Ethanol Quantification 29
3.1.6 Nitrite and Nitrate Quantification 29
3.1.7 Ammonia Quantification 29
3.1.8 Gases Quantifications 30
3.1.9 Scanning Electron Microscopy Visualization and Elemental Analysis 30
3.1.10 Optical Density and Dry Weight 31
3.1.11 Biomass Precursors 31
3.1.12 Biomass macromolecules Quantification 33

3.2 Reconstruction of *N. vulgaris GSM* Model **35**
3.2.1 *N. vulgaris* Phylogenetic Analysis 35
3.2.2 Semi-automatic *merlin* Annotation 36
3.2.3 Thresholds Calculation using *SamPler* 37
3.2.4 Manual Annotation 38

3.2.5 Integration of the Annotation in the Model 42
3.2.6 Transport Proteins Annotation 42
3.2.7 Compartments Prediction and Annotation 42

3.3 *GSM Model* Curation 43
3.3.1 Gap Filling and Dead-end Removal Process 43
3.3.2 Reactions Reversibility 46
3.3.3 Balancing Stoichiometry 46
3.3.4 Biomass Equation 47
3.3.5 Gene-protein-reaction Rules 47

3.4 Community Model Reconstruction 48
3.4.1 Merging Tool Description 48

3.5 Simulation Methods 48
3.5.1 *N. vulgaris* Model Simulations 49
3.5.2 Simulations on the Community Model 49

4. Results and Discussion 50

4.1 Wet-lab Results 50
4.1.1 Chemostat Results 50
4.1.2 Macroscopy and Scanning Electron Microscopy Visualization 50
4.1.3 Optical Density and Dry weight 52
4.1.4 Atomic quantification 53
4.1.5 Gases Quantifications 54
4.1.6 Ethanol Quantification 55

4.2 Computational Results 55
4.2.1 *N. vulgaris* Taxonomy 56
4.2.2 Determining the Thresholds 57
4.2.3 Genome Annotation 57
4.2.4 Compartments Annotation 59
4.2.5 Transporters Annotation 59
4.2.6 Gap Filling and Dead-end Removal 59
4.2.7 Reactions Balancing 59
4.2.8 Biomass Composition Quantification 60
4.2.9 *GSM* Model of *N. vulgaris* 60
4.2.10 Simulation and *in vivo* Values 61
4.2.11 Simulation Results of *N. vulgaris* Model 62
4.2.12 Validation of *N. vulgaris* Model 63

4.3 *N. vulgaris* Metabolism and Physicochemical Results 64
4.3.1 Nitrogen Metabolism 65
4.3.2 Carbon Fixation 66
4.3.3 ATP Production 67
4.3.4 Biomass Production 67
4.3.5 pH Analysis 68

4.4 *N. vulgaris* - *N. europaea* Community Model 69
4.4.1 Simulations of the Community Model 70

4.5 Additional Results 72
4.5.1 *N. vulgaris* Light Sensitivity 72
4.5.2 Compounds of Interest 73
4.5.3 Nitrate Commercial Use 73

5. Conclusion 74

6. Future Work *76*

Annex *77*

References *76*

List of Figures

FIG. 1. SCHEME REPRESENTING THE SYSTEM COMPOSED BY *N. EUROPAEA* AND *N. VULGARIS* AND N REACTIVE SPECIES TRANSFORMATIONS. 12

FIG. 2. SCHEMATIC REPRESENTATION OF THE CHEMOSTAT SETUP FOR *N. VULGARIS*. M- FLASK, WITH FRESH MEDIUM; R- REACTOR, WITH THE CULTURE OF *N. VULGARIS*; S- FLASK S, WITH THE SEWAGE REPOSITORY. THE REPOSITORIES HEIGHT REPRESENTS THE REAL POSITION (NOT AT SCALE). 28

FIG. 3. SCHEME REPRESENTING THE METHODOLOGY USED IN THE RECONSTRUCTION OF *THE N. VULGARIS GSM* MODEL, BASED ON DIAS AND COLLEAGUES, 2014. 36

FIG. 4. EXAMPLE OF THE RESULTS DISPLAYED BY SAMPLER WHERE THE USER SELECTED THE ALPHA VALUE OF 0.8. 38

FIG. 5. MANUAL ANNOTATION PIPELINE USED FOR GENOMIC CURATION. 39

FIG. 6. DEAD-END REMOVAL PIPELINE AND RESULTS. GREEN BOXES REPRESENT DESIRABLE OUTCOMES, RED ONES REPRESENT UNDESIRABLE OUTCOMES. 45

FIG. 7. GAP FILLING PIPELINE AND RESULTS. GREEN BOXES REPRESENT DESIRABLE OUTCOMES, RED ONES REPRESENT AVOIDABLE OUTCOMES THAT SHOULD BE AVOIDED. 46

FIG. 8. IMAGES CAPTURED IN THE LABORATORY. A – CORRESPONDS TO THE REACTOR ON ITS EARLY STAGES. B - SHOWS THE REACTOR IN ITS LATER STAGES. C – DISPLAYS THE SAMPLES COLLECTED OVER TIME, FAR LEFT (DAY 16) TO FAR RIGHT (DAY 118). 51

FIG. 9. SEM IMAGES OF FREEZE DRIED *N. VULGARIS*. MEASUREMENTS DISPLAYED AT GREEN OF SINGULAR BACTERIUM. 52

FIG. 10. RELATION BETWEEN BIOMASS (MG L-1) AND OPTICAL DENSITY (Λ = 600NM). 53

FIG. 11. EDS RESULTS OF ATOMIC PERCENTAGES OF *N. VULGARIS*. 54

FIG. 12. GRAPHICAL REPRESENTATION OF THE ATOMIC PERCENTAGES OF *N. VULGARIS*. 54

FIG. 13. RESULTS FROM MUSCLE. SHOWING THE CLOSEST ORGANISMS TO *N. VULGARIS* AND THEIR RESPECTIVE SIMILARITY SCORE. 56

FIG. 14 SCHEME OF *N. VULGARIS* WITH PRINCIPAL METABOLITES, REACTIONS A PATHWAYS. 75

FIG. 15 NITROGEN REACTIVE SPECIES OVER TIME. 66

FIG. 16 ETHANOL CONCENTRATION OVER TIME. 67

FIG. 17 BIOMASS IN THE REACTOR OVER TIME. 68

FIG. 18 PH LEVEL OVER TIME. 69

List of Tables

TABLE 1 RECONSTRUCTION SOFTWARE COMPARISON. 18
TABLE 2 TRACE AND STOCK SOLUTIONS COMPOSITION FOR 1 L. 26
TABLE 3 CONFIDENCE LEVEL OF HOMOLOGUE GENES FOUND IN SWISS-PROT. 40
TABLE 4 CONFIDENCE LEVEL OF HOMOLOGUE GENES FOUND IN TREMBL. 41
TABLE 5 O_2 AND CO_2 CHARACTERISTICS AND REACTORS DIMENSIONS. 55
TABLE 6 A SCORE AND THRESHOLDS SELECTED AFTER THE SAMPLER SELECTED GENES CURATION PROCESS. 57
TABLE 7 PROTEIN CLASSIFICATION AND RELATIVE FREQUENCY ON *N. VULGARIS*. 58
TABLE 8 BIOMASS CONSTITUTION. 60
TABLE 9 *IN VIVO* AND *IN SILICO* METABOLITES RATES OBTAINED. [1]MAXIMUM RATE POSSIBLE. 62
TABLE 10 COMMUNITY MODEL PROPERTIES. 70
TABLE 11 METABOLITE RATES ON THE COMMUNITY MODEL. 72
TABLE 12 POSSIBLE PRODUCTS OF INTEREST OF *N. VULGARIS* AND RESPECTIVE COMMERCIAL PRICE. 73

TABLE S1 MANUAL ANNOTATION OF THE 50 SAMPLER GENES SELECTED FOR SWISS-PROT, FINAL SCORES ATTRIBUTED BY *MERLIN* AND RESPECTIVE CONFIDENCE LEVEL. 77
TABLE S2 COMPLETE SWISS-PROT GENE ANNOTATION, FINAL SCORES ATTRIBUTED BY *MERLIN* AND RESPECTIVE CONFIDENCE LEVEL. 78
TABLE S3 COMPLETE TREMBL GENE ANNOTATION, FINAL SCORES ATTRIBUTED BY *MERLIN* AND RESPECTIVE CONFIDENCE LEVEL. 82
TABLE S4 MANUALLY INSERTED REACTIONS. 84

List of Equations

Equation 1 $\mathrm{NH_3 + O_2 + 2H^+ \; + 2e^- \; \rightarrow NH_2OH + H_2O \; \rightarrow NO_2^- \; + 5H^+ \; + \; 4e^-}$ 3

Equation 2 $\mathrm{NH_3 + O_2 + 2H^+ \; + 2e^- \; + H_2O \rightarrow NO_2^- \; + 5H^+ \; + \; 4e^- \; + H_2O \rightarrow NO_3^- + \; 7H^+ + 6e^-}$ 3

Equation 3 $CH_4 \; + \; H_2O \; \rightarrow CO + 3H_2$ 8

Equation 4 $\frac{1}{2} N_2 \; + \frac{3}{2} H_2 \leftrightarrow NH_3$ 8

Equation 5 $\sum_{k=1}^{P} \mathrm{c_k X_{k \rightarrow biomass}}$ 23

Equation 6 $[N - NH_3] = \frac{[N - NH_4^+] \; \times \; 10^{pH}}{\exp\left(\frac{6344}{T}\right) + \; 10^{pH}}$ 28

Equation 7 $\dot{\mathrm{V}} = \frac{(P_1 \; - \; P_2) \; \times \; A \; \times \; D}{T}$ 29

Equation 8 $Score = \; \alpha \; \times Score_{frequency} + (1 - \; \alpha) \; \times \; Score_{taxonomy}$ 35

Equation 9 $Area = 3.8102 \times [Ethanol] - 0.4006$ 53

Acronyms

AOB	Ammonia-oxidizing bacteria
ATP	Adenosine Triphosphate
BLAST	Basic Local Alignment Search Tool
BRENDA	BRaunschweig Enzyme Database
CBM	Constraint based modelling toolkit
CEB	Center of Biological Engineering
COBRA	COnstraint-Based Reconstruction and Analysis
CoReCo	Comparative Reconstruction
DAPI	4,6-Diamidino-2-Phenylindole Dihydrochloride
DNA	Deoxyribonucleic acid
EC	Enzyme commission
EDS	Energy Dispersive X-ray Spectroscopy
EDTA	Ethylenediaminetetraacetic Acid
FAME	Flux Analysis and Modelling Environment
FBA	Flux Balance Analysis
FVA	Flux Variability Analysis
GEMSiRV	Genome-scale Metabolic model Simulation, Reconstruction and Visualization
GPR	Gene-Protein-Reaction
GSM	Genome-Scale Metabolic
GUI	Graphical User Interface
HPLC	High-Performance Liquid Chromatography
IFA	International Fertilizer Industry Association
KEGG	Kyoto Encyclopaedia of Genes and Genomes
ME	Metabolic engineering
MEMOSys	Metabolic Model research and development System
merlin	Metabolic Models Reconstruction Using Genome-Scale Information
MFA	Metabolic Flux Analysis
MNA	Metabolic Network Analysis
MOMA	Minimization Of Metabolic Adjustments

MUSCLE	MUltiple Sequence Comparison by Log-Expectation
N	Nitrogen
N_2	Di-molecular nitrogen
NADH	Nicotinamide Adenine Dinucleotide
NCBI	National Center for Biotechnology Information
NOB	Nitrite-oxidizing bacteria
PBS	Phosphate-Buffered Saline
PGDBs	Pathways/genome databases
PHP	Hypertext Preprocessor
PySCeS	Python Simulator for Cellular Systems
RAS	Recirculating Aquaculture System
RAVEN	Recombination, Analysis and Visualization of Metabolic Networks
rDNA	Recombinant DNA
RNA	Ribonucleic acid
ROOM	Regulatory On/Off Minimization
SB	Systems Biology
SEM	Scanning Electron Microscope
TC	Transporter Classification
TCA	Citric Acid Cycle
TCDB	Transporter Classification DataBase
TrEMBL	Translated EMBL Nucleotide Sequence Data Library
TRIAGE	Transport Reactions Annotation and Generation
UniProt	Universal Protein Resource Knowledgebase
USEPA	United States Environmental Protection Agency
WHO	World Health Organization

1. Introduction

1.1 Context and Motivation

Demographic growth has been an important issue to society since the beginning of civilization, especially in its extremes, displaying concern when either too low or considerably high. When demographic growth is too scarce, the community development is weak and may represent its disappearance. On the other hand, an overwhelmingly high demographic growth creates problems, as overpopulated areas usually have lower quality of life. Even though some countries currently experience a low population growth (mostly European countries), in a global perspective this growth has increased rapidly since the industrial revolution to this day [1].

From the moment it was developed and established, agriculture has been the main source of food to the human population. Since the beginning of the 20th century, humans recognized the importance of nitrogen (N) for mass production of vegetables. Hence, nitrogen-rich compounds have been used, such as guano deposits or chemically generated fertilizers, to improve crops [2].

The exponential increase of the world population lead to the intensive use of chemical fertilizers to fulfil the global food requirements, creating a high demand of energy to obtain N [3].

Agriculture has the highest N demand, nevertheless a vast plethora of industries require large quantities of this element. Ammonia (NH_3) is used in the manufacture of fibres, plastics, explosives, paper, rubber, as a coolant in the metal industry, as a cleansing agent, as food additive or a drug ingredient [4].

An excessive use of fertilizers or waste disposal from industry activities into the environment will lead to an increased concentration of N based molecules in the soil and water, severely damaging their quality and aggravating the sustainability problem related to the need and usage of N in agriculture [5].

To effectively address this issue, the establishment of a circular economy strategy is required. Circular economy consists in a regenerative cycle of resources, in which the input, wastes and energy leaks are reduced to the maximum by decreasing energy and by developing material loops [6].

Modern methods to synthetize NH_3 employ the Haber-Bosch synthesis, this process is largely used in industry today and consists in the use atmospheric di-molecular nitrogen (N_2) of to produce NH_3. This procedure requires large amounts of energy, since the reaction can only take place when temperature reaches 450 °C and 20 260 kPa [7].

On the other hand, NH_3 is available in large quantities in wastewater, especially in aquaculture ponds, where fish excretion and mineralization of organic matter can lead to toxic levels of NH_3 in the system [8]. Making aquaculture a good candidate for establishing the above mentioned circular economy strategy. However, the overall process of removing NH_3 from wastewater and reutilizing it by the existing methods has low efficiency and low profitability [8].

Biotechnology, the integration of natural science and organisms for products and services, can provide feasible solutions to these problems. The usage of tools provided by biotechnological advances has increased the energy supply, farming yield and exponentially benefitted the pharmaceutical industry in the past. In addition, biotechnology continues to establish novel approaches to various problems, usually requiring mild operating conditions and exhibiting enhanced yields [9].

This work will focus on a new approach to the recycling of N reactive species, using biotechnology, having in mind a circular economy strategy. We propose the use of microorganisms, which require mild conditions (considerably less energy intensive), to recycle undesirable products into products with economic interest. To optimize “wet lab” resources, bioinformatics methodologies will be implemented and validated [10].

Such strategy requires a deep understanding of the metabolism of the involved microorganisms, what they consume, produce, at which rates and in which conditions. This knowledge will allow manipulating variables to maximize the removal or production of desirable compounds. The reconstruction of genome-scale metabolic (*GSM*) models allows the execution of *in silico* simulations to determine reaction fluxes, growth rates and other factors in different environmental and genetic conditions. This network system provides information about which biochemical reactions are active in the microorganism, in such conditions [11].

Bacteria from the *Nitrosomonas* genus, such as *Nitrosomonas europaea* are ammonia-oxidizing bacteria (AOB), these bacteria can oxidize NH_3 into hydroxylamine (NH_2OH), which rapidly decomposes into nitrite (NO_2^-), as seen in Equation 1 [12]. Bacteria like *Nitrobacter vulgaris* , *Nitrobacter winogradskyi* and *Nitrospira spp.* are aerobic chemoautotrophic bacteria known for oxidizing NO_2^- into nitrate (NO_3^-) [12], [13]. Therefore, these three main nitrogen-based ions can be bio-converted into only one, NO_3^-, through a microbial community (presented on Equation 2) [12]. NO_3^- is a very useful substance, as it can be used in the production of highly soluble fertilizers, explosives and other products of interest [14].

$$\mathrm{NH_3 + O_2 + 2H^+ + 2e^- \rightarrow NH_2OH + H_2O \rightarrow NO_2^- + 5H^+ + 4e^-} \qquad \text{(Eq. 1)}$$

$$\mathrm{NH_3 + O_2 + 2H^+ + 2e^- + H_2O \rightarrow NO_2^- + 5H^+ + 4e^- + H_2O \rightarrow NO_3^- + 7H^+ + 6e^-} \qquad \text{(Eq. 2)}$$

1.2 Goals

The objective of this project is to reconstruct the *N. vulgaris GSM* model using *merlin*. This process requires performing the genome annotation and the reactions identification, verifying the stoichiometry of reactions, determining the reactions compartmentalization, assemble the biomass abstraction and add constrains to the model. Moreover, experimental data will be collected from *N. vulgaris* cultures that will be analysed *in vivo*. Finally, the accuracy of the model will be evaluated by comparing the *in silico* data with the *in vivo* data [15]–[17].

After validation, the *N. vulgaris* model will be integrated with an already existing model for *N. europaea*. This will result in a complex new type of model, able to predict the metabolic behaviour of the bacteria community.

Finally, a validation of the model comprising the community is required to evaluate if it correctly represents the community. This process will consist in a series of robustness tests to the model, to assess whether it can predict *in silico* the results obtained *in vivo*.

The construction of a community model will provide a better understanding of the interactions between these two species, which will not only provide a cost-effective solution to nitrogen-contaminated environments, but will also allow generating profit.

This project will contain a series of stages:

- Research concerning the current problem with the excessive use of fertilizers to the environment and humans. Study of the bacteria *N. europaea*, *N. vulgaris* and their interactions.
- Reviewing *merlin* documentation.
- Analysis of the previously generated *N. europaea GSM* model.
- Reconstruction of the *N. vulgaris GSM* model and the community *GSM* model. Which involves performing the genome annotation, identification of metabolic reactions and components, validate the stoichiometry of the reactions, identify the reactions locations in the organism (or system), construction of the biomass abstraction, and apply possible restrains to both models.
- Validation of the models, comparing the *in vivo* results with the results obtained *in silico*.
- Utilize the new community model to discover the optimal conditions for the consumption of NH_3 and production of NO_3^-.

1.3 Structure of the Document

This document will be divided in the following sections:

Chapter 2 - State of Art

Nitrogen contamination impact on the environment, causes and consequences. Current approaches to the nitrogen pollution issue, its disadvantages and costs. Nitrification process and its constituents. Metabolic Engineering and Systems Biology introduction, uses and solutions offered. *GSM* model reconstruction processes, validation, databases and tools available.

Chapter 3 - Methods and Materials

Reconstruction of *N. vulgaris GSM* model and genome annotation. *N. vulgaris - N. europaea* community model construction and analysis. *In vivo* materials and methodology, compounds quantification, biomass composition and models validation.

Chapter 4 - Results and Discussion

In silico results from simulations of both models and comparison with *in vivo* data. Predicted results. Discussion on the validity of the results.

Chapter 5 - Conclusion

Summary of the work and goals achieved.

Chapter 6 – Future Work

A brief enumeration of improvements, to this work, that may be performed in a recent future. Brief introduction of possible following projects.

2. State of Art

2.1 Environmental and Energetic Problem

Nitrogen is a relatively common element on Earth being, in addition to oxygen, carbon, phosphorus and sulphur, essential to all life forms currently known [18]. Its di-molecular form (N_2) is the major component of Earth atmosphere, constituting approximately 78.08 % of its total, thus making it very easy to obtain [19]. Even though N_2 can have its uses, its current value lies in its reactive species, being the most commons NH_3, ammonium (NH_4^+), nitric acid (HNO_3), nitrous oxide (N_2O), NO_2^-, NO_3^-, and organic compounds as urea, proteins and nucleic acids [19].

The high availability of these N reactive species in water and soil is responsible for the increase acidification of the environment. This is problematic since even the slightest pH change can disturb an entire ecosystem or even be directly prejudicial to humans that consume agricultural goods derivatives from these areas [20]. Also, the presence of these molecules in water bodies may lead to eutrophication of water ecosystems [21], [22].

Despite its low concentration in water reserves destined for human consumption, N reactive species have an associated toxicity level. The World Health Organization (WHO) estimates that doses of NH_4^+ of 100 mg kg^{-1} of body weight per day may cause lung oedema, nervous system dysfunction and kidney damage due to the inability of the body to detoxify such intake [4], [23].

Besides direct effect on humans, N may also causes environmental problems. Eutrophication is a phenomenon that can happen to water systems, when these are loaded with excessive amounts of nutrients (particularly N and phosphorous), causing a rapid increase in the biomass of algae communities. This phenomena, commonly named as “Algae Boom”, will deprive the entire aquatic system from light, including algae that initially caused the problem. The deprivation from light will force algae to consume O_2 and consequently produce CO_2, leading to an eventual depletion of the O_2 available in the system, and simultaneously will lead to a pH drop. In this conditions only a limited number of resistant organisms can survive, such as microalgae and anaerobic bacteria. Eutrophication will leave a considerable amount of organic matter from dead organisms that anaerobic bacteria metabolize into toxic gases, such as methane and hydrogen sulphide. This chain of events inevitably leads to an abrupt reduction of biodiversity and consequentially a negative impact into the aquatic and peripheral ecosystems [20], [21], [24]. Eutrophication is expensive to revert and usually requires the use of algaecides that can be prejudicial to other local species. The best solution to the problem would be to massively reduce the input of nutrients in waters, specially phosphorus and N, however, this may be unfeasible given the current widely spread anthropogenic agricultural and industrial activities [24], [25]. Therefore, the development of a novel and efficient strategy to remove such components from the water is of paramount importance.

The rapid decline of marine life represents a problem to the growing seafood demand. RAS are currently being used to sustain this gap in seafood supply [26]. RAS relies in the conversion of NH_4^+ and NH_3, that at concentrations higher than 0.02 mg L^{-1} are toxic to finfish, into NO_3^- ,toxic at 100 mg L^{-1} [27], [28]. The N removal process in this industry relies heavily in biofilters with electrical and maintenance demands [26]. If the aquaculture industry continues with its growth, a more efficient way to remove this contaminant from RAS could be beneficial.

The production of fertilizers is a process that requires considerable amounts of energy. According to the International Fertilizer Industry Association (IFA), approximately 1.2 % of the global annual energy consumption is destined to the production of fertilizers [29]. NH_3 is usually produced by three different methods: Steam reforming of natural gas, partial oxidation of heavy fuel oil or coal gasification [30]. Steam reforming, which is the most used technique, consists in the use of hydrogen, a gas produced using methane (CH_4) present in natural gas reserves, to produce NH_3. However, the production of hydrogen (H_2) is a rather costly process, as it requires pressurizing water and methane between 300 kPa to 2 530 kPa and temperatures of 700 °C to 1000 °C, to obtain H_2. Finally, H_2 is combined with N, easily retrieved from the atmosphere, resulting in the synthesis of NH_3. This process is called the Haber-Bosh process. Even though the reaction is exothermic, it requires temperatures of 450 °C and pressure of 20 260 kPa to be time efficient [2], [7], [31].

The following formulas describe the principal stages to produce NH_3 through the Steaming reforming gas method (Equation 3), and using the Haber ammonia synthesis (Equation 4) [7], [31]:

$$CH_4 + H_2O \rightarrow CO + 3H_2 \qquad \Delta H^\circ = 165\ \text{kJ mol}^{-1} \qquad \text{(Eq. 3)}$$

$$\frac{1}{2} N_2 + \frac{3}{2} H_2 \leftrightarrow NH_3 \qquad \Delta H^\circ = -45.8\ \text{kJ mol}^{-1} \qquad \text{(Eq. 4)}$$

2.2 Causes

The N cycle represents the totality of the transferences of N and its reactive species between Earth sub-systems. The major portion of N is in its diatomic state in the atmosphere and can be fixed by ammonia-reducing bacteria [32]. Once fixed, plants can absorb N through symbiotic relationships with bacteria, or ammonia-oxidizing and nitrite-oxidizing bacteria can convert it into NO_2^- and later NO_3^-, respectively [2], [32]. Once in NO_3^-, plants can directly absorb N into its system or denitrifying bacteria can return it to the atmosphere in the form of N_2. Transfers between Earth sub-systems were approximately constants throughout time, until anthropogenic activities largely increased N availability during, food production processes and combustion of fossil fuel [2].

Industrialization wastes, resulting from the intense application of fertilizers and pesticides in recent agriculture, are the most common cause of nitrification of water and soil, representing a large portion of this problem. These practices comprise the main sources of inorganic ions like NH_4^+, NO_2^- and NO_3^-, that otherwise would be present only in small amounts due to atmospheric deposition or organic decomposition [33], [34].

