Universidade do Minho
Escola de Engenharia
Departamento de Informática

Pedro Miguel Brígida Raposo

Reconstruction of the genome-scale metabolic model of *Nitrosomonas europaea*

October 2017

Universidade do Minho
Escola de Engenharia
Departamento de Informática

Pedro Miguel Brígida Raposo

Reconstruction of the genome-scale metabolic model of *Nitrosomonas europaea*

Master dissertation
Master Degree in Bioinformatics - Technologies of Information

Dissertation supervised by
Oscar Manuel Lima Dias
Jorge Manuel Padrão Ribeiro

October 2017

ACKNOWLEDGMENTS

Quero agradecer ao Oscar Dias por me ter aceitado como orientando, e por me ter acompanhado durante todo este percurso. Este trabalho exigiu tanto esforço do meu lado assim como paciência do orientador para ensinar-me novos conceitos e metodologias. Agradeço por ter sido a pessoa com quem mais aprendi este ano, não só sobre bioinformática, mas também sobre pensar por si próprio.

Agradeço também ao Jorge Padrão por me ter integrado no laboratório, assim como ter-me ensinado vários conceitos biológicos. Agradeço por ter sido uma pessoa sempre aberta a novas ideias e perspetivas, ao mesmo tempo que inicialmente me sentia inseguro nesse local de trabalho. Aprendi também a valorizar os detalhes na prática laboratorial, assim como a ter um maior rigor cientifico. E mais importante, ter sempre pensamento positivo!

Agradeço à Sophia Santos e à Aline Barros por terem despendido tempo a ensinar-me variadas técnicas laboratoriais. Também por serem pessoa muito amigáveis!

Agradeço ainda à Vitória Maciel, à Maura Guimarães e à Diana Vilas Boas por terem sido atenciosas e por ajudarem-me em várias situações no laboratório.

Agradeço aos professores que fazem deste curso de Bioinformática uma experiência altamente positiva.

Agradeço às pessoas que mais indiretamente me ajudaram no trabalho, tanto através da conversa sobre este ou por algum acompanhamento mútuo: António Dias, Daniela Azevedo, Bruna Mendes, Fernando Cruz, José Dias, Bruna Soares, e Emanuel Oliveira.

Agradeço aos meus pais e irmão por sempre me apoiarem na vida académica.

Gostava ainda de agradecer às pessoas que de uma maneira ou outra que foram um pouco responsáveis pela posição em que me encontro, e que moldaram a pessoa que sou hoje. Agradeço a Sara Salazar por saber valorizar-me e por saber que sem ela, não teria seguido estudos científicos. Agradeço ao Felipe Luz, ao Ivo Rodrigues e ao André Batista pela amizade duradoura. Agradeço ao Ruben Gonçalves, André Lúcio, Vítor Silva, Bernardo Nunes por terem um grande impacto na minha vida, durante um grande período. Um especial agradecimento ao Luís Cordeiro por me ter sempre ajudado e por me fazer sentir compreendido, não só antes, mas também, e especialmente, nestes 2 anos de Mestrado.

ABSTRACT

Nitrogen is one of the four most common elements in any cell and thus, it is needed to sustain all kinds of life, making the nitrogen cycle crucial to life on Earth. However human activities have doubled the transfer of the reactive nitrogen into the biosphere, largely through the excessive use of fertilizers. This lead to eutrophication of aquatic systems, a negative ecosystem response usually associated with reduction of the biodiversity in it.

This work is set to improve the removal technique of reactive nitrogen by transforming it into non-reactive nitrogen - through *Nitrosomonas europaea*, an essential and ubiquitous bacteria in the nitrogen cycle. By using it in wastewater treatment plants, it is possible to overcome a limiting step of this transformation, which ultimately helps to stop eutrophication.

N. europaea is the most studied ammonia-oxidizing bacteria to date and has various pathways that involve different compounds of nitrogen, making it metabolically versatile and, therefore, suitable for wastewater treatments. In this work, it was reconstructed a genome-scale metabolic model of *N. europaea*, using *merlin* (a specialized software for this task), to allow performing *in silico* simulations with different environmental conditions, providing knowledge of its underlying metabolic fluxes.

This reconstruction was made through computational means (including several iterative steps such as automatic and manual annotation of the genome, curation of the metabolic pathways, among others), was validated through laboratorial means (by growing the organism in a chemostat and quantifying the compounds of its biomass), and was supported by literature in many cases.

This validation was represented by the accuracy of the model (a comparison between the *in vivo* with the *in silico* data), and was equal to $98,36$ %.

Now, with a metabolic model of the organism, a guided approach may be developed to optimize the conversion of ammonia into nitrite, to be later metabolized by other organisms to produce molecular diatomic nitrogen (inactive nitrogen), thus providing a solution to eutrophication.

Keywords: *Nitrosomonas europaea*, Genome-scale metabolic model, Nitrogen, Eutrophication

RESUMO

O Azoto é um dos quarto elementos mais comuns na célula, e por isso é necessário para sustentar qualquer tipo de vida, tornando o ciclo do azoto crucial para a vida na Terra. Mas as actividades humanas duplicaram a transferência de azoto reactivo para a biosfera, maioritariamente através do uso excessivo de fertilizantes. Isto conduziu à eutrofização de sistemas aquáticos, uma resposta negativa do ecossistema normalmente associada à sua redução da sua biodiversidade.

Este trabalho está focado em melhorar a técnica de remoção de azoto reactivo ao transformá-lo na sua forma inactiva - através da *Nitrosomonas europaea*, uma bactéria essencial e ubíqua no ciclo do azoto. Ao usá-la em plantas de águas residuais, é possível ultrapassar um passo limitante desta conversão, que por sua vez ajuda a parar a eutrofização.

N. europaea é a bactéria oxidante de amoníaco mais estudada e esta contém várias vias metabólicas que envolvem diferentes compostos de azoto, tornando-a metabolicamente versátil, e assim é adequada para tratamento de águas residuais. Foi reconstruído, neste trabalho, um modelo metabólico à escala genómica da *N. europaea*, usando *merlin* (um *software* especializado para esta tarefa), para permitir a realização de simulações *in silico* sujeitadas a diferentes condições ambientais, fornecendo informação sobre os seus fluxos metabólicos.

Esta reconstrução foi feita através de meios computacionais (incluindo vários passos iterativos, como a anotação automática e manual do genoma, a curação das vias metabólicas, entre outros), foi validada através de meios laboratoriais (ao crescer este organismo num quimiostato e ao quantificar os compostos da sua biomassa), e foi apoiada e justificada através da literatura, em muitos casos.

Esta validação foi representada através da exatidão do modelo (uma comparação entre informação *in vivo* e *in silico*), e foi igual a $98,36$ %.

Agora, com o modelo metabólico deste organismo, uma abordagem orientada para a optimização da conversão de amoníaco para nitrito poderá ser desenvolvida, para este composto ser metabolizado por outros organismos para ser produzido azoto diatómico molecular (azoto na sua forma inactiva), e assim, fornecer uma solução para a eutrofização.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ABHT Amonibacteriohopanetriol.

AIBENCH Artificial Intelligent workBench.

AMO Ammonia Monooxygenase.

ANAMMOX Anaerobic Ammonium Oxidation.

AOB Ammonia-Oxidizing Bacteria.

ATP Adenosine Triphosphate.

BLAST Basic Local Alignment Search Tool.

BRENDA Braunschweig Enzyme Database.

BSA Bovine Serum Albumin.

CA Correctly annotated and Accepted.

CHR Charge Reduction Sample Holder.

CORECO Comparative Reconstruction.

CR Correctly annotated and Rejected.

CSS Cascading Style Sheets.

DAPI 4,6-Diamidino-2-Phenylindole Dihydrochloride.

DNA Deoxyribonucleic Acid.

DO Dissolved Oxygen.

DTDPA 3.3 Dithiodipropionic Acid.

EC Enzyme Commission.

EDS Energy-Dispersive X-ray Spectroscopy.

EDTA Ethylenediaminetetraacetic Acid.

ETC Electron Transport Chain.

EXPASY Expert Protein Analysis System.

FAME Flux Analysis and Modeling Environment.

FBA Flux Balance Analysis.

FN False Negative.

FP False Positive.

FVA Flux Variability Analysis.

GEMSIRV GEnome-scale Metabolic model Simulation, Reconstruction and Visualization.

GLPK GNU Linear Programming Kit.

GPR Gene-Protein-Reaction.

GSM Genomic-Scale Metabolic.

HAMAP High-quality Automated and Manual Annotation of Proteins.

HAO Hydroxylamine Oxidoreductase.

HPLC High-Performance Liquid Chromatography.

HTML HyperText Markup Language.

IA Incorrectly annotated and Accepted.

ICP-MS Inductively Coupled Plasma Mass Spectrometry.

IR Incorrectly annotated and Rejected.

KEGG Kyoto Encyclopedia of Genes and Genomes.

MATLAB Matrix Laboratory.

ME Metabolic Engineering.

MEMOSYS Metabolic Model research and development System.

MERLIN Metabolic Models Reconstruction Using Genome-Scale Information.

MIRIAM Minimal Information Required In the Annotation of Models.

MOMA Method Of Minimization of Metabolic Adjustment.

MUSCLE Multiple Sequence Comparison by Log-Expectation.

N Nitrogen.

NA Non-annotable and Accepted.

NAD Nicotinamide Adenine Dinucleotide.

NADH Reduced Nicotinamide Adenine Dinucleotide.

NCBI National Center for Biotechnology Information.

NOB Nitrite-Oxidizing Bacteria.

NPV Negative Predictive Value.

NR Non-annotable and Rejected.

NR non-redundant.

ORF Open Reading Frames.

P Phosphorus.

PBS Phosphate-Buffered Saline.

PC Phosphatidylcholine.

PDME Phosphatidyl-N,N-dimethylethanolamine.

PE Phosphatidylethanolamine.

PFBA Parsimonious Flux Balance Analysis.

PG Phosphatidylglycerol.

PGDB Pathway or Genome DataBases.

PHP Personal Home Page.

RAVEN Reconstruction, Analysis and Visualization of Metabolic Networks.

RNA Ribonucleic Acid.

ROOM Regulatory On/Off Minimization.

RRNA Ribosomal Ribonucleic Acid.

SBML Systems Biology Markup Language.

SEM Scanning Electron Microscope.

TC Transporter Classification.

TCDB Transporter Classification Database.

TMHMM Transmembrane Helices prediction by Hidden Markov Models.

TN True Negative.

TP True Positive.

TREMBL Translated EMBL Nucleotide Sequence Data Library.

TRIAGE Transport Proteins Annotation and Reactions Generation.

TRNA Transfer Ribonucleic Acid.

UNIPROT Universal Protein Resource.

UNIPROTKB Universal Protein Resource Knowledgebase.

1

INTRODUCTION

1.1 CONTEXT AND MOTIVATION

Nitrogen (N) is required by all organisms and it is one of the four most common elements in the cell (Tamm, 1991; Galloway, 1998; Conley and Likens, 2009). Thus, the distribution of this element in the biosphere is required to sustain all kinds of life, making the N cycle crucial to survival on Earth (Sprent, 1987).

Prior to the anthropogenic input of bioavailable N, this cycle was balanced, however, human activities have doubled the transfer of the reactive N into the biosphere, largely through the misuse use of fertilizers, setting in motion a variety of environmental consequences (Chindler and Ilman, 1997; Galloway, 1998). The excessive disposal of nutrients into an aquatic system lowers the quality of the water and promotes the overabundance of algae, aquatic plants and phytoplankton (Smith, 2003; Kemp and Stevenson, 2005; Conley and Likens, 2009). This leads to a lower concentration of dissolved oxygen in the aquatic system and consequent death of many species from all trophic levels, drastically reducing the biodiversity in the habitat (Chindler and Ilman, 1997; Conley and Likens, 2009). The reduction of biodiversity in aquatic systems is a serious issue that has ecologic, economic, societal and health impacts (Chapin et al., 2000; Hooper et al., 2005). For example, reduction of algae diversity alone, diminishes the removal of pollutants in the water, promoting waterborne pathogens (Cardinale et al., 2012). The reduction and reduction of fish diversity promotes the spread of human diseases and loss of water quality either to drinking or to irrigate (Moyle and Leidy, 1992).

The impact of over-addition on nutrients in the ecosystem is called eutrophication (Schindler and Vallentyne, 2008). To stop eutrophication, improve the quality of water for human consumption, and to sustain the species diversity in the aquatic systems, an action must be made to reduce the input and concentration levels of nutrients (Chase and Leibold, 2002; Smith, 2003; Galloway, 2005). There are many approaches to decrease the concentration of the different nutrients in a water body. In this work we set to improve a removal technique of one of those nutrients - the N by transforming this element from reactive into non-reactive - using anaerobic ammonia-oxidizing bacteria (Bagchi and Nandy, 2012). This technique is one of the most used in the industry and is implemented in various wastewater treatment plants (Bagchi and Nandy, 2012).

Nitrosomonas europaea, a gram-negative bacterium, is the most studied Ammonia-Oxidizing Bacteria (AOB) to date, especially at the molecular level (Chain and Whittaker, 2003; Bagchi and Nandy, 2012). This organism has various pathways that involve different compounds of N, making it metabolically versatile and, therefore, suitable for wastewater treatments (Schmidt and Jetten, 2002; Bagchi and Nandy, 2012). To optimally use this organism in wastewater treatments, the underlining reactions of it must be fully understood.

Reconstructing a Genomic-Scale Metabolic (GSM) model of *N. europaea*, will allow performing *in silico* simulations with different environmental conditions, providing knowledge of the underlying metabolic fluxes (Villadsen, 2009; Dias et al., 2015). This work envisages the use of various bioinformatic tools, including *Metabolic Models Reconstruction Using Genome-Scale Information (merlin)*, a software to reconstruct the GSM model of *N. europaea*. Then, a guided approach may be developed to optimize the conversion of ammonia into nitrite, to be later metabolized by other organisms to produce molecular diatomic N (non-reactive N), thus providing a solution to eutrophication (Conley and Likens, 2009).

By studying a relevant microorganism, this work may represent an important step for the solution of the eutrophication problem, a serious issue for Mankind. The current knowledge on the N on eutrophication, as well as, lack of it on *N. europaea* is a great opportunity to tackle the problem in a new angle, providing new perspectives to scientific research by using this organism more efficiently. Reconstructing a functional GSM model of this organism allows a better management of *in vivo* experiments, by reducing wet-lab costs and time spent designing. Hence, these models prove to be an advantage to research departments as budgetary restrictions are increasing, constraining the capital spent on wet-lab resources and studies. This relative new approach in science, has proven to be reliable, as Bioinformatics is an area that stands out for being accurate and for keeping up with the ever-growing knowledge of Biology, using many state-of-the-art methods available today (Samish and Najmanovich, 2015).

1.2 GOALS

The goal of this work is the reconstruction of the GSM model of *N. europaea*, and the definition of strategies for improving production of nitrite. This work consists of several technical objectives:

- Studying of the background of the eutrophication problem and the organism is case, and reviewing of the state-of-the-art GSM model reconstruction and validation software.
- Drafting of the metabolic model by performing genome annotation, identifying the metabolic reactions, verifying the stoichiometry of the reactions, confirming the localization of the reactions, constructing the biomass equation and adding other constraints to the model.
- Validating of the metabolic model, iteratively, by comparing *in silico* simulations with *in vivo* experiments under specific conditions.

1.3 STRUCTURE OF THE DOCUMENT

This document is organized in the followings parts:

Chapter 2

State-of-the-art

Environmental background on eutrophication. Current solutions to the eutrophication problem. Introduction to the Metabolic Engineering, its goals and strategies. Nature of metabolic models and its mathematical representation. Presentation of the reconstruction and the validation process of the GSM model, and the available software.

Chapter 3

Methods and Development

N. europaea draft GSM model reconstruction, including the automatic and manual annotation of the genome. Experimental material and methods for the growth the organism and for the quantification its compounds.

Chapter 4

Results and Discussion

Results from the reconstruction of the GSM model, from the growth of the organism and from the quantification of its compounds, by *in vivo* data. Results from simulations, by *in silico* data. Discussion of the validity of the GSM model. Comparison with other GSM models.

Chapter 5

Conclusions and Future work

Summary of the work accomplished and next tasks for future efforts to solve the eutrophication problem.

2

STATE OF THE ART

2.1 ENVIRONMENTAL BACKGROUND

The Biosphere is the ”peripherial envelope” of the planet Earth, along with the organisms that naturally inhabit it (Lieth and Whittaker, 1975; Veziroglu, 1984; Kumar, 1998). Humans and all of the other species draw the essential resources such as air, water and food from the Biosphere, and thus are consequently dependent on its ecological integrity and great cycles, for survivability (Lieth and Whittaker, 1975; Kumar, 1998). However, Mankind has a powerful influence over it, and can disturb the Biosphere balance - even though Mankind is in absolute dependence of it (Veziroglu, 1984; Vernadsky, 1998). One of the most serious damaging activities is the pollution (Kumar, 1998).

Pollution is the introduction of substances, in the environment, where their distribution, concentration or physical behavior have undesirable or deleterious consequences (National Research Council, 1978; Harrison, 2001). This can occur on the atmosphere, the hydrosphere or on the lithosphere and all of these three systems are equally important on this issue, meaning that dumping the residues into one of them cannot solve the problem, as it is only shifting the problem from one element to another (Kumar, 1998). However this work focus on the hydrosphere, the water system.

Water is one of the most important molecules for life - in fact all organisms require it and biological evolution could not be possible without it (Agarwal, 2005). It is also one of the most important resources, which Mankind has exploited for the sustenance of life. However, only about 3 percent of water can be used for human daily requirements, being this portion distributed in the form of ice sheets, underground sources, lakes, rivers, ponds, atmosphere and biological water contained in the living organisms (Agarwal, 2005). This fact makes it more important to preserve the quality of the water overall (the physical, chemical and biological characteristics able to fit for some use) by not polluting it (Agarwal, 2005).

The input of pollutants, chemicals causing environmental harm, leads to an array of effects of water pollution, such as: (1) Aesthetic disruption - Causes visual nuisances; (2)

Temperature issues - Usually, causes the aquatic system to overheat; (3) Deoxygenation - Removes oxygen from the water; (4) Toxicity - Leads to acute or chronic toxicity, causing damage to aquatic or human life; (5) Sublethal toxicity - Disrupts or changes the biodiversity; (6) Acidity/Alkalinity: Disturbs the pH regime of the water; (7) Eutrophication - Nutrients give rise to excessive growth of some organisms (Harrison, 2001). This set of problems triggers a cascade of events, leading to other problems in the aquatic system (Harrison, 2001). The urgent development and enhancement of biotechnological to solve the eutrophication problem is the main motivation of this work.

2.2 EUTROPHICATION

Eutrophication is a form of pollution that can be described as the natural or artificial addition of nutrients to bodies of water, its increased production associated with it, and the environmental effects that come as a result (National Academy of Sciences, 1969; Pfafflin and Ziegler, 2006). The called eutrophication *sensu stricto* is the addition into a body of water nutrient supply and does not affect the quality of the water *per se*, but the effects of eutrophication can induce a course of action leading to undesirable consequences (National Academy of Sciences, 1969).

Eutrophication is a natural process, and it reflects the aging of the body of water (this differs according to the characteristics of each body of water, *i.e.*, lakes, rivers and others) - the more aged, the more is effected by eutrophication (National Academy of Sciences, 1969). However, when accelerated by man-made activities, it can result an a serious problem to the environment and human health (Pfafflin and Ziegler, 2006).

2.2.1 *History and relevance*

The earliest scientific reports of eutrophication dates to 1907, however it only got attention from the scientific community to develop the management of this ecological problem and to lead the course of future research in 1967, with the International Symposium on Eutrophication (National Academy of Sciences, 1969). In this symposium, speakers agreed that prevention of further damage to water resources was a matter of great urgency, and other participants agreed that there should be a greater acquirement of knowledge about the processes involved in eutrophication (National Academy of Sciences, 1969).

Eutrophication has been accelerating in the last decades due human activities that increase the rate of nutrient input in a water body, due to rapid urbanization, industrialization and intensified agricultural production (Yang et al., 2008).

This issue is so aggravating for the fact that it can affect different important areas: by detracting natural beauty, reducing of property values, destructing of water resources, making the water effected non-potable and increasing the cost of filtration - making it a political, societal and economical problem (Committee on Environmental and Natural Resources, 2003; Pfafflin and Ziegler, 2006).

2.2.2 *Mechanisms of effect*

One of the various ways eutrophication can affect the aquatic system is by promoting the productivity of some species of the phytoplankton community and its derived organic matter (Kemp and Stevenson, 2005). In lakes, the increasing productivity makes a succession of stages on the body of water: starting with the oligotrophic stage (low productivity), becoming mesotrophic (medium productivity) eventually eutrophic (highly productive) and finally dystrophic - a stage in which the lake has almost been filled in by weeds and the productivity has been greatly decreased (Pfafflin and Ziegler, 2006). The progression of this body of water, along with other typed of aquatic systems, can be explained by a cascade of chemical and biological processes (Committee on Environmental and Natural Resources, 2003; Yang et al., 2008).

First, the input of nutrients into a aquatic system leads to a direct primary production (Rabalais et al., 2002). In some cases, even when the fluxes of nutrients is often high in a system, biological uptake is frequently comparable to its magnitude, meaning that the nutrients were rapidly transformed into organic matter in this system because there is, in these instances, a great nutrient demand from phytoplankton to its primary production (Lohrenz et al., 1997). The nutrient loaded into the system, controls the size and the species composition by selecting the species with lower requirements of that specific nutrient (Dortch et al., 2001). Aside from each species population growth, the species composition is influenced by the cellular sinking rate of each specie (leading the cells into their death) - this rate is caused by environmental stress and it varies markedly between taxa (Harper, 2012). The sedimentation of this cells eventually leads to accumulation of it in the bottom of the body of water (Harper, 2012).

This growth of population of some phytoplankton species, but not others, changes dramatically the dynamic of the body of water and its characteristics (Yang et al., 2008). The rapid changes in phytoplankton community composition, where it becomes dominated by a single (or a few) species, over the course of days, is called algal bloom (Egerton and Mulholland, 2014). This overabundance of one species not only reduces diversity in the phytoplankton community but, associated with the nutrient input, it also increases the primary production of the body of water for the fact that the prevalent specie has a great capability of growth with an excessive quantity of its limitant nutrient (Lohrenz et al., 1997; Egerton and Mulholland, 2014). Studies done by Egerton and Mulholland (2014) demonstrated a correlation between increased phytoplankton biomass overall and decreased diversity in the phytoplankton community, in these circumstances.

Eventually, this overproduction of primary production in the lake, along with its sedimentation, leads to decay of this biomass (Dortch et al., 2001). This decomposition is done by aerobic bacteria, that use Dissolved Oxygen (DO) of the the water, as a electron acceptor in the electron transport chain (Rabalais et al., 2002). Also, massive quantities of phytoplankton may act as a barrier to the penetration of oxygen into the water as well as a barrier to sunlight, reducing the transparency of the water and weakeing or even stopping the photosynthesis of plants underwater (Pfafflin and Ziegler, 2006; Yang et al., 2008). Because of this, the aquatic system starts lacking DO (this condition is called hypoxia and the most severe form of is called anoxia, where there is no DO in the system) (Chindler and Ilman, 1997). Studies by Kemp and Stevenson (2005) have confirmed a correlation between number of algae, DO and nutrient loads into aquatic systems.

Ultimately, the hypoxia state of the aquatic system contributes substantially to the destabilization of food chains by killing fish, shellfish, and benthic organisms (Committee on Environmental and Natural Resources, 2003; Kemp and Stevenson, 2005). The blooming algae can even start releasing toxins and render the organic matters in water to be decomposed into harmful gases, which will also poison the fish and seashell (Yang et al., 2008). At last, the DO levels becomes so aggravating that the environment kills every species except a few anaerobic bacteria, converting the ecosystem into a nearly sterile state (National Academy of Sciences, 1969; Othman et al., 2014).

Eutrophication can also alter further the quality of the water, by releasing harmful toxins to the human health when the blooming algae die and attracting waterfowl which contribute to the pollution of the water, among many other ways (Pfafflin and Ziegler, 2006; Yang et al., 2008). The massive death of the species of different trophic levels has a great negative impact on the habitat and on the environment in general (Committee on Environmental and Natural Resources, 2003).

2.2.3 *Causes*

There are many factors that lead to the eutrophication of the general aquatic system (Chindler and Ilman, 1997; Pfafflin and Ziegler, 2006). The common sources of the most relevant nutrients are rainfall, ground water, urban runoff, rural runoff, agricultural runoff, industrial wastes, municipal water treatment, waterfowl and domestic sewage effluent (Pfafflin and Ziegler, 2006). The domestic sewage effluent (being the origins, the nutrients derived from human wastes, waste food and synthetic detergents) is one of the greatest contributors to eutrophication, however, the agricultural runoff is considered by the scientific community to be the most influential one (being the origins, the wastes from farm animals and the overuse of fertilizers) (Pfafflin and Ziegler, 2006; Yang et al., 2008).

Although vitamins, growth hormones, amino acids and trace elements can all contribute to eutrophication, N and Phosphorus (P) have been assign as the most influential in this issue (only a mere $0,3$ to $0,015$ ppm of nitrates, in the water, is enough to produce blooms of certain species of algae, in some conditions) (National Academy of Sciences, 1969). N and P are considered the limiting factor that triggers the algal bloom and are the only ones thoroughly studied in the field and in the laboratory (Conley and Likens, 2009; Pfafflin and Ziegler, 2006).

This work focused on N, as one of the most important factors leading to eutrophication.

2.2.4 *Possible solutions*

Since the alarming consequences of eutrophication has been sensibilized to the scientific community, measures were taken by namely research on the subject and developing various technological solutions to this problem (National Academy of Sciences, 1969).

There are two ways to stop eutrophication and its effects from occurring: prevention of introduction of nutrients resulted from Man activities into the aquatic systems, and enhanced the removal of those nutrients present in it (National Academy of Sciences, 1969). While prevention is generally perceived as a greater solution, it is necessary to apply techniques to remove the reactive N from the current bodies of water subjected by it (National Academy of Sciences, 1969; Bagchi and Nandy, 2012). There is a plethora of removal techniques used in the modern day: Drenching; Removal of algae, aquatic weeds and other organisms; and other techniques (Pfafflin and Ziegler, 2006).

Specifically about the N, there are an array of removal techniques that can be applied in a water treatment: (1) Land application - a method to make ammonium in the water to be adsorbed when its soil is percolated; (2) Ammonia Stripping - an aeration process, making the N (as non-reactive molecule) liberate into the atmosphere; (3) Anaerobic Ammonium Oxidation (ANAMMOX) - a biological process performed by bacteria that converts nitrite and ammonium into diatomic N; (4) Anaerobic denitrification - process that reduces nitrate to diatomic N and nitrous oxide by denitrifying bacteria; (5) Nitrification and Denitrification - process that reduces ammonia into nitrate, and then nitrate to diatomic N (Pfafflin and Ziegler, 2006; Bagchi and Nandy, 2012). The technique proposed in this work is the latter process.

Nitrification and Denitrification

Nitrification and denitrification are two processes that can be coupled to be a solution to remove ammonia (reactive N) from water to the atmosphere as diatomic N (non-reactive N) (Bagchi and Nandy, 2012). Nitrification is the process that can be done by AOB that transforms NH_3 (rather than NH_4^+, for the fact that this molecule is not permeable to cell membrane) into nitrite (NO_2), and by Nitrite-Oxidizing Bacteria (NOB), that transforms nitrite into nitrate (NO_3) (Baribeau, 2006). Then this nitrate is transformed by denitrifier bacteria into diatomic N, in a process called denitrification (Baker and Irvin, 2007; Bagchi and Nandy, 2012). A simplified pathway of these processes is presented in Figure 1.

There has been studies concerning these communities of bacteria, which established taxonomic groups based on their capability to perform each of the denitrification process (Baribeau, 2006). The fact that AOB and NOB have low growth rates, and therefore have doubling times (time needed for the biomass or population to double in number) ranging between 11 and 50 hours, slows the study of these organisms (Baribeau, 2006). However, the bottleneck of the N removal is at the AOB performance, because usually NOB are facultative chemolitoautotrofic bacteria which can use organic carbon as substrate and AOB are often chemolithoautotrophic bacteria, which means the only source of carbon is derived from CO_2 fixation (Mandana and Tahmurespour, 2012). Unfortunately, this fixation requires a great amount of energy, which further slows down the growth rate of AOB comparatively to NOB. For this, there is a great interest for studying the AOB to improve their growth and optimize the whole nitrification process.

And so, this work is focused on *N. europaea* ATCC 19718, for the reason that it is the most studied AOB, especially at the molecular level and because it has a rich and complex network of N pathways (Chain and Whittaker, 2003; Bagchi and Nandy, 2012).

Also, there was previously made a plethora of research on N on the eutrophication issue, however, there has not been relevant study on the *N. europaea* using the strategy proposed by this work. Even though there is a sufficient amount of information on scientific articles and databases to make this work possible.

2.2.5 *Nitrosomonas europaea*

N. europaea, a gram-negative bacterium, is considered an obligate chemolithoautotroph that obtains energy from ammonia oxidation and assimilate carbon from atmospheric carbon dioxide fixation (Chain and Whittaker, 2003; Bagchi and Nandy, 2012). This organism has various pathways that evolve many different compounds of N, which makes it metabolh

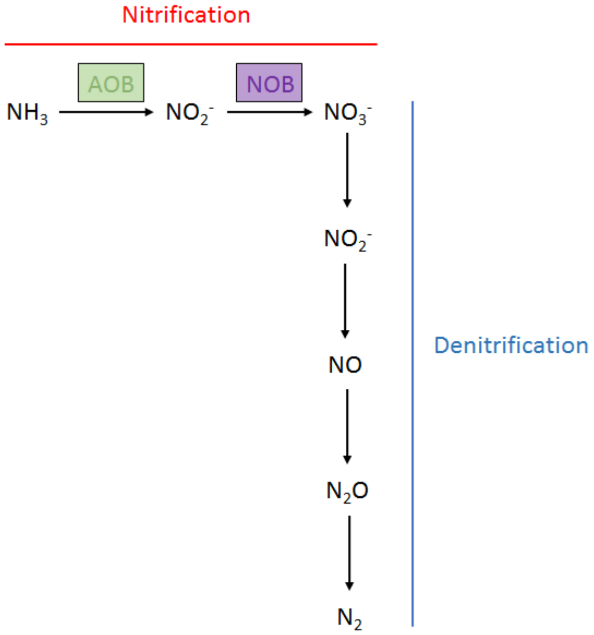


Figure 1.: Nitrification and denitrification simplified pathways. Representation of AOB and NOB activities in the nitrification pathway. Figure based on Peng and Zhu (2006).

ically versatile, and therefore suitable for wastewater treatments (Schmidt and Jetten, 2002; Bagchi and Nandy, 2012). The main metabolic features are the ability to oxidize ammonia to nitrite (via aerobic and anaerobic pathway), and reduce this nitrite generated to nitrous oxide through nitrifier denitrification, in anoxia conditions (Bagchi and Nandy, 2012).

N. europaea uses the advantageous pathway depending on the environmental conditions. This bacterium, through the use of the different enzymes it can synthesize, it can metabolically produce or degrade an array of N compounds for different purposes. There are even strains of *N. europaea* that metabolize NH_3 into N_2, however, this work is trying to elaborate on a proof of concept that states that the N removal is more efficient through a series of operations, each one performed by a species, than the whole process performed by a single organism (Shrestha et al., 2001).

In laboratory, *N. europaea* is a difficult organism to work with. Additionally, it has a low specific growth rate, which makes the develop of new approaches and improvements of N removal techniques even more inefficient. The main reason for this inefficiency is due to the chemolithoautotrophic possess a growth efficiency ranging between 4,4 to 21,3 %, due to a part of the energy generated by autotrophs is used to fix CO_2 (Baribeau, 2006). To understand and improve the ways *N. europaea* can be grown and multiplied, as well as metabolically studied, it has to be approached at a molecular level.

This work uses metabolic engineering to do this, and the final objective is to reconstruct a metabolic model of *N. europaea*. The following sections will detail on what are metabolic models, and how, ultimately, it can help improve on N removal techniques to stop eutrophication.

2.3 METABOLIC ENGINEERING

Metabolic Engineering (ME) can be defined as the construction, redirection and manipulation of cellular metabolism through the introduction, deletion, and/or modification of metabolic pathways, commanded by enzymes, to achieve biosynthesis or biocatalysis of desired natural and non-natural compounds (Lee and Papoutsakis, 1999; Smolke, 2009). The roots of ME have started being developed in the 1970s with the advancement of one important tool in the field of genetic engineering - recombinant Deoxyribonucleic Acid (DNA) technology (Nielsen, 2003; Smolke, 2009). With the advent of this technique, ME has differed from the initial applications of genetic engineering, in the way that this new approach could not only work in one or a few reactions in an organism, but in multiple reactions (through genes) to create entirely new pathways that could produce a wide range of compounds (Nielsen and Arnold, 2005). This allowed the manipulation of the organisms as a whole, as one of the defining aspects of the ME is the focus on integrated metabolic pathways instead of individual reactions - this means that it examines entire biochemical reaction networks, concerning itself the pathway fluxes and its control (Stephanopoulos et al., 1998). Another defining aspect of ME is the strongly directed effort to cell improvement, compared to random mutagenesis (Stephanopoulos et al., 1998; Nielsen and Arnold, 2005).