Agriculture is an activity that depends on N availability, since N is a macronutrient essential to plants. N can be absorbed primarily in two of its reactive forms: NH_4^+ or NO_3^-

.

Currently, NH_3 is the main component of most fertilizes representing up to 82 % of its composition, which will convert into NH_4^+ when dissolved in water, allowing plants to absorb N into their system [35].

The growth of population requires increasing food production to sustain such growth, forcing the use of fertilizers in crops. Consequently, as fertilizers use increases, the contamination of soil with nitrogen-based compounds, damaging the soil microbial communities and ultimately render it infertile [36].

2.3 Solutions

Current solutions to the N concentration in water problem consists in the removal of the majority of pollutants in residual waters in wastewater treatment plants. This involves using chemical, mechanical and biological methods [37]. This process requires substantial amounts of energy. The United States Environmental Protection Agency (USEPA) estimates that approximately 4 % of the energy consumption in the United States of America is used in the transferring and treating of contaminated water [38]. In addition to its high-energy demand, it is an ineffective procedure to decontaminate water, as it does not reuse the waste [38].

The possibility of stagnating completely the supply of N or its reactive species to the soil and water is unfeasible. Moreover, the current concentration of eutrophic inducing elements in soil and water require immediate answers.

The above mentioned possible solution of using algaecides to decrease the number of algae, which can cause damage to other species, will be temporary as the population of algae might recover or gain resistance.

Implementation of fast reproducing and omnivore fish species is a proposed solution, although such species might not be able to adapt to such harsh habitat or it may perturb the already fragile ecosystem due to their reproductive rate and vast feeding options [39].

Current RAS NH_4^+ and NH_3 removal processes rely in the use of mechanical biofilters that require electricity to operate and must be replaced or maintained to assure its functionality [28]. Recently, combinations between biofilters and AOB are being studied [26]. The use of a system that completely circumvents the use of mechanical means to remove NH_4^+ and NH_3 and relies solemnly in bacteria could improve the efficiency of RAS.

The solution presented in this thesis involves using two microorganisms to, not only decrease concentration of NH_3 and NO_2^- in water, but also profit from it. Using *N. europaea* to remove NH_3, together with *N. vulgaris* to remove NO_2^-, from water is a plausible solution when comparing to those mentioned above. These bacteria do not require permanent supervision nor high-energy, making them a good “tool” to use in wastewater treatment plants, contaminated water bodies and in RAS [40]–[42].

2.4 *Nitrosomonas europaea*

N. europaea is a Gram-negative bacteria, regarding its trophic nature, *N. europaea* can be described as obligatory chemolithoautotroph, as it is only capable of generating energy by oxidizing NH_3 into NO_2^- [43]. *Nitrosomonas eutrophus* and *Nitrosococcus oceanus*, are among the most studied AOB, nevertheless *N. europaea* is the most studied AOB, hence its importance for this work [13].

This bacterium tolerates pH levels ranging between 6.0 and 9.0, has an aerobic metabolism, favours temperatures between 20 and 30 °C and its only carbon source for biomass growth is provided through CO_2 fixation [43], [44]. Due to its chemolithoautotrophic characteristics, *N. europaea* colonies have a slow growth rate, which is described as its theoretical maximum as 0.05 h^{-1}, due to the significant amount of energy required to fixate CO_2 [40], [41], [45].

2.5 *Nitrobacter vulgaris*

Unlike *N. europaea* (with approximately 452 articles where it was the main focus of the study), *N. vulgaris* was not the subject to a high number of studies, as the main studies about this organism were written during the 20th century, and only 6 have *N. vulgaris* as its main focus in total (data obtained from Google Scholar). This gram-negative bacterium is mobile due to its single sub-polar flagellum, and can be found in soil, water or rocks (thus the species nomenclature *"vulgaris"*, which means common). *N. vulgaris* can grow heterotrophically, lithoautotrophically or both (mixotrophically). Therefore, it is classified as a facultative lithoautotrophic [46]. Though being aerobic, this bacterium can survive in anaerobic conditions under specific conditions [47], [48].

N. vulgaris is a preferential nitrite-oxidizing bacterium (NOB) by using NO_2^- to produce energy (when O_2 is present), excreting NO_3^- in the process, even though it can also oxidize organic matter. Studies have found that *N. vulgaris* can produce energy and fix carbon utilizing acetate, pyruvate and formate in anaerobic conditions [48]–[50].

Optimum lithoautotrophic conditions comprehend a temperature between 23 and 28 °C and pH levels between 7.5 and 8.0 [46].

Depending on the environmental conditions, doubling time may reach 140 h hours in lithoautotrophic environments and between 25 h and 27 h hours in mixotrophic and heterotrophic conditions respectively [46], [48]. Its reproducing methods are budding or binary fission.

This species can be found on mineral rich mediums containing NO_2^- or in light deprived locations, with organic carbon and N without the need of NO_2^-. Previous studies have found that *N. vulgaris* aggregate in small colonies with yellowish-white colours although some other species might have orange colour [48].

2.6 Community of *N. europaea* and *N. vulgaris*

Studies regarding *N. europaea* and *N. vulgaris* community growth were performed before by Grunditz and Dalhammar and within the Bioresources, Bioremediation and Biorefinery (BRIDGE) research group at the University of Minho [51].

Unspecified strains of *Nitrosomonas spp.* and *Nitrobacter spp.* were grown on agar plates and then transferred to liquid media in fed-batch conditions, by Grunditz and Dalhammar. Growth was measured according to the substrate consumption rate. Physical conditions were: temperature of 30 °C, pH level of 8.0 and agitation of 200 rpm. These and other researchers found that temperatures of 35 °C were ideal for *Nitrosomonas spp.* and 38 °C for *Nitrobacter spp*. Ideal pH levels for both strains were relatively similar, although *Nitrosomonas spp.* had shown higher activity at pH of 8.1 and *Nitrobacter spp.* at 7.9 [51]–[54].

As mentioned before, *N. europaea* and *N. vulgaris* have different energy production methods. *N. europaea* is an AOB, whereas *N. vulgaris* is a NOB. Theoretically, the system would only need the regular feeding of NH_3, a C source, O_2 and the essential minerals, to sustain *N. europaea*, which would produce NO_2^- that *N. vulgaris* would oxidize. Considering the chain reaction, it would be expected that the final N reactive agent produced by the culture would be NO_3^-, represented in Figure 1.

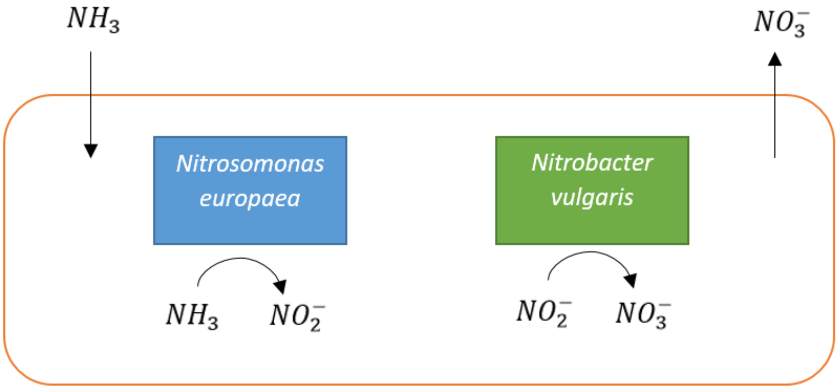


Fig. 1. Scheme representing the system composed by *N. europaea* and *N. vulgaris* and N reactive species transformations.

N compounds can be removed from water throughout a variety of physicochemical and biological ways. Due to its favourable efficiency and associated low costs, biological methods have been adopted over physicochemical ones, even at the cost of a slower pace [55], [56]. However, Metabolic Engineering can prove to be a valuable option when trying to surpass these obstacles.

2.7 Metabolic Engineering

Metabolic engineering (ME) is the process of modulating the metabolic functions of an organism to produce a desired metabolite by performing genetic, environmental or other manipulations [9]. This manipulations may involve the insertion, deletion and/or modification of metabolic pathways [57].

The ability to manipulate recombinant DNA (rDNA) outside living cells, developed in the 70s, allowed scientists to realize that such processes enabled the bioproduction of compounds that usually derived from chemical reactions, turning ME in to one of the most promising fields. ME showed, for the first time, its real potential when *Escherichia coli* was used to produce "human-like" insulin, a process that at the time was almost unable to fulfil the demand due to its complex and arduous collection method. This first medicine to be produced through ME methods consisted in the insertion of human genes in *E. coli*, promoting the synthesise of this hormone in a safe and effective way, ultimately leading to the acceptance of this technology [58].

Utilization of ME to produce desirable products has multiple advantages when compared with traditional methods. Many chemicals are still too hard or expensive to obtain throughout other methods, as their production/harvesting processes require extreme conditions. All these advantages favourably contribute to our modern days challenges of energy and environmental sustainability [58].

The development and understanding of this technique allowed the transition from insertion or deletion of a single pathway to the total manipulation of the metabolic system of an organism, rendering organisms as "factories" for industrial production of commodities of interest [58].

ME requires the understanding of multiple areas of knowledge:

1. Biosynthetic pathway. When the objective of the ME process is to overexpress the pathway that leads to the desirable product, creating surplus that can later be collected.
2. Genes that encode related enzymes. Another way to over-produce a certain compound is to constrain other pathways that might consume the original

substrate that leads to the desirable product. These inhibitions must be thoroughly controlled, because they can be essential for the organism, and thus may compromise its viability.

3. Regulation of enzymes. This comprehends the expression of a set of genes that encode the most efficient enzymes for the production of a certain metabolite that otherwise would have low production values.
4. Transfer and expression or suppression of the gene on the host. This process consists on the use of mutations to alter the genome of an organism to produce a substance that otherwise could not be produced naturally. This may be achieved by inducing random mutations and selection of desirable ones or through computational modelling of a specific reaction (altering active sites, enzyme structures or available substrates).
5. Mutate genes *in vivo* and *in vivo.* This process involves the use of mutations to alter enzymes characteristics [9].

ME is still in its early phases since many organisms have not yet been characterized or do not have their genome sequenced. Meaning that, many metabolic systems are still undiscovered, some of which might be valuable, as is the case of some plants that could produce certain medical substances that due to their complexity could not yet be modified to produce higher quantities of the desirable element [58]. Similarly, even though *N. vulgaris* and *N. europaea* have their genome sequenced, there are no effective methodologies for their ME reported yet.

2.8 Systems Biology

Systems Biology (SB) is a field of study that made the project of sequencing the first genomes an achievable task. This field consists in using biological data, computational capabilities and mathematic functions to understand biological systems at system-level [59]. This understanding requires a set of principles and methodologies that link behaviours of molecules to system functions [59].

SB has led to great advances in medicine and biology since the reconstruction of the first *GSM* models, for familiar organisms, 18 years ago [60]. Methods for reconstructing these models were developed together with algorithms to analyse the models properties [61]. Most of the SB efforts were focused on cell metabolism as the synthesis of specific substances, which otherwise would be hard to obtain, were now possible, utilizing microorganisms that had their genomes sequenced. Insulin and ethanol are some examples of substances that could now be synthesised at higher rates by *E. coli* and *Saccharomyces cerevisiae,* respectively [58], [62].

To reconstruct a model that accurately mimics the metabolic potential of an organism, a great amount of data is required, such as, responses to genetic and environment stressors [63]. Thus, the set of reactions, metabolites and transporters present within the organism must be collected [16].

As Dias and co-workers described in [15] the collection of this data can be divided in four steps:

1. Performing the functional genome annotation, which includes enzyme commissions (EC) numbers, transporter classification (TC) numbers, associated genes and product names are also important.
2. Assembling the metabolic network. This involves the collecting of biochemical reactions to form a network. This involves the collection of genes, proteins and reactions and their associations, collecting spontaneous reactions, stoichiometry revision, reactions compartmentalization and finally perform manual curation [64].
3. Convert the metabolic network into a stoichiometric model, add constrains to the model and biomass equation abstraction [64].
4. Validation of the metabolic model [64].

The reconstruction of a *GSM* model following these steps can be performed manually. However, it is a time consuming procedure that can take over an year to achieve [65]. There are multiple steps to reconstruct a model, including checking databases for reactions and metabolites. All these steps increase the models quality, as all information is curated. However, the time needed to construct a *GSM* model manually calls for a faster reconstruction method. Therefore, the use of an automated software together with manual curation allow decreasing error when reconstructing the model, while guaranteeing models quality.

2.9 Biologic Databases

The reconstruction of a *GSM* model implies the collection of information from various biological fields of study. A thorough and detailed collection of the information of the organism is imperative for the accuracy of its model [15]. Online databases are the main source of information. The following section briefly describes the principal databases that hold relevant information for the reconstruction of *GSM* models.

BioCyc is a collection of pathways/genome databases (PGDBs) that have information relative to genomes a cellular networks. BioCyc allows a computational analysis and exploitation of the database. The information in this database is manually curated [15], [66].

Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database that contains information regarding genes, proteins, pathways and reactions. KEGG provides a set of tools that allow users to browse genome maps, compare genome maps and other functions. The information in this database is however not curated [67].

Universal Protein Resource Knowledgebase (UniProt) is a collection of protein sequences and their annotations. The information on UniProt is manually annotated [68].

MetaCyc is a metabolic pathway database that holds information about organisms enzymes and pathways involved in primary and secondary metabolism. Information in this database is manually curated [66].

BRaunschweig Enzyme Database (BRENDA) is an enzyme/enzyme-ligand database. Its data is curated from literature and text mining processes [69].

The National Center for Biotechnology Information (NCBI) is a set of different databases including PubMed, PubMed Central, and GenBank. NCBI does not provide fully curated information [70].

2.10 Simulation Methods

Once a *GSM* model is reconstructed, a stoichiometry matrix is obtained. This matrix connects metabolites consumptions and productions with the reactions present within the organism [71]. This matrix can be subjected to various simulations that can provide additional information about the organisms metabolism [71].

Information obtained with these simulations is important not only for the validation of the model, but may also provide data regarding the better approaches for increasing growth rates or for the production of certain metabolites.

Flux Balance Analysis (FBA) is a mathematical method for simulating of *GSM models*. It calculates the flow of metabolites though the metabolic network, allowing the prediction of growth or production rates [71].

Flux Variability Analysis (FVA) is a computational tool used to determine the robustness of metabolic models. It is used to find the minimum and maximum flux reactions in the network while maintaining its functionality and still satisfying the constrains imposed [72].

Metabolic Flux Analysis (MFA) is a technique used for the accurate quantification and analysis of metabolic fluxes comparing fluxes distributions throughout the metabolic network [73].

Metabolic Network Analysis (MNA) is a tool for the analysis of features that identify the topology of a metabolic network. The main focus of this tool is to investigate the metabolic network structure [74].

Minimization Of Metabolic Adjustments (MOMA) uses a quadratic programming formulation to calculate a minimum distance or a minimum number of cuts required in a metabolic network when conducting a simulation [75], [76] .

Regulatory On/Off Minimization (ROOM) is an algorithm for predicting the metabolic steady-state when performing gene knockouts, by minimizing the number of significant flux change regarding the wild type [77].

2.11 Available Software for *GSM Models* Reconstruction and/or Simulation

The following section will describe some software tools that can aid the reconstruction of *GSM models* or to perform simulations, as well as a brief description and analysis of their capabilities and limitations. Table 1 summarizes the software currently available for the reconstruction of a *GSM* model and their capabilities [15].

Table 1 Reconstruction software comparison.

Software	*CoReCo*	*MEMOSys*	*FAME*	*Pathway Tools*	*SuBliMinal Toolbox*	*merlin*	*ModelSEED and KBase*	*RAVEN*	*GEMSiRV*	*MicrobesFlux*
Enzyme annotation	X					X	X	X	X	
Transporters annotation				X		X	X			
Compartmentalization		X	X			X	X	X		
Pathway Visualization			X	X		X	X	X	X	X
Biomass Abstraction	X		X			X	X	X		
Highlight Metabolic Dead-ends			X	X		X			X	X
GUI for Manual Curation				X		X			X	
Runs locally	X	X		X		X		X		
GPRs						X	X		X	
Prokariotic models	X	X	X	X	X	X	X	X	X	X
Eukaryotic models	X			X		X		X		
Free	X	X	X	X	X	X	X		X	X

CoReCo

Comparative Reconstruction (CoReCo) is a software released in 2014, able to reconstruct *GSM models* semi-automatically, for prokaryotic and eukaryotic organisms. Its main use consists in the refining of already existing metabolic models that have a low level of curation [78].

MEMOSys

Metabolic Model research and development System (MEMOSys) is a platform designed for management, storage and development of *GSM models*. Released in 2011, this tool was developed in Java™ and uses the JBoss® Seam framework [79].

FAME

Flux Analysis and Modelling Environment (FAME) was released in 2012 and at the time was the only software for *GSM models* reconstruction that allowed creating, editing, running and analysing/visualizing stoichiometric models within a single program. FAME was developed in Hypertext Preprocessor (PHP©) and used the Python Simulator for Cellular Systems (PySCeS), a constraint based modelling toolkit (CBM) for linear solving capabilities [80]–[82].

Pathway Tools

Released in 2012, Pathway Tools is a software environment for creating Pathway/Genome Databases. These Pathway/Genome Databases compile information regarding genes and the respective proteins and metabolic network of organisms through the PathoLogic component [83].

SuBliMinal Toolbox

The SuBliMinal Toolbox, released in 2011, was designed to automate the steps required for reconstructing *GSM* networks and has already reconstructed a *GSM model* for *Saccharomyces cerevisiae* using data from KEGG and MetaCyc [84].

merlin

Metabolic Models Reconstruction Using Genome-Scale Information (*merlin*), first released in 2010, is a user-friendly software built on top of the AIBench software development framework written in Java™. This tool was designed to automate most of the steps necessaries to reconstruct a reliable and complete *GSM model* [15].

merlin has multiple features, such as: enzymes and transporters annotation, loading intracellular compartments predictions, can be run locally, does not require a commercial software (all functionalities are free), allows manual curation of the models and annotations through its GUI. *merlin* also allows pathways analysis/visualisation, infers gene-protein-reaction rules, highlights metabolic dead-ends, allows performing the validation of reactions stoichiometry and can work with prokaryotic and eukaryotic models [15].

Moreover, *merlin* allows inferring some biomass components (i.e. average protein, average DNA and average Ribonucleic acid (RNA) compositions) from the genome sequence [15].

ModelSEED and *KBase*

ModelSEED is a resource for the reconstruction, exploration, analysis and optimization of *GSM models*. It was built upon the SEED framework and released in 2010 as a web-based resource, which automates most steps required to reconstruct a *GSM model* [85].

KBase is an open-source integrated software platform, developed in Python, designed to support large-scale bioinformatics analysis and model building [86], [87].

This tool allows takes advantage of the features offered by ModelSEED to reconstruct *GSM models.* This tool also has simulation capabilities such as FBA [86], [87].

The main limitations of this tool are associated with the inability to perform manual curation intuitively through the GUI [86], [87].

RAVEN

Recombination, Analysis and Visualization of Metabolic Networks (RAVEN) is a MatLab® toolbox able to reconstruct *GSM models* semi-automatically [88].

GEMSiRV

Genome-scale Metabolic model Simulation, Reconstruction and Visualization (GEMSiRV) is a software released in 2012, which allows the reconstruction analysis and visualisation of *GSM models*. It is a software written in Java™ that uses the GNU Linear Programming Kit® for calculations [89].

MicrobesFlux

MicrobesFlux was released in 2012 and is a user-friendly, web-based platform for the reconstruction of *GSM models*. It was developed with Google Web toolkit™ and Python over the Django™ web framework [90], [91].

The following software are designed to perform simulations and analyse results of *GSM* models:

CellNetAnalyzer

Released in 2006, CellNetAnalyzer is a software designed for analysis of cellular networks. This application works as a MATLAB® toolbox, which features several tools for metabolic engineering such as FBA, FVA and gene deletion analysis [92].

This software also provides an user-friendly GUI and a flux visualization system [92].

COBRA

Some of features of the COnstraint-Based Reconstruction and Analysis (COBRA) MATLAB® toolbox, first released in 2011, include: FBA, MFA, regulatory network simulations, MNA and has a built-in visualisation. However, this toolbox does not feature a user-friendly interface [93].

OptFlux

Released in 2009, OptFlux is an open-source and modular software, written in Java™ and built over the AIBench framework [75].

OptFlux features various metabolic engineering tools such as phenotype simulations, FBA, FVA, ROOM of metabolic flux changes, MOMA, MFA, gene-reaction associations and MNA such as minimal cut sets[75].

It was the first software to implement the OptKnock algorithm, which determines the optimal cut sets required to optimize the production of a certain metabolite and OptGene that is as an extension of OptKnock, which uses genetic algorithms to increase prediction capability [75].

OptFlux has a built-in visualization that allows the user to visualize and analyse the results obtained and a graphical user interface [75].

This software has the advantage of being user-friendly and is not associated to any commercial software, rendering all its features free [75], [94].

As shown before, there are several tools available for reconstructing *GSM models*, most of them in continuous development. However, *merlin* will be used in this work, for multiple reasons.

Features such as the enzymes and transporters annotation, a GUI for manual curation and compartments predictions loading are available in various software tools such as RAVEN, Model SEED and CoReCo. However, only *merlin* provides these capabilities simultaneously, making it the most practical and timesaving choice for this project.

The ability to reconstruct prokaryotic models is, of course, essential since *Nitrobacter spp.* and *Nitrosomonas spp.* belong to this biological category. Other features such as, highlighting metabolic dead-ends or the manual curation capabilities are also features that will improve the model quality.

Additionally, *merlin* is able to generate the biomass abstraction. This is an important feature since different organisms have different biomass constitutions, therefore, their growth precursors and rates differ from each other. A correct biomass abstraction is essential for an accurate model. *merlin* utilizes Equation 5 to determine biomass formation [64].

$$\sum_{k=1}^{P} \mathrm{c_k X_k} \rightarrow \text{biomass} \qquad \text{(Eq. 5)}$$

Lastly, the structure in which *merlin* was built on was the AIBench framework, which allows an easier comprehension of the code. Moreover, *merlin* was developed within the BioSystems Group at the University of Minho which facilitates the implementation of new features to the current software, which may be required to simulate a community model.

Other bioinformatics tools, like Model SEED and KBase, might be used as these provide a vast number of features that could be used to consolidate the results obtained with *merlin*.

Currently there is a *GSM* model for *N. europaea*, developed by Raposo and colleagues, which will be used to develop the community model together with the model that will be developed for *N. vulgaris* in this work [17]. *N. europaea* model was developed using *merlin* and a similar set of methods to those mentioned above, obtaining an *in silico* model that describes *N. europaea* metabolism, regarding growth rate, when compared with *in vivo* data. Hence, some of the methods and materials used in the development of the former model will be used throughout this work.

OptFlux will be the software used to perform the *GSM* analysis and simulations. Its user-friendly approach, built-in visualization and graphical interface make this tool easy to use. OptKnock will be an essential tool to optimize cell growth or NO_2^- and NO_3^- rates. In addition to its complete kit for phenotype simulation, OptFlux is currently, the best simulation software to use in this project. Finally, the author of this thesis has previously developed a tool that connects OptFlux to *merlin* internal database, which expedites the process of model validation.

3. Methods and Materials

This chapter will cover the methodology, materials and tools used both *in vivo* and *in silico*. The reconstruction of the *N. vulgaris GSM* model will be performed with *merlin,* whereas OptFlux will be used to perform the simulations. All data used to validate and measure the accuracy of the model will be generated through wet-lab procedures and retrieved from literature data. The community model will be assessed using OptFlux, which provides a plug-in for this task.

If the data obtained through simulations and data obtained *in vivo* is dissimilar, the *GSM* model will be iteratively curated.

The reconstruction and validation process can better be perceived through Figure 3 based on Dias and colleagues, 2014 [95].

3.1 Wet-lab Materials and Methods

Laboratory work comprehends the establishment of a steady-state culture of *N. vulgaris*, and kinetic parameters analysis. This section also comprehends the quantification of all biomass macromolecules.

3.1.1 Organisms

For the experimental work, *N. vulgaris* strain DSM 10236 was used as its genome was identical to the *N. vulgaris* Ab_1 genome [96].

For the community, the same *N. vulgaris* strain mentioned above was used in combination with *N. europaea* strain NCIMB 11850. This strain was selected since it was the phylogenetically closer strain available to the ATCC 19718, which was used to reconstruct the *N. europaea* model on [17].

3.1.2 Medium Preparation

The medium used to feed the *N. vulgaris* culture was the 756c. autotrophic medium with a slight modification. The medium is constituted by 2 solutions (Trace element and Stock solution), described in Table 2, ethanol and sodium nitrite ($NaNO_2$) dissolved in deionized water.

Table 2 Trace and Stock solutions composition for 1 L.