The major aspect of ME, along with its goal, is the large-scale production of useful chemicals (Kulkarni, 2016). This is achieved by controlling the biochemical pathways fluxes, that determine the cell physiology, leading to the overproduction of different compounds (Stephanopoulos et al., 1998). There are different approaches of ME for the production of these biochemicals: (1) by overexpressing the gene encoding the enzyme that participates in the biosynthetic pathway of the desired product; (2) by inhibiting the competing metabolic reactions that use the same precursor of the desired product for another biochemical; (3) by carrying out the production of the desired biochemical to a non-native organism, *i.e.*, by isolating the gene responsible for the production of the desired biochemical and import it to another organism that can do the same biosynthesis more easily - this is called heterologous expression; (4) by mutating genes, altering the resultant amino acids, to synthesize non-natural chemicals not found in Nature (Kulkarni, 2016). However, the traditional method of ME have focused on the modification of enzymatic pathways near to the end-product (Lee and Papoutsakis, 1999).

Although there are competitive areas of study to provide compounds of interest, there are many advantages of synthesizing chemicals, materials or energy via ME, in contrast to the traditional chemical synthesis methods: Firstly, many chemicals remain difficult to synthesize with the latter strategy, whereas the former one has demonstrated to provide adaptability to be enable the production of complex molecules. Secondly, unlike the latter strategy, the former is often conducted under mild conditions, enabling the production of fewer toxic-byproducts. And thirdly, ME strategies use the cell natural ability to replenish enzymes, and to provide precursors from inexpensive and renewable starting materials, which ultimately softens the global environmental impact (Smolke, 2009). The presented reasons and the fact that the ME is multidisciplinary, are advantages in meeting goals of this field (Stephanopoulos et al., 1998).

ME contributes in the measurement and understanding of the control of fluxes *in vivo* by revealing the degree of pathway engagement in the metabolic process (Stephanopoulos et al., 1998). This provides insights into yield optimization (through optimal flux distribution), being it important to approach some common challenges of bioprocessing (Nielsen and Arnold, 2005). Besides biotechnology, medicine and agriculture can also benefit from advances in ME which can contribute to overcome their challenges (Nielsen and Arnold, 2005). More specifically, ME can improve the processes of fermentation engineering, drug target identification, and microbial engineering (Jing and Alashwal, 2014). ME will continue to progress, considering the incorporation of new experimental and computational tools (Nielsen, 2003). The implementation of these tools may expand ME into new areas of application, being the industrial an important one, as the economic potential of biotechnology is increasing (Nielsen and Arnold, 2005).

ME focuses on the development of new cell factories and on the improvement of existing cell factories and is an application of Systems Biology (Nielsen and Jewett, 2007).

2.3.1 *Systems Biology*

Systems biology can be described as a multidisciplinary science that works with mathematical modelling, global analysis, mapping of interactions between cellular components and quantification dynamic responses on living cells (Nielsen and Jewett, 2007). The goal of Systems Biology is to describe quantitatively biological systems in the form of a mathematical model to predict the behavior of the biological system under specific circumstances, providing new insight into the molecular mechanisms occurring in living cells (Nielsen and Jewett, 2007) . This is only possible through the combination of mathematical modeling and experimental biology (Nielsen and Jewett, 2007).

2.3.2 *Metabolic models*

It has been determined that genes inherently can determine one or more functions in the cell, by producing its respective protein or proteins, and as a consequence changing the cell phenotype (Crick, 1970; Malcolm and Goodship, 2001). These enzymes (proteins with biological catalytic activity) promote biochemical reactions by converting reactant metabolites into product metabolites (Champe et al., 2005; Benner et al., 2014). The set of these reactions is defined as the metabolic network which is responsible for the uptake and degradation of substrates and for the synthesis of building blocks and energy that the whole cell requires (Baart and Martens, 2009; Benner et al., 2014). In other words, enzymes are responsible for catalyzing biochemical reactions within the cell, and the complete set of these reactions represents the metabolism of the cell (Baart and Martens, 2009).

Understanding of the metabolic networks has become an important aspect of biology, as more knowledge is being generated, and allows studying how the system responds to the ever-changing external environment (Wittmann and Lee, 2012). Metabolic modeling aims at the quantitative understanding of the cell reaction networks to predict the intracellular dynamic behavior with reasonable precision (Kholodenko and Westerhoff, 2004; Rocha et al., 2008). Metabolic networks are largely studied, among the different biological network, and have been used in many ways, such as drug target identification, gene deletion predictions, and cellular regulatory network elucidation and industrial production (Wittmann and Lee, 2012; Jing and Alashwal, 2014).

To elucidate a better understanding of the response of systems to various environmental *stimuli*, a gathering of three different networks are optimal to have in a model: (1) Metabolic networks - system of biochemical reactions; (2) Transcriptional networks - system of the genome expression; (3) Signaling networks - system of proteins that transduce information, leading to changes in the transcriptional state of the cell (Villadsen, 2009; Wittmann and Lee, 2012). However, this work is focused only on the metabolic networks, considering it is the reconstruction of a metabolic model.

Metabolic models are representations of the full or part of the metabolism of a cell (Baart and Martens, 2009; Terzer and Stelling, 2009). They can be represented using a formal mathematical description: (1) A stoichiometric matrix is build according to the metabolic reactions; (2) Each row of the matrix represents one metabolite and each column represents one reaction; (3) Each element of the matrix is the stoichiometric coefficient of the correspondent reaction and metabolite(s) - if negative the metabolite is consumed, if positive the metabolite is produced, and if it is zero the metabolite is not produced nor consumed (Savinell and Palsson, 1992; Benner et al., 2014). This matrix allows performing several analyses to identify flux distributions, such as Flux Balance Analysis (FBA), Parsimonious Flux Balance Analysis (pFBA), Flux Variability Analysis (FVA) or others methods (Wittmann and

Lee, 2012; Benner et al., 2014). These analyses provide an alternative to the more expensive and time-consuming ”wet-lab” experimental work, by performing *in silico* simulations which have been shown to provide reliable results (Villadsen, 2009; Dias et al., 2015). These simulations can be used to predict the phenotypical behavior of the organism from growth on different substrates to changes associated to gene knockouts, providing essential information for a better understanding of the overall metabolism to identify metabolic genes that can be manipulated and to understand complex biochemical processes (Smolke, 2009; Wittmann and Lee, 2012; Hartmann and Schreiber, 2014). With the full genome sequence, it is possible to reconstruct models at the genome-scale (Wittmann and Lee, 2012).

2.3.3 *Genome-scale metabolic models*

GSM model are used to analyze and characterize their respective organisms, and to investigate their physiological characteristics or to suggest engineering strategies for improving the overproduction of a desired compound (Wittmann and Lee, 2012). GSM models are metabolic models with, usually, several hundred up to several thousand reactions and metabolites within it, leading to a more accurate representation of the cell (Benner et al., 2014).

Reconstruction of genome-scale metabolic models

Reconstructing a GSM model involves collecting information from various fields of study, from genomics and metabolomics to cellular physiology, thus it is important to retrieve as much relevant information as possible to obtain a more accurate representation of the organism (Dias and Rocha, 2015). Online databases, along with literature, are the data source that provide most of this information (Terzer and Stelling, 2009). A collection of different databases that are usually used for this purpose is shown in Table 1.

The reconstruction of a GSM model is a very complex process in which various fields of study are involved and an array of computational tools and laboratorial methods have to be applicated (Stephanopoulos et al., 1998; Thiele and Palsson, 2010). This process may take weeks to years, depending on its complexity, hence automated, or at least semi-automated tools developed with the aim of generating high-quality GSM model should be considered (Thiele and Palsson, 2010). The reconstruction of GSM models comprises hundreds of steps distributed by four stages (Thiele and Palsson, 2010; Dias and Rocha, 2015).

Each stage presented above have a plethora of intricacies, making the reconstruction of GSM model a complicated process, with different subjects of study, types information and details. To grasp the big picture, it is here presented next the most important concepts

Table 1.: Set of useful databases to reconstruct GSMMs, based on information by Jing and Alashwal (2014) and Dias and Rocha (2015).

Database	Data Type	Description	Curated	Reference
BioCyc	Genomic, metabolic	BioCyc is a group of Pathway or Genome DataBases (PGDB) that presents information on genomes and cellular networks, as well as allows powerful computational analysis and exploitation of the database.	Yes	Caspi et al. (2012)
MetaCyc	Metabolic	MetaCyc is a well-known metabolic pathway database that contains information on organism enzymes and pathways involved in primary and secondary metabolism and its associated compounds, enzymes, and genes.	Yes	Caspi et al. (2012)
BRENDA	Metabolic	Braunschweig Enzyme Database (BRENDA) is a protein function database, which contains a huge amount of enzymatic and metabolic data and is updated and evaluated by extracting information from primary literature.	Yes	Schomburg et al. (2002)
KEGG	Genomic, metabolic	Kyoto Encyclopedia of Genes and Genomes (KEGG) is a bioinformatics database that contains information such as proteins, genes, pathways, and reactions.	No	Kanehisa and Goto (2000)
BKM	Metabolic	BRENDA-KEGG-MetaCyc reactions (BKM-react) is an integrated and non-redundant database containing known biological reactions collected from BRENDA, KEGG, and MetaCyc.	No	Lang et al. (2011)
NCBI	Genomic	The National Center for Biotechnology Information (NCBI) is a collection of databases that can provide analysis, visualization, and is a source for biological data.	No	Sayers et al. (2013)
TCDB	Genomic, netabolic	Transporter Classification Database (TCDB) has a classification system for the membrane transporter proteins known as the transporter classification system.	Yes	Saier et al. (2006)
UniProt	Genomic, metabolic	Universal Protein Resource Knowledgebase (UniProtKB) is a collection of accurate and rich information on proteins. It consists of two sections: the first section contains manually annotated proteins with information extracted from literature and computational analysis (referred to as UniProtKB/Swiss-Prot) and the second section with computationally analyzed proteins to be fully manually annotated yet (UniProtKB/TrEMBL).	Yes	Apweiler et al. (2011)

of each stage based on Dias and Rocha, explaining the different aspects of bioinformatics within it (Dias and Rocha, 2015).

According to Dias and Rocha (2015), the reconstruction of GSM models can be summarized on four stages: (1) Genome annotation; (2) Assemblage of a metabolic network from the genome; (3) Conversion of the network to a stoichiometric model; (4) Validation of the metabolic model. These stages are repeated iteratively to increase the accuracy of the GSM model, and thus better represent the organism *in vivo*. As these steps require compiling information from distinct data sources, the reconstruction process involves curating the retrieved data to improve the GSM model.

Initially, the genome of the organism is retrieved from public repositories of genomic data, such as NCBI or KEGG, in which a manual curation may have been performed, however, if the genome annotation of the organism is not available, it can be performed by a series of processes. Regarding the latter case, is important to retrieve certain data for each gene namely the gene or Open Reading Frames (ORF) names, product names and Enzyme Commission (EC) numbers (if possible). Some tools use Basic Local Alignment Search Tool (BLAST) and HMMER to perform an automatic annotation of the whole genome.

Genes with scores above a user determined threshold (upper threshold) are automatically annotated and considered correct, whereas gene annotations with scores below a user defined lower threshold should be rejected and considered incorrect. The gene annotations with scores between the two thresholds should be reviewed manually to identify which metabolic genes should be integrated in the GSM model. The manual curation of these genes annotation involves determining and following a series of steps to assign a function to each gene and to determine the confidence level of that assignment, called an annotation workflow.

Genes annotated as metabolic should be identified with EC numbers, whenever possible. The EC number of each gene can be confirmed through BRENDA to eliminate possible errors due to: transfer of the EC number to other EC code, deletion or mismatch between EC number and enzymatic function. This step is crucial to the genes found between the two thresholds and therefore present in the annotation workflow.

When the gene annotation is completed, it is possible to assemble the metabolic network.

The first step on this stage is to collect the reactions promoted by the proteins encoded in the genome (called the Gene-Protein-Reaction (GPR) association). This is performed by searching databases with the protein name, EC number(s) and other identifiers (such as KEGG reaction number) identified in the previous step to find the promoted reaction. Likewise, the transport proteins, *i.e.*, proteins that transport metabolites through biomembranes, identified with a Transporter Classification (TC) number can also be used to infer transport reactions. TCDB can offer information about transport proteins. Both enzymatic and transport reactions must be added to the draft network.

The reactions that do not need enzymes or other external factors to occur, called spontaneous reactions, should be automatically inserted into the metabolic network.

Afterwards, the stoichiometry of the reactions in the network should be revised with the help of BRENDA, KEGG, and/or MetaCyc.

Then, the next step is to investigate where de reactions occur in the cell. As cell have distinct compartments (prokaryotes: cytosol, periplasmic space and extracellular space; eukaryotes: Golgi apparatus, the lysosome, the mitochondrion, the endoplasmic reticulum and the glyoxysome among several others) consecutive reactions may take place in different compartments. This step is called the compartmentalization of the model, and is determined by the localization of the enzymes. To predict the localization of the enzymes, different tools such as the ones from the PSort family, TargetP, SignalP, ChloroP or Transmembrane Helices prediction by Hidden Markov Models (TMHMM). Also, this information can be found in literature and some online databases, such as Universal Protein Resource (UniProt).

The last step of the assembly of the metabolic network is the manual curation. Since automated steps are fallible, the GSM model should be revised with the help of organism-specific databases, expert researchers and literature, including publications and textbooks. This validation allows gap-filling of the metabolic networks. The problems to fix can vary in its nature: the protein and function identifier may have inconsistencies; the addition of new organism-specific reactions can be unavailable in the queried data sources; or the assignment of reactions can have ambiguous identifiers (Dias and Rocha, 2015).

Another thing to consider revisiting is the reversibility of the reactions, determined by the standard Gibbs free energy of formation and of reactions.

This reconstruction can be supported by another curated model from closely related organisms, to fix de gaps in the network.

There are also problems with its own nature that can only be solved by a case-to-case method, therefore, with the help of the available biological and chemical databases it is hoped to build a bug free GSM model able to have the metabolic network consistent with the literature and current knowledge of the organism, using information retrieved from curated sources and from deduction.

The next stage is to convert the metabolic network into a stoichiometric matrix through the addition of constraints to the model and an abstraction, in the form of a reaction, which represents the drain of biomolecules (such as amino acids, nucleotides, lipids and others) to the formation of biomass. This reaction can be represented by the following Equation 1:

$$\sum_{k=1}^{P} c_k X_k \to biomass \tag{1}$$

in which c_k represents the stoichiometry of metabolite X_k. As this equation represents the growth rate of the organism, it should also include growth energy requirements, represented in the reaction as the depletion of Adenosine Triphosphate (ATP). Whenever wet-lab experiments cannot be performed to determine the amount of each biomolecule present in the biomass formulation, the biomass equation from a related organism may be used, though the simulation results are approximations of the correct results.

Once the metabolic model is completed with the biomass equation, all reactions can be transposed into a stoichiometry matrix. By representing each metabolite concentration, in face of the rate of its production or degradation (by each of its fluxes), with ordinary differential equations, it is obtained the following Equation 2:

$$\frac{dX_i}{dt} = \sum_{j=0}^{N} S_{ij} \cdot v_j + \mu X, \quad i = 1, ..., M \tag{2}$$

where the rate of variation of the concentration of metabolite i with time t is represented. X_i is then the concentrations of metabolite i, v_j is the rate of reaction j (*i.e.*, its metabolic flux), and S_{ij} is the stoichiometric coefficient of metabolite i in reaction j. The growth rate of the system is represented by μX.

As these models are stoichiometric, steady-state condition on the whole system are considered, meaning that all fluxes are constant, as the concentration of the metabolites, through time.

By considering this state, the rates of production and degradation of all metabolites are equal, providing the following equation:

$$S \cdot v = 0 \tag{3}$$

where v is the flux vector and S is the stoichiometric matrix, where the columns represent the reactions and the rows represent the metabolites.

This mathematical representation should be saved with Minimal Information Required In the Annotation of Models (MIRIAM) annotations into the Systems Biology Markup Language (SBML), which is the standard format for exporting GSM models (Hucka and Wang, 2003; Juty et al., 2012).

Validation of Genome-scale Metabolic Models

The GSM model should be iteratively validated by comparing computer simulations to laboratory data (Kholodenko and Westerhoff, 2004; Smolke, 2009). This assessment allows improving the model as it measures the accuracy of the model (Dias et al., 2015). The GSM model does not intend to represent all the characteristics of the *in vivo* cell, but to effectively explain certain metabolic properties under given conditions (Kholodenko and Westerhoff, 2004; Smolke, 2009).

There is an array of methods used to measure, *in vivo* and *in silico*, fluxes of the metabolites in the cell (Kholodenko and Westerhoff, 2004; Smolke, 2009). Wet laboratory experiments may involve: (1) Gas chromatography/mass spectrometry or Nuclear magnetic resonance spectroscopy and enzyme assays to monitor the metabolic response, performing a non-stationary (steady state, being the status of the sys-tem when the exchange of energy and/or matter with its environment is at a constant rate, stated by Tschoegl (2000) pulse experiment. However, this method is limited to relatively small-size central metabolisms; (2) ^{13}C flux analysis to determine the metabolic fluxes in central metabolism, using different stationary experiments under different conditions to picture a metabolic regulation; (3) Performing cell extracts to determine enzyme activities; (4) Most kinetic parameters of enzymes can be found in databases, and are usually reliable (Kholodenko and Westerhoff, 2004; Smolke, 2009).

Regarding software for measuring fluxes, there are algorithms that can determine the fluxes of the metabolic networks of the GSM model: (1) FBA identifies the flux distributions responsible for making the network in steady state. This algorithm usually uses the maximization of the biomass flux as objective function to simulate the cellular growth, or the maximization of ATP to simulate the natural objectives of the cell (Benner et al., 2014); (2) Method Of Minimization of Metabolic Adjustment (MOMA) is based on the same stoichiometric constraints as FBA, but has a relaxed display for optimal growth flux when genes are deleted. Therefore, MOMA mimics with higher accuracy the metabolic networks of a mutant cell to determine its growth than FBA. This is due to its sub-optimal flux distribution that reasult from the minimization of adjustments after the gene knockout. Consequently, MOMA performs better than an optimal algorithm like FBA (Segrè et al., 2002); (3) Regulatory On/Off Minimization (ROOM) predicts the flux of metabolic networks of cells with knockout genes such as MOMA (with the minimization of changes in the cell, minimizing the adaptation costs). However ROOM implicitly prefers high growth-rate solutions, leading to enhanced predictions (Shlomi et al., 2005). (4) pFBA is an improved approach that considers that there is a selection for the strains that minimize the production of enzymes, and therefore require the lowest overall flux in the metabolic network (Lewis et al., 2010);

(5) The FVA even further supports the suppression of non-functional metabolic reactions (Lewis et al., 2010). There are other algorithms used to predict phenotypes with different approaches as the ones above, yet FBA is the most widely used (Maranas and Zomorrodi, 2016). Although the algorithms previously presented allow performing simulations with the GSM model, which provide predictions of the phenotypical behavior of the organism on different substrates and/or gene knock-outs, there are other methodologies which can be used to analyze the GSM models, such as pathway analysis, elementary flux analysis, gene expression analysis and adaptive evolution analysis (Wittmann and Lee, 2012).

The number of reconstructed GSM model is expected to increase rapidly, as recent advances in high-throughput technologies provide substantial quantities of data. Hence, the automation of this process with computer sowftware became a requirement for the novel ME approaches. There are several software available, with different tools and characteristics, for reconstruction and validation of GSM model (Dias et al., 2015).

2.3.4 *Structure of the model*

Pathways are abstract concepts of networks of reactions, that represent part of the metabolism of the cell. This metabolism is enclousered by the frontier that separates the inside from the outside of the cell, the membrane. From the membrane the transport reactions occur, however, aside from metabolic reactions and transporters, another concept to consider is the drain, which are reactions used to unbalance the model allowing to perform simulations. Drains are abstractions of the environmental conditions, when simulating the GSM model. These simulate both the growth medium and the by-products.

The biomass equation is an abstraction of organism biomass, which simulates the drain of building blocks required for replicating the organism. These concepts are represented in Figure 2.

2.4 SYSTEMS BIOLOGY FRAMEWORKS

In this section, different software developed for reconstructing and validating GSM model are presented and compared. Several of these software are specific for a group of species, or may have peculiar and unique characteristics, turning the selection of the best framework a case-by-case scenario (Jing and Alashwal, 2014; Weber and Kim, 2016).

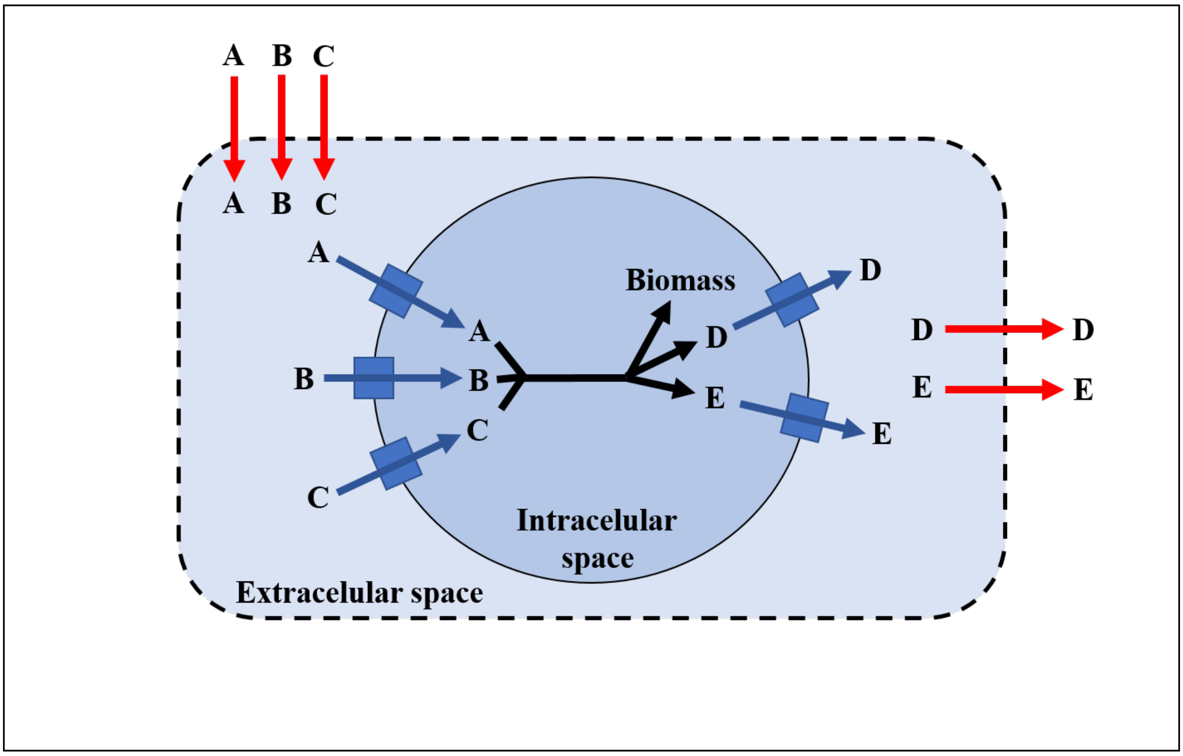


Figure 2.: Representation of the GSM model structure. The red colored arrows represent the drains, the blue arrows represent transport reactions and the black arrows represent the metabolic reactions inside the cell. Compounds A, B and C are available in the extracellular space through a drain, with a certain flux defined for the simulation. These compounds are imported inside the intracellular space through a transporter and metabolized into biomass and into compounds D and E.

2.4.1 *Software for genome-scale metabolic model reconstruction*

Since 2010, 11 high-throughput software programs were developed for the reconstruction of GSM models (Weber and Kim, 2016). In this section, the main characteristics, the advantages and disadvantages of each of the software programs able to reconstruct GSM models are presented. Nevertheless, there are other programs that may help in the reconstruction protocol, but do not provide the actual model.

Metabolic model research and development System

Released in 2011, Metabolic Model research and development System (MEMOSys) is a versatile platform for the management, storage, and development of genome-scale metabolic models. MEMOSys was implemented in Java uses the JBoss Seam framework (Pabinger et al., 2011).

ADVANTAGES

1. Support for the laborious reconstruction process with accessible intermediate versions of the GSM model to enable iterative changes, enhancing the reconstruction process;
2. It uses SBML to represent the different models using unambiguous identification of components, such as EC numbers on enzymes and KEGG Compound Database on compounds , to enable the comparison with other models (Pabinger et al., 2011).

DISADVANTAGES

1. Does not perform the enzymes annotation;
2. Does not perform the transport annotation;
3. Cannot create the biomass reaction;
4. Lacks a graphical interface for manual curation;
5. Does not provide pathway visualization;
6. Cannot create GPR rules;
7. Does not highlight metabolic dead-ends;
8. Does not reconstruct eukaryotic models;
9. It only has the feature to manually insert the compartmentalization;
10. Does not generate automatically the biomass equation (Dias et al., 2015).

Flux analysis and modeling environment

Released in 2012, "Flux Analysis and Modeling Environment (FAME) is the first web-based modeling tool that combines the tasks of creating, editing, running, and analyzing/visualizing stoichiometric models into a single program". Apart from the visible HyperText Markup Language (HTML) and Cascading Style Sheets (CSS), FAME was implemented in Personal Home Page (PHP) 5 and Java-Script. PHP was chosen because it is a fast browser that possess independent language which works well with other programs (Boele et al., 2012).

ADVANTAGES

1. No installation procedures;
2. Does not have requirement of proprietary software;

3. Supports the interpretation of results with an user-friendly environment that allows biologists to ask questions;
4. Has an incorporated simulator (Boele et al., 2012).

DISADVANTAGES

1. Does not perform the enzyme annotation;
2. Does not perform the transport annotation;
3. Inability to run locally;
4. Lacks a graphical interface for manual curation;
5. Cannot create GPR rules;
6. Reaction stoichiometry validation;
7. Does not reconstruct eukaryotic models;
8. It only has the feature to manually insert the compartmentalization;
9. Does not generate automatically the biomass equation (Dias et al., 2015).

MicrobesFlux

Released in 2012, ”MicrobesFlux is a semi-automatic, web-based platform for generating and reconstructing metabolic models for annotated microorganisms”. The front end of MicrobesFlux is written in Google Web Toolkit technology and the back end is written in Python using the Django web framework (Feng et al., 2012).

ADVANTAGES

1. Has a high-throughput metabolic models generation;
2. Customizes metabolic models drafting;
3. Does constraint-based flux analyses in steady and dynamic metabolic states (Feng et al., 2012).

DISADVANTAGES

1. Does not perform the enzyme annotation;
2. Does not perform the transport annotation;

3. Inability to run locally;

4. Lacks a graphical interface for manual curation;

5. Cannot create GPR rules;

6. Does not reconstruct eukaryotic models;

7. Does not generate automatically the biomass equation (Dias et al., 2015).

Pathway Tools

Released in 2012, "The Pathway Tools is a reusable, production-quality software environment for creating a type of model-organism database called a PGDB". The PGDB gathers knowledge about genes, the respective proteins and metabolic network that is inserted into, and the genetic network of the organism (Karp et al., 2002).

ADVANTAGES

1. Visualizes and interacts with contents of PGDB;

2. Performs complex queries, symbolic computations, and data mining on the contents of PGDB;

3. Able of Web publishing in PGDB (Karp et al., 2002).

DISADVANTAGES

1. Does not perform the enzyme annotation;

2. Does not export model in SBML format;

3. Does not predict compartmentalization;

4. Cannot create GPR rules;

5. Does not generate automatically the biomass equation (Dias et al., 2015).

Comparative reconstruction

Released in 2014, Comparative Reconstruction (CoReCo) is a computational approach for comparative metabolic reconstruction. It generates reliable GSM model semi-automatically for a series of organisms, with minimal amount of curation. This comparative reconstruction can help to refine already existing metabolic models when genomes of related species have been sequenced. It also can help in the reconstruction of species with distant but extensively studied model species (Pitkä nen et al., 2014).

ADVANTAGES

1. Reconstructs GSM models of a large number of related species;
2. Fills automatically the gaps of GSM models;
3. Reconstructs GSM models with high accuracy (Pitkä nen et al., 2014).

DISADVANTAGES

1. Does not perform the transport annotation;
2. Does not predict compartmentalization;
3. Lacks a graphical interface for manual curation;
4. Does not provide pathway visualization;
5. Cannot create GPR rules;
6. Does not highlight metabolic dead-ends;
7. Does not reconstruct eukaryotic models;
8. Does not generate automatically the biomass equation (Dias et al., 2015).

Reconstruction, analysis and visualization of metabolic networks

Released in 2013, Reconstruction, Analysis and Visualization of Metabolic Networks (RAVEN) toolbox is a software that allows reconstructing genome-scale models automatically. This toolkit allows analyzing, simulating and visualizing GSM model within Matrix Laboratory (MATLAB) (Agren et al., 2013).

ADVANTAGES

1. Reconstruct eukaryotic models;
2. Reconstructs GSM model based on the orthology between its proteins sequences and the target model proteins sequences;
3. Fills automatically the gaps of GSM models;
4. Has an incorporated simulator (Agren et al., 2013).

DISADVANTAGES

1. Does not perform the transport annotation;
2. Requires commercial software;
3. Lacks a graphical interface for manual curation;
4. Does not provide pathway visualization;
5. Cannot create GPR rules;
6. Does not validate reaction stoichiometry;
7. Does not generate automatically the biomass equation (Agren et al., 2013; Dias et al., 2015).

Model SEED

Released in 2010, Model SEED is a web-based resource for high-throughput generation, optimization and analysis of GSM models. It is built upon the genome annotation provided by SEED (Henry et al., 2010).

ADVANTAGES

1. Performs the enzyme annotation;
2. Perform the transport annotation;
3. Predict compartmentalization;
4. Export model in SBML format;
5. Provides pathway visualization;
6. Creates GPR rules (Henry et al., 2010).

DISADVANTAGES

1. Inability to run locally;
2. Lacks a graphical interface for manual curation;
3. Does not highlight metabolic dead-ends;
4. Does not reconstruct eukaryotic models.
5. It only has the feature to manually insert the reactions stoichiometry validation (Dias et al., 2015).

SuBliMinaL Toolbox

Released on 2011, the SuBliMinaL Toolbox facilitates the reconstruction process with features such as generating draft reconstructions, adding transport reactions and a biomass function, and many more. It is written in Java, yet it has third-party dependencies Swainston et al. (2011).

ADVANTAGES

1. Performs the enzyme annotation;
2. Perform the transport annotation;
3. Predict compartmentalization;
4. Has a graphical interface for manual curation;
5. Provides pathway visualization;
6. Creates GPR rules;
7. Merges models;
8. Determines metabolite protonation state;
9. Does pre-draft reconstructions Swainston et al. (2011).

DISADVANTAGES

1. Relies on various third-party packages to run all the features;
2. Does not highlight metabolic dead-ends (Swainston et al., 2011; Hamilton and Reed, 2014).

Genome-scale metabolic model simulation, reconstruction and visualization

Released in 2012, GEnome-scale Metabolic model Simulation, Reconstruction and Visualization (GEMSiRV) is a user-friendly software able to reconstruct, simulate and visualize GSM model. It is written in Java and it uses the solver GNU Linear Programming Kit (GLPK) (Liao et al., 2012).

ADVANTAGES

1. Has a graphical interface for manual curation;
2. Provides pathway visualization;

3. Creates GPR rules;

4. Highlights metabolic dead-ends;

5. Has an incorporated simulator;

6. Generates robust images for presentations (Liao et al., 2012).

DISADVANTAGES

1. Does not perform the enzyme annotation;

2. Does not perform the transport annotation;

3. Does not predict compartmentalization;

4. Does not validate reaction stoichiometry;

5. Does not generate automatically the biomass equation Liao et al. (2012).

Metabolic models reconstruction using genome-scale information

Released in 2015, *merlin* is a friendly user software that help the GSM model reconstruction. Is an open-source application implemented in Java and built on top of the Artificial Intelligent workBench (AIBench) software development framework (Dias et al., 2015).

ADVANTAGES

1. Performs the enzyme annotation;

2. Perform the transport annotation;

3. Predict compartmentalization;

4. Export model in SBML format;

5. Runs locally;

6. Does not require software;

7. Has a graphical interface for manual curation;

8. Provides pathway visualization;

9. Creates the GPR rules;

10. Highlights metabolic dead-ends;

11. Reconstructs eukaryotic models (Dias et al., 2015).

DISADVANTAGES

1. Does not generate automatically the biomass equation (Dias et al., 2015).

Most of these softwares are still growing and are still being implemented with innovative tools able to diminish their weaknesses, including *merlin*, with its great features. *merlin* is one of the best choices for a reconstruction from scratch or based on a relative close GSM model, being for prokaryotic or eukaryotic organism. Furthermore, this work is executed with the help of the staff responsible for producing and maintaining *merlin*, which may give support when needed. Therefore, in this work it is going to be used *merlin* as the GSM model reconstruction tool for the reason that it can make most of the step of this reconstruction automatically, however still being able to manually cure the model contents - making this semi-automated software a great tool to facilitate and quicken all of the process.

One of the major aspects of *merlin* is helping in the curation of the model, by updating the KEGG pathways maps, which allows to understand better the status of the current reactions integrated in the model or the missing link in the network. *merlin* colors each reaction in the map with one of four color, which have different meanings.

When reactions are colored in green, all metabolites involved in them are being consumed and produced in the network and the EC number promoting the reaction is available in the model. If the EC number was unavailable in the model, the reaction would be colored in blue. The third case is when one or more compounds of the reaction are not being synthesized or consumed (colored red in the maps). These compounds are called dead ends. The fourth is when a reaction is connected to an unconnected reaction (colored cyan in the maps). This mechanism eases the detection and traction of dead end metabolites, as a chain sequence of cyan colored reactions will eventually lead to an unconnected reaction.