Trace element solution (for 1 L)	**Stock solution (for 1 L)**
33.80 mg manganese sulphate heptahydrate	0.07 g calcium carbonate
49.40 mg boric acid	5.00 g sodium chloride
43.10 mg zinc sulphate heptahydrate	0.5 g magnesium sulphate
37.10 mg ammonium heptamolybdate	1.50 g potassium dihydrogen phosphate
97.30 mg iron(II) sulphate heptahydrate	
25.00 mg copper(II) sulphate pentahydrate	
1 L Distilled water	1 L Distilled water

Both solution were autoclaved, and mixed at room temperature. The final mixture, consisted in 4.93 g of ethanol, 2 g of $NaNO_2$, 1 ml of Trace Element, 100 ml of Stock Solution for 1 L of aqueous solution. The pH level was adjusted with sodium hydroxide to pH of 8.6. Approximately, three days after, pH levels spontaneously adjusts to 7.5.

The medium used for *N. europaea* growth was the same used in P. Raposo and colleagues, 2018 [17]. It consists in 4 solutions diluted in deionized water. Solution A: 35.68 mM Ammonium Sulphate, 62.99 mM Potassium Dihydrogen Phosphate, 59.54 mM Sodium Dihydrogen Phosphate. Solution B: 1.80 mM Calcium Chloride. Solution C: 37.74 mM Sodium Carbonate. Solution D: 0.22 µM Iron(II) Sulphate, 5.26 µM Copper(II) Sulphate. All solutions were autoclaved individually and mixed at room temperature.

The medium used in the community was a mixture of both mediums described before, since it must sustain both bacteria simultaneously. It is constituted by the Trace Element solution (same proportions), Solution A (1.3 times concentrated), Solution B (10 times diluted) and 107.01 mM ethanol. To ensure the medium sterilization, ethanol filtration was performed with a 0.2 µm filter and all the solutions were autoclaved individually.

3.1.3 Organisms Compatibility

All the constituents present in the community medium are present in the *N. vulgaris* medium, however ethanol is not present in the *N. europaea* medium. This could be problematic since *N. europaea* reaction to this compound is not documented.

In order to discover *N. europaea* response to ethanol, a pure culture of *N. europaea* was grown in a medium with 107.01 mM ethanol. NH_4^+ concentration was measured regularly in order to determine the cultures activity.

If *N. europaea* does not grow in this medium, a community medium with no added ethanol might be considered.

3.1.4 Chemostat Setup

A 420 mL (working volume) reactor was used for the steady-state culture. The reactor connected to the medium repository and to the waste repository.

The reactors cotton lids prevent the contamination while allowing gas transfer. This was essential since bacteria need O_2 to generate energy and can use CO_2 to fixate carbon.

All the chemostat system was sterilized prior to inoculation of the bacteria in an autoclave. The sterilization process conditions were 121 °C for 20 minutes.

The feeding was thoroughly controlled to obtain a steady-state culture, in a permanent exponential phase. Feeding was slow at first, due to the bacterial lag phase, and slowly increased in order to maintain the exponential growth phase. This control was achieved through a pump connected to a timer. The timer was schedule diary, assuring a steady feeding rate.

To maintain the reactor homogeneity, magnetic stirrer was added. Rotation was constant at approximately 120 rpm. Also, *N. vulgaris* and *N. europaea* are described to halt all metabolic functions when exposed to light, therefore, the reactor was maintained in the dark [97].

The reactors were temporary opened, under dim light, from which the samples were collected in aseptic conditions. Samples were used to measure pH levels, ethanol, NH_3, NH_4^+, NO_2^-, NO_3^- and biomass. This procedure was performed in the *N. vulgaris* culture and community.

The chemostat system is displayed in Figure 2. The system comprises three repositories: Reactor R, flask M and flask S. The medium is located in flask M and is transferred to reactor R, where the culture is established. When reactor R reaches its full capacity, the surplus is transferred to the sewer system, on flask S. Medium was transferred using a peristaltic pump, however, surplus removal was gravitationally induced.

These conditions were applied to the *N. vulgaris* and the community systems.

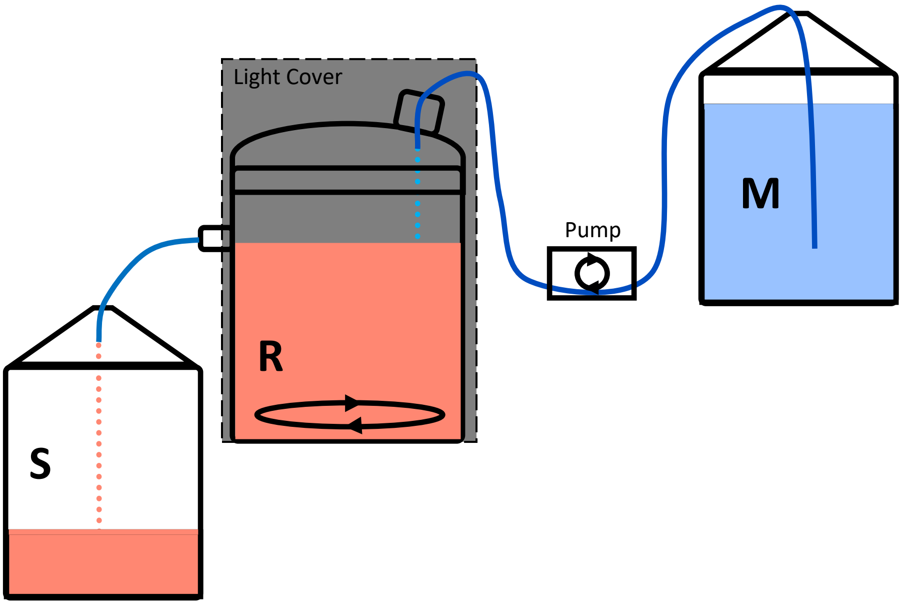


Fig. 2. Schematic representation of the chemostat setup for *N. vulgaris*. M- Flask, with fresh medium; R- Reactor, with the culture of *N. vulgaris*; S- Flask S, with the sewage repository. The repositories height represents the real position (not at scale).

3.1.5 Ethanol Quantification

Ethanol quantification is used to determine *N. vulgaris* ethanol consumption. To measure ethanol concentration in the medium a High-Performance Liquid Chromatography (HPLC) was performed regularly. A calibration curve was generated from ethanol solutions with different concentrations. The curve represents the proportion of ethanol in relation to the absorption values. This curve was then used to determine the ethanol concentration of the samples.

The HPLC system used to perform all chromatographic runs consisted of a Jasco PU2085 pump combined with a refractive index detector Jasco RI4030 and was equipped with an autosampler Jasco AS4050. The column used was an Aminex® HPX-87H 300 x 7.8 mm with 8 µm of particle size. The pump rate was 0.6 mL min^{-1} of 0.005 M solution of H_2SO_4, previously filtered and degassed, at 60 °C for 30 minutes, and injection volume of 20 µL. The software used was the Chrompass version 1.8.6.1, developed by Jasco.

3.1.6 Nitrite and Nitrate Quantification

NO_2^- and NO_3^- were measured using the LCK 342 HACH and LCK 339 HACH test cuvettes, respectively.

3.1.7 Ammonia Quantification

Ammonia concentration was calculated using the Nessler procedure [98]. First, 50 µl of Nessler reagent was added to 1 ml of sample. The mixture was then vortexed for 10 seconds and let to rest for 15 minutes. Then, the mixture was placed in 96-well plates, and the optical density was registered at wavelength of 425 nm. Optical density level was used to calculate the NH_4^+.

A calibration curve that correlates the absorbance level and NH_4^+ concentration was determined.

Finally, this value can be used to determine NH_3 concentration thorough Equation 6 [99].

$$[N - NH_3] = \frac{[N - NH_4^+] \times 10^{pH}}{\exp\left(\frac{6344}{T}\right) + 10^{pH}} \qquad \text{(Eq. 6)}$$

Where **[N- NH_3]** represents the ammonia N concentration (mg L^{-1}), **[N- NH_4^+]** represents the ammonium N concentration (mg L^{-1}), **pH** represents the pH and **T** represents the temperature (K).

3.1.8 Gases Quantifications

The two gases that will be used in the model are CO_2 and O_2. Equation 7, resembling Fick's Law, was used to discover the amount of CO_2 and O_2 available in the reactor [100].

$$\dot{V} = \frac{(P_1 - P_2) \times A \times D}{T} \qquad \text{(Eq. 7)}$$

Where **$\dot{V}$** (mol m^{-2} s^{-1}) represents the rate at which a gas enters the medium. **P_1** (mmHg) and **P_2** (mmHg) are the partial pressures of the gases in the atmosphere and water, respectively, at 25 °C. **A** (m^2) is the area of contact between the medium and the atmosphere, is this case, the medium surface area. **D** (m^2 s^{-1}) is the diffusion constant of the gas. **T** (m) is the thickness of the layer. In this work, the height of the medium was considered to be the layer, due to its homogeneity given its constant stirring.

3.1.9 Scanning Electron Microscopy Visualization and Elemental Analysis

Scanning Electron Microscope (SEM) is a technique used to scan a surface using a focused beam of electrons [101]. SEM was used to obtain an image of *N. vulgaris*. This technique will also allow measurement of individualized bacterium.

SEM was coupled with energy-dispersive X-ray spectroscopy (EDS) analysis (Phenom ProX with EDS detector (Phenom-World BV, Netherlands)). The acquired results were obtained with the ProSuite software integrated with Phenom Element Identification software, allowing the quantification of the concentration of the elements present in the samples, expressed in atomic concentration.

The sample was added to aluminium pin stubs with electrically conductive carbon adhesive tape (PELCO Tabs™). The aluminium pin stub was then placed on a Phenom Charge Reduction Sample Holder (CRH) at 5 Kv and a spot size of 3.3. Samples were imaged without coating. Different points for each sample were analysed for elemental composition. EDS analysis was conducted at 15 kV.

Elemental analysis was determined using EDS. This technique comprises the use of X-rays to excite the molecules of a sample. The unique emission spectrum of atoms allows their respective quantification [102].

3.1.10 Optical Density and Dry Weight

Optical density determines the radiance absorption of a material. This value can be used to determine the bacterium density in a medium [103]. The wavelength used was 600 nm. The dry weight was determined using constant volumes of culture that were freeze-dried and then weighted.

A correlation between dry weight and optical density was achieved.

3.1.11 Biomass Precursors

For the reconstruction of this *GSM* model seven constituents were considered as biomass precursors: carbohydrates, lipids, proteins, cofactors, inorganic ions, DNA and RNA. In the model these molecules will be treated as metabolites, with their own name, mass and formula. These metabolites were named with the prefix "**e-**" (e-DNA for example). All these constituents specifications will be described in the following section. Simpler biomass precursors, such as, H_2O and Adenosine Triphosphate (ATP) represent the water and energy required for the production of biomass.

The source of information for this section of the work was *in vivo* data obtained from the chemostat growth, biomass samples analysis and literature. Due to the lack of literature data about *N. vulgaris*, some of the data was retrieved from the *E. coli* iAF1260 *GSM* model [104].

Carbohydrates

The carbohydrates composition and quantities were based on the *E. coli* iAF1260 *GSM* model and experimental data. No information regarding carbohydrates composition on *Nitrobacter* were found from any other source [104].

Lipids

Lipids composition and quantities were retrieved from literature regarding other *Nitrobacter*, namely *Nitrobacter agilis* and *N. winogradskyi* [105], [106].

Proteins

Proteins composition and quantities were calculated experimentally.

Cofactors

Cofactors composition and quantities were based on the *E. coli* iAF1260 *GSM* model and adapted with literature data [104].

Inorganic Ions

Inorganic ions composition and quantities were based on the medium used for *N. vulgaris* growth and the *E. coli* iAF1260 *GSM* model [104].

DNA and RNA

Nucleic acids composition and quantities were determined experimentally. Their deoxyribonucleotide composition was estimated within *merlin* (with the e-biomass equation tool) from the genome sequence of *N. vulgaris*.

Biomass equation

After all the biomass precursors were determined, the biomass equation was constructed. All the precursors and requirements to produce one gram of biomass were added in their respective quantities.

3.1.12 Biomass macromolecules Quantification

Biomass samples for macromolecules quantification were collected from the sewage repository.

Protein quantification

In order to determine the Protein content in biomass a modified Biuret method was used [107]:

Samples of *N. vulgaris* were freeze dried and dissolved in 2 g L^{-1} Phosphate-Buffered Saline (PBS) and mixed with 0.5 mL of 1 M NaOH. The samples were then incubated in 100 °C for 10 min and then cooled into 25 °C with ice. 0.3 mL of 0.1 M cooper sulphate was mixed with 0.9 mL of the sample. The samples were centrifuged for 5 min at 10000 rpm. Finally, the samples absorbance was measured at 540 nm. The samples used to create the calibration curve were made using Bovine Serum Albumin as standard solutions.

Carbohydrates quantification

Carbohydrates quantification was determined using the phenol-sulphuric method described by Herbert and Strange [108].

Samples of *N. vulgaris* were freeze dried and dissolved in PBS in concentration of 0.1 g L^{-1}. Then, the solution was mixed with 200 µL of phenol 5 % (v/v) and 1 mL of sulfuric acid 96 % (v/v). The samples were left to rest for 25 min, then, absorbance was measured at 490 nm. Glycose solutions at different concentrations were used as standards.

DNA quantification

In order to determine the DNA content in biomass the Mey and Vandamme methodology was used [109]:

Samples of *N. vulgaris* were freeze dried and dissolved in 5 g L^{-1} TNE buffer (1 M NaCl, 10 mM Ethylenediaminetetraacetic Acid (EDTA)), 10 mM Tris). 33 µL of the sample solution was mixed with 1 mL of DAPI dye solution (4, 6-Diamidino-2-Phenylindole Dihydrochloride DAPI 0.25 g L^{-1} in TNE buffer) and incubated for 30 min. The samples fluorescence were then measured with wavelengths of 350/460. A calibration curve with a standard constituted with calf thymus DNA allowed the quantification process in the samples.

RNA quantification

For the RNA quantification, the Benthin and Villadsen methodology was used [110]:

Samples of *N. vulgaris* were freeze dried, 10 mg of cells were washed three times in 1 mL 0.7 M HClO4 and resuspended in 1 mL 0.3 M KOH. Both solutions were kept cold with ice. The resuspended biomass was maintained in 37 °C for 1 h. 100 µl of 3 M $HClO_4$ was added to the samples and centrifuged for 2 min in 14000 rpm. After the centrifugation the supernatants was collected and this process was repeated two more times. The supernatants collected were mixed and its absorbance was measured using a Micro-Spectrophotometer Nanodrop. The samples dilution was used to determine the RNA percentage in biomass.

3.2 Reconstruction of *N. vulgaris GSM* Model

The first step taken to reconstruct a *GSM model* is retrieving the genome of the organism, and to perform its genome annotation. For *N. vulgaris* (strain AB_1), the genome used in this work, was retrieved from NCBI [96].

3.2.1 *N. vulgaris* Phylogenetic Analysis

The genome annotation process requires information about the genome of an organism. Organisms that are phylogenetically close to *N. vulgaris* should have similar genomes. The phylogenetic proximity of the *Nitrobacter* genus was already determined based on the nucleotide sequence of 16S rRNA in [111], [112].

MUltiple Sequence Comparison by Log-Expectation (MUSCLE) is a tool designed to align multiple sequences. MUSCLE was used to construct phylogenetic trees from the 16SrRNA of a group of NOBs, confirming the information retrieved from literature.

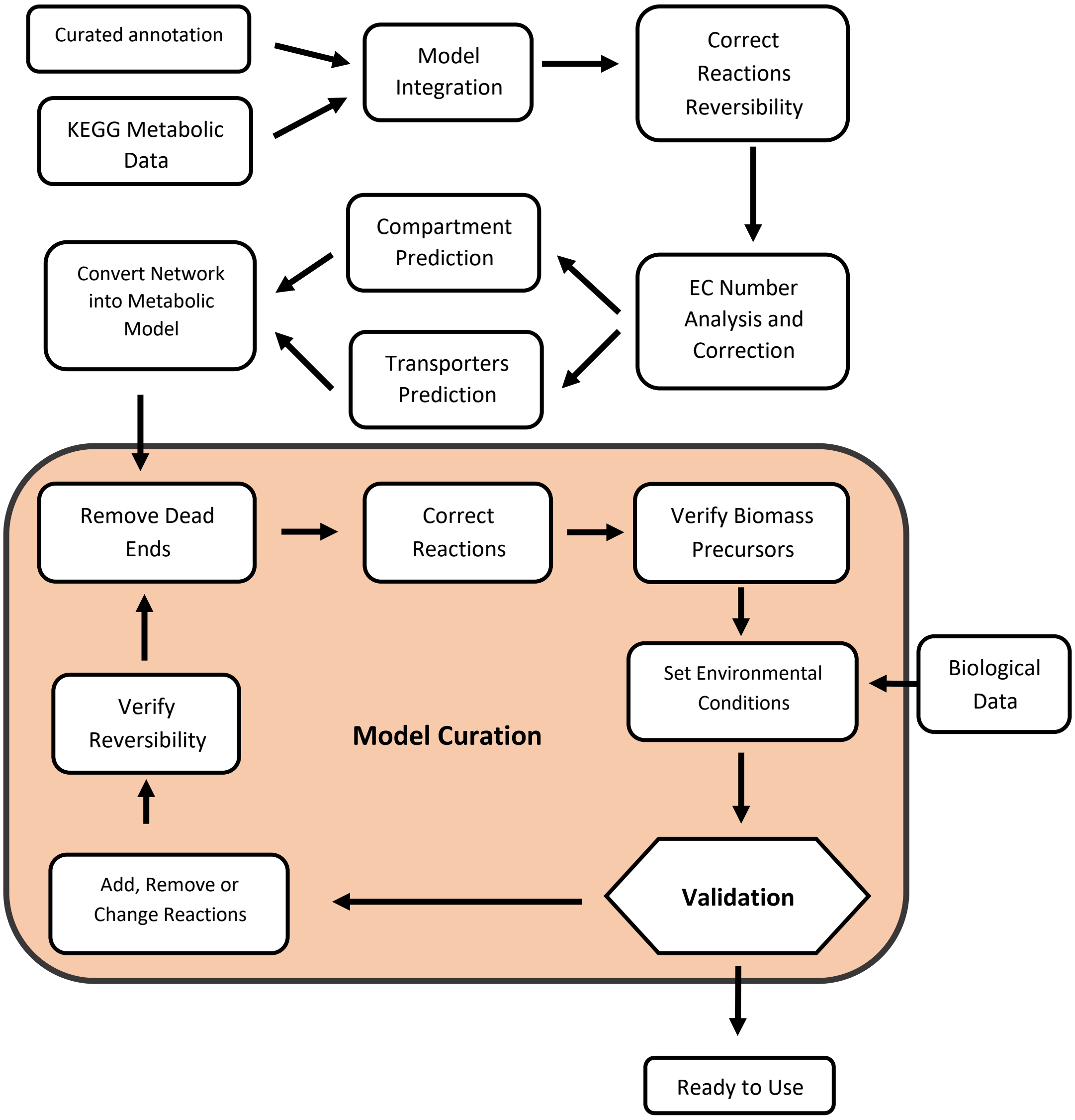


Fig. 3. Scheme representing the methodology used in the reconstruction of *the N. vulgaris GSM* model, based on Dias and colleagues, 2014.

3.2.2 Semi-automatic *merlin* Annotation

merlin is able to retrieve an organism genome annotation directly from KEGG. However, if the genome annotation is not available, it can be performed by a series of processes. Collecting data regarding the genes ORF name, product name and EC numbers is the first step. This is achieved within *merlin,* through a Basic Local Alignment Search Tool (BLAST) which will annotate all homologue genes available. To each gene, a score will be attributed by *merlin*. The score of each gene is obtained through Equation 8 which assigns EC numbers and product numbers to each gene.

$$Score = \alpha \times Score_{frequency} + (1 - \alpha) \times Score_{taxonomy} \qquad \text{(Eq. 8)}$$

Where **Score** is the final score attributed to the annotation, **Score$_{frequency}$** is a value that represents the number of times the annotation is present in the BLAST results for that gene, **Score$_{taxonomy}$** is determined from the taxonomy proximity of the BLAST hits organisms for such annotation, and **α** is a parameter used to leverage both **Score$_{frequency}$** and **Score$_{taxonomy}$** in the for the final result.

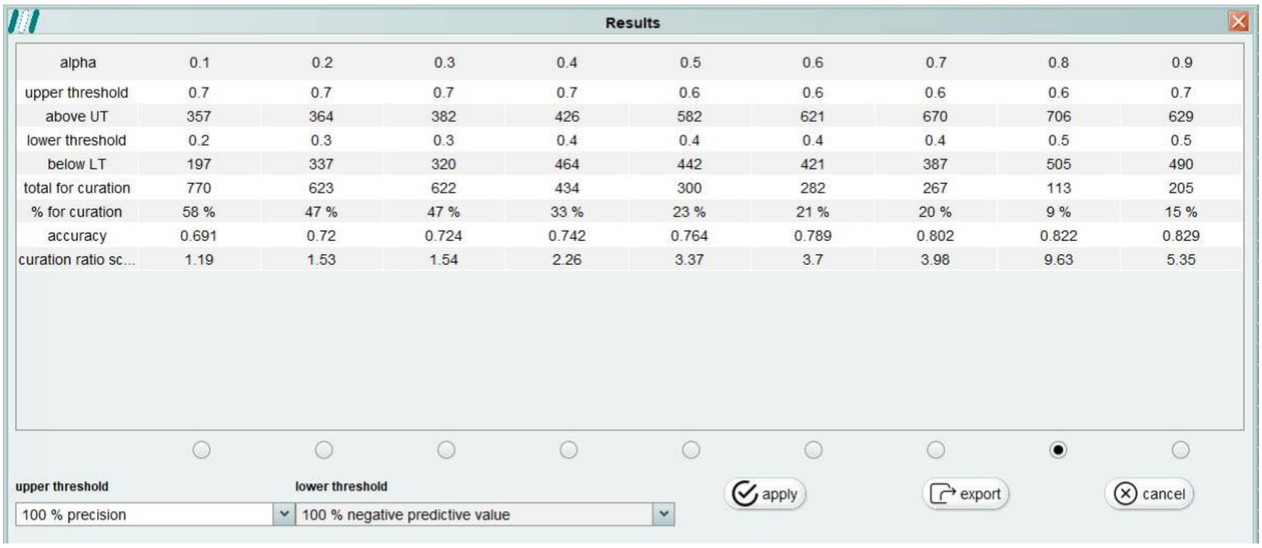
In summary, these assignments are performed by the number of times each EC number is found (frequency) and the taxonomy of the organisms to which such records belong.

3.2.3 Thresholds Calculation using *SamPler*

The score mentioned in Equation 8 is used determine if a certain gene is part of the organism metabolism and what its function. Usually, this threshold was empirically determined and all genes above this score should be ideally curated to increase the reliability of the annotation.

SamPler, is a new plugin developed for merlin that allows semi-automating the annotation process. Genes with scores above an upper threshold will be automatically annotated, while genes below a lower threshold will be automatically rejected as metabolic. Genes that have score values in-between these thresholds should be curated manually [15], [17].

SamPler receives a number of genes that the user choses as a sample size. The sample will contain genes with various score values. This genes must be manually annotated, as *SamPler* uses these annotations to calculate the best combination of parameters (alpha, upper and lower thresholds), as seen on Figure 4. SamPler calculates the precision, negative predictive value and accuracy, and uses these to maximize the confidence of the annotation while minimizing the number of genes to be curated. Proximate upper and lower thresholds will have fewer genes to be manually annotated.



alpha	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
upper threshold	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.7
above UT	357	364	382	426	582	621	670	706	629
lower threshold	0.2	0.3	0.3	0.4	0.4	0.4	0.4	0.5	0.5
below LT	197	337	320	464	442	421	387	505	490
total for curation	770	623	622	434	300	282	267	113	205
% for curation	58 %	47 %	47 %	33 %	23 %	21 %	20 %	9 %	15 %
accuracy	0.691	0.72	0.724	0.742	0.764	0.789	0.802	0.822	0.829
curation ratio sc...	1.19	1.53	1.54	2.26	3.37	3.7	3.98	9.63	5.35
	○	○	○	○	○	○	○	●	○

upper threshold
100 % precision

lower threshold
100 % negative predictive value

Fig. 4. Example of the results displayed by SamPler where the user selected the alpha value of 0.8.

3.2.4 Manual Annotation

A good manual annotation is an essential part of this model reconstruction, since it will determine if a substantial number of genes are accepted or rejected.

The manual annotation was performed using a pipeline. This pipeline consists in a series of operations that assign a function to a gene and determine the confidence level of that assignment (Figure 5).