All of these four types of reactions are exemplified in Figure 3.

As mentioned before, a functional GSM model, must have each compound metabolized in equal rates of production and degradation. Hence, reactions associated with dead-end metabolites cannot be part of the model. Therefore, it is important to detect these reactions, conveniently hinted by the cyan colored reactions. This system was one of the major reasons why *merlin* was favorable for this work.

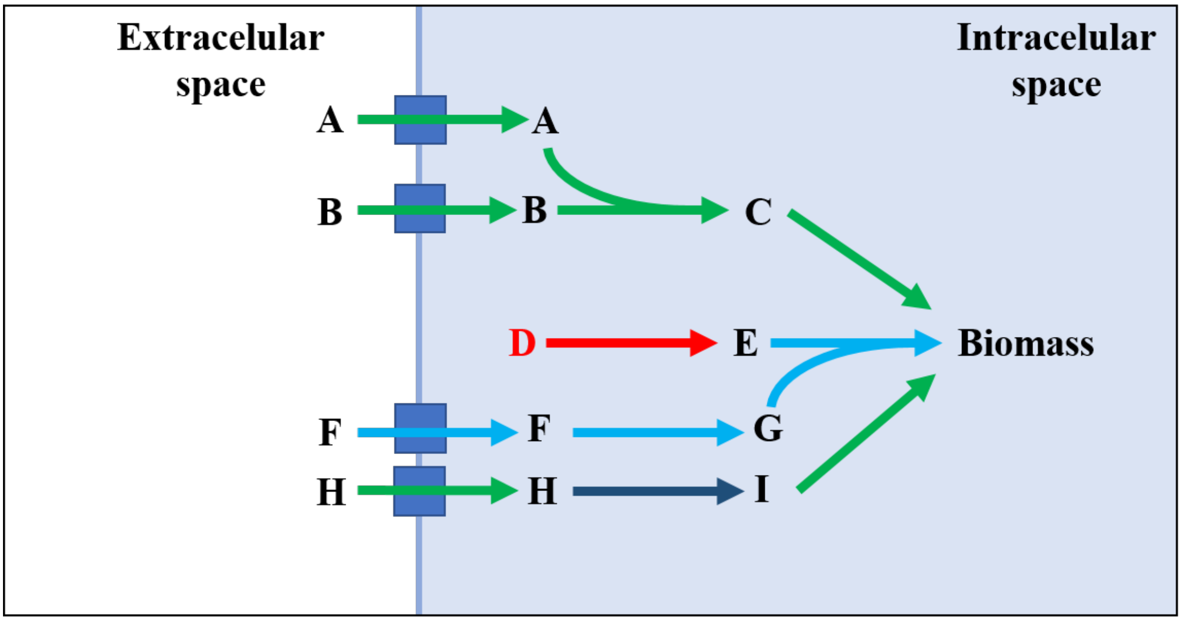


Figure 3.: Scheme of the four different status of reactions in the GSM model. The green colored arrows represent reactions associated with compounds that are always consumed or produced by other reactions. The blue colored reactions not catalyzed by the enzymes indicated in the pathway. The red colored arrows represent reactions with dead-end metabolites. The cyan colored arrows represent reactions affected indirectly by dead end metabolites. While the compounds A, B, C, H and I and correctly metabolized, the compounds E, F and G cannot be because they are associated, directly or indirectly, with the dead-end metabolite D. Metabolite D is a dead-end because there is no reaction to synthesize it.

2.4.2 *Software for genome-scale metabolic model simulation*

In silico metabolic network analysis software can be used to perform simulations with GSM model, to predict the phenotype of the modeled organism (Weber and Kim, 2016). There is a plethora of standalone tools, toolbox-based tools, and web-based tools that are able to perform this analysis (Jing and Alashwal, 2014). Reportedly, there were at least 21 different tools in 2014, and each of these are different in terms of the features and usability (Jing and Alashwal, 2014). Each tool has its own advantages as well as disadvantages that can become a limitation, thus for each particular research project should select the one that allows reaching its objective (Jing and Alashwal, 2014). For this work, Optflux was chosen, as it fulfilled the project requirements as shown below.

OptFlux

Optflux is an open-source user-friendly software, implemented in Java, that applies steady-state stoichiometric models to study the phenotype of microorganisms, under different environmental and genetic conditions (Rocha et al., 2010). Optflux is also the first tool that aims to be the reference platform for the ME community. This platform makes available the GSM models from both academia and industry, to their further development and exploitation. It is also a modular software, meaning that facilitates the addition of specific features by computer scientists, and is compatible with the SBML (Rocha et al., 2010).

Along with the tools researched and presented by Jing and Alashwal (2014), Optflux is the only tool that has the three metabolic network analysis algorithms presented earlier (FBA, MOMA and ROOM) plus two other algorithms (Optknock, to identify best set of genes to be knockout in order to increase the production of metabolites; and OptGene, an extension of OptKnock which utilize genetic algorithm to increase the prediction capability) (Rocha et al., 2010; Jing and Alashwal, 2014). It also provides a build-in visualization to facilitate the interpretation of the results, a graphical user interface and many other features, making it of the best choices to simulate accurately the GSM model (Rocha et al., 2010; Jing and Alashwal, 2014).

Transport proteins annotation and reactions generation

Transport Proteins Annotation and Reactions Generation (TRIAGE) is a tool developed in Java with MySQL relational databases that detects and classifies potential transport proteins upon analyzing genes that encode transmembrane proteins (Dias and Rocha, 2017). The TCDB classification system for transport proteins are used in this work, for that fact that this database is the most comprehensive one, in storage of the manually annotated cellular transport systems (Dias and Rocha, 2017). This tool was chosen because of to the lack of good transporters annotation (Dias and Rocha, 2017). Also, *merlin* has TRIAGE included in its core.

This tool detects and classifies potential transport proteins with a TC numbers, which are identifiers similar to EC numbers though also including phylogenetic information. TC numbers are associated with proteins that transport one or more substrates, in a certain direction, using a certain mechanism and most of the times associated to a single organism. Hence, these proteins should not be directly classified with TC numbers from homology alone, unlike the classification with EC numbers.

To detect potential transport genes, initially TRIAGE identifies genes with transmembrane helices using TMHMM or Phobius (Krogh and Sonnhammer, 2001; Käll and Sonnhammer, 2007).

These tools can predict transmembrane domains in amino acid sequences. Genes' amino acid sequences with at least one transmembrane helix are considered potential transport proteins. Then TRIAGE compares these potential transport systems to all the proteins available in TCDB, using the Smith-Waterman algorithm to perform local alignment (Smith and Waterman, 1981).

TRIAGE's algorithm creates new transport reactions, associated with the potential transport systems. This algorithm is similar to the one used to annotate enzymes, requiring an α value leverage the frequency and the taxonomy and a cut-off threshold.

3

METHODS AND DEVELOPMENT

In this chapter, methods used in both the *in silico* and the *in vivo* approaches will be presented. *merlin* will be used to reconstruct the GSM model of *N. europaea*, as mentioned before. Through an iterative process, with validation from data obtained through wet-lab experiments, the model is going to be increasingly improved until these results match with the simulations performed with the model.

The methodology for reconstructing this GSM model, as described in the previous chapter are shown in Figure 4. First, the GSM model draft has to be assembled and then a series of iterations to validate the model, comparing simulations to wet-lab experiments, are performed.

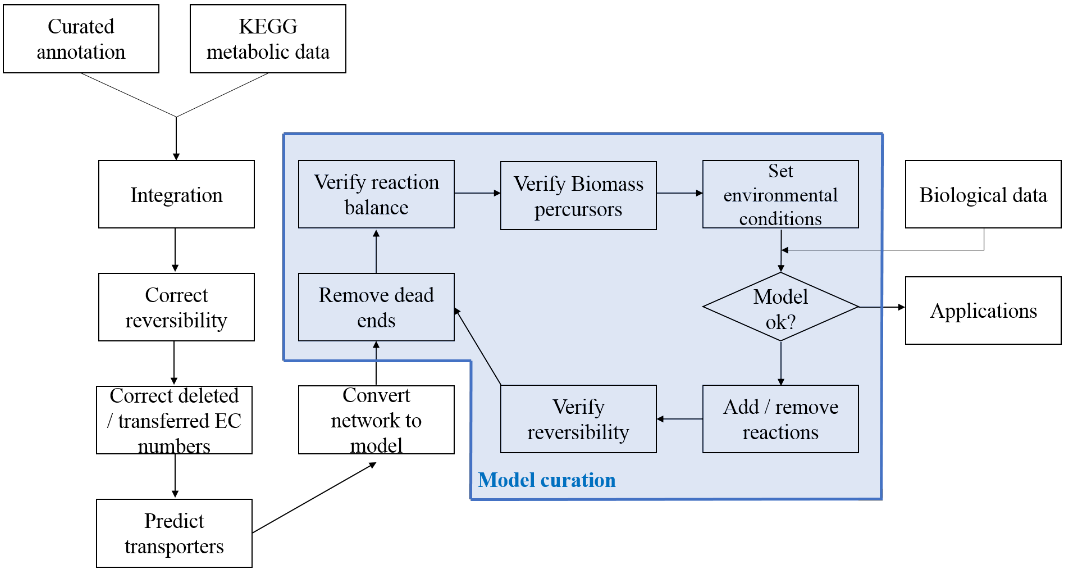


Figure 4.: Scheme of the methodology used to reconstruct the GSM model of *N. europaea* based on (Dias et al., 2014).

Each step of this process will presented in detail in the following sections.

3.1 RECONSTRUCTION OF THE DRAFT MODEL

The first step to reconstruct the draft GSM model is the genome annotation. The genome sequenced by Chain and Whittaker (2003) was used as the base of this reconstruction.

3.1.1 *Finding organisms for comparison*

This process involved determining a taxonomically close organism for which a GSM model was already reconstructed, thus accelerating the process of reconstructing a model for *N. europaea*. A phylogenetic tree was built using Multiple Sequence Comparison by Log-Expectation (MUSCLE), with the nucleotide sequence of 16S Ribosomal Ribonucleic Acid (rRNA) of each bacteria on three different GSM models databases (`http://darwin.di.uminho.pt/models`, `http://systemsbiology.ucsd.edu/InSilicoOrganisms/OtherOrganisms` and `http://www.maranasgroup.com/models.htm`).

Another method to find organisms for comparison consists in searching in the genus, for species with a high percentage of curated data. Each species of the genus *Nitrosomonas* was assessed in UniProt to find the better organism for this purpose (*i.e.*, with the highest number of reviewed proteins). The organism with the most curated genes was selected.

3.1.2 *merlin automatic annotation*

The first step was to use merlin's remote BLAST to compare the genome of *N. europaea* to all sequences available in the non-redundant (nr) database. *merlin* annotates each gene balancing two factors: the frequency and the taxonomy of the homologous genes' functions. Both of these factors are calculated independently. However, a third parameter (the α value) is used to leverage the weight of each of these scores, according to Equation 4:

$$Score = \alpha \times Score_{frequency} + (1 - \alpha) \times Score_{taxonomy} \tag{4}$$

3.1.3 *Determining the α value*

The first stage was to determine the α value, to annotate correctly the majority of the genes automatically.

Initially, a random sample of 50 genes (10 samples of 5 genes) was selected. Each sample was composed of genes with scores from 0 to 1, with intervals of 0.1, providing for an uniformly random sample of genes throughout the score range.

The 50 genes' annotation was manually curated and how many of the number of correct *merlin* automatic annotations was determined. The manual curation was performed with the standard α value of 0.5, as it is the mean between 0 and 1.

The automatic annotation of each gene by *merlin*, can be classified accordingly to *truthfulness* and *status*. *Truthfulness* indicates whether *merlin*'s automatic annotation was correct, when compared to the manual curation. The *status* indicates if the automatic annotation was accepted or rejected by *merlin*, by setting a threshold to reject annotations. For instance, a gene whose annotation is correct (*truthfulness*) and accepted (*status*) means that the automatic annotation is the same as the manual annotation and its score is above the threshold. The combination of these classifications provides six classes:

1. Correctly annotated and Accepted (CA);
2. Incorrectly annotated and Accepted (IA);
3. Correctly annotated and Rejected (CR);
4. Incorrectly annotated and Rejected (IR);
5. Non-annotable and Accepted (NA);
6. Non-annotable and Rejected (NR).

The Non-annotable (NA and NR) are genes with no manual annotation, retrieved from information gathered from databases and literature.

The second step was to verify which α had better overall *Accuracy* (described in Equation 5). To ease data analysis, a confusion matrix based of each instance of the α value, with the following four classifications: (1) True Positive (TP), (2) True Negative (TN), (3) False Positive (FP) and (4) False Negative (FN), was created. TN in the confusion matrix are NR and not IR genes. The latter genes annotation should be revised, as these may yet provide useful information for the model, whereas the former are non-metabolic genes. The classification of genes for the *Accuracy* confusion matrix are presented in Table 2.

A confusion matrix was assembled for each α ranging between 0 and 1 with 0.1 increments. However, annotation *status* are as important as its *truthfulness*, thus for each α, changes in the annotations threshold between 0.1 and 0.9 with 0.1 increments were also considered.

Table 2.: Classification of genes on the Accuracy confusion matrix, based on their description.

Confusion matrix classification	Genes description	Observations
TP	CA	automatically integrated into the model
TN	NR	automatically discarded from the model
FP	IA + NA	
FN	IR + CR	

After calculating the *Accuracy* for each α-threshold pair, the average of each α's *Accuracy* was determined.

$$Accuracy = \frac{TP + TN}{Total} = \frac{CA + NR}{Total} \tag{5}$$

3.1.4 *Setting the thresholds*

The *Accuracy* indicated *which* α had the most CA and NR genes, overall. The thresholds indicate *which* annotations should be revised . Thus, the third step was to set a lower and an upper threshold. As shown in Figure 5, the lower threshold sets a value for discarding genes as metabolic. Whereas, the upper threshold separates genes can be automatically integrated into the model. Genes with scores in between the thresholds are should be manually annotated.

The lower threshold was calculated by the NPV, which is described in Equation 6. A higher NPV allows automatically discarding genes with (NR) annotations. For calculating this parameter, TN are NR genes and TP are CA and IA genes. Conversely, FP are NA and FN are IR and CR genes. The classification of genes for the lower threshold confusion matrix are presented in Table 3.

Table 3.: Classification of genes on the lower threshold confusion matrix, based on their description.

Confusion matrix classification	Genes description	Observations
TP	CA + IA	
TN	NR	Pretended to be automatically discarded from the model
FP	NA	
FN	IR + CR	

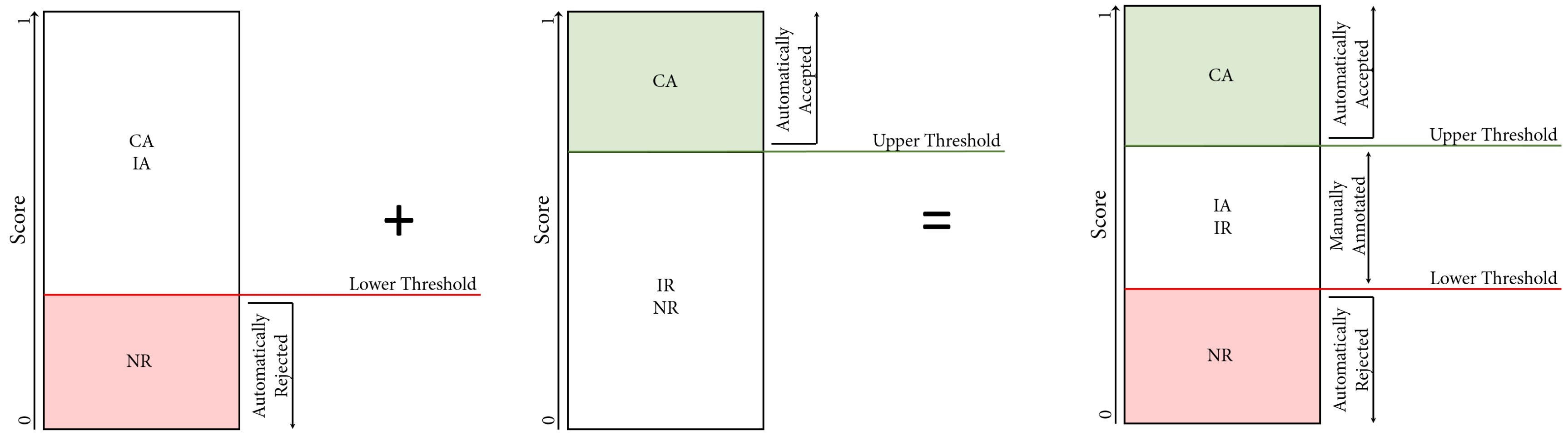


Figure 5.: Scheme representing how *merlin* would *ideally* distributes genes, by score. (Right to Left) The first panel demonstrates how the lower threshold separates the NR genes from all of the other ones - the red area is where the genes are automatically rejected from the model; the second panel demonstrates how the threshold upper separates the CA genes from all of the other ones - the green area is where the genes are automatically accepted from the model; and the third panel shows how the two thresholds combined separate the rest of the genes (IA and IR) from the automatically accepted and reject genes - the white area, in this last panel, is where the genes are to be manually curated. Note that this scheme represents an ideal scenario where Negative Predictive Value (NPV) and Precision would be 1, meaning that there were no NA, IR or CR genes in the first panel, and IA, NA or CR genes in the second panel. In this scenario, both thresholds combined would accept all the corrected genes, reject all the Non-annotable genes and let the user manually annotate all the incorrect genes (IA + IR). The highest NPV and Precision are wanted to be chosen, to reach this scenario as much as possible.