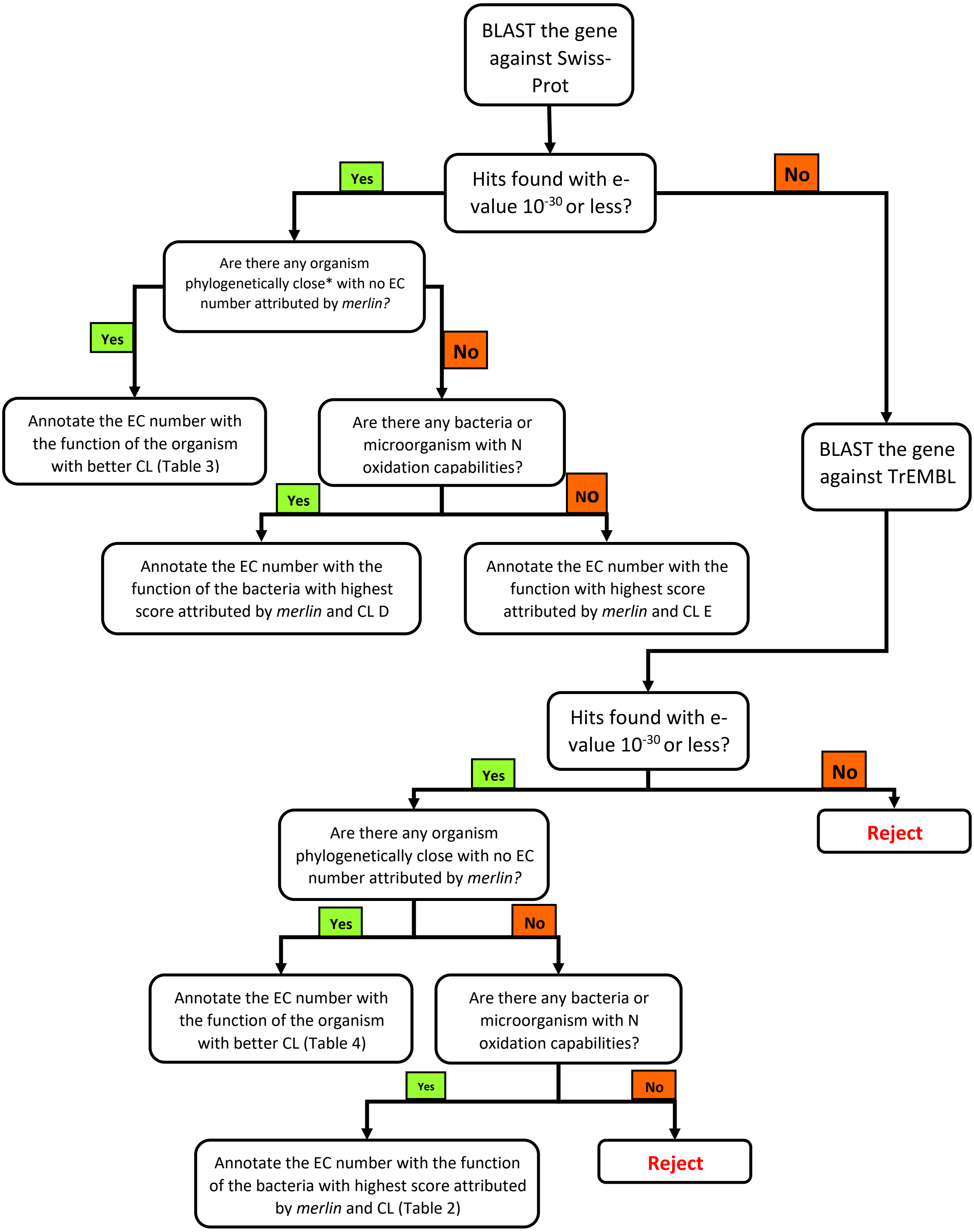

BLAST the gene against Swiss-Prot
Hits found with e-value 10^{-30} or less?
Yes
No
Are there any organism phylogenetically close* with no EC number attributed by *merlin?*
Yes
No
Annotate the EC number with the function of the organism with better CL (Table 3)
Are there any bacteria or microorganism with N oxidation capabilities?
Yes
NO
Annotate the EC number with the function of the bacteria with highest score attributed by *merlin* and CL D
Annotate the EC number with the function with highest score attributed by *merlin* and CL E
BLAST the gene against TrEMBL
Hits found with e-value 10^{-30} or less?
Yes
No
Reject
Are there any organism phylogenetically close with no EC number attributed by *merlin?*
Yes
No
Annotate the EC number with the function of the organism with better CL (Table 4)
Are there any bacteria or microorganism with N oxidation capabilities?
Yes
No
Annotate the EC number with the function of the bacteria with highest score attributed by *merlin* and CL (Table 2)
Reject

Fig. 5. Manual annotation pipeline used for genomic curation.

N. vulgaris is prone to have similar protein functions to organisms that are phylogenetic close to it [113]. Therefore, the pipeline developed to manually annotate genes was based in the phylogenetic distance that *N. vulgaris* has with selected organisms. Curation status of homologous genes will also determine the confidence level attributed to a gene.

The confidence level attributed favours the curation status over the phylogenetic proximity, for example, a curated homologous gene of a distant organism will have a greater confidence level than a gene of a closer organism with no curation.

The first step of the pipeline is to perform a BLAST alignment to a curated database. For this process Swiss-Prot was the database selected. *merlin* automatically annotates all genes and selects the EC number of each gene with the higher score according to Equation 8 [15].

Since Swiss-Prot only contains curated information, the score attributed to these genes will be inevitably high. Confidence levels, according to phylogenetic proximity to *N. vulgaris,* can be consulted in Table 3.

Table 3 Confidence level of homologue genes found in Swiss-Prot.

Gene function similar to:	***Confidence Level***
N. winogradskyi and *N. hamburgensis*	A
N. winogradskyi or *N. hamburgensis*	B
E. coli	C
Bacteria with NOB characteristics	D
More than 4 other bacteria	E

Since there is no curated data regarding *N. vulgaris* genes, these were not considered on Table 3, otherwise they would have the maximum score.

A considerable number of *N. vulgaris* genes do not have homologous genes available in Swiss-Prot. For these genes, BLAST alignments against TrEMBL, a non-curated database which will decrease confidence levels, were performed.

Unlike the results from Swiss-Prot, TrEMBL has information regarding *N. vulgaris,* to which will be attributed the higher score.

The confidence level was attributed regarding the phylogenetic proximity as well as its associated function. Table 4 contains information regarding the attributed confidence levels.

Table 4 Confidence level of homologue genes found in TrEMBL.

Gene function similar to:	***Confidence Level***
N. winogradskyi and *N. hamburgensis*	F
N. winogradskyi or *N. hamburgensis*	G
Nitrobacter spp.	H
E. coli	I
More than 4 other bacteria	J

During the manual curation, all genes were revised. This revision implied the use of literature and/or databases, such as BRENDA, to confirm the function of the genes. Due to the lack of studies on *N. vulgaris*, some of the literature consulted was based on other NOB bacteria, such as *N. winogradskyi* or *Nitrobacter hamburgensis* [69].

All manual annotated genes are prone to be changed in future steps of the work, if such changes are justified.

This whole process should attribute to each gene its most likely EC number(s) and function(s).

3.2.5 Integration of the Annotation in the Model

The next step is the integration process of the manually curated genes into the model. *merlin* does this process automatically with an internal algorithm that determines which reactions will be included in the model. Then, *merlin* uses the annotation results to construct a network with metabolic data retrieved from KEGG [67].

3.2.6 Transport Proteins Annotation

The Transporter Classification DataBase (TCDB) provides a classification system for transport proteins, which are proteins responsible for promoting the transport of metabolites across biomembranes, with TC numbers [114].

Transporters annotation was performed using *merlin* Transport Reactions Annotation and Generation (TRIAGE) tool. TRIAGE identifies transporters encoded in the genome, and determines which metabolites are transported, the stoichiometry of the reaction and between which compartments it takes place. The similarity threshold used for this analysis was 0.1. The created transport reactions were then integrated in the model.

3.2.7 Compartments Prediction and Annotation

The compartments prediction allows assigning each reaction to one or more subcellular compartments.

PSortb (v 3.0), a subcellular localization prediction tool, was designed to predict protein localization sites in cells, using information of an amino acid sequence and its source origin [115], [116]. Ultimately, Psort3 generates a report that can be imported into *merlin*.

The integration of this information with the model will allow assigning reactions to compartments, according to the location of the proteins catalysing such reactions. This can be done automatically through *merlin* [15]. Reactions inserted manually were automatically inserted into the cytoplasm and then curated accordingly.

3.3 *GSM Model* Curation

Manual curation is the next step in the metabolic network reconstruction. This step is crucial, as automatic methods are fallible and compromise the accuracy of the model.

This step consists in revising the *GSM model* with the help of literature, organism-specific databases and consulting expert researchers.

This section comprises detecting inconsistencies in the model, adding new organism-specific reactions, solving connectivity problems in the network and verifying the reactions reversibility.

A series of operations were performed to gradually improve the quality of the model. This is an iterative process, that involves revisiting the genome annotation and adding/removing reactions from the network, which continued until the model accurately depicted the metabolism of the organism.

The model was compared to other existing models of organisms with some degree of similarity (e. g. *N. winogradskyi, N. hamburgensis* and *E. coli*), to assess the metabolic network.

3.3.1 Gap Filling and Dead-end Removal Process

Draft genome-scale networks usually have blocked metabolites. That is, some metabolites may not have a reaction that consumes or produces them. These metabolites are labelled dead-ends and may disrupt the network, if involved in essential reactions. In other cases, dead-ends include reactions involved in a specific via; however, at least one reaction is missing from the path. The process of filling this openings is designated as gap filling.

If a metabolite is not consumed by any reaction, it could be because one or more reactions that should consume the metabolite are not present in the network. When this was the case, the missing reactions were identified. This was performed by consulting all reactions that could consume the metabolite and by selecting those that better fitted the problem. Then, the enzymes responsible for those reactions were identified, if the enzymes were associated with a gene, a literature research was performed to determine its purpose in the organism. If the enzyme presence was justified, it would be added and integrated in the model. Lastly, if a metabolite is not consumed by any reaction, it could be hypothesised that it is a waste product. If this was not the case, the reactions that produce the metabolite were revised. Some reactions were integrated because they were associated to an ancestral gene that could have lost its function. Finally, if none of the cases above could solve the dead-end, the reaction was identified and added manually. This procedure was carefully and critically performed and only used as a final course of action, since such reactions were not supported and could hinder the models accuracy.

If a metabolite was not produced, the reaction that needed this metabolite was revised and removed if not essential or justified. Otherwise, the set of reactions that could produce these metabolites were selected and determine their associated enzymes. If the enzymes presence in the model was plausible, they were added.

When essential reactions were missing from a cycle or via, the process to solve this problem was similar to those posed before. The reactions were selected, and the enzymes were identified. The enzymes were integrated into the model if it were justifiable by the literature or other data source. Other required reactions were manually added.

Finally, when a metabolite was isolated from a compartment that had a reaction that needed the metabolite, the reaction compartmentalization was manually curated and changed if required. If the compartmentalization was correct, a transporter would be added if confirmed by literature.

Figure 6 depict schematically the dead-end removal process and Figure 7 displays the gap filling schematic process.

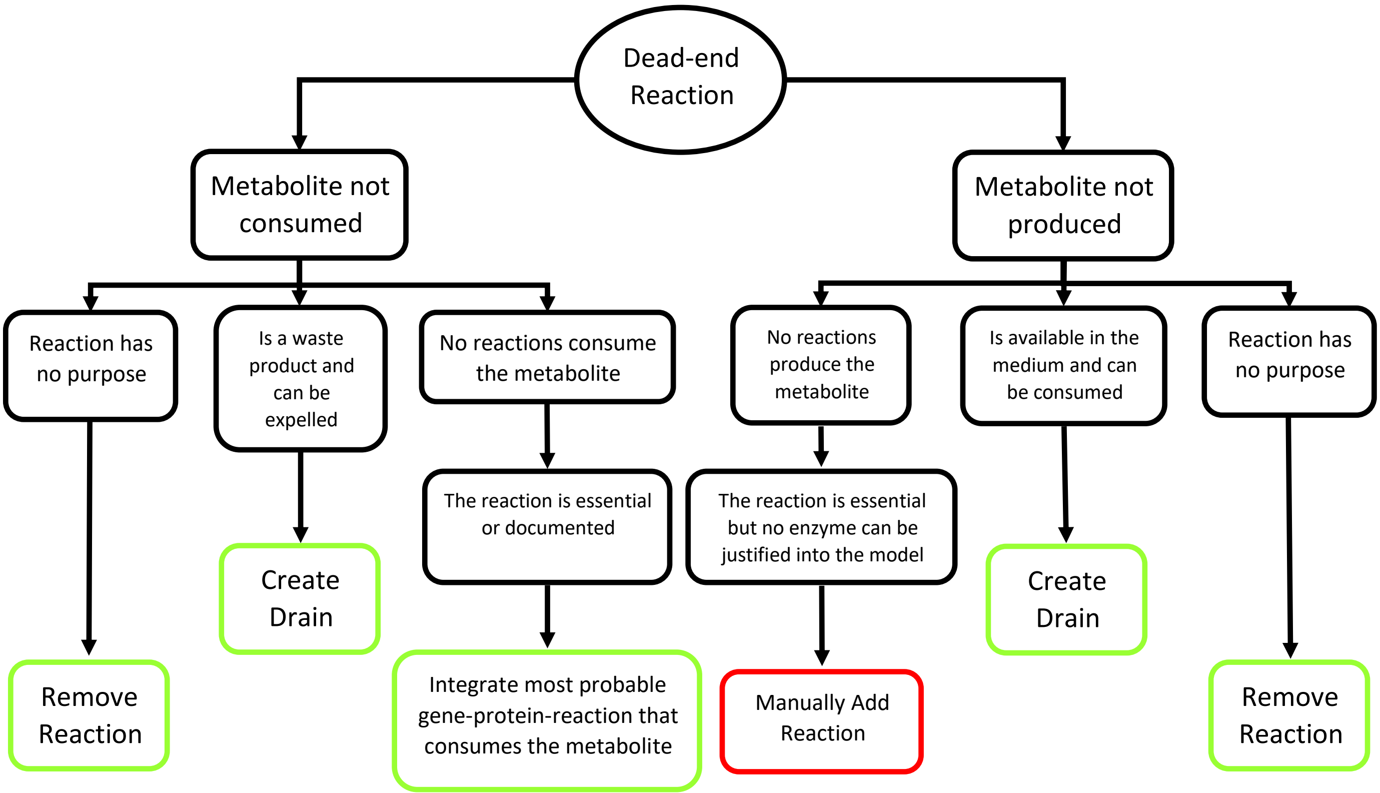

Dead-end Reaction
Metabolite not consumed
Metabolite not produced
Reaction has no purpose
Is a waste product and can be expelled
No reactions consume the metabolite
No reactions produce the metabolite
Is available in the medium and can be consumed
Reaction has no purpose
The reaction is essential or documented
The reaction is essential but no enzyme can be justified into the model
Create Drain
Create Drain
Remove Reaction
Integrate most probable gene-protein-reaction that consumes the metabolite
Manually Add Reaction
Remove Reaction

Fig. 6. Dead-end removal pipeline and results. Green boxes represent desirable outcomes, red ones represent undesirable outcomes.

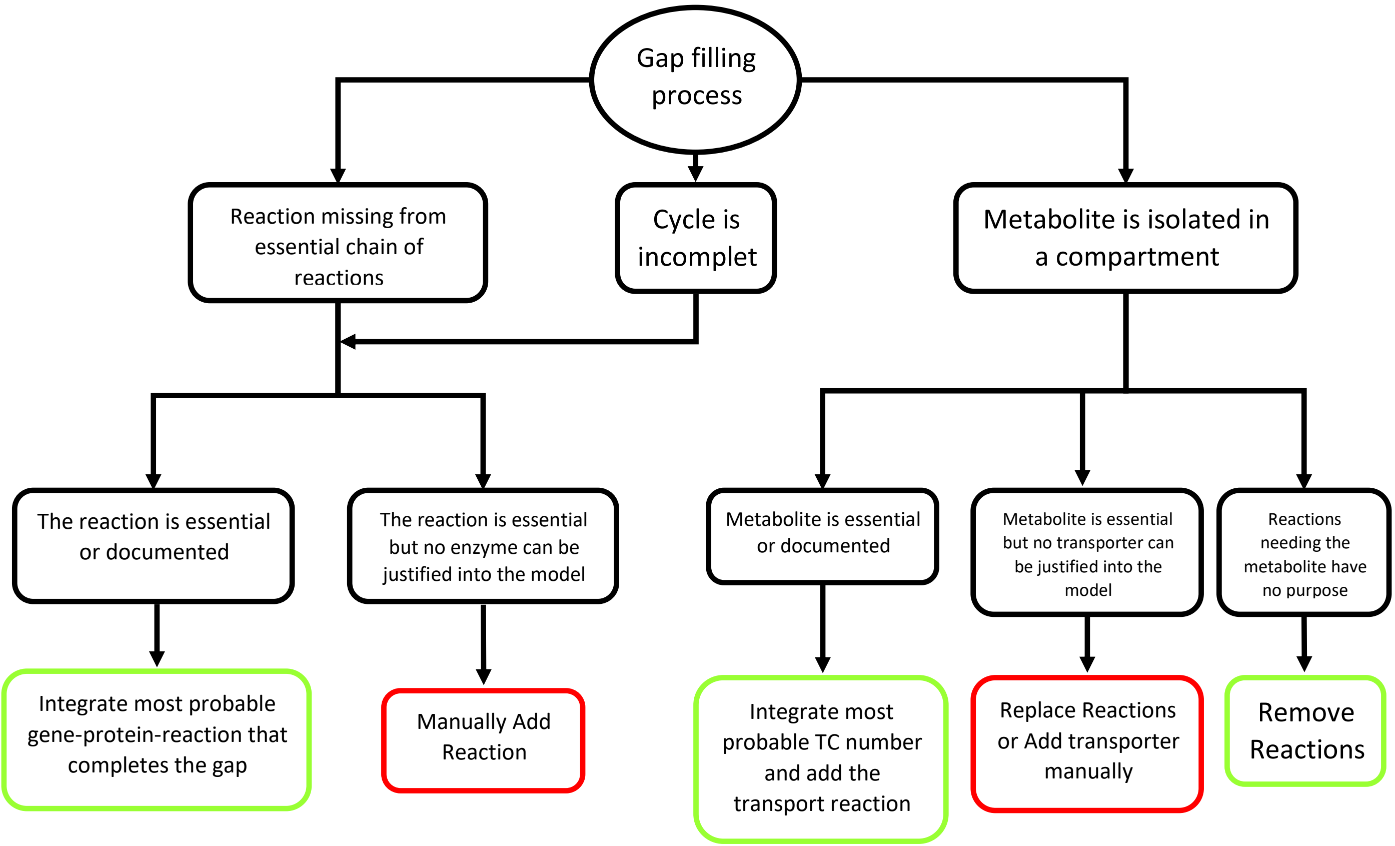

Gap filling process
Reaction missing from essential chain of reactions
Cycle is incomplet
Metabolite is isolated in a compartment
The reaction is essential or documented
The reaction is essential but no enzyme can be justified into the model
Metabolite is essential or documented
Metabolite is essential but no transporter can be justified into the model
Reactions needing the metabolite have no purpose
Integrate most probable gene-protein-reaction that completes the gap
Manually Add Reaction
Integrate most probable TC number and add the transport reaction
Replace Reactions or Add transporter manually
Remove Reactions

Fig. 7. Gap Filling pipeline and results. Green boxes represent desirable outcomes, red ones represent avoidable outcomes that should be avoided.

3.3.2 Reactions Reversibility

Reaction reversibility was initially performed by *merlin*. However, some reactions reversibility were incorrect, and were manually revised. This corrections were based mainly on KEGG pathways, MetaCyc and eQuilibrator (a web interface designed to enable easy thermodynamic analysis of biochemical systems) [66], [67], [117].

3.3.3 Balancing Stoichiometry

This *GSM model* represents a steady-state metabolism. This means that every metabolite is consumed and produced at the same rate. A steady-state model has the advantage of being simpler and does not require knowledge about reactions kinetics. However, it does not represent the real metabolic system of the organism [118].

merlin is able to detect unbalanced reactions. These reactions were properly balanced consulting KEGG and MetaCyc [66], [67].

Unbalanced reactions were usually caused by misplaced protons (H^+) and unspecific metabolites (i.e. fatty acids, electron acceptors and donors, among others). Unspecific metabolites usually have repetitive monomers that disrupt the reactions balance. These were later specified using literature support or manual stoichiometry analysis.

3.3.4 Biomass Equation

The biomass abstraction must be included and will represent the macromolecular composition of the cell and the building blocks used to generate these molecules. To perform simulations on the model, it is necessary to include a reaction that represents a drain of biomolecules into biomass [64]. The biomass formation will follow the structure of Equation 5.

Where $\mathbf{c}_k$ represents the coefficient of the metabolite $\mathbf{X}_k$ [64]. The growth rate of the organism is represented by the flux of this reaction. The abstraction will include growth-associated energy requirements. If the biomass abstraction cannot be determined, it will be used one from an organism with similar characteristics.

After the biomass abstraction is added to the metabolic network, all the reactions can be represented as a stoichiometric matrix, finishing the reconstruction process.

3.3.5 Gene-protein-reaction Rules

The draft network is reconstructed associating enzymes and transport proteins to reactions and metabolites.

The next step will consist in assembling the metabolic network. Here, each gene must be associated to its protein and consequently to its reaction. This can be achieved checking databases and search which reaction is associated to which gene/protein.

3.4 Community Model Reconstruction

After reconstructing *GSM* models for both *N. vulgaris* and *N. europaea*, community simulations can be performed. This section will describe the process and tools used to merge the models.

The community model will be assembled with a tool designed specifically for this purpose.

3.4.1 Merging Tool Description

The community model was assembled with a tool was developed in-house, within the BioSystems group in the Centre of Biological Engineering (CEB) by Sophia Santos [119].

This tool requires a specific nomenclature for drain reactions of both models and the same identification number for all common metabolites and compartments between models.

Finally, the models were merged into a community model that can be used as a regular *GSM* model to perform simulations.

3.5 Simulation Methods

All metabolites required by both models were provided as extracellular drains directly on *merlin*. Their fluxes were set according to their rates in the cultures. Oxygen and CO_2 were limited by its diffusion rates into the medium. The gases diffusion rates were determined using Ficks Law (Equation 7) [100].

From this point on, *N. vulgaris* and *N. europaea* models can be simulated, using FBA and pFBA, to perceive their behaviour in different media.

Simulations were performed with OptFlux, using CPLEX® (linear programing solver exclusively) as solver [120].

3.5.1 *N. vulgaris* Model Simulations

Initial simulations consisted in the determination of the minimum medium. These simulations determined which metabolites are essential to *N. vulgaris*. The simulations implied the insertion and removal of drains for metabolites existing in the medium and other possible micronutrients. These simulations will allow us to define which metabolites must be supplied for the production of biomass.

Most of the simulations consisted in the maximization of biomass production. These simulations tested the models functionality and took place as the model was being reconstructed.

When the model was able to produce biomass, another set of simulations took place. Essential metabolites consumptions and productions rates were calculated and adjusted simultaneously.

When the model was able to closely simulate the metabolism of *N. vulgaris*, the final set of simulations took place. These were used mostly to determine the maximum growth (maximizing biomass production) since this implied the maximum NO_3^- production.

3.5.2 Simulations on the Community Model

As soon as the *N. vulgaris* and *N. europaea* models accurately simulate the expected results, they were merged.

In this case, the simulation and model reconstruction process were not simultaneous, since the models were already reconstructed. The simulations were performed for the generation of results.

Most simulations focused in maximizing NH_3 and NO_2^- consumption and/or biomass and NO_3^- production.

4. Results and Discussion

In this chapter, all results obtained throughout the project will be presented and explained. *In vivo* results will be implemented into the *GSM* model and *in silico* results, such as simulation results, will be analysed and used to validate the model.

4.1 Wet-lab Results

In this section all laboratorial results obtained will be presented. The results obtained in this section were used to reconstruct the models and to validate its data. Some undocumented findings about *N. vulgaris* will also be stated in this section.

4.1.1 Chemostat Results

As mentioned before, *N. vulgaris* synthesizes ATP oxidizing NO_2^- into NO_3^-. Therefore, monetarizing the NO_3^- will depict the bacterial activity. Note that NO_2^- cannot be used to accurately depict the bacterial activity, since *in silico* results show that a part of NO_2^- could be used as a N source. NO_3^- was measured for a total of 160 days. Showing a minimum concentration value of NO_3^- of 64.4 mg L^{-1} on day 0 and 2085 mg L^{-1} on day 23. From then on, the bacterium showed constant values for NO_3^- of 1864 ± 171 mg L^{-1}. The medium used has no NO_3^-, the initial value of 64.4 mg L^{-1} was most likely been due to the NO_3^- present in the bacterium inoculum.

The concentration of NO_2^- was also monetarized for 160 days. The results show that the negative exponential phase lasted for 23 days. This is expected, since NO_3^- had the same exponential phase duration. NO_2^- consumption halted and maintained a steady concentration value of 7.31 ± 1.93 mg L^{-1}.

4.1.2 Macroscopy and Scanning Electron Microscopy Visualization

The *N. vulgaris* culture was established in a cylindrical reactor and it included the medium described before and *N. vulgaris* inoculum. The culture slowly took a rose tone as the cellular density increased, ultimately leading to a strong coral colour when the culture achieved its maximum concentration. *N. vulgaris* was responsible for this colouration, contrary to the brown colour that is attributed to other *Nitrobacter* [12]. Figure 8 is a picture of the reactor appearance at its early and later stages as well as the samples collected over time.

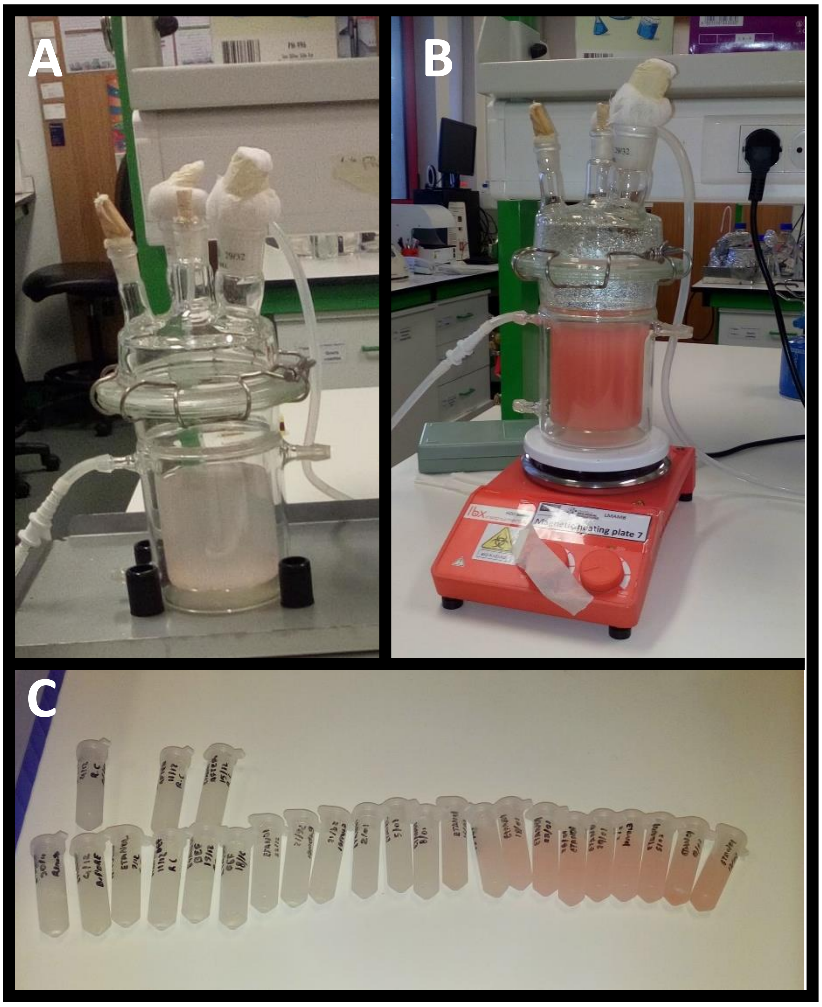


Fig. 8. Images captured in the laboratory. A – Corresponds to the reactor on its early stages. B - shows the reactor in its later stages. C – displays the samples collected over time, far left (day 16) to far right (day 118).

To visualise *N. vulgaris* microscopically, we resorted to SEM. This allowed us to view the cell structure in detail. In Figure 9 it is possible to visualise cells or cell agglomerates of *N. vulgaris*. All cells appear to have rod-shaped structure and similar size. Individualised cells measured from 1.55 µm to 1.95 µm of diameter between poles. *Nitrobacter* was described to have 1.0-2.0 µm between poles and having a rod-shaped form [121].

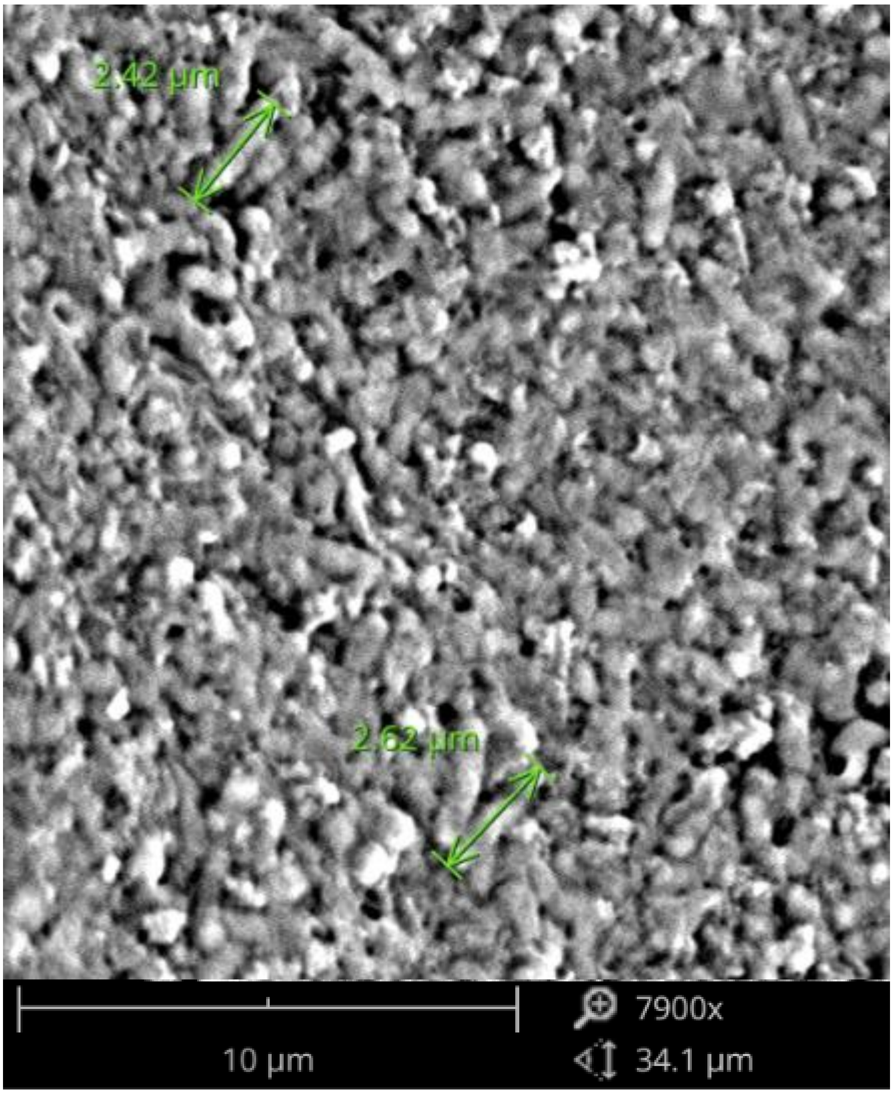


Fig. 9. SEM images of freeze dried *N. vulgaris*. Measurements displayed at green of singular bacterium.

4.1.3 Optical Density and Dry weight

The relation between optical density and dry weight of *N. vulgaris* was determined to be 993.5 ± 37.46 mg L^{-1} at OD600. Figure 10 depicts the graphical relation between these two variable. Equation 9 represents the relation between biomass and optical density.

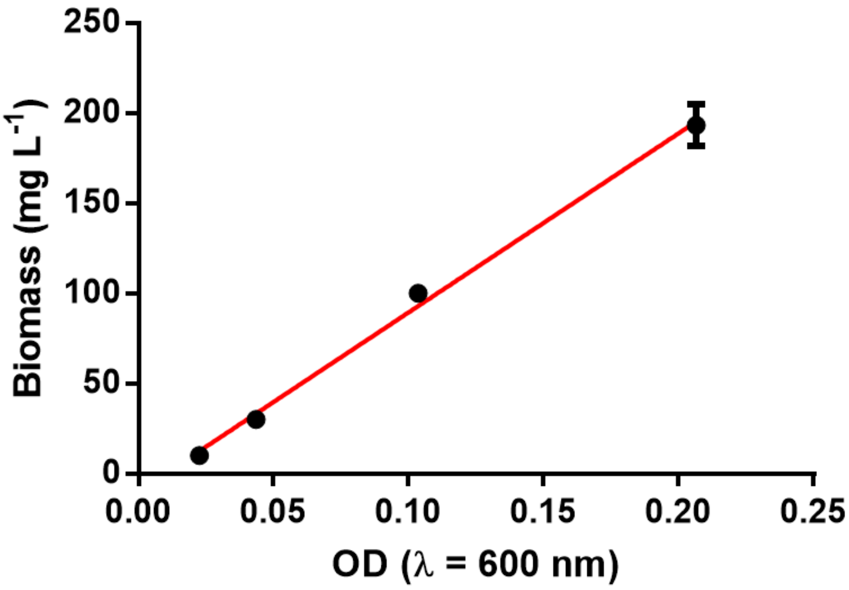


Fig. 10. Relation between Biomass (mg L-1) and Optical Density (λ = 600nm).

$$Biomass = 525.6 \times OD - 10 \qquad \text{(Eq. 9)}$$

4.1.4 Atomic quantification

EDS was used to determine the atomic composition of *N. vulgaris*. Unfortunately, hydrogen quantification is not possible using this technology. The results show a predominance of carbon, oxygen and nitrogen, expected from a biological source. Lower quantities of potassium, phosphorous, sodium, chlorine, sulphur and magnesium were also detected. Figure 11 shows the atomic percentages.

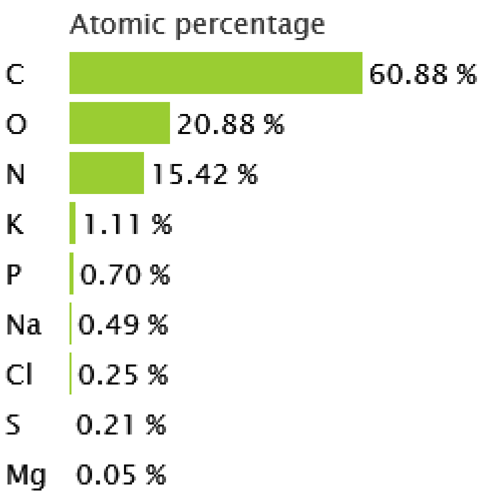


Fig. 11. EDS results of Atomic percentages of *N. vulgaris*.

Figure 12 shows the spectrum generated by the EDS.

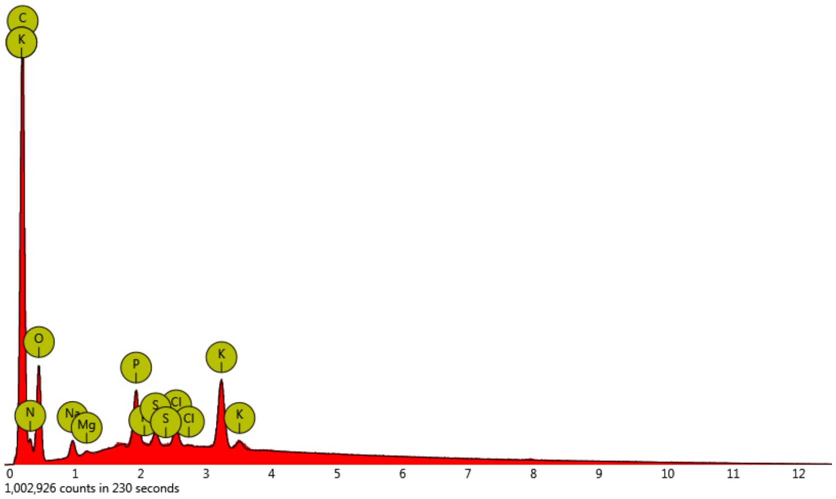


Fig. 12. Graphical representation of the Atomic percentages of *N. vulgaris*.

4.1.5 Gases Quantifications

Following Equation 7, the O_2 available in the reactor is approximately 6.740×10^{-2} mmol m^{-2} s^{-1} and CO_2 is 1.366×10^{-2} mmol m^{-2} s^{-1}. Table 5 shows the values used.

Table 5 O_2 and CO_2 characteristics and reactors dimensions.

Gas	*O_2*	*CO_2*
Partial pressure in air (mmHg)	159.3	23.75
Partial pressure in water (mmHg)	30.4	0.29412
Diffusion constant ($m^2\ h^{-1}$)	7.56×10^{-6}	6.912×10^{-6}
Surface area (m^2)	5.25×10^{-3}	
Thickness (m)	0.08	

4.1.6 Ethanol Quantification

As previously mentioned, a HPLC was used to determine the ethanol concentration in the medium over time. A curve that relates the area measured by the HPLC and the real concentration of ethanol is represented by Equation 10.

$$Area = 3.8102 \times [Ethanol] - 0.4006 \qquad \text{(Eq. 10)}$$

Where **Area** (mV/min) is the area of the peak represeting the ethanol molecule and **[Ethanol]** (g L^{-1}) is the concentration of ethanol. The retention time detected ranged between 21.058 and 21.083 min.

4.2 Computational Results

In this section all computational results obtained will be presented. First, results regarding the *N. vulgaris* culture and then the results regarding the *N. vulgaris / N. europaea* community model. The results obtained in this section were used determine the models accuracy. Explanations for biological responses about *N. vulgaris* will be stated in this section.

4.2.1 *N. vulgaris* Taxonomy

It was concluded that the organisms phylogenetically closer to *N. vulgaris* are other *Nitrobacter* species [111], [112]. Namely, *N. winogradskyi*, *N. hamburgensis* and *Nitrobacter alkalicus*.

Results from MUSCLE, confirm that *N. winogradskyi* is the closest NOB (with its genome sequenced) to *N. vulgaris* as seen in Figure 13.

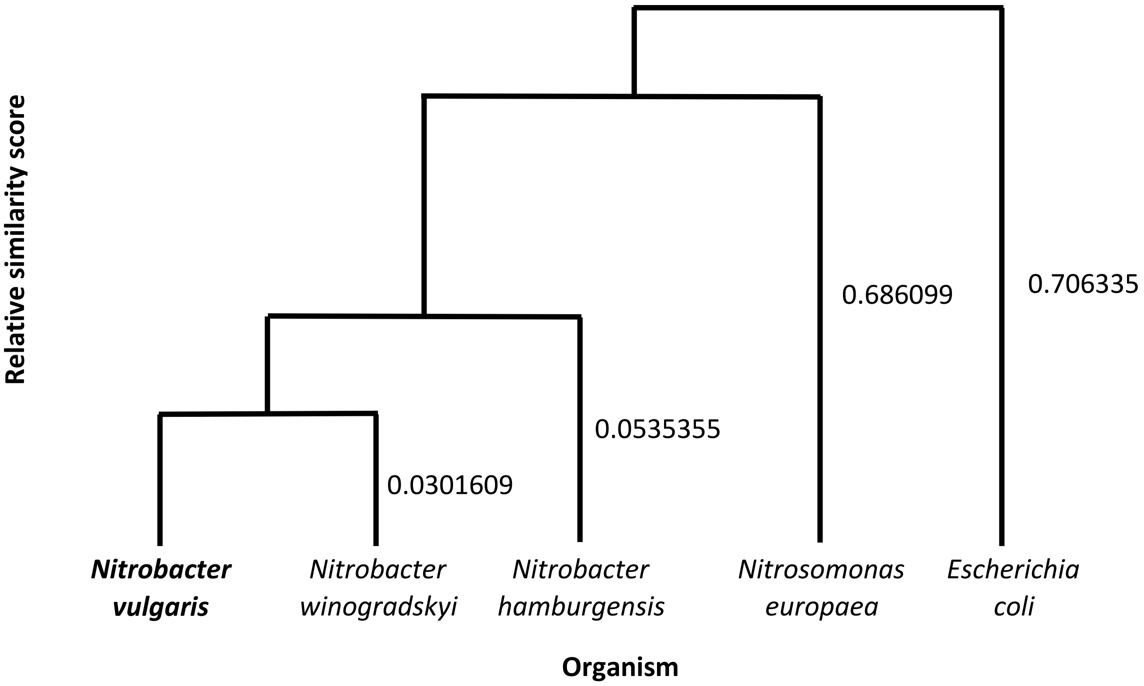


Fig. 13. Results from MUSCLE. Showing the closest organisms to *N. vulgaris* and their respective similarity score.

N. winogradskyi was the closes organism with a *GSM* model reconstructed. The model (BMID000000141943) retrieved from the BioModels Database and was automatically generated by SuBliMinal Toolbox [84]. Since this model is not curated, it was not used as a template, but rather as support.

4.2.2 Determining the Thresholds

The upper and lower thresholds that determine if a gene is automatically accepted or rejected were calculated by SamPler. This calculation required the manual annotation of initial 50 genes. The annotation of these 50 genes can be consulted in Table S1 (on annex). After this, the α values were presented with its associated thresholds. This procedure was made twice, the first for curated genes and the second for non-curated genes.

The options selected can be consulted in Table 6.

Table 6 α score and thresholds selected after the SamPler selected genes curation process.

Curated Genes (BLAST against Swiss-Prot)		*Non-curated Genes (BLAST against TrEMBL)*	
Parameter	Value	*Parameter*	Value
α	0.5	*α*	0.1
Upper Threshold	0.6	*Upper Threshold*	0.5
Lower Threshold	0.5	*Lower Threshold*	0
Total genes for curation	139	*Total genes for curation*	86
Accuracy	0.756	*Accuracy*	0.556
Ratio	6.7	*Ratio*	0.59

4.2.3 Genome Annotation

As previously mentioned, this process involved the phylogenetically closest organisms to *N. vulgaris*, namely *N. winogradskyi* and *N. hamburgensis,* and *E. coli* since it is a well-documented bacteria.

The 50 genes automatically selected by SamPler were manually annotated, this annotation is presented in Table S1 (on annex). From the options presented from SamPler, the one with the best accuracy-total number for curation ratio (6.7) was picked, where α = 0.5, upper thresholds = 0.6, lower threshold = 0.5. The total number of genes that were manually curated was 139. All these genes are presented in Table S2 (on annex).

Although the highest accuracy was provided for an alpha of 0.9, for this project the selected alpha was 0.5. Selecting the former alpha would involve curating 15 % of the genome annotation, whereas the latter only requires the curation of 11 % of the genome. This difference decreases the curation efforts by 92 genes, demanding the curation of only 139 gene annotations. The decrease in accuracy is negligible, as it goes down from 0.829 to 0.756. Thence, *SamPler* proposed 0.5 as the best alpha value, together with upper and lower thresholds of 0.6 and 0.5, respectively.

SamPler results for the non-curated database were picked with the same criteria mentioned above, ratio = 0.59. $\alpha = 0.1$, upper thresholds = 0.5, lower threshold = 0. The total number of genes that were manually curated was 86. All manually annotated genes for TrEMBL are available in Table S3 (on annex).

A total of 363 proteins were detected in *N. vulgaris,* all of them are named. 359 proteins are described as enzymes (approximately 99 %) and 4 (approximately 1 %) are transporters.

Table 7 summarizes the protein classifications.

Table 7 Protein classification and relative frequency on *N. vulgaris*.

Protein Class	*Identifier*	*Percentage*
Oxidoreductases	EC 1	18%
Transferases	EC 2	37%
Hydrolases	EC 3	15%
Lyases	EC 4	11%
Isomerases	EC 5	5%
Ligases	EC 6	14%

4.2.4 Compartments Annotation

N. vulgaris is a Gram-negative bacteria, thus it contains two subcellular compartments: cytoplasm and periplasm. A third compartment was considered that represents the space outside the bacteria: extracellular space [122]. The extracellular space does not represent a real biological compartment, but represents the outside of the cell. This compartment also serves as a connecting point between the *N. vulgaris* and the *N. europaea* model.

4.2.5 Transporters Annotation

TRIAGE has detected 193 genes responsible for transporters on *N. vulgaris*, with 57 associated reactions.

Only 25 transport reactions were kept in the model. All the other reactions were either duplicates or had no function in the model were removed.

Transport reactions also supplied information concerning metabolism movements within the cell, such as the flux of glycerol, acetate, fumarate. This reactions point to the possibility of other possible sources of carbon or energy for *N. vulgaris*.

4.2.6 Gap Filling and Dead-end Removal

The *N. vulgaris* draft model described the bacteria as a simple system, with one compartment and had no transporters associated. After the gap filling and dead-end removal processes, the final model, comprising all compartments, transport reactions and drains has no unconnected reactions nor dead-end metabolites.

4.2.7 Reactions Balancing

Unbalanced reactions in the model were mostly caused by misplaced protons or by generic compounds. No more than 20 % of the reactions were unbalanced, however, many were disrupting to the model, since they could create an endless supply/demand of metabolites.

These reactions were corrected.

4.2.8 Biomass Composition Quantification

The biomass composition was determined using laboratorial methods for DNA, RNA, carbohydrates and proteins. Amino acids were determined using a specific *merlin* tool (e-biomass). Lipids, cofactors and inorganic ions quantification was estimated from the *E. coli* iAF1260 model [104]. Table 6 shows the overall biomass composition.

Table 8 Biomass constitution.

Constituent	*Percentage (%)*
DNA	0.526
RNA	5.304
Proteins	52.548
Carbohydrates	28.622
Lipids	9.100
Cofactors	1.00
Inorganic Ions	2.90

4.2.9 *GSM* Model of *N. vulgaris*

The *N. vulgaris* model has a total of 410 genes, from which 75 have no associated name. There are 170 genes encoding more than one protein, 193 genes encoding transporters and 19 genes encoding both enzymes and transporters.

The number of unique metabolites in the model is 579. Some of these metabolite are macromolecules, such as **e-Protein,** that represents the average composition of the proteins biosynthesized in this organism. Some metabolites represent generic reactants that could not yet be identified, such as (i.e.: Holo-[carboxylase] and [Enzyme]-cysteine).

Metabolite analysis show that 41 metabolites are present in the extracellular compartment, from which, one is exclusive to this compartment (Nitric Oxide). 15 metabolites are present in the periplasmic compartment, from which, two are exclusive to this compartment (ATP-L-glycero-β-D-manno-heptose and Di[3-deoxy-D-manno-octulosonyl]-lipid A). Finally, 575 metabolites are present in the cytoplasmic compartment.

From a total of 711 reactions, 56 (approximately 8% total) metabolic reactions have no gene association. These reactions were inserted or manually created to complete the model or represent the polymerization of macromolecules involved in biomass assembly. Most have literature support; however, others had to be inserted to ensure the connectivity of the model. Table S3 (on annex) shows manually inserted reactions and the respective reason for their insertion.

Several drains were created, though the default environmental conditions set in the model only requires ten. Additionally, 29 transport reactions were manually added to the model.

The model comprises 83 metabolic pathways, though 72 reactions have no metabolic pathway associated.

4.2.10 Simulation and *in vivo* Values

The real rates of metabolites and gases available to the culture when biomass was in a stationary phase were determined and are available in Table 8. The NO_2^- consumption rate was set to the rate obtained *in vivo* when biomass production was in a stationary phase. Biomass production rate was determined to be 5.33×10^{-4} gDW h^{-1} in this condition.

Table 9 *In vivo* and *in silico* metabolites rates obtained. [1]Maximum rate possible.

Metabolite	*in vivo Rates (mmol gDW^{-1} h^{-1})*	*Simulation results Rates (mmol gDW^{-1} h^{-1})*
NO_2^-	-0.034981936	-0.034
O_2	-18.46020229 [1]	-0.045
CO_2	-3.740331315[1]	-0.006
Ethanol	-0.086342467	-0.0034
NO_3^-	0.023325908	0.03194

NH_3, Fe^{2+}, H_3PO_4, H_2SO_4, H_2O and H^+ were not measured during this experiment, therefore their rates will not be considered to evaluate the models validity. Since NH_3, Fe^{2+}, H_3PO_4 and H_2SO_4 are metabolites with low consumption associated, it is expected that their concentration in the model will not limit *N. vulgaris* growth. H_2O and H^+ will be given an unbound availability rate in the model due to their high disposal in the medium.

4.2.11 Simulation Results of *N. vulgaris* Model

The first simulations had the objective to test the minimum medium.The results obtained showed that O_2, NO_2^-, orthophosphate (H_3PO_4), sulphate (H_2SO_4), iron (II) (Fe^{2+}), H^+ and CO_2 or ethanol or both had to be present to produce biomass. NO_3^- and H_2O were obligatory waste products. No additional waste products were detected.

These results are in accordance with the expected outcome, where O_2 is mandatory for the production of ATP and other biological functions.

NO_2^- is used as the primary energy source of the proton pump and produces NO_3^- in this process. NO_2^- is also used as the N source, but can be replaced by NH_3. Even if NO_2^- could sustain both energy and N demands, *N. vulgaris* will consume NH_3 as it requires energy to be produced from NO_2^-.

H_3PO_4 and H_2SO_4 are essential metabolites since they are the phosphorus and sulphur source, respectively. Iron is essential since it is used as a catalyser of the electron transport chain.

The H^+ into H_2O flux within the organism are involved in the production of energy.

Either CO_2 or ethanol must be available since *N. vulgaris* must use one as a carbon source, or in case of ethanol, carbon and energy source.

When only one carbon source is available, the biomass produced is higher when ethanol is the precursor, with approximately 5 % more biomass produced, compared to when only CO_2 is available. The fact that ethanol can be used as a carbon and energy source simultaneously might be the cause for these results.

If a limit to biomass production is imposed, and no ethanol is available as a carbon source, the model will show an increased consumption of NO_2^- in 17 % and NH_3 in 18 %, when compared to the previous case. This is indicative of the usage of ATP synthase, which requires NO_2^- to produce ATP and since ethanol is not available, ATP requirements must be fulfilled only by NO_2^- reduction, increasing its consumption. NH_3 consumption increase should be related to its energy production costs. Since ethanol is not available, NH_3 cannot be synthesised due to the lower ATP availability.