Table 4.: Classification of genes on the upper threshold confusion matrix, based on their description.

Confusion matrix classification	Genes description	Observations
TP	CA	automatically integrated into the model
TN	IR + NR	
FP	IA + NA	
FN	CR	

$$NPV = \frac{TN}{TN + FN} = \frac{NR}{NR + IR + CR} \tag{6}$$

The upper threshold was calculated by the *Precision* metric, which is described in Equation 7. A higher *Precision* allows automatically annotating genes (CA). TP are CA genes and TN are IR and NR genes. Conversely, FP are IA and NA genes and FN are CR genes. The classification of genes for the upper threshold confusion matrix is presented in Table 4.

$$Precision = \frac{TP}{TP + FP} = \frac{CA}{CA + IA + NA} \tag{7}$$

This information is concisely presented in Table 4.

The upper and lower *thresholds* are selected after selecting the best α (highest). Selecting the highest NPV and *Precision*, allows decreasing the number of genes to be manually curated.

3.1.5 *Setting a new metric*

The ideal NPV and Precision values are 1, as both threshold separate correctly the genes in question with 100 % efficiency. The best *thresholds* selected for this projects were 0.2 and 0.8, lower and upper respectively, for an α of 0.0.

However, the number of genes to be manually annotated was overwhelming (around 50% of the genes of the genome). Hence, a new metric that would take into compromise *Accuracy* to the number of genes to be manually curated was implemented. This metric is described in Equation 8 and the higher it gets, the more CA and NR genes (automatically accepted and rejected genes, respectively) whilst decreasing the number of genes to be curated.

$$\begin{aligned} x = \frac{Accuracy}{\text{\% of genes to annotate}} = \frac{Accuracy}{\frac{\text{No. of genes to annotate}}{Total}} = \frac{Accuracy \times Total}{\text{No. of genes to annotate}} = \\ = \frac{\frac{TP+TN}{\cancel{Total}} \times \cancel{Total}}{\text{No. of genes to annotate}} = \frac{TP+TN}{\text{No. of genes to annotate}} = \frac{CA+NR}{\text{No. of genes to annotate}} \end{aligned} \qquad (8)$$

3.1.6 *Manual annotation*

Phylogenetically close organisms are more prone to have similar protein structure-function than distant ones. From this premise, an annotation workflow based on the phylogeny of some organisms relatively close to *N. europaea*, was developed The phylogenetic distance between *N. europaea* and other organisms will influence the curation confidence level. Likewise, the curation status of the homologous gene was also taken into account for the confidence level. In summary, manually curated homologous genes from organisms phylogenetically close to *N. europaea* have a greater confidence level.

This selection was of paramount importance, as this workflow emphasizes first on the gene's curation status and afterwards on the organisms of comparison (*curation over phylogeny*).

Initially, a BLAST search against Swiss-Prot (curated) was performed. Using the hit list from BLAST, *merlin* annotates automatically each gene, by selecting the EC number with the highest score according to Equation 4. The level of confidence was assigned to the gene according to which organism the homologous genes belong to, after curation.

If none of the four organisms was available in the hit list, *merlin* annotation was accepted and the confidence level was lowered. In Table 5, the confidence levels of the organisms are presented.

Table 5.: Confidence level of homologues found in Swiss-Prot.

Species	**Confidence Level**
Nitrosomonas europaea **ATCC 19718**	A
Nitrosomonas europaea	B
Nitrosomonas eutropha	C
Escherichia coli	D
Other organism	E

Genes without homologies to Swiss-Prot entries were annotated against TrEMBL, which is a non-curated database. The resulting annotations will therefore have lower confidence levels. Regarding these genes, the number of hits with EC numbers associated are scarce, thus EC numbers automatically selected by *merlin* are meaningless most of the times.

The same methodology (giving importance to phylogeny) was used, thus the order of the confidence levels remains the same for the four organisms, except on *N. europaea* ATCC 19718, as TrEMBL includes all entries for this organism, which unbalances merlin's scorer. Yet, it was still used as a reference to compare with the function of other homologous genes.

The annotation and the respective level of confidence assignment was performed by comparing the homologous genes' functions. Although EC numbers were scarce in the hit list, the function description allowed inferring these with help of other databases. The number of homologous genes with the same function and EC number of these three organisms: (*N. europaea* ATCC 19718, *N. europaea* and *N. eutropha*) dictated the level of confidence of the annotation. The levels of confidence are clearly represented in Table 6.

This workflow will always try to annotate the function and EC number of the gene, except when is not possible (when it is an uncharacterized protein, or a non-metabolic annotation). These cases are described before as NA for their nature of not being possible to infer a EC number in any way.

Table 6.: Confidence level of homologues found in TrEMBL.

	Nitrosomonas europaea	*Nitrosomonas eutropha*	**Confidence level**
Case 1	Equal	Equal	F
Case 2	Equal	Different	G
Case 3	Different	Equal	H
Case 4	Different	Different	I

For the manual curation, all genes classified between both thresholds, of the case study organism and respective homologous genes were carefully revised.

Their function and EC number proposed by UniProt (Swiss-Prot and TrEMBL) were confirmed using databases such as Expert Protein Analysis System (ExPASy) (including ENZIME, PROSITE and High-quality Automated and Manual Annotation of Proteins (HAMAP)) and BRENDA. However, it is important to note that some of the manual annotations may be changed in further steps of the reconstruction of the GSM model, and for different reasons that may lead to an annotation associated with a different function and EC number.

The Figure 6 describes all of the workflow, simplified, without depicting all of the confidence levels.

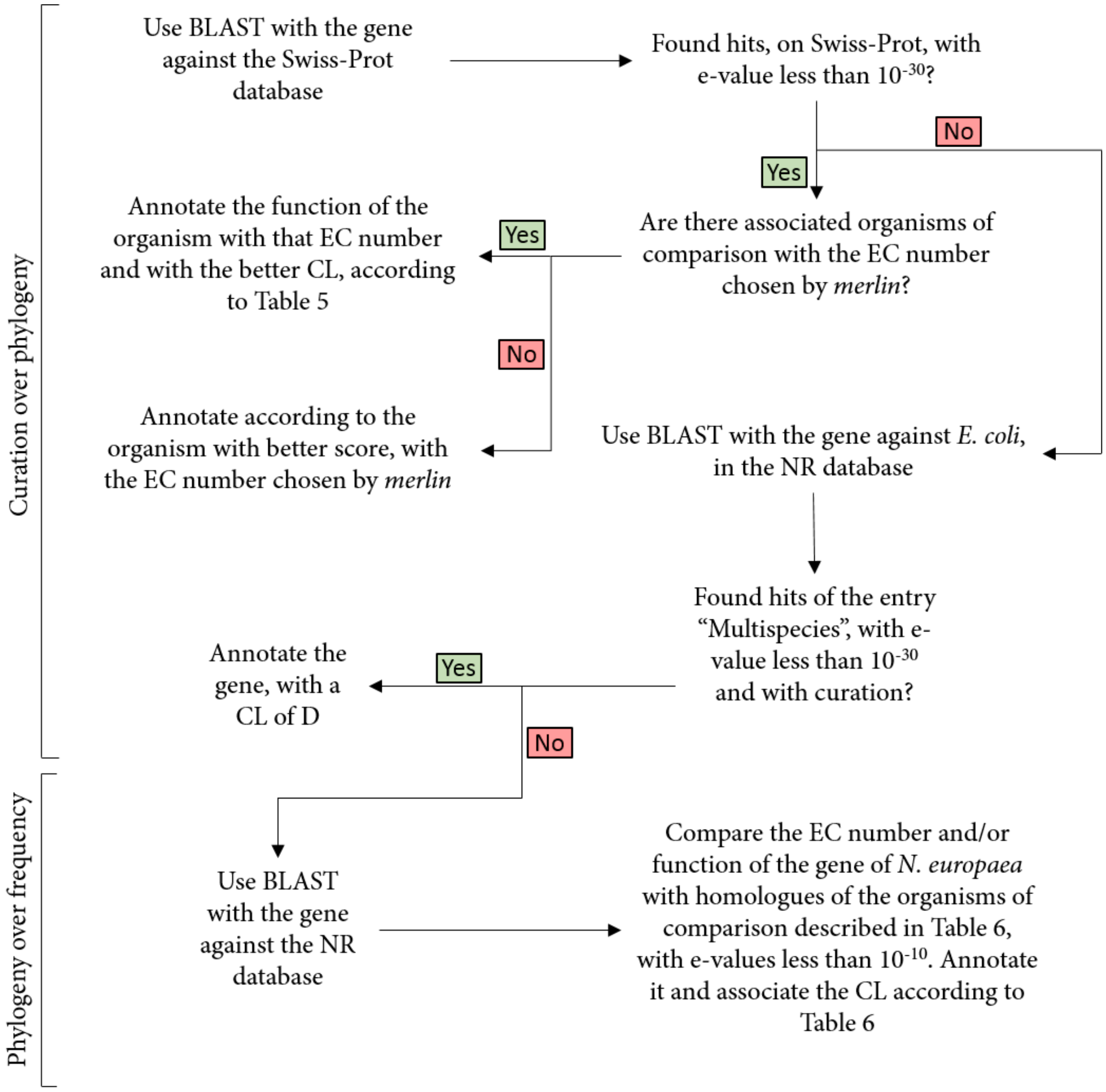

Curation over phylogeny
Use BLAST with the gene against the Swiss-Prot database
Found hits, on Swiss-Prot, with e-value less than 10^{-30}?
No
Yes
Are there associated organisms of comparison with the EC number chosen by *merlin*?
Yes
Annotate the function of the organism with that EC number and with the better CL, according to Table 5
No
Annotate according to the organism with better score, with the EC number chosen by *merlin*
Use BLAST with the gene against *E. coli*, in the NR database
Found hits of the entry "Multispecies", with e-value less than 10^{-30} and with curation?
Yes
Annotate the gene, with a CL of D
No
Phylogeny over frequency
Use BLAST with the gene against the NR database
Compare the EC number and/or function of the gene of *N. europaea* with homologues of the organisms of comparison described in Table 6, with e-values less than 10^{-10}. Annotate it and associate the CL according to Table 6

Figure 6.: Workflow of the manual annotation.

3.1.7 *Integration with the model*

The next step was to integrate the annotation in the model database. Hence, *merlin* retrieved all metabolic information from KEGG, namely enzymes, pathways, metabolites and reactions. Then, *merlin* uses these data, together with the annotation, to assemble the draft network. This way, it was possible to use KEGG pathways as a template to ease this reconstruction.

3.1.8 *Correct the reversibility of reactions*

The reversibility of the reactions in the model was automatically corrected using *merlin*. However, some reactions were needed to be manually confirmed by comparing with the direction exhibited in the KEGG pathways.

Other tools used encompass MetaCyc and eQuilibrator (an web interface for thermodynamic analysis of biochemical reactions), although occasionally it was necessary additional evidence by other sources (Caspi and Karp, 2016; Flamholz and Milo, 2012).

3.1.9 *Correct EC numbers*

Using different databases as data sources for the annotation can lead to contradictions because EC numbers may have been updated since they were firstly described, may have been deleted or transferred to other EC numbers. EC number entries not available in KEGG were manually updated, by inspecting carefully their function using other databases, such as BRENDA or ExPASy.

3.1.10 *Predict transporters*

TRIAGE was used to predict transporter proteins. *merlin*'s default α value of 0.3 and cut-off threshold of 0.2 were used, for this performance. The alignment results were used to determine, for each transport system, which metabolites are transported, the transport type and direction, using a workflow presented in Figure 7.

Then, the transport reactions were directly integrated into the model.

3.1.11 *Convert the network to model*

With the reactions set, these were integrated as a stoichiometric matrix, as described in the previous chapter. This matrix represents the model fluxes, with all the metabolites

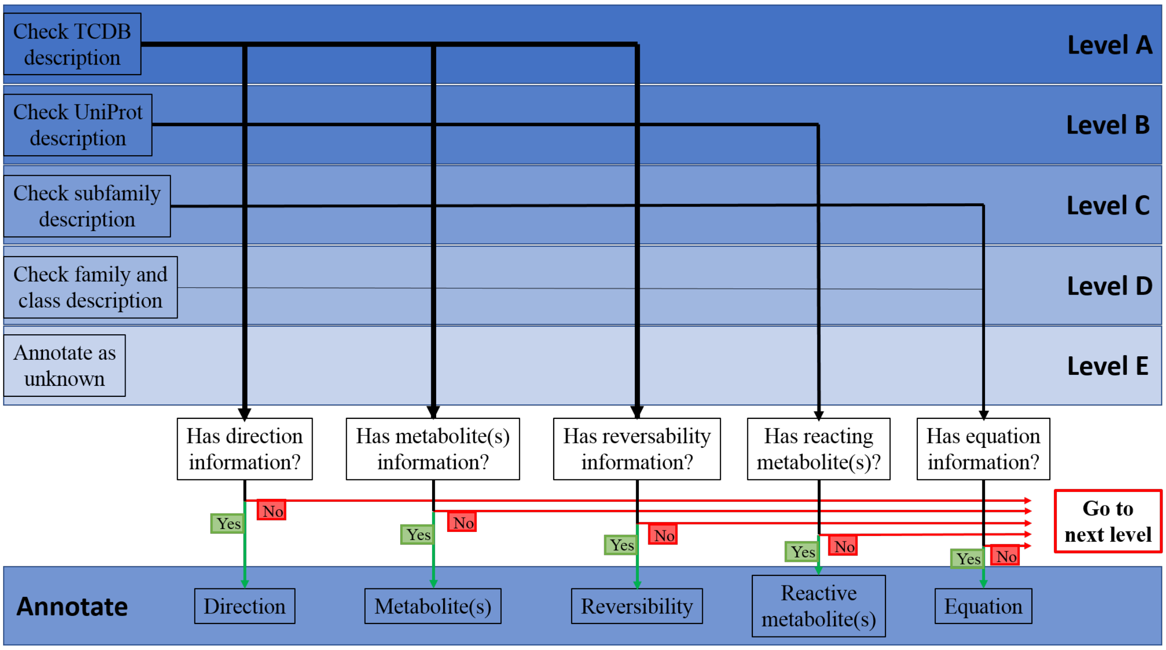

Check TCDB description
Level A
Check UniProt description
Level B
Check subfamily description
Level C
Check family and class description
Level D
Annotate as unknown
Level E
Has direction information?
Has metabolite(s) information?
Has reversability information?
Has reacting metabolite(s)?
Has equation information?
Yes
No
Go to next level
Annotate
Direction
Metabolite(s)
Reversibility
Reactive metabolite(s)
Equation

Figure 7.: Workflow for the characterization of transport systems. Each characteristic such as direction, metabolites involved, the reversibility and its equation are annotated through the TCDB or the UniProt.

and reactions in it, associated with their constraints, and it formulates biomass as a set of reactions associated by GPR rules.

3.2 CURATION OF THE GENOMIC-SCALE METABOLIC MODEL

The model curation, is an iterative process, that constantly updates the GSM model. Each iteration comprises several steps that improve the model qualitatively and quantitatively.

3.2.1 *Remove dead ends*

Metabolic pathways may be incomplete when some compounds that should be produced cannot be synthesized, *i.e.*, dead-end metabolites. This might happen due to the absence of one or more reactions in the network. The problem is exemplified in Figure 3, at the previous chapter.

If the dead end metabolite is only synthesized, all reactions that can potentially consume it have to be assessed and *vice-versa*. This assessment involves verifying missing EC numbers and incorrect annotations.

There are several cases in which the integration of reactions had to be performed without gene associations:

1. Some metabolic functions are described by Metacyc to exist in the organism, though without gene associations;
2. Some reactions were imported from *Escherichia coli* GSM model iAF1260;
3. Evidences were found in literature regarding the production of specific metabolites, though no EC numbers were available in the annotation or databases;

3.2.2 *Reactions balance*

The stoichiometric balance of every reaction in the GSM model must be ensured as, in a steady-state, every metabolite has to be synthesized and consumed at the same rate. *merlin* has a tool that allows to detect unbalanced reactions in the model. KEGG and MetaNetX were used to assess the reason for the unbalancing of the reactions and to fix these issues (Martin and Pagni, 2016). Often, the problem was associated with generic compounds that have a repetitive monomers. In these cases, the problem was solved by adjusting the stoichiometry of such monomers in the reactants or in the products of the reaction. Other common problem lied in reactions that have unspecific acceptors and donors. This was usually solved by searching in databases and literature for which specific compound the organism uses. Still, most reactions were unbalanced by a single proton and for many of these reactions MetaNetX was used to correct the reactions.