When only ethanol is available as a carbon source, nitric oxide is expected to be released by the model. *In vivo* tests could not validate nor contradict this result.

Finally, a total of 185 genes (5.3 % of the total genome) were determined as critical. In comparison, *E. coli* has 302 critical genes (9 % of its total genome) [123].

4.2.12 Validation of *N. vulgaris* Model

Previous tests from Carvalho *et al*. [124] on *N. vulgaris* show that 71.7 % of the NO_2^- consumed is converted into NO_3^-. The results obtained *in vivo* in this work show a rate 65.7 % conversion rate, whereas the obtained through model simulations show that this rate is approximately 94 %, with variations when NH_3 is supplied or not, or which carbon source is available.

NO_2^- is the main source of energy to *N. vulgaris* and represents its only source when no organic matter is available. Simulation results show that when ethanol and CO_2 are available, NO_2^- consumption is at its maximum and when only ethanol is present NO_2^- consumption is higher than when CO_2 is the sole carbon source.

Values for ethanol consumption differ slightly between *in vivo* and *in silico* results. Ethanol consumption rates *in vivo* are 26 fold higher than *in silico*. The explanation for this discrepancy lies in the fact that consumption rates *in vivo* could be compromised by evaporation of ethanol throughout time, resulting in higher consumption rates values or measurement errors.

pFBA simulations show that biomass production varies slightly when any of the environmental conditions change. The model reacts differently depending on the carbon source available. The model achieves its maximum growth when both sources of carbon are available. We can then conclude that *N. vulgaris* growth is directly correlated with the amount of carbon that is available for consumption.

4.3 *N. vulgaris* Metabolism and Physicochemical Results

Analysis on the metabolic model, together with literature, allowed a better comprehension of *N. vulgaris* internal metabolism.

This section comprises the physicochemical analysis of the results obtained throughout the work.

A scheme representing *N. vulgaris* metabolism is shown on Figure 14.

Table 8 shows all the rate values obtained.

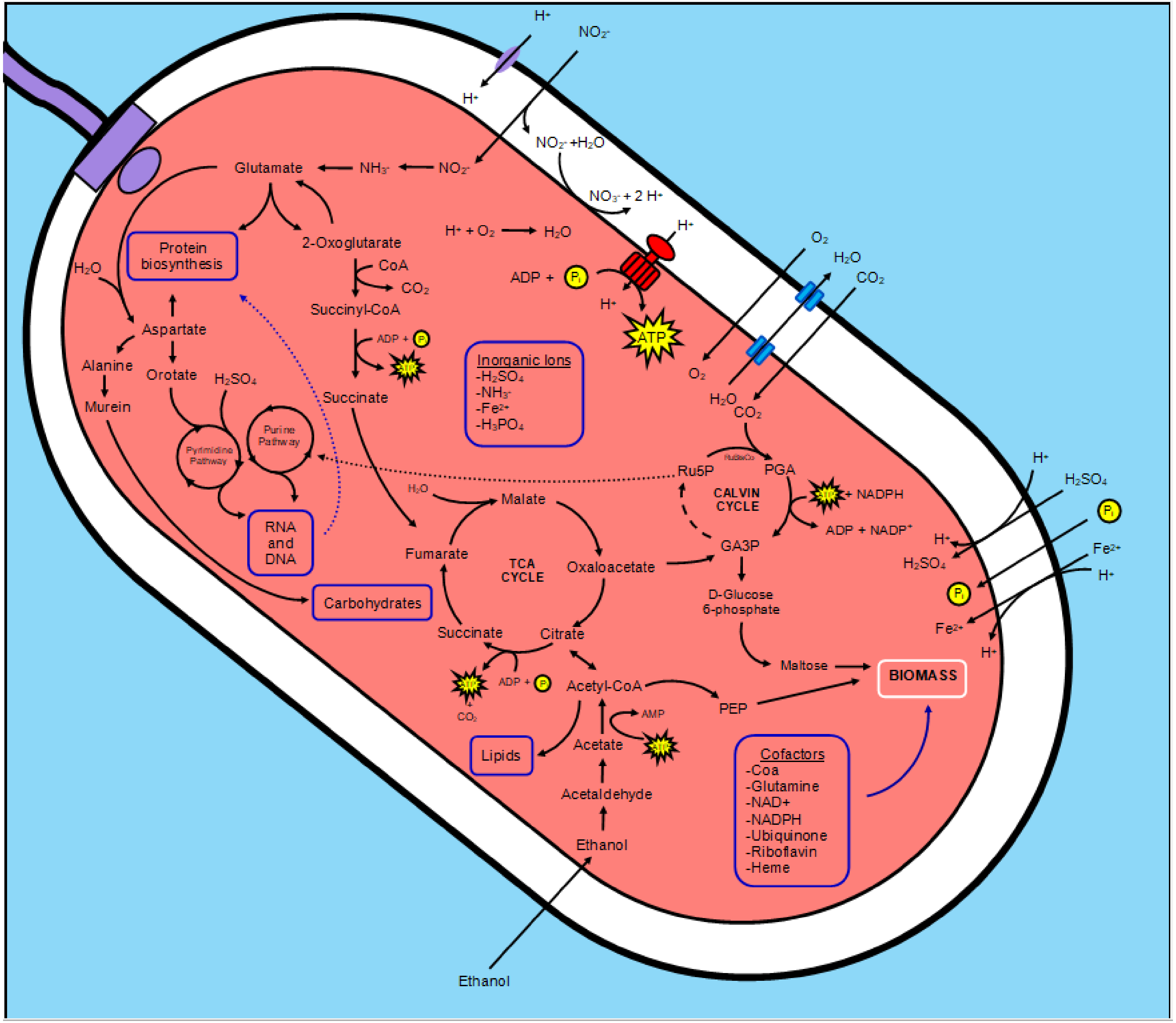


Fig. 14 Scheme of *N. vulgaris* with principal metabolites, reactions a pathways.

4.3.1 Nitrogen Metabolism

NO_2^- oxidation into NO_3^- is done in the periplasm and it is expected to produce 3 ATP for every molecule of NO_2^- that is oxidized. 0.5 ATP are produced in the ATP synthase from 2 H^+ and 2.5 ATP from the electron transport chain using the Nicotinamide Adenine Dinucleotide (NADH) produced during this process.

NO_2^- was found to be the limiting metabolite for *N. vulgaris*.

It is expected that *N. vulgaris* can grow in the absence of NH_3, producing this metabolite from NO_2^-, however this was not tested *in vivo*. This is plausible due to the discrepancy between NO_2^- consumed and NO_3^- produced by *N. vulgaris*. It is expected that the totality of the NH_3 produced from NO_2^- is converted into amino acids or nucleic acids.

Additionally, simulations show that *N. vulgaris* will only convert NO_2^- into NH_3 if NH_3 availability is low and is restricting *N. vulgaris* growth. This is expected to be the observed case since the NH_3 production through NO_2^- is energy dependent (approximately 7.5 ATP for every NH_3 produced). NO_2^- oxidation into NO_3^- has a rate of 100 % conversion when NH_3 availability satisfies *N. vulgaris* demands. When no NH_3 is available the NO_2^- to NO_3^- conversion is reduced to approximately 80 %, using up to 20 % of the NO_2^- to produce NH_3, to subsequently be used as N source.

Figure 14 depicts the NH_3 pathway into these macromolecules.

Nucleic acids are shown to be produced through two essential cycles: Pyrimidine and Purine pathways.

Figure 15 shows the N reactive species concentration through time. NO_2^- consumption and NO_3^- production stabilize after approximately 23 days.

NO_2^- consumption rate when in the biomass stationary phase are determined to be $0.0345 \pm 4.0 \times 10^{-4}$ mmol gDW^{-1} h^{-1}. NO_3^- production rate when in the biomass stationary phase is determined to be 0.023 ± 0.014 mmol gDW^{-1} h^{-1}.

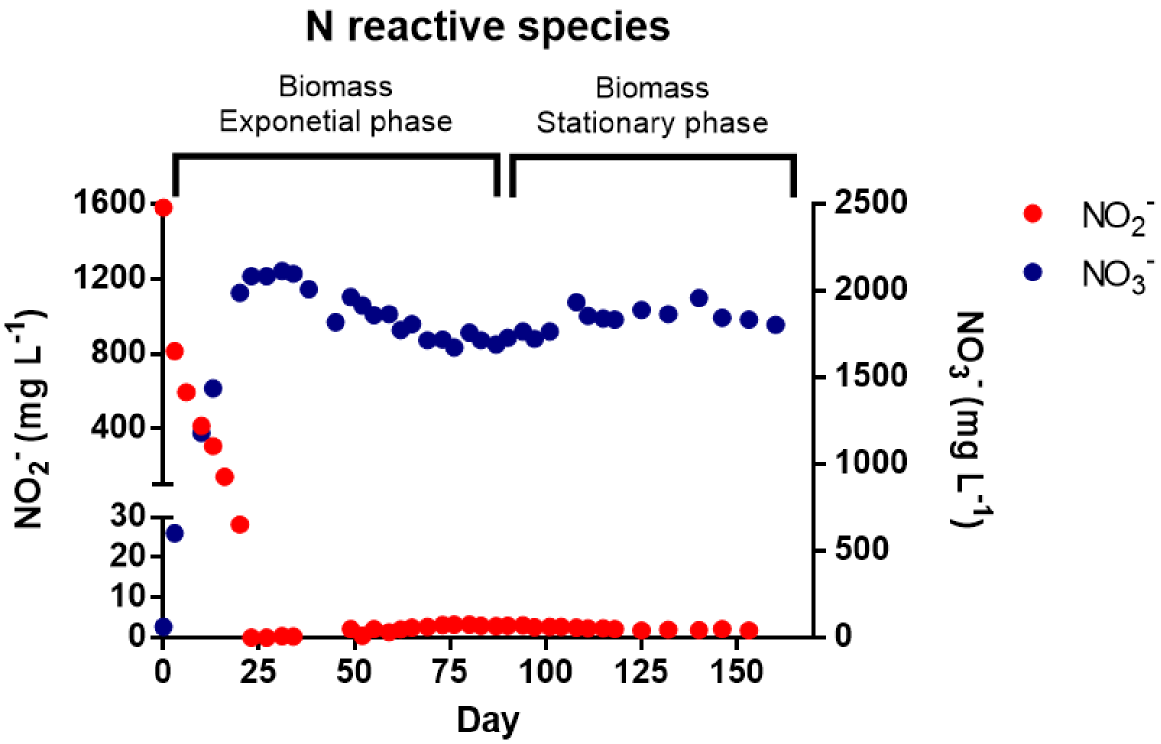


Fig. 15 Nitrogen reactive species over time.

4.3.2 Carbon Fixation

For carbon metabolism two precursors were used: CO_2 and ethanol. Simulations and literature data show that organic compounds are the preferable source of carbon in ideal conditions. Through transporters and reactions analysis, it was expected that degradation of other organic molecules to obtain carbon is plausible. Some of the possible alternative substrates are fructose, fumarate, malate, acetate and glycerol [49], [125].

Figure 16 is the graphical representation of the ethanol concentration over time. The ethanol consumption rate stabilizes after 80 days and maintains an estimated value of 0.0219 g L^{-1}.

The ethanol consumption rate was determined to be 0.0888 ± 0 mmol gDW^{-1} h^{-1}.

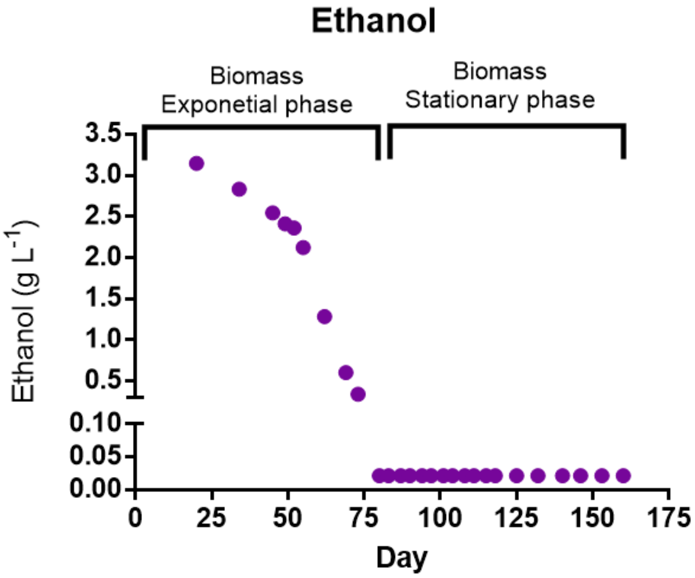


Fig. 16 Ethanol concentration over time.

4.3.3 ATP Production

Energy production can be achieved through H^+ gradient difference between the periplasm and the cytoplasm or through degradation of ethanol that enables the TCA cycle. ATP levels were not directly measured in this experiment, but it is expected that energy production values are correlated to NO_2^- and ethanol consumption rates.

4.3.4 Biomass Production

Biomass production was calculated using optical density. Figure 17 shows the biomass concentration through time in the reactor. Biomass stabilizes after 94 days and maintains an approximate value of 0.695 ± 0.113 g of dry weight in the reactor.

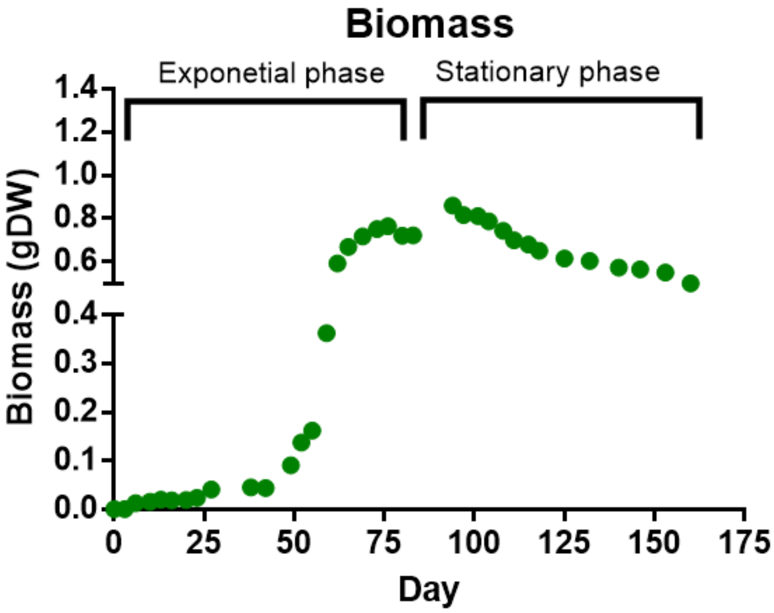


Fig. 17 Biomass in the reactor over time.

The estimated mixotrophic growth rate is 27 h [46], [48]. The results obtained *in vivo* in this work point to a duplication time of 17 h. Since the studies on growth of *N. vulgaris* do not utilize ethanol as a carbon source and the duplication time determined in this work is 10 h lower than other studies, this could imply that ethanol is a better carbon source, than fructose, fumarate, malate, acetate or glycerol for *N. vulgaris* growth [49], [125].

4.3.5 pH Analysis

pH level was monitored through time to ensure an optimal growth rate. Figure 18 depicts the pH level of the reactor. Figure 18 shows a slight increase in pH. The pH level was maintained relatively stable state throughout the experiment.

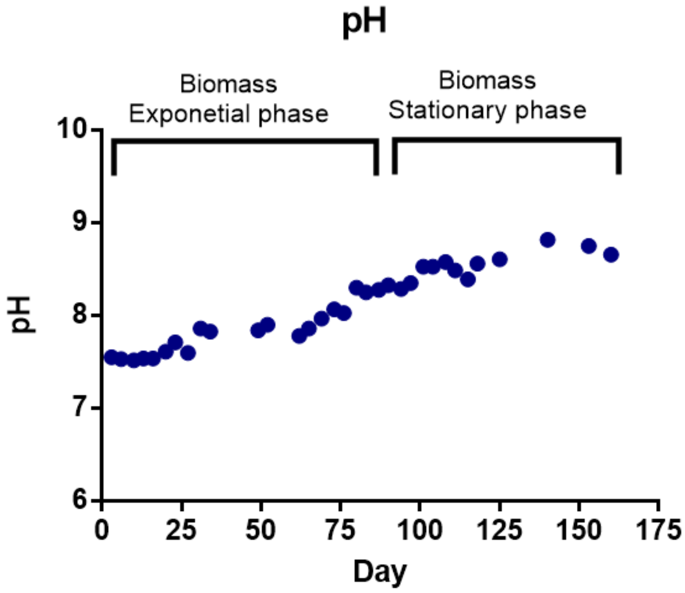


Fig. 18 pH level over time.

4.4 *N. vulgaris* - *N. europaea* Community Model

In this section, we will analyse results obtained from the *N. europaea* - *N. vulgaris* community model.

The community model comprises a total of 1297 internal reactions and 23 drains (2 exclusive to *N. europaea* and 3 to *N. vulgaris*). It is estimated that 1142 metabolites are present in the community model, although some might be duplicates with different identification numbers. 798 genes are present in the model being 378 (47 %) essential for biomass production of both bacteria. Table 9 summarises these properties.

Table 10 Community model properties.

Property	*Community model*	*Exclusive to N. europaea model*	*Exclusive to N. vulgaris model*
Reactions	1297	589	708
Essential Reactions	832	430	402
Drains	23	2	3
Genes	798	388	410
Critical Genes	378	193	185
Metabolites	1142	-	-

4.4.1 Simulations of the Community Model

Simulations on the community model show that both bacteria can live in community, as expected, corroborating the already observed results obtained during *in vivo* experiments performed by members of Environmental Microbiology Laboratory (data not shown).

All simulations performed show that O_2 is the most consumed metabolite from the community model. This result is expected, since both bacteria require this metabolite to survive.

The second most consumed metabolite is NH_3. This metabolite is essential for *N. europaea* but is also consumed by *N. vulgaris*. *N. europaea* consumes 97 % of the available NH_3. This was expected as an essential metabolite for *N. vulgaris* (NO_2^-) can only be produced from NH_3 through *N. europaea* in this system. This reveals NH_3 as the limiting metabolite in the community model.

CO_2 and ethanol are also consumed in significant proportions, as 100 % of the consumed ethanol is consumed by *N. vulgaris* and 76 % of the CO_2 consumption is attributed to *N. europaea*, its only carbon source. *N. vulgaris* consumes 1 % more ethanol and 1 % less

CO_2 in community in comparison to its solo conditions. These changes should be regarded as residual oscillations since *N. vulgaris* carbon sources are not limited in any case (solo or community model).

The combined biomass production was determined to be 0.0147 gDW h^{-1}. *N. vulgaris* is directly dependent of the NO_2^- produced by *N. europaea*, thus it is expected that most of the biomass produced by the model belongs to from *N. europaea*. This possibility, however, requires *in vivo* validation.

Table 10 shows all metabolites rates obtained from the community model, when NH_3 consumption rate was set to its theoretical consumption rate.

These results show that there is a virtually no accumulation of NO_2^- in the system. This is a good result, as the main focus of the work is to remove both NH_3 and NO_2^- from biological systems.

Currently, the model shows an oxidation rate of approximately 48 % of NH_3 into NO_3^- with a leftover of approximately 0 % in the form of NO_2^-. The remaining N is expected to be converted in to urea exclusively produced by *N. europaea* (residual quantities), nitric oxide (47 %) and in the biomass (5 %) of the two bacteria.

Table 11 Metabolite rates on the community model.

Metabolite	*Pool*
NH_3	-3.83
O_2	-9.70
CO_2	-0.58
H_2SO_4	-0.003
H_3PO_4	-0.005
Fe^{2+}	-0.00037
H_2O	-11.86
Ethanol	-0.93
Urea	0.0025
H^+	-1.06
NO_2^-	0.0001
NO	1.81
NO_3^-	1.84

4.5 Additional Results

In this section, we will discuss additional results that were discovered during this project.

4.5.1 *N. vulgaris* Light Sensitivity

During later stages of the laboratory work, *N. vulgaris* was maintained in flasks that were not protected from light and demonstrated slight growth. This might indicate that *N. vulgaris* does not completely halt its growth in the presence of light or it is able to activate and deactivate metabolic function within a day cycle. This goes in accordance with Guerrero and colleagues, 1996 and Vanzella and colleagues, 1989, that report a slight sunlight-resistance of NOB bacteria [97], [126].

4.5.2 Compounds of Interest

Some of the compounds found within the model may contain a significant economic value. These metabolites, produced by *N. vulgaris,* could potentially be produced using ME techniques that over or under express genes in combination with large scale growth cultures. Table 11 shows some of these metabolites and their current commercial values (retrieved from Sigma-Aldrich®)[127].

Table 12 Possible products of interest of *N. vulgaris* and respective commercial price.

Name	*Formula*	*Price (€ g^{-1})*
Octadecanoic acid	$C_{18}H_{36}O_2$	23.75
Hexadecanoid acid	$C_{16}H_{30}O_2$	6.25
Methylmalic acid	$C_5H_8O_5$	9 450

4.5.3 Nitrate Commercial Use

Major fertilizers producers like ©Yara, ©Agrium and ©The Mosaic Company all use ammonium nitrate as a constituent of their products [128]–[130]. The prices of these fertilisers range from 0.78 to 1.30 € kg^{-1}.

Given that the current price of ethanol is estimated to be 0.85 € L^{-1} (data from Markets Insider [131]) and the price of ammonia is (approximately 0.26 € kg^{-1} (data from Market Realist [132]). An optimized NO_3^- production chain through the *N. europaea – N. vulgaris* community would be very advantageous if the production rate is high enough. The profits could be even higher if wastewater was the source of ammonia and the carbon source was CO_2 exclusively.

5. Conclusion

N. vulgaris is not intensively described in literature studies and its metabolic characterization is widely unknown. This model is focused, primarily, in its utility for NO_2^- oxidation into NO_3^-. This can be of interest for denitrification of contaminated soils, water bodies or wastewater in a more efficient way. Therefore, this work describes a novel insight on *N. vulgaris* metabolism and provides a better comprehension of *N. vulgaris* growth rates and kinetic behaviour. Specifically, *N. vulgaris* consumption rates for NO_2^-, CO_2, O_2 and ethanol, and production rates for NO_3^- were accurately determined *in vivo* experiments through the establishment of a steady-state culture. To our knowledge, this is the first work successfully describing the establishment of a *N. vulgaris* steady-state culture.

Moreover, this work unravelled details of the mixotrophic behaviour of *N. vulgaris* through the use of an organic carbon source ethanol. *N. vulgaris* shows a greater growth rate when CO_2 and ethanol are available. However, when only one source of carbon is available, *N. vulgaris* shows 5 % more biomass production when ethanol is present than when only CO_2 is present.

Finally, the use of ethanol in the culture medium was crucial to establish a steady-state culture, since *N. vulgaris* exhibited a doubling time approximately 1.6 fold shorter in the presence of this carbon source, when compared to carbon sources used in other studies.

The kinetic parameters of this NOB bacterium showed that approximately 80 % of the N consumed is released in the form of NO_3^- and NO, whereas 20 % of the N is used to produce biomass precursors, mainly proteins and nucleic acids.

In this work, a thorough genome annotation of *N. vulgaris* was also performed. This allowed the reconstruction of a *GSM* model that accurately represents the metabolic functions of *N. vulgaris*. Data from *in vivo* experiments and from *in silico* simulations was used to enhance the accuracy of the *GSM* model in mimicking the organism metabolism. The model can be used to optimize the production of desirable compounds or the consumption of waste products.

The results obtained in both processes were similar regarding the N species consumption and production rates; however, ethanol consumption rates differ in 0.0854 mmol gDW^{-1} h^{-1}, with a higher consumption rate detected *in vivo* than *in silico*. These are positive results, since simulation results should match *in vivo* measures. The ethanol value discrepancy might be related to the evaporation of ethanol within the reactor or measuring errors.

Simulations also implied that part of the N in the system can be released as a gas in the form of NO.

Finally, this work also demonstrates that a *N. europaea – N. vulgaris* community system can be achieved and can in theory consume all NH_3 and NO_2^- and produce NO_3^- with 48 % efficiency. This is a good result since this is the main objective of the work and indicates that the model and the bacteria system can be used as an efficient N reactive species removal system.

6. Future Work

A steady-state community of *N. europaea* and *N. vulgaris* should be established in order to validate the simulation results obtained *in silico*. This was not possible to obtain in current thesis due to the extreme slow growth of *N. europaea*, which did not allow the use of a strong inoculum required for this experiment. Pedro and co-workers took approximately two months to establish a steady-state culture of *N. europaea*.

Even though the model is functional and represents *N. vulgaris* accurately, improvements can still be preformed. First, model accuracy could be enhanced by analysing the essential pathways of the organism and rearranging their reaction in order to improve the *in vivo* results fitting.

A more extensive search for additional compounds with potential economic interest should be performed.

Finally, *N. vulgaris* should be cultured in culture media with different formulations (specifically with different carbon sources), and using different environmental conditions to understand the adaptability of this bacterium.