3.2.3 *Verify biomass precursors*

Seven different entities were considered biomass precursors in this GSM model: proteins, DNA, Ribonucleic Acid (RNA), lipid, carbohydrate, inorganic ions and cofactors. Each of these entities are formed by a set of metabolites identified as their basic elements, for instance, e-Protein (average protein) is composed by aminoacids. Each of these entities that are presented here, and all of their compounds, were carefully analyzed and described in detail to develop a precise GSM model.

Another important components incorporated in the biomass is ATP, though in this case it represents the amount of energy required to biosynthesize one gram of biomass.

When compiling the biomass composition, the first data source to be considered was the experimental data acquired with this work, as the biomass used in the experiments is a result of a growth in chemostat and which will be used in the validation of this GSM model. The second was the data retrieved from literature, though not being as accurate as the experimental data from this work as such experiments use other strains or different environmental conditions. Finally, *E. coli* iAF1260 GSM model was used as third data source for being a extremely detailed and studied Gram-negative bacteria.

Average protein composition

The amount of Protein per gram of biomass was determined experimentally, as well as the amount of each aminoacid in it.

Average DNA composition

The amount of DNA per gram of biomass was determined experimentally. The contribution of each deoxyribonucleotide was estimated from the genome sequence, using a tool developed for that purpose available in *merlin*.

Average RNA composition

The amount of RNA per gram of biomass was determined experimentally. Likewise, the contribution of each ribonucleotide was estimated from the genome sequence, using a tool developed for that purpose available in *merlin*.

Average lipid composition

The amount of lipids per gram of biomass was also determined experimentally. And it composition was solely based on literature.

Average Carbohydrate composition

The carbohydrates composition was based in *E. coli*'s iAF1260 GSM model, as no information concerning this macromolecule was found for *N. europaea*. The amount of carbohydrates was experimentally determined and is presented in that section.

Average Cofactors composition

The composition of the cofactors were based on literature. The universal cofactors were integrated into the model, as well as conditional ones, if those were the case for this organism. The amount of these were based in *E. coli*'s iAF1260 GSM model.

Average Inorganic ions composition

Regarding inorganic ions, their composition is based on the medium used for *N. europaea* growth experiments. The amount of these ions in the biomass equation was based on *E. coli* GSM model iAF1260. The quantities of iron and copper were determined experimentally. The quantities of ammonium and orthophosphate were based on literature, and sulfate was based on the relative quantity in *E. coli* GSM model iAF1260.

Lipopolysaccharide composition

Although this macromolecule is present in almost every Gram-negative bacteria, it is not available in this GSM model. However, this molecule is a complex structure constructed from other two types of molecules (carbohydrates and lipids), which were quantified as separated macromolecules in this work. Thus, even though this entity was not integrated into the model, lipopolysaccharide contents were taken into account in the model.

Finalizing the model

After manually curating the GSM model until it provides reliable predictions, there can be still dead end metabolites and unconnected reactions. Usually, this dead end metabolites are not intermediate compounds to the synthesis of the biomass, thus not impairing model predictions. Hence, at this stage, these compounds can be removed to simplify the GSM model. Nevertheless, dead-end metabolites which are considered compounds of interest should be kept in the model, making them available for further studies.

3.2.4 *Environmental conditions*

The medium used for the growth of the organism is presented at the experimental section of this work. Metabolites available in the medium used in the laboratory are defined in the model as drains and their fluxes, *i.e.*, the rates at which the compounds enter the cell to be metabolized, were calculated from experimental data.

At this point, the model can be simulated, through FBA, by using OptFlux (with the IBM ®CPLEX solver), simulating the maximization of the biomass reaction, to assess the GSM model behavior. This operation is repeated several times until the GSM model is able to mimic the the organism's *in vivo* behavior.

3.3 EXPERIMENTAL MATERIAL AND METHODS

In this section, each of macromolecules such as DNA, RNA, Proteins, Carbohydrates, Lipids are quantified, as well as other individual compounds. An average of each of these metabolite groups are synthesized relatively to the biomass produced by the organism.

Therefore, a culture of *N. europaea* was continuously maintained in a chemostat to obtain the biomass required to those quantifications.

3.3.1 *Organism*

The experiments were performed with *N. europaea* strain NCIMB 11850, since its culture was already established at the laboratory where the experiments took place. This strain is genomically identical to the strain ATCC 19718, so the validation in the GSM model is considered viable.

3.3.2 *Medium and growth conditions*

The mineral medium used in the wet lab experiments comprised the preparation and sterilization of three distinct solutions that were mixed previously to their use. The medium uses the following constituents dissolved in deionized water: Solution A - 25 mM ammonium sulphate, 43 mM monopotassium phosphate, 3.9 mM monosodium phosphate, 1 M iron (II) sulphate (dissolved in 8.4 M ethylenediaminetetraacetic acid, pH 7.0) and 8.4 M copper (II) sulphate, with the pH adjusted dropwise to 8.0 with sodium hydroxide 10 M; Solution B - 2.4 mM calcium chloride; and Solution C - 3.8 mM sodium carbonate. All solutions were autoclaved at 121 ^{o}C for 20 minutes, and were aseptically mixed once chilled. These solutions were prepared separately to prevent spontaneous inorganic precipitation, namely due to the reaction between phosphate, ammonia and calcium ions, making them inaccessible to the organisms metabolization(Fattah, 2012).

3.3.3 *Chemostat setup*

The growth of the organism was performed in a chemostat attached to two consecutive bioreactors to collect biomass. The chemostat had the capacity of 395 mL, the first bioreactor, the capacity of 430 mL and the second one with the capacity of 600 mL, and all were protected from light by a capsule. The chemostat and the first bioreactor were continuously stirred at 120 rpm. The lids of the chemostat and of the bioreactors allowed air transfer (including CO_2 and O_2 transfers) from the inside out and vice-versa, without compromising the aseptic conditions inside the reactor. Room temperature (25 ^{o}C) was the experimental temperature. The medium feeding rate was controlled to promote a healthy growth of the culture, thus the pH, concentration of nitrite and ammonium in the solution were regularly monitored. At the beginning of the experiment, the growth was slow due to the lag phase of the organism. However, once the growth was exponential, it was added more medium until all the referenced conditions were stabilized, to make a constant flow in the chemostat with productive and healthy conditions, where the bacteria could grew in a continuous exponential phase.

The chemostat structure, along with other components in the system are presented in Figure 8. Fresh medium was introduced, by pump E, from the flask A to the chemostat (B). The culture medium volume in it rose until the maximum volume of 395 mL. Beyond that threshold, the medium was transferred from the chemostat to the bioreactor C through gravitational potential. The same happened between bioreactor C and D, when the volume of the former rose to the volume of 430 mL. In reactor D, the maximum culture volume was 600 mL. When the operation volume of the bioreactor D was near its maximum, the culture medium was aseptically collected through filtration using a Whatman membrane filter with a pore size of 0.2 μm, in an aseptic environment. The collected biomass was resuspended in sterile deionized water, and then the resultant solution was lyophilized to proceed in the quantification of the components of the organism. The lyophilization system (Alpha 1-4 LD, by Christ) was set with the cooling system subjected to -57 ^{o}C.

To analyse the concentration of nitrogen from nitrite and ammonia, and the values of pH, a sample was regularly extracted from the chemostat in aseptic conditions. The sampling time points were usually between 48 to 72 hours and the feeding rate was regulated whenever seemed necessary to provide an appropriate environment for *N. europaea*, to promote the maximum growth rate of the microorganism.

3.3.4 *N-compounds analysis*

The concentration of nitrogen from nitrite was measured through the use of commercial test cuvettes (LCK 342 HACH). The concentration of nitrogen from ammonia was measured by the Nessler procedure (Arthur, 1979).

3.3.5 *Quantification of macromolecules*

This section comprises the protocols used to quantify the macromolecules of *N. europaea*.

Protein quantification

Protein content was determined by using the Biuret method according to Verduyn and Dijken (1990). One milliliter of resuspended biomass in Phosphate-Buffered Saline (PBS) (2 g/L) was mixed with 0.5 mL 1 M NaOH, incubated at 100 ^{o}C for 10 min and subsequently cooled on ice. Then, 0.9 mL of the solution were mixed with 0.3 mL of 0.1 M copper sulfate solution and incubated for 5 min at room temperature. Ended that time

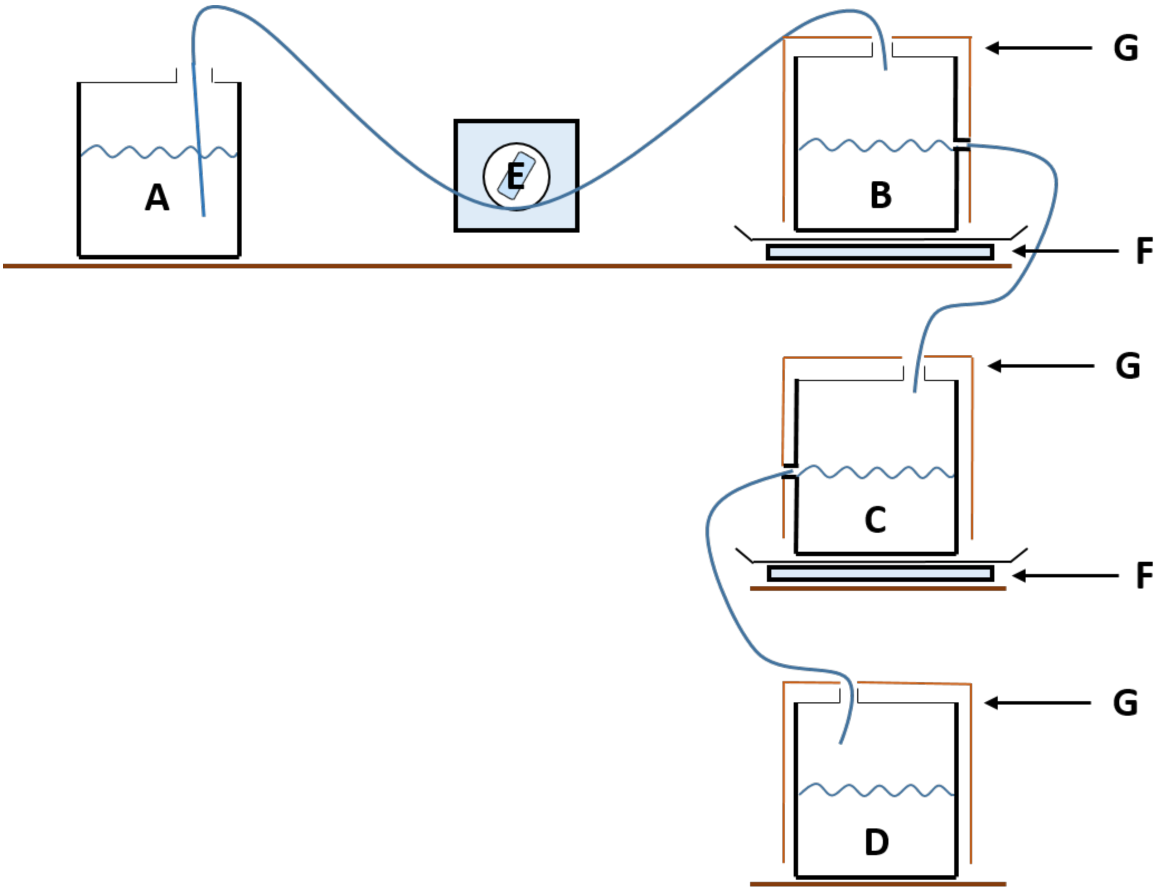


Figure 8.: Schematics of the chemostat used for *N. europaea* culture. A - Fresh medium; B - Chemostat; C and D - Bioreactors; E - Pump; F - Orbital shaker; G - Capsule. The chemostat and bioreactor C were submitted to agitation by the orbital shakers, and B, C and D were protected from light by the use of an opaque capsule. The chemostat design envisaged the creation of a gravity potential between the chemostat and the bioreactor C and D, in order to achieve a spontaneous medium transfer.

the solution was centrifuged for 5 min at 7378 *g*. The absorbance was measured at 540 nm in a 96-well microtiter plate in a Microplate reader for ELISA Bio-Tek Synergy HT. Protein concentration was calculated by interpolation in a calibration curve using Bovine Serum Albumin (BSA) as standard.

DNA quantification

Biomass macromolecular content determination DNA content of biomass was determined according to Mey and Vandamme (2006) with some modifications. 4,6-Diamidino-2-Phenylindole Dihydrochloride (DAPI) was used instead of Hoechst as fluorescent dye solution. Freeze dried biomass samples were dissolved in TNE buffer (1 M NaCl, 10 mM Ethylenediaminetetraacetic Acid (EDTA), 10 mM Tris, pH 7.4) at a concentration of

5 mg/mL. Then 33 μL of the sample solution was mixed with 1 mL of DAPI dye solution (DAPI 0.25 μg/mL in TNE buffer) and incubated for 30 min. Fluorescence was measured using the excitation/emission wavelengths of 350/460 in a black 96-well microtiter in a Spectrofluorimeter Jasco FP-6200. DNA content was calculated by interpolation in a calibration curve performed using as standard calf thymus DNA.

RNA quantification

RNA content of biomass was determined according to Benthin and Villadsen (1991) with some modifications. Freeze dried biomass samples, with 10 mg, were washed twice in 1 mL of cold 0.7 M $HClO_4$ and resuspended in 1 mL 0.3 M KOH. The resuspended biomass was then incubated at 37 ^{o}C for 1 h. To the pellets was added 100 μL of 3 M $HClO_4$ and the samples were centrifuged at 14462 *g* for 2 minutes. The supernatant was collected and the pellet was washed twice with 450 μL 0.5 M $HClO_4$. The three supernatants collected were mixed and absorbance was measured at 260 nm using a Micro-Spectrophotometer Nanodrop. RNA content was calculated taking into account the sample dilution.

Carbohydrate quantification

Total carbohydrates were determined by the phenol-sulphuric method as described according to Herbert and Strange (1971). Briefly, 200 μL sample containing freeze dried biomass (0.1 mg dry weight/mL ofPBS) was mixed with 200 μL phenol 5 % (v/v) and 1 ml 96 % (v/v) sulphuric acid in glass tubes. After 25 min, absorbance at 490 nm was measured using glucose solutions as standard.

Lipid quantification

Total lipids were determined by the sulpho-phospho-vanillin method as described by Izard and Limberger (2003) with some modifications. Briefly, 10 mg of freeze dried biomass was mixed with 2 mL of a mixture of chlorophorm and methanol (1:1) in glass tubes. After 15 min, 250 μL of the resultant mixture is placed in other glass tube and the mixture is evaporated. When all the liquid is evaporated, 100 μL of H_2SO_4 are added and the tubes are incubated at 100 ^{o}C for 10 min. The tubes are then cooled at room temperature and 2.4 mL of phosphoric acidvanillin reagent were added to the tubes and incubated at room temperature for 15 min. To prepare the phosphoric acidvanillin reagent, 0.120 g of vanillin was added to 20 mL of water, and the volume adjusted to 100 mL with 85 % (v/v) phosphoric acid. Absorbance at 490 nm was measured using olive oil as standard.

3.3.6 *Quantification of aminoacids*

In this section, the protocols used to quantify each aminoacid of the macromolecule protein are shown.

The glycogen content was analyzed according to Smolders (Smolders et al., 1994). Glycogen was hydrolyzed adding 5 ml of 0.6 M HCl to 10 mg of freeze dried biomass and incubating over-night at 105 ^{o}C. The glucose produced was quantified using a glucose quantification Kit (D-Glucose - BOEHRINGER MANNHEIM/R-BIOPHARM from Roche Yellow Line).

Biomass hydrolysis for amino acid content determination was performed according to Tuan and Dove (1999). To biomass samples of 40 mg were added 150 μL of 2 % (v/v) of 3.3 Dithiodipropionic Acid (DTDPA) in 0.2 M NaOH, 245 μL of HCl 12 M, 75 μL of H_2O, 25 μL of internal standard mixture (sarcosine and norvaline 100 mM) and 5 μL of thiaglycolic acid (1 %) (v/v). A N_2 stream was used to remove all O_2 in order to prevent the oxidation during hydrolysis. The samples were then incubated at 105 ^{o}C for 24 hours. To neutralize the samples 300 μL of NaOH (10 M) was added to the samples. Amino acid analysis was performed using High-Performance Liquid Chromatography (HPLC) a Nexera X2 HPLC system from Shimadzu with a diode-array detector and a SIL–30AC autosampler. The column is a Zorbax Eclipse-Amino acid analysis with dimensions 4.6 x 150 mm. All the procedures of the HPLC method were made according to the column manufacture instructions.