Annex

Table S1 Manual annotation of the 50 Sampler genes selected for Swiss-Prot, final scores attributed by *merlin* and respective confidence level.

GENE	NAME	FUNCTION	EC NUMBER	SCORE	CL
B2M20_00340	pheT	Phenylalanine--tRNA ligase beta subunit	6.1.1.20	0.93	A
B2M20_00495		Glutathione S-transferase	-	-	-
B2M20_00830	gpmA	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase	5.4.2.11	0.95	A
B2M20_00855	cmk	Cytidylate kinase	2.7.4.25	0.95	A
B2M20_01310		Threonine synthase	4.2.3.1	0.42	C
B2M20_01330	atpF	ATP synthase subunit b 1	-	-	A
B2M20_01680		RNA helicase	3.6.4.13	0.63	C
B2M20_02430		Arylsulfatase	-	-	-
B2M20_03320	cheB	Chemotaxis response regulator protein-glutamate methylesterase	3.1.1.61	0.95	A
B2M20_04485		Aspartate aminotransferase	2.6.1.1	0.44	E
B2M20_04605		UTP--glucose-1-phosphate uridylyltransferase	2.7.7.9	0.63	C
B2M20_05805		D-glycero-beta-D-manno-heptose-1,7-bisphosphate 7-phosphatase	3.1.3.82	0.42	C
B2M20_05995		Cytochrome D ubiquinol oxidase subunit I	-	-	-
B2M20_06495	glgE	Alpha-1,4-glucan:maltose-1-phosphate maltosyltransferase	2.4.99.16	0.81	C
B2M20_06735		Elongation factor Tu	-	-	-
B2M20_08025		Carbonic anhydrase	4.2.1.1	0.74	F
B2M20_10415		Glycosyl transferase family 2	-	-	-
B2M20_10660	trpC	Indole-3-glycerol phosphate synthase	4.1.1.48	0.95	A
B2M20_10700		Arylsulfatase	-	-	-
B2M20_10885		Phosphoglucosamine mutase	5.4.2.10	0.63	C
B2M20_10910		Glycosyl transferase	-	-	-
B2M20_10995		Putative pre-16S rRNA nuclease	3.1.-.-	0.95	A
B2M20_11035		SAM-dependent methyltransferase	-	-	-
B2M20_11515		Serine O-acetyltransferase	-	-	-
B2M20_11580		Error-prone DNA polymerase	2.7.7.7	0.9	C
B2M20_11810		XdhC/CoxI family protein	-	-	-
B2M20_12090		Threonine ammonia-lyase	-	-	-
B2M20_12295		Phosphatidylserine decarboxylase proenzyme	4.1.1.65	0.64	C
B2M20_12560		DNA helicase	3.6.4.12	0.55	C
B2M20_12690		Acetyltransferase	-	0.39	-
B2M20_12695		Ornithine monooxygenase	-	-	-
B2M20_12975		ATP-dependent 6-phosphofructokinase	2.7.1.11	0.74	F
B2M20_13025		8-amino-7-oxononanoate synthase	2.3.1.47	0.78	C
B2M20_13165		DNA protection during starvation protein	1.16.-.-	0.53	F
B2M20_13765		Undecaprenyl-phosphate 4-deoxy-4-formamido-L-arabinose transferase	2.4.2.53	0.56	C
B2M20_13980		Protease HtpX homolog	3.4.24.-	0.9	C
B2M20_14470		FMN reductase (NADH) RutF	1.5.1.42	0.67	C

CONTINUES ON NEXT PAGE

TABLE S1 CONTINUED

B2M20_15095		Fatty acid oxidation complex subunit alpha	-	-	-
B2M20_15450		GGDEF domain-containing protein	-	-	-
B2M20_15650		DNA topoisomerase I	-	-	-
B2M20_15810		Acyl-CoA thioester hydrolase YbgC	3.1.2.-	0.58	C
B2M20_15900		Fructose-bisphosphate aldolase	4.1.2.13	0.76	C
B2M20_16195		Superoxide dismutase [Cu-Zn]	1.15.1.1	0.78	C
B2M20_17500		Chemotaxis response regulator protein-glutamate methylesterase	3.1.1.61	0.72	C
B2M20_17615		Uncaracterized Protein	-	-	-
B2M20_18325		Glucoamylase	-	-	-
B2M20_18350		MFS transporter	-	-	-
B2M20_18370		4-cresol dehydrogenase	-	-	-
B2M20_18570		Uncaracterized Protein	-	-	-
BM20_18635		DUF2309 domain-containing protein	-	-	-

Table S2 Complete Swiss-Prot gene annotation, final scores attributed by *merlin* and respective confidence level.

GENE	**NAME**	**FUNCTION**	**EC NUMBER**	**SCORE**	**CL**
B2M20_00010		Metalloprotease TldD	3.4.-.-	0.58	C
B2M20_00065		Ribosomal-protein-alanine acetyltransferase	2.3.1.128	0.57	C
B2M20_00310		Protease 4	3.4.21.-	0.57	D
B2M20_00595		Membrane-bound lytic murein transglycosylase A	4.2.2.n1	0.56	C
B2M20_00775		Oxygen-independent coproporphyrinogen III oxidase	1.3.99.-	0.5	C
B2M20_00920		Lipopolysaccharide export system ATP-binding protein LptB	3.6.3.-	0.53	C
B2M20_01030		Ribosomal RNA small subunit methyltransferase B	2.1.1.176	0.56	D
B2M20_01050		Probable L,D-transpeptidase ErfK/SrfK	2.-.-.-	0.59	D
B2M20_01165		Uncharacterized protein			-
B2M20_01365		8-amino-7-oxononanoate synthase	2.3.1.37	0.58	C
B2M20_01525		3-ketoacyl-CoA thiolase	2.3.1.16	0.54	C
B2M20_01590		Glutathione import ATP-binding protein GsiA	3.6.3.-	0.5	D
B2M20_01915		PTS system mannose-specific EIIAB component	2.7.1.191	0.54	D
B2M20_01955		DNA ligase D	6.5.1.1	0.58	D
B2M20_01970	polA	DNA polymerase I	2.7.7.7	0.57	C
B2M20_01995		Phosphomannomutase	5.4.2.8, 5.4.2.2	0.18	C
B2M20_02125		CTP pyrophosphohydrolase	3.6.1.65	0.16	E
B2M20_02790		Dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase complex	2.3.1.61	0.57	C
B2M20_03025		Type III restriction-modification system EcoP15I enzyme mod	2.1.1.72	0.56	D
B2M20_03325		Chemotaxis protein methyltransferase	2.1.1.80	0.6	C
B2M20_03435		UDP-glucose 6-dehydrogenase	1.1.1.22	0.56	C

CONTINUES ON NEXT PAGE

TABLE S2 CONTINUED

B2M20_03490		Elongation factor 4	3.6.5.n1	0.57	C
B2M20_03600		Methylated-DNA--protein-cysteine methyltransferase	2.1.1.63	0.59	C
B2M20_03650		Prephenate dehydrogenase	1.3.1.12	0.5	D
B2M20_03665	metXA	Homoserine O-acetyltransferase	2.3.1.31	0.6	A
B2M20_03690		Sulfite reductase [NADPH] flavoprotein alpha-component	1.8.1.2	0.57	C
B2M20_03780		Riboflavin biosynthesis protein RibBA	4.1.99.12, 3.5.4.25	0.54	E
B2M20_03805		Quercetin 2,3-dioxygenase	1.13.11.24	0.5	C
B2M20_03825		Probable L,D-transpeptidase ErfK/SrfK	2.-.-.-	0.58	C
B2M20_03865		Uncharacterized protein			-
B2M20_03890		23S rRNA (uracil(1939)-C(5))-methyltransferase RlmD	2.1.1.190	0.58	C
B2M20_04025		Uncharacterized protein			-
B2M20_04080		Type 4 prepilin-like proteins leader peptide-processing enzyme	3.4.23.43, 2.1.1.-	0.58	C
B2M20_04175		Nitrate reductase subunit alpha	1.7.99.4	0.51	D
B2M20_04360		Uncharacterized protein			-
B2M20_04365		Nitronate monooxygenase	1.13.12.16	0.54	D
B2M20_04405		3-ketoacyl-CoA thiolase	2.3.1.16	0.59	C
B2M20_04410		Fatty acid oxidation complex subunit alpha	5.1.2.3, 4.2.1.17, 1.1.1.35	0.52	C
B2M20_04545		Phosphoserine aminotransferase	2.6.1.52	0.54	D
B2M20_04640		Lysophospholipase L2	3.1.1.5	0.52	C
B2M20_04645		Inositol-1-monophosphatase	3.1.3.15	0.12	C
B2M20_05210		Sensor protein KdpD	2.7.13.3	0.58	C
B2M20_05425		Oligoendopeptidase F	3.4.24.-	0.53	D
B2M20_05525		Succinyl-CoA--3-ketoacid-CoA transferase	2.8.3.5	0.51	C
B2M20_05685		Peptidoglycan D,D-transpeptidase FtsI	3.4.16.4	0.53	C
B2M20_05815		ADP-heptose--LPS heptosyltransferase 2	2.-.-.-	0.52	C
B2M20_05845		Mannose-1-phosphate guanylyltransferase/mannose-6-phosphate isomerase	2.7.7.13	0.51	C
B2M20_05880		Hydroxymethylpyrimidine/phosphomethylpyrimidine kinase	2.7.1.49, 2.7.4.7	0.51	C
B2M20_06430		Probable periplasmic serine endoprotease DegP-like	3.4.21.107	0.54	C
B2M20_06575	nadE	NH(3)-dependent NAD(+) synthetase	6.3.1.5	0.56	C
B2M20_06665		3-oxoacyl-[acyl-carrier-protein] synthase 2	2.3.1.179	0.56	C
B2M20_06735		Elongation factor Tu			-
B2M20_06735		Elongation factor Tu			-
B2M20_06985		Probable periplasmic serine endoprotease DegP-like	3.4.21.107	0.55	C
B2M20_07340		DNA polymerase III subunit delta	2.7.7.7	0.6	C
B2M20_07350		D-alanyl-D-alanine carboxypeptidase DacA	3.4.16.4	0.57	C
B2M20_07490		Cytochrome	1.14.-.-	0.5	D
B2M20_07765		Uncharacterized protein			-
B2M20_08005		Very short patch repair protein	3.1.-.-	0.57	C

CONTINUES ON NEXT PAGE

B2M20_08085		2-octaprenyl-6-methoxyphenyl hydroxylase	1.14.13.-	0.53	C
B2M20_08115	mfd	Transcription-repair-coupling factor	3.6.4.-	0.58	C
B2M20_08245		Undecaprenyl-phosphate 4-deoxy-4-formamido-L-arabinose transferase	2.4.2.53	0.56	C
B2M20_08255		Penicillin-binding protein 1C	2.4.1.129	0.54	C
B2M20_08405		Guanosine-5'-triphosphate,3'-diphosphate pyrophosphatase	3.6.1.40	0.54	C
B2M20_08430	purN	Phosphoribosylglycinamide formyltransferase	2.1.2.2	0.52	C
B2M20_08770		Bifunctional ligase/repressor BirA	6.3.4.15	0.54	C
B2M20_08820		NADH-quinone oxidoreductase subunit E	1.6.5.11	0.58	C
B2M20_09040		Uncharacterized protein			-
B2M20_09065		Soluble lytic murein transglycosylase	4.2.2.n1	0.54	C
B2M20_09400		Single-stranded-DNA-specific exonuclease RecJ	3.1.-.-	0.59	C
B2M20_09410		Homoserine dehydrogenase	1.1.1.3	0.6	C
B2M20_09685		D-alanyl-D-alanine carboxypeptidase DacA	3.4.16.4, 3.5.2.6	0.14	C
B2M20_09785		Penicillin-binding protein 1A	3.4.16.4, 2.4.1.129	0.42	C
B2M20_09790		N-acetylmuramoyl-L-alanine amidase	3.5.1.28	0.54	C
B2M20_09875		Peroxidase	1.11.1.15	0.6	D
B2M20_10295		Type I restriction enzyme EcoEI M protein	2.1.1.72	0.59	C
B2M20_10300		Type I restriction enzyme EcoKI R protein	3.1.21.3	0.52	C
B2M20_10455		Cysteine synthase	2.5.1.47	0.52	C
B2M20_10570		Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex	2.3.1.12	0.58	C
B2M20_10950		NAD(P)H-quinone oxidoreductase subunit 5, chloroplastic	1.6.5.-	0.53	C
B2M20_11145		Putative NAD(P)H nitroreductase YdjA	1.-.-.-	0.54	C
B2M20_11385		Guanosine-5'-triphosphate,3'-diphosphate pyrophosphatase	3.6.1.40	0.56	C
B2M20_11395		Uncharacterized protein			-
B2M20_11425		Uncharacterized protein			-
B2M20_11475		Argininosuccinate lyase	4.3.2.2	0.47	C
B2M20_11785		Nod factor export ATP-binding protein I	3.6.3.-	0.52	D
B2M20_11980		Dihydropteroate synthase	2.5.1.15	0.6	C
B2M20_11985		Dihydroneopterin aldolase	4.1.2.25	0.51	C
B2M20_11990		2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase	2.7.6.3	0.5	C
B2M20_11995		Methylmalonyl-CoA mutase	5.4.99.2	0.58	C
B2M20_12005		Methylmalonyl-CoA mutase	5.4.99.2	0.53	C
B2M20_12055		Uncharacterized protein			-
B2M20_12145		GTP 3',8-cyclase	4.1.99.22	0.51	E
B2M20_12190		Uncharacterized protein			-
B2M20_12315		GDP-mannose-dependent alpha-mannosyltransferase	2.4.1.-	0.51	E
B2M20_12445		Sulfate/thiosulfate import ATP-binding protein CysA	3.6.3.25	0.56	E
B2M20_12465		Acetolactate synthase small subunit	2.2.1.6	0.54	C

CONTINUES ON NEXT PAGE

TABLE S2 CONTINUED

B2M20_12470		Acetolactate synthase	2.2.1.6	0.5	C
B2M20_12485		Probable periplasmic serine endoprotease DegP-like	3.4.21.107	0.55	C
B2M20_12560		DNA helicase	3.6.4.12	0.55	C
B2M20_12565		Uncharacterized protein			-
B2M20_12640		UDP-glucose 6-dehydrogenase	1.1.1.22	0.56	C
B2M20_12685		Lipid A export ATP-binding/permease protein MsbA	3.6.3.-	0.58	C
B2M20_12855		UDP-N-acetyl-D-glucosamine dehydrogenase	1.1.1.336	0.24	C
B2M20_12900		Prophage bactoprenol glucosyl transferase homolog	2.4.2.53	0.56	D
B2M20_13045		L-aspartate oxidase	1.4.3.16	0.58	C
B2M20_13165		DNA protection during starvation protein	1.16.-.-	0.53	E
B2M20_13305		Protease 2	3.4.21.83	0.16	E
B2M20_13365		L,D-transpeptidase	2.-.-.-	0.58	C
B2M20_13375		3-mercaptopyruvate sulfurtransferase	2.8.1.2	0.23	C
B2M20_13500		NAD/NADP-dependent betaine aldehyde dehydrogenase	1.2.1.3	0.06	D
B2M20_13510		Osmolarity sensor protein EnvZ	2.7.13.3	0.53	C
B2M20_13655		Uncharacterized protein			-
B2M20_13725		Metalloprotease PmbA homolog	3.4.-.-	0.58	C
B2M20_13765		Undecaprenyl-phosphate 4-deoxy-4-formamido-L-arabinose transferase	2.4.2.53	0.56	C
B2M20_13870		Probable protease SohB	3.4.21.-	0.56	C
B2M20_13945		Adenine DNA glycosylase	3.2.2.-	0.59	C
B2M20_14030		6-carboxy-5,6,7,8-tetrahydropterin synthase	4.1.2.50	0.54	C
B2M20_14040		CCA-adding enzyme	2.7.7.72	0.55	E
B2M20_14125	tal	Transaldolase	5.3.1.9	0.4	E
B2M20_14130		6-phosphogluconate dehydrogenase, decarboxylating	1.1.1.44	0.6	C
B2M20_14150		Sugar phosphatase YidA	3.1.3.104	0.09	C
B2M20_14205		NADH-quinone oxidoreductase subunit N	1.6.5.11	0.57	D
B2M20_14285		Sensor protein QseC	2.7.13.3	0.51	C
B2M20_14330		Anthranilate synthase component 1	2.6.1.85	0.22	C
B2M20_14495		Uncharacterized protein			-
B2M20_14510		Arabinose 5-phosphate isomerase KdsD	5.3.1.13	0.5	C
B2M20_14565		Homospermidine synthase	2.5.1.44	0.58	C
B2M20_14945		Nitrate reductase-like protein NarX	1.7.99.4	0.57	C
B2M20_14965		Nitrate reductase subunit alpha	1.7.99.4	0.51	C
B2M20_15060	lysA	Diaminopimelate decarboxylase	4.1.1.20	0.59	C
B2M20_15110		ATP:cob(I)alamin adenosyltransferase	2.5.1.17	0.5	E
B2M20_15720		Uncharacterized protein			-
B2M20_15725		Peroxiredoxin	1.11.1.15	0.53	E
B2M20_15810		Acyl-CoA thioester hydrolase YbgC	3.1.2.-	0.58	C
B2M20_15950		Lipid A export ATP-binding/permease protein MsbA	3.6.3.-	0.55	C
B2M20_16045	cysC	Sulfate adenylyltransferase subunit 1	2.7.1.25, 2.7.7.4	0.34	C
B2M20_16225		L-aspartate oxidase	1.4.3.16	0.5	C
B2M20_16315	dusA	tRNA-dihydrouridine(20/20a) synthase	1.3.1.-	0.59	C
B2M20_17060		Uncharacterized protein			-

CONTINUES ON NEXT PAGE

TABLE S2 CONTINUED

B2M20_17120	ndvA	Beta-(1-->2)glucan export ATP-binding/permease protein NdvA	3.6.3.42	0.57	A
B2M20_17125		Peptidase M15	3.4.16.4	0.54	C
B2M20_17550		Glycogen debranching enzyme	3.2.1.-	0.12	C
B2M20_17615		Uncharacterized protein			-
B2M20_17765		Alpha-1,4 glucan phosphorylase	2.4.1.1	0.6	E
B2M20_17895		Nod factor export ATP-binding protein I	3.6.3.-	0.5	C
B2M20_18170		Toxin	3.1.-.-	0.52	C
B2M20_18375		NAD/NADP-dependent betaine aldehyde dehydrogenase	1.2.1.8	0.6	D
B2M20_18510		Sensor protein VraS	2.7.13.3	0.58	E

Table S3 Complete TrEMBL gene annotation, final scores attributed by *merlin* and respective confidence level.

GENE	**FUNCTION**	**EC NUMBER**	**SCORE**	**CL**
B2M20_00320	Alkyl hydroperoxide reductase		0.77	H
B2M20_00600	DNA mismatch repair protein MutS		0.69	H
B2M20_00910	DNA-directed RNA polymerase subunit N		0.5	H
B2M20_00990	Glutathione S-transferase		0.86	F
B2M20_01120	GDSL family lipase		0.69	H
B2M20_01155	Metal-dependent hydrolase		0.69	H
B2M20_01220	Phosphoesterase		0.72	H
B2M20_01415	Uncharacterized protein		0.74	-
B2M20_01450	RecA-family ATPase		0.54	G
B2M20_01465	Uncharacterized protein		0.52	-
B2M20_01470	Uncharacterized protein		0.53	-
B2M20_01920	Serine kinase		0.69	H
B2M20_02645	Glucokinase		0.68	H
B2M20_02900	Uncharacterized protein			F
B2M20_02985	Uncharacterized protein			-
B2M20_03145	Methylase involved in ubiquinone/menaquinone biosynthesis		0.43	F
B2M20_03150	Alpha/beta hydrolase		0.7	-
B2M20_03465	Glycosyl transferase, family 2		0.47	F
B2M20_03480	Glycosyl transferase family 1		0.64	F
B2M20_03525	Uncharacterized protein		0.82	-
B2M20_03660	Chorismate mutase	5.4.99.5	0.57	G
B2M20_03940	Uncharacterized protein			-
B2M20_04035	Thioredoxin family protein		0.52	H
B2M20_04150	Sel1-like protein		0.3	G
B2M20_04315	Serine/threonine protein phosphatase		0.66	F
B2M20_04555	Peptidase S10, serine carboxypeptidase		0.46	-

CONTINUES ON NEXT PAGE

B2M20_04655	N-formylglutamate amidohydrolase	0.86	F
B2M20_04990	Carboxysome shell carbonic anhydrase	0.84	-
B2M20_05045	Uncharacterized protein	0.4	-
B2M20_05155	Methyltransferase type 11	0.57	G
B2M20_05160	Pseudaminic acid biosynthesis-associated methylase	0.81	H
B2M20_05180	Glyoxalase	0.61	J
B2M20_05255	Uncharacterized protein	0.51	-
B2M20_05270	Uncharacterized protein	0.68	-
B2M20_05885	UDP-2,3-diacylglucosamine hydrolase	0.59	H
B2M20_06355	Peptidase P60	0.68	-
B2M20_07365	Uncharacterized protein		-
B2M20_08105	AMP-dependent synthetase	0.69	F
B2M20_08240	Lipid A biosynthesis	0.42	H
B2M20_08315	ROK family protein	0.7	-
B2M20_08575	RecA-family ATPase	0.55	G
B2M20_09235	Carboxysome shell carbonic anhydrase	0.82	-
B2M20_09360	Phosphoesterase, PA-phosphatase related protein	0.38	F
B2M20_09835	Serine hydroxymethyltransferase	0.87	H
B2M20_10035	Serine/threonine protein kinase	0.86	F
B2M20_10320	Uncharacterized protein	0.88	-
B2M20_10970	Patatin	0.73	F
B2M20_11195	Peptidase	0.67	H
B2M20_11575	Uncharacterized protein	0.88	-
B2M20_11675	Uncharacterized protein	0.61	-
B2M20_12150	Phosphorylase	0.51	H
B2M20_12830	dTDP-6-deoxy-L-hexose 3-O-methyltransferase	0.65	J
B2M20_12860	Uncharacterized protein	0.58	-
B2M20_14335	Putative glycosyl transferase	0.3	H
B2M20_14340	Glycosyltransferase	0.49	H
B2M20_14345	Glycosyl transferase	0.48	H
B2M20_14430	Uncharacterized protein	0.88	-
B2M20_14550	GCN5-related N-acetyltransferase	0.48	G
B2M20_14950	Nitrate reductase molybdenum cofactor assembly chaperone	0.57	H
B2M20_14980	Glycosyl transferase	0.51	H
B2M20_15210	Uncharacterized protein	0.6	-
B2M20_15230	Uncharacterized protein	0.89	-
B2M20_15245	Uncharacterized protein	0.38	-
B2M20_15335	Uncharacterized protein	0.92	-
B2M20_15385	Amino acid/amide ABC transporter substrate-binding protein, HAAT family	0.46	I
B2M20_15490	Extradiol dioxygenase	0.5	I
B2M20_15575	Uncharacterized protein	0.49	I
B2M20_15845	DNA mismatch repair protein MutT	0.64	H

CONTINUES ON NEXT PAGE

TABLE S3 CONTINUED

B2M20_15945	GNAT family N-acetyltransferase	0.53	-
B2M20_16900	Phenylacetic acid catabolic	0.61	H
B2M20_17290	Uncharacterized protein	0.68	-
B2M20_17345	Uncharacterized protein	0.89	-
B2M20_17405	Uncharacterized protein	0.47	-
B2M20_17630	Uncharacterized protein	0.33	-
B2M20_17735	Transposase	0.52	I
B2M20_17760	Glycogen debranching enzyme GlgX	0.37	I
B2M20_17920	Helix-turn-helix domain-containing protein, Fis-type	0.27	G
B2M20_18295	Uncharacterized protein	0.75	-
B2M20_18395	LysR family transcriptional regulator	0.6	I
B2M20_18405	Uncharacterized protein	0.42	-

Table S4 Manually inserted reactions.

Reaction (KEGG id)	***Reason for insertion***
R00134	Formate degradation; close carbon fixation pathway
R03314	Non-enzymatic; Proline biosynthesis
R07621	Thiamine Metabolism; Dead-end removal - not essential for the model
R02324	Dead-end removal - not essential for the model
R09977	Thiamine Metabolism; Dead-end removal - not essential for the model
R04558	Histidine Pathway closure; Imidazole-glycerol-3P production
R04638	Folate biosynthesis;
R05311	Benzoate degradation; ; Dead-end removal - not essential for the model

The current versions of the *GSM* models are available in the following links:

https://www.dropbox.com/sh/lofwvlhxxne0kn0/AADrmN34Di76gknGgOs4GScna?dl=0

or

https://nextcloud.bio.di.uminho.pt/s/a95KfKgEm8Bpdf3

References

[1] J. P. Ehrlich, Paul R; Holdren, “Impact of Population Growth,” *Science (80-.).*, vol. 171, no. 1977, pp. 1212–1217, 1971.

[2] J. N. Galloway *et al.*, “Nitrogen cycles: Past, present, and future,” *Biogeochemistry*, vol. 70, no. 2, pp. 153–226, 2004.

[3] D. Cordell, J. O. Drangert, and S. White, “The story of phosphorus: Global food security and food for thought,” *Glob. Environ. Chang.*, vol. 19, no. 2, pp. 292–305, 2009.

[4] WHO, “Guidelines for Drinking-Water Quality - Second Edition - Volume 2 - Health Criteria and Other Supporting Information,” *Who 1996*, vol. 2, no. 1152404, p. 15, 1996.

[5] M. Singh, A. K. Tripathi, K. S. Reddy, and K. N. Singh, “Soil phosphorus dynamics in a Vertisol as affected by cattle manure and nitrogen fertilization in soybean-wheat system,” *J. Plant Nutr. Soil Sci.*, vol. 164, no. 6, pp. 691–696, 2001.

[6] M. Geissdoerfer, P. Savaget, N. M. P. Bocken, and E. J. Hultink, “The Circular Economy – A new sustainability paradigm?,” *Journal of Cleaner Production*, vol. 143. pp. 757–768, 2017.

[7] J. M. Modak, “Haber process for ammonia synthesis,” *Resonance*, vol. 16, no. 12, pp. 1159–1167, 2011.

[8] J. A. Hargreaves, “Nitrogen biogeochemistry of aquaculture ponds,” *Aquaculture*, vol. 166, no. 3–4. pp. 181–212, 1998.

[9] R. Kulkarni, “Metabolic engineering: Biological art of producing useful chemicals,” *Resonance*, vol. 21, no. 3, pp. 233–237, 2016.

[10] S. Saeidnia, A. Manayi, and M. Abdollahi, “From in vitro Experiments to in vivo and Clinical Studies; Pros and Cons.,” *Curr. Drug Discov. Technol.*, 2015.

[11] R. Mahadevan and C. H. Schilling, “The effects of alternate optimal solutions in constraint-based genome-scale metabolic models,” *Metab. Eng.*, vol. 5, no. 4, pp. 264–276, 2003.

[12] J. I. Prosser, *Nitrification*. Academic Press, 1986.

[13] T. Yamanaka, *Chemolithoautotrophic Bacteria*. 2008.

[14] S. Bagchi, R. Biswas, and T. Nandy, "Autotrophic ammonia removal processes: Ecology to technology," *Critical Reviews in Environmental Science and Technology*, vol. 42, no. 13. pp. 1353–1418, 2012.

[15] O. Dias, M. Rocha, E. C. Ferreira, and I. Rocha, "Reconstructing genome-scale metabolic models with merlin," *Nucleic Acids Res.*, vol. 43, no. 8, pp. 3899–3910, 2015.

[16] C. Francke, R. J. Siezen, and B. Teusink, "Reconstructing the metabolic network of a bacterium from its genome," *Trends in Microbiology*, vol. 13, no. 11. pp. 550–558, 2005.

[17] P. Raposo, *"Reconstruction of the genome-scale metabolic model of Nitrosomonas europaea,"* no. September. 2017.

[18] N. Gruber and J. N. Galloway, "An Earth-system perspective of the global nitrogen cycle," *Nature*, vol. 451, no. 7176. pp. 293–296, 2008.

[19] J. N. GALLOWAY *et al.*, "The Nitrogen Cascade," *Bioscience*, vol. 53, no. 4, p. 341, 2003.

[20] J. a Camargo and A. Alonso, "Ecological and toxicological effects of inorganic nitrogen pollution in aquatic ecosystems: A global assessment.," *Environment international*, vol. 32, no. 6. pp. 831–849, 2006.

[21] D. J. Conley *et al.*, "Ecology - Controlling eutrophication: Nitrogen and phosphorus," *Science*, vol. 323, no. 5917. pp. 1014–1015, 2009.

[22] J. H. Ryther and W. M. Dunstan, "Nitrogen, Phosphorus, and Eutrophication in the Coastal Marine Environment," *Science (80-.).*, vol. 171, no. 3975, pp. 1008–1013, 1971.

[23] H. G. Gorchev and G. Ozolins, *WHO guidelines for drinking-water quality.*, vol. 38, no. 3. 2011.

[24] W. K. Dodds *et al.*, "Eutrophication of U.S. Freshwaters: Analysis of Potential Economic Damages," *Environ. Sci. Technol.*, vol. 43, no. 1, pp. 12–19, 2009.

[25] B. Moss, "Allied attack: climate change and eutrophication," *Inl. Waters*, vol. 1, no. 2, pp. 101–105, 2011.

[26] Y. Tal, J. E. M. Watts, and H. J. Schreier, "Anaerobic ammonium-oxidizing (Anammox) bacteria and associated activity in fixed-film biofilters of a marine recirculating aquaculture system," *Appl. Environ. Microbiol.*, 2006.

[27] R. R. Stickney, *Aquaculture: An Introductory Text. By Robert R Stickney.* 2005.

[28] A. G. Hall, "A Comparative Analysis of Three Biofilter Types Treating Wastewater Produced in Recirculating Aquaculture Systems," *Thesis Submitt. to Fac. Virginia Polytech. Inst. State Univ. Partial fulfillment Requir. degree MASTER Sci.*, 1999.

[29] IFA, "Feeding the Earth: Fertilizers and Global Food Security, Market Drivers and Fertilizer Economics.," *Int. Fertil. Ind. Assoc.*, 2008.

[30] I. Rafiqul, C. Weber, B. Lehmann, and A. Voss, "Energy efficiency improvements in ammonia production - Perspectives and uncertainties," *Energy*, vol. 30, no. 13. pp. 2487–2504, 2005.

[31] M. Oertel, J. Schmitz, W. Weirich, D. Jendryssek???Neumann, and R. Schulten, "Steam reforming of natural gas with intergrated hydrogen separation for hydrogen production," *Chem. Eng. Technol.*, vol. 10, no. 1, pp. 248–255, 1987.

[32] J. R. Postgate, "Nitrogen Fixation," *Nature*, vol. 766, pp. 165–176, 1998.

[33] C. A. Lembi, "Limnology, Lake and River Ecosystems," *J. Phycol.*, vol. 37, no. 6, pp. 1146–1147, 2001.

[34] N. N. Rabalais, "Nitrogen in Aquatic Ecosystems," *AMBIO A J. Hum. Environ.*, vol. 31, no. 2, pp. 102–112, 2002.

[35] K. Hasler, S. Bröring, S. W. F. Omta, and H. W. Olfs, "Life cycle assessment (LCA) of different fertilizer product types," *Eur. J. Agron.*, vol. 69, pp. 41–51, 2015.

[36] P. Marschner, E. Kandeler, and B. Marschner, "Structure and function of the soil

microbial community in a long-term fertilizer experiment,” *Soil Biol. Biochem.*, 2003.

[37] Y. Zhang, N. Love, and M. Edwards, “Nitrification in drinking water systems,” *Critical Reviews in Environmental Science and Technology*, vol. 39, no. 3. pp. 153–208, 2009.

[38] G. Klein, M. Krebs, V. Hall, T. O’Brien, and B. B. Blevins, “California’s Water – Energy Relationship,” 2005.

[39] H. W. Paerl and J. Huisman, “Blooms like it hot,” *Science*, vol. 320, no. 5872, pp. 57–58, 2008.

[40] H. . Painter, “A review of literature on inorganic nitrogen metabolism in microorganisms,” *Water Res.*, vol. 4, no. 6, pp. 393–450, 1970.

[41] R. H. Wijffels and J. Tramper, “Performance of growing Nitrosomonas europaea cells immobilized in κ-carrageenan,” *Appl. Microbiol. Biotechnol.*, vol. 32, no. 1, pp. 108–112, 1989.

[42] H. Daims, J. L. Nielsen, P. H. Nielsen, K. H. Schleifer, and M. Wagner, “In Situ Characterization of Nitrospira-Like Nitrite-Oxidizing Bacteria Active in Wastewater Treatment Plants,” *Appl. Environ. Microbiol.*, vol. 67, no. 3–12, pp. 5273–5284, 2001.

[43] P. Chain *et al.*, “Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph Nitrosomonas europaea,” *J. Bacteriol.*, vol. 185, no. 9, pp. 2759–2773, 2003.

[44] C. Clark and E. L. Schmidt, “Effect of mixed culture on Nitrosomonas europaea simulated by uptake and utilization of pyruvate.,” *J. Bacteriol.*, vol. 91, no. 1, pp. 367–373, 1966.

[45] H. Baribeau, “Microbiology and Isolation of Nitrifying Bacteria,” in *Fundamentals and Control of Nitrification in Chloraminated Drinking Water Distribution Systems*, vol. 6, 2006, p. 270.

[46] E. Bock, H. P. Koops, U. C. Möller, and M. Rudert, “A new facultatively nitrite

oxidizing bacterium, Nitrobacter vulgaris sp. nov.," *Arch. Microbiol.*, vol. 153, no. 2, pp. 105–110, 1990.

[47] R. J. Maier, "Rhizobium japonicum mutant strains unable to grow chemoautotrophically with H2," *J. Bacteriol.*, vol. 145, no. 1, pp. 533–590, 1981.

[48] E. Rosenberg, *The prokaryotes: Alphaproteobacteria and betaproteobacteria*. 2013.

[49] W. Steinmuller and E. Bock, "Growth of Nitrobacter in the presence of organic matter.," *Arch. Microbiol.*, vol. 108, pp. 299–304, 1976.

[50] A. J. Smith and D. S. Hoare, "Acetate assimilation by Nitrobacter agilis in relation to its 'obligate autotrophy'.," *J. Bacteriol.*, vol. 95, no. 3, pp. 844–855, 1968.

[51] C. Grunditz and G. Dalhammar, "Development of nitrification inhibition assays using pure cultures of Nitrosomonas and Nitrobacter," *Water Res.*, vol. 35, no. 2, pp. 433–440, 2001.

[52] H. LAUDELOUT and L. VAN TICHELEN, "Kinetics of the nitrite oxidation by Nitrobacter winogradskyi.," *J. Bacteriol.*, vol. 79, pp. 39–42, 1960.

[53] A. M. BUSWELL, T. SHIOTA, N. LAWRENCE, and I. VAN METER, "Laboratory studies on the kinetics of the growth of Nitrosomonas with relation to the nitrification phase of the B.O.D. test.," *Appl. Microbiol.*, vol. 2, no. 1, pp. 21–5, 1954.

[54] T. Hofman and H. Lees, "The biochemistry of the nitrifying organisms. III. Composition of Nitrosomonas.," *Biochem. J.*, vol. 54, pp. 293–295, 1953.

[55] US and EPA, "Process Design Manual for Nitrogen Control," *U.S. Environ. Prot. Agency*, p. 476, 1975.

[56] Y. H. Ahn, "Sustainable nitrogen elimination biotechnologies: A review," *Process Biochemistry*, vol. 41, no. 8. pp. 1709–1721, 2006.

[57] S. Y. Lee, G. N. Bennett, and E. T. Papoutsakis, "Construction of Escherichia coli-Clostridium acetobutylicum shuttle vectors and transformation of Clostridium acetobutylicum strains," *Biotechnol. Lett.*, vol. 14, no. 5, pp. 427–432, 1992.

[58] C. Smolke, *The Metabolic Pathway Engineering Handbook: Fundamentals*. 2009.

[59] H. Kitano, “Computational systems biology,” *Nature*, vol. 420, no. 6912, pp. 206–210, 2002.

[60] R. Fleischmann *et al.*, “Whole-genome random sequencing and assembly of Haemophilus influenzae Rd,” *Science (80-.).*, vol. 269, no. 5223, pp. 496–512, 1995.

[61] B. Palsson, “Metabolic systems biology,” *FEBS Letters*, vol. 583, no. 24. pp. 3900–3904, 2009.

[62] J. Förster, I. Famili, P. Fu, B. Palsson, and J. Nielsen, “Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network,” *Genome Res.*, vol. 13, no. 2, pp. 244–253, 2003.

[63] I. Rocha, J. Förster, and J. Nielsen, “Design and Application of Genome-Scale Reconstructed Metabolic Models,” *Methods Mol. Biol. vol. 416 Microb. Gene Essentiality*, vol. 416, pp. 409–431, 2007.

[64] O. Dias and I. Rocha, “Systems Biology in Fungi,” in *Molecular Biology of Food and Water Borne Mycotoxigenic and Mycotic Fungi*, 2015, pp. 69–92.

[65] I. Thiele *et al.*, “A protocol for generating a high-quality genome-scale metabolic reconstruction,” *Nat. Protoc.*, vol. 5, no. 1, pp. 93–121, 2010.

[66] R. Caspi *et al.*, “The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D471–D480, 2016.

[67] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, “KEGG: Kyoto encyclopedia of genes and genomes,” *Nucleic Acids Research*, vol. 27, no. 1. pp. 29–34, 1999.

[68] The UniProt Consortium, “UniProt: a hub for protein information.,” *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D204-12, 2015.

[69] S. Placzek *et al.*, “BRENDA in 2017: New perspectives and new tools in BRENDA,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D380–D388, 2017.

[70] J. Ostell and J. McEntyre, “The NCBI Handbook,” *NCBI Bookshelf*, pp. 1–8, 2007.

[71] J. D. Orth, I. Thiele, and B. Ø. Palsson, “What is flux balance analysis?,” *Nat. Biotechnol.*, vol. 28, no. 3, pp. 245–248, 2010.

[72] S. Gudmundsson and I. Thiele, “Computationally efficient flux variability analysis,” *BMC Bioinformatics*, vol. 11, 2010.

[73] K. Schügerl, “Development of bioreaction engineering.,” *Adv. Biochem. Eng. Biotechnol.*, vol. 70, pp. 41–76, 2000.

[74] B. Christensen and J. Nielsen, “Metabolic network analysis. A powerful tool in metabolic engineering.,” *Adv. Biochem. Eng. Biotechnol.*, vol. 66, pp. 209–31, 2000.

[75] I. Rocha *et al.*, “OptFlux: an open-source software platform for in silico metabolic engineering.,” *BMC Syst. Biol.*, vol. 4, no. 1, p. 45, 2010.

[76] D. Segre, D. Vitkup, and G. M. Church, “Analysis of optimality in natural and perturbed metabolic networks,” *Proc. Natl. Acad. Sci.*, vol. 99, no. 23, pp. 15112–15117, 2002.

[77] T. Shlomi, O. Berkman, and E. Ruppin, “Regulatory on‾off minimization of metabolic flux,” *Pnas*, vol. 102, no. 21, pp. 7695–7700, 2005.

[78] E. Pitkänen *et al.*, “Comparative Genome-Scale Reconstruction of Gapless Metabolic Networks for Present and Ancestral Species,” *PLoS Comput. Biol.*, vol. 10, no. 2, 2014.

[79] S. Pabinger, R. Rader, R. Agren, J. Nielsen, and Z. Trajanoski, “MEMOSys: Bioinformatics platform for genome-scale metabolic models,” *BMC Syst. Biol.*, vol. 5, 2011.

[80] J. Boele, B. G. Olivier, and B. Teusink, “FAME, the Flux Analysis and Modeling Environment.,” *BMC Syst. Biol.*, vol. 6, p. 8, 2012.

[81] R. Lerdorf, K. Tatroe, and P. MacIntyre, *Programming PHP*, vol. 37. 2006.

[82] B. G. Olivier, J. M. Rohwer, and J. H. S. Hofmeyr, “Modelling cellular systems with

PySCeS,” *Bioinformatics*, vol. 21, no. 4, pp. 560–561, 2005.

[83] P. D. Karp *et al.*, “Pathway tools version 19.0 update: Software for pathway/genome informatics and systems biology,” *Brief. Bioinform.*, vol. 17, no. 5, pp. 877–890, 2016.

[84] N. Swainston, K. Smallbone, P. Mendes, D. Kell, and N. Paton, “The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks.,” *J. Integr. Bioinform.*, vol. 8, no. 2, p. 186, 2011.

[85] C. S. Henry, M. Dejongh, A. A. Best, P. M. Frybarger, B. Linsay, and R. L. Stevens, “High-throughput generation, optimization and analysis of genome-scale metabolic models,” *Nat. Biotechnol.*, vol. 28, no. 9, pp. 977–982, 2010.

[86] J. N. Edirisinghe, J. P. Faria, N. L. Harris, B. H. Allen, and C. S. Henry, “Reconstruction and Analysis of Central Metabolism in Microbes,” *Humana Press. New York, NY*, vol. 1716, pp. 111–129, 2018.

[87] R. W. Cottingham, “The DOE systems biology knowledgebase (KBase),” in *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics - BCB ’15*, 2015, pp. 510–510.

[88] R. Agren, L. Liu, S. Shoaie, W. Vongsangnak, I. Nookaew, and J. Nielsen, “The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for Penicillium chrysogenum,” *PLoS Comput. Biol.*, vol. 9, no. 3, 2013.

[89] Y. C. Liao, M. H. Tsai, F. C. Chen, and C. A. Hsiung, “GEMSiRV: A software platform for GEnome-scale metabolic model simulation, reconstruction and visualization,” *Bioinformatics*, vol. 28, no. 13, pp. 1752–1758, 2012.

[90] X. Feng, Y. Xu, Y. Chen, and Y. J. Tang, “MicrobesFlux: a web platform for drafting metabolic models from the KEGG database,” *BMC Syst. Biol.*, vol. 6, 2012.

[91] A. Holovaty and J. Kaplan-Moss, *The Definitive Guide to Django: Web Development Done Right*. 2009.

[92] S. Klamt, J. Saez-Rodriguez, and E. D. Gilles, “Structural and functional analysis of cellular networks with CellNetAnalyzer,” *BMC Syst. Biol.*, vol. 1, 2007.

[93] S. G. Thorleifsson and I. Thiele, “rBioNet: A COBRA toolbox extension for reconstructing high-quality biochemical networks,” *Bioinformatics*, vol. 27, no. 14, pp. 2009–2010, 2011.

[94] L. S. Jing *et al.*, “Database and tools for metabolic network analysis,” *Biotechnol. Bioprocess Eng.*, vol. 19, no. 4, pp. 568–585, 2014.

[95] O. Dias, R. Pereira, A. K. Gombert, E. C. Ferreira, and I. Rocha, “iOD907, the first genome-scale metabolic model for the milk yeast Kluyveromyces lactis,” *Biotechnol. J.*, 2014.

[96] B. L. Mellbye, E. W. Davis, E. Spieck, J. H. Chang, P. J. Bottomley, and L. A. Sayavedra-Soto, “Draft Genome Sequence of Nitrobacter vulgaris Strain Ab 1 , a Nitrite-Oxidizing Bacterium,” *Genome Announc.*, 2017.

[97] a Vanzella, M. Guerrero, and R. Jones, “Effect of CO and light on ammonium and nitrite oxidation by chemolithotrophic bacteria ,” *Mar. Ecol. Prog. Ser.*, vol. 57, no. 1971, pp. 69–76, 1989.

[98] A. I. Vogel and G. Svehla, “Vogel’s Textbook of Macro and semimicro qualitative inorganic analysis.,” in *Vogel’s Textbook of Macro and semimicro qualitative inorganic analysis.*, 1979.

[99] R. Chang, *Chemistry 10th ed*. 2010.

[100] D. B. Jaynes and A. S. Rogowski, “Applicability of Fick’s law to gas diffusion,” *Soil Sci. Soc. Am. J.*, 1983.

[101] D. J. Stokes, *Principles and Practice of Variable Pressure/Environmental Scanning Electron Microscopy (VP-ESEM)*. 2008.

[102] J. Goldstein *et al.*, *Scanning Electron Microscopy and X-ray Microanalysis*. 2003.

[103] S. Sutton, “Measurement of microbial cells by optical density,” *J. Valid. Technol.*, 2011.

[104] A. M. Feist *et al.*, “A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information,” *Mol. Syst. Biol.*, 2007.

[105] A. Lipski, E. Spieck, A. Makolla, and K. Altendorf, “Fatty acid profiles of nitrite-oxidizing bacteria reflect their phylogenetic heterogeneity,” *Syst. Appl. Microbiol.*, 2001.

[106] T. B. Auran and E. L. Schmidt, “Lipids of Nitrobacter and effects of cultural conditions on fatty acid composition,” *Biochim. Biophys. Acta (BBA)/Lipids Lipid Metab.*, 1976.

[107] C. Verduyn, E. Postma, W. A. Scheffers, and J. P. van Dijken, “Physiology of Saccharomyces Cerevisiae in Anaerobic Glucose-Limited Chemostat Culturesx,” *J. Gen. Microbiol.*, 1990.

[108] D. Herbert, P. J. Phipps, and R. E. Strange, “Chemical Analysis of Microbial Cells,” *Methods Microbiol.*, 1971.

[109] M. De Mey, G. Lequeux, J. Maertens, S. De Maeseneire, W. Soetaert, and E. Vandamme, “Comparison of DNA and RNA quantification methods suitable for parameter estimation in metabolic modeling of microorganisms,” *Anal. Biochem.*, 2006.

[110] S. Benthin, J. Nielsen, and J. Villadsen, “A simple and reliable method for the determination of cellular RNA content,” *Biotechnol. Tech.*, 1991.

[111] B. Vanparys *et al.*, “The phylogeny of the genus Nitrobacter based on comparative rep-PCR, 16S rRNA and nitrite oxidoreductase gene sequence analysis,” *Syst. Appl. Microbiol.*, 2007.

[112] J. Caliz, M. Montes-Borrego, X. Triadó-Margarit, M. Metsis, B. B. Landa, and E. O. Casamayor, “Influence of edaphic, climatic, and agronomic factors on the composition and abundance of nitrifying microorganisms in the rhizosphere of commercial olive crops,” *PLoS One*, 2015.

[113] J. Skolnick and J. S. Fetrow, “From genes to protein structure and function: Novel applications of computational approaches in the genomic era,” *Trends in Biotechnology*. 2000.

[114] M. H. Saier, “A Functional-Phylogenetic Classification System for Transmembrane

Solute Transporters," *Microbiol. Mol. Biol. Rev.*, 2000.

[115] P. Horton *et al.*, "WoLF PSORT: Protein localization predictor," *Nucleic Acids Res.*, vol. 35, no. SUPPL.2, 2007.

[116] N. Y. Yu *et al.*, "PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes," *Bioinformatics*, 2010.

[117] A. Flamholz, E. Noor, A. Bar-Even, and R. Milo, "EQuilibrator - The biochemical thermodynamics calculator," *Nucleic Acids Res.*, 2012.

[118] A. CAKMAK *et al.*, "A NEW METABOLOMICS ANALYSIS TECHNIQUE: STEADY-STATE METABOLIC NETWORK DYNAMICS ANALYSIS," *J. Bioinform. Comput. Biol.*, 2012.

[119] D. Machado, S. Andrejev, M. Tramontano, and K. R. Patil, "Fast automated reconstruction of genome-scale metabolic models for microbial species and communities," *Nucleic Acids Res.*, 2018.

[120] IBM Corp. and IBM, "V12. 1: User's Manual for CPLEX," *Int. Bus. Mach. Corp.*, 2009.

[121] B. Pillay, G. Roth, and R. A. Oellermann, "Cultural characteristics and identification of marine nitrifying bacteria from a closed prawn-culture system in Durban," *South African J. Mar. Sci.*, 1989.

[122] D. Schüler, "Molecular analysis of a subcellular compartment: The magnetosome membrane in Magnetospirillum gryphiswaldense," *Archives of Microbiology*. 2004.

[123] I. King Jordan, I. B. Rogozin, Y. I. Wolf, and E. V. Koonin, "Essential genes are more evolutionarily conserved than are nonessential genes in bacteria," *Genome Res.*, 2002.

[124] L. Carvalho, A. Nobre, M. Mota, and J. Padrão, "Effect of different carbon sources on biomass production of Nitrobacter vulgaris," Universidade do Minho, 2016.

[125] S. R. Starkenburg *et al.*, "Complete genome sequence of Nitrobacter hamburgensis X14 and comparative genomic analysis of species within the genus

Nitrobacter," *Appl. Environ. Microbiol.*, 2008.

[126] M. A. Guerrero and R. D. Jones, "Photoinhibition of marine nitrifying bacteria. I. Wavelength-dependent response," *Mar. Ecol. Prog. Ser.*, 1996.

[127] "Portugal | Sigma-Aldrich." [Online]. Available: https://www.sigmaaldrich.com/portugal.html?gclid=EAIaIQobChMI8YPfsrPj3QIVV4fVCh0v8QAXEAAYASAAEgKHrfD_BwE. [Accessed: 30-Sep-2018].

[128] "Yara International." [Online]. Available: https://www.yara.com/. [Accessed: 30-Sep-2018].

[129] "The Mosaic Company: Concentrated Phosphate and Potash Crop Nutrition." [Online]. Available: http://www.mosaicco.com/. [Accessed: 30-Sep-2018].

[130] "Home | Nutrien." [Online]. Available: https://www.nutrien.com/. [Accessed: 30-Sep-2018].

[131] "Markets Insider: Stock Market News, Realtime Quotes and Charts." [Online]. Available: https://markets.businessinsider.com/. [Accessed: 30-Sep-2018].

[132] "Market Realist - Stock Market News, Financial Research and Analysis." [Online]. Available: https://marketrealist.com/. [Accessed: 30-Sep-2018].