3.3.7 *Quantification of iron and copper*

The quantification was performed using a digestion (performed with a microwave digestion system speedwave 4, by BERGHOF) of a biomass sample followed by the determination of metals using Inductively Coupled Plasma Mass Spectrometry (ICP-MS) performed with Optima 8000, by PerkinElmer. The digestion was performed in two different conditions to ensure the validity of the values. The first digestion was done by submitting the biomass sample to a 5 % (v/v) of nitric acid, for 10 minutes at 200 ^{o}C. And the second digestion was made by submitting the sample to the same solution, for 10 minutes at 100 ^{o}C. The measures were done in axial view, with a wavelength of 238.204 nm to measure iron, and 324.752 nm and 327.393 nm to measure to isotopes of copper.

3.3.8 *Scanning electron microscope*

The samples were characterized using a desktop Scanning Electron Microscope (SEM) coupled with Energy-Dispersive X-ray Spectroscopy (EDS) analysis (Phenom ProX with EDS detector (Phenom-World BV, Netherlands)). All results were acquired using the ProSuite software integrated with Phenom Element Identification software, allowed for the quantification of the concentration of the elements present in the samples, expressed in either weight or atomic concentration. The *N. europaea* samples were added to an aluminium pin stubs with electrically conductive carbon adhesive tape (PELCO Tabs^{TM}).Samples were imaged without coating. The aluminum pin stub was then placed inside a Phenom Charge Reduction Sample Holder (CHR),and different points were analyzed for elemental composition. EDS analysis was conducted at 15 kV with intensity map.

4

RESULTS AND DISCUSSION

In this chapter, the results from the laboratorial experiments are presented, to be implemented in the GSM model. And, thus, the model is presented in a simplified manner, along with the simulations. As mentioned before, the cross-validation from both of these sources serves to raise the accuracy of the model (Dias et al., 2014).

4.1 LABORATORIAL RESULTS

Here, the data retrieved from all the laboratorial experiments, in this work, is presented. As explained before, *N. europaea* was grown in a chemostat under controlled conditions, and from this, it was obtained biomass used to quantify compounds (or groups of them) in the cell. The quantity of each of this compounds were introduced into the GSM model, to consider the growth of the biomass in the simulations with the same conditions used in the laboratory.

4.1.1 *Analytical analysis preformed in the chemostat*

This organism oxidizes ammonia into nitrite to synthesize ATP, the energy used by the cell. The nitrite was measures regularly, over a period of 2941 hours, to particularly study the growth in the exponential phase (correspondent to the maximum growth rate). About 1313 hours into this experiment, the chemostat growth began to stabilize, *i.e.*, the rate of production and sinking of nitrite in the chemostat were the same, meaning that the organism was constantly in exponential phase. Also the pH started to stabilize, in approximately the same period. Considering that the feeding rate volume was stabilized, it was possible to make the calculations for the biomass within the chemostat, as well, as the biomass produced by each hour, since the correlation between the conversion of ammonia to nitrite and biomass production available in the literature is 0.146 mg biomass/mg NH_4^+-N (Grady and Daigger, 1999). Note that the ammonia consumed is the same amount as the nitrite produced (Perez-garcia and Singhal, 2014). The concentrations of pH, the volume of medium added per day, and the concentrations of ammonia and nitrite are presented in Graphs 9A, 9B, 9C and 9D, respectively.

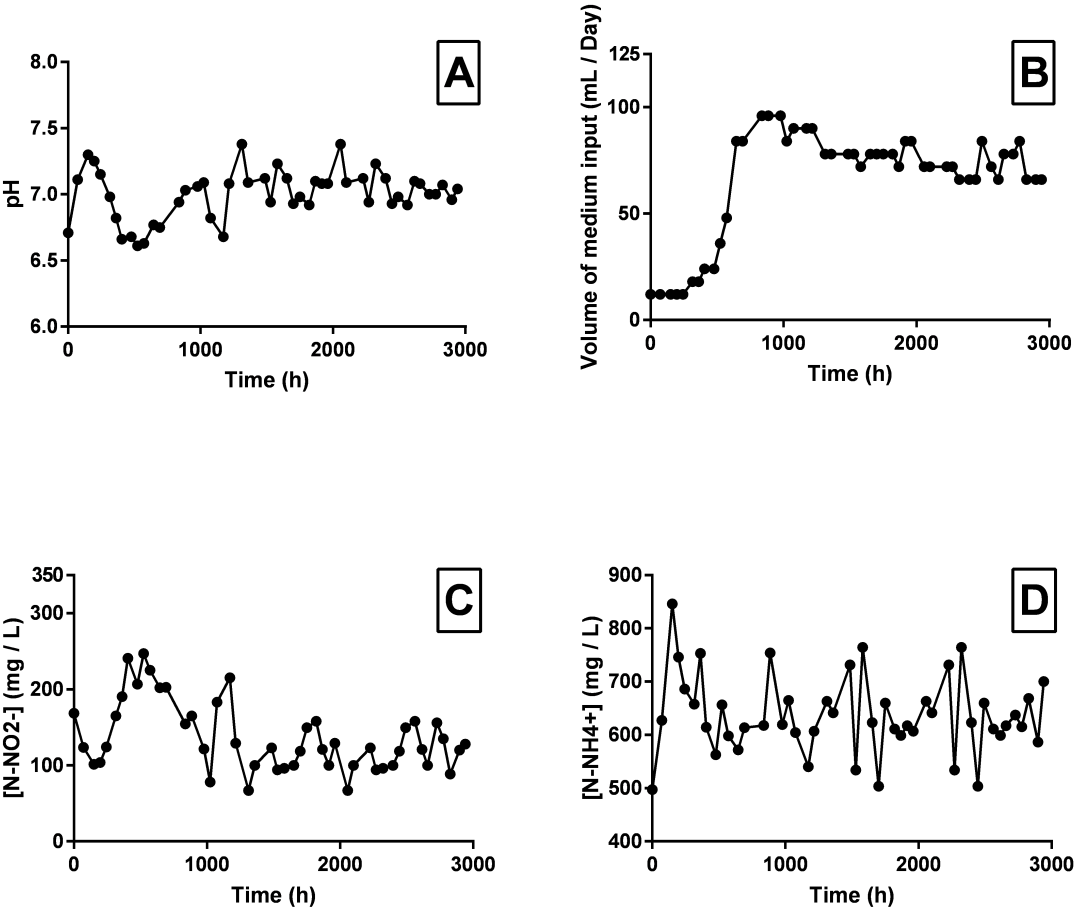


Figure 9.: Graphs of *N. europaea* chemostat parameters. A - pH value variation; B - Feed volume; C - Nitrite nitrogen concentration; D - Ammonia nitrogen concentration.

4.1.2 *Quantities of metabolized compounds*

The quantities of compounds of this GSM model were directly measured in laboratory, retrieved from literature, deduced with basis in the *E. coli* GSM model iAF1260 as the reference model, or calculated through *merlin*.

Determined in laboratory in this work are the quantities of macromolecules, including the Protein, DNA, RNA, Carbohydrate and Lipid ones. Also the quantity of each aminoacid, iron and copper were measured in laboratory. The iron quantity was in the range of expected values, for the medium used (Vajrala and Arp, 2011). However, the copper was undetectable experimentally because ICP-MS could not read concentrations under a certain value. It is known that copper activates Ammonia Monooxygenase (AMO), so it is necessary, however, because the quantity of it was not determined, it was not included in the biomass of the model (Ensign and Arp, 1993).

4.1.3 *Quantities of transported compounds*

As mentioned before, the transport of compounds inward or outward of the cell, is constrained by the drains. After the chemostat stabilized, it was possible to calculate the absorption of nitrite, through only experimental data. Unfortunately, because ammonia is volatile and the chemostat has transfers of gases, the measures made experimentally were compromised. The same is applicable to carbon dioxide and oxygen. And so, for these three compounds, the values of absorption were extrapolated by the literature and through the nitrite absorption in the chemostat.

4.1.4 *Scanning electron microscope*

With the images obtained from SEM, it is possible to make a qualitative analysis at the viability of the cells, after their collective process. In Figure 10 it is possible to observe that an extensive majority of cells present an intact appearance. In other words, no evident extensive cell lyses was observed. Moreover, the cells present a similar size and form as presented in literature (Grady and Daigger, 1999; Yu et al., 2015). Thus, this indicates that all of the experimental quantifications performed in this work are viable, since all the compounds were maintained within the cells, avoiding their degradation.

4.2 COMPUTATIONAL RESULTS

In this section, all the results regarding the reconstruction of the GSM model, as well as the simulations performed with it, are presented.

4.2.1 *Finding organisms for comparison*

The closest organism, with a GSM model described, to *N. europaea* was found to be *Neisseria meningitidis*. However these organisms only have their class in common, thus they are still relatively distant. Therefore, the genome of *N. europaea* was compared to *N. meningitis* using BLAST to assess this comparison, which confirmed that these organisms had very

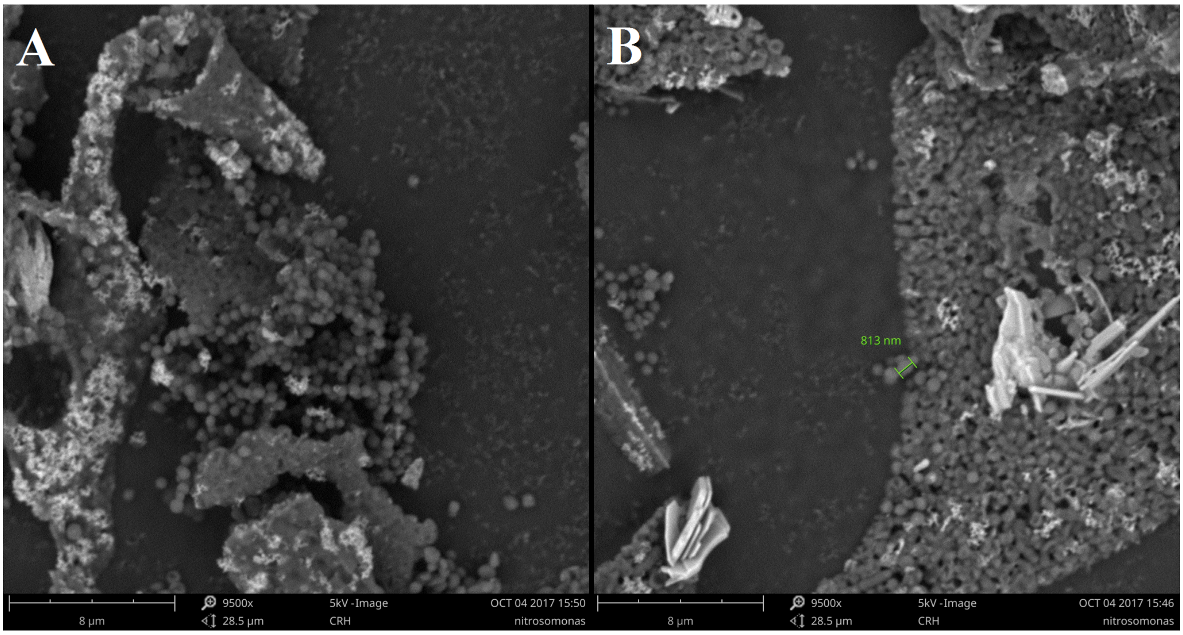


Figure 10.: *N. europaea* freeze dried biomass SEM images. A - Agreggation of cells; B - Measure of one cell of *N. europaea*, marked its size in green

few matches between the two genomes. The phylogenetic tree is represented in Figure 13, in Annex A.

Because a GSM model to accelerate this model reconstruction was not obtained, annotated organisms taxonomically close to *N. europaea* were sought to ease the gene annotation process.

The data comparing *N. europaea* genes with the species sharing it genus is presented in Table 7 and the organism selected for comparison was *N. eutropha* because it is the one with the most curated genes (except *N. europaea*).

4.2.2 *Determining the α and thresholds with new metric*

The new metric was measured with the α in each instance, presented in Table 16, combined with the NPV, presented in Table 18, and with the Precision, presented in Table 17, in Annex A. These tables were constructed through their respective confusion matrices, presented in Tables 13, 14 and 15 in Annex A, of the 50 genes sample. The manual annotation of this set of genes, is presented in Table 11, in Annex A.

This information is demonstrated in Figure 11. One aspect to note is that the accuracy decreases as the α increases, meaning that the homology of the genome values the taxonomy

Table 7.: Number of genes annotated on TrEMBL (non-curated database) and Swiss-Prot (curated database) of all the species of the genus *Nitrosomonas*.

Species	TrEMBL	Swiss-Prot
Nitrosomonas aestuarii	4	0
Nitrosomonas communis	3066	0
Nitrosomonas cryotolerans	9	0
Nitrosomonas halophila	4	0
Nitrosomonas marina	25	0
Nitrosomonas nitrosa	4	0
Nitrosomonas oligotropha	16	0
Nitrosomonas stercoris	1	0
Nitrosomonas ureae	2819	0
Nitrosomonas eutropha	2147	335
Nitrosomonas europaea	4597	436

score over the frequency one. This suggests that the classification is overall more accurate by giving more importance to closer organisms than to more homology hits.

The highest *Accuracy* per number of entries to be curated (1.39) was associated with an α of $0,9$ and the lower and upper *thresholds* of $0,1$ and $0,5$, respectively. This α had an *Accuracy* of about $0,433$ and every gene with the score between $0,1$ (inclusively) and $0,5$ (exclusively) were manually curated, about 348 genes.

4.2.3 *Genome annotation*

Four organisms were selected to develop this approach: The closest organism *N. europaea* was selected because, in the NCBI database *N. europaea* and *N. europaea* ATCC 19718 (the strain that the GSM model is based on) have two different entries. The second one was *N. eutropha*, as previously described. The third one was *E. coli*, as it is the most studied bacteria and it has several curated GSM models.

The manual annotation of the sample of 50 genes, is presented in Table 11, in Annex A. And the results of the the manual curation of 385 genes (including the previously mentioned set of 50 genes, and other genes considered important) are shown in Table 12, in Annex A. About 70 % of the genes EC number were altered, suggesting that this step was significant to a better reconstruction of the model. In most of the cases the confidence level of the genes were D or F, meaning that most of the homologies from a curated source came from *E. coli*, and from a non-curated source came from the same strain for the reconstruction of this model.

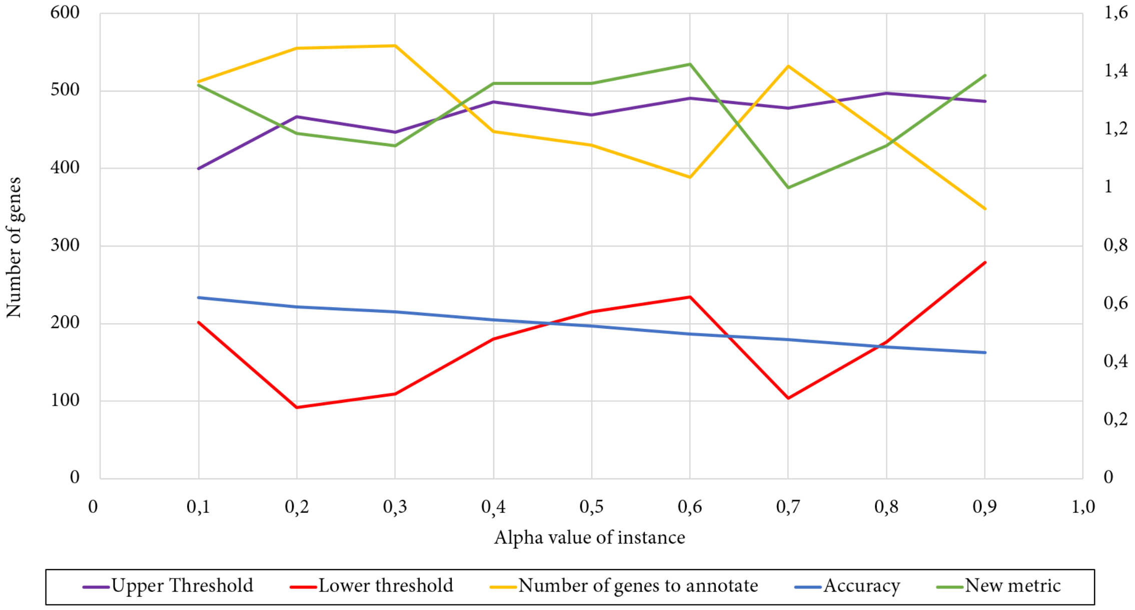


Figure 11.: Representation of the number of genes automatically accepted by the upper threshold, the number of genes automatically rejected by the lower threshold, and the number of genes to be manually curated (left vertical axis). Representation of the*Accuracy* and the *Accuracy* per number of entries to be curated (right vertical axis). All of these parameters are presented for every α.

4.2.4 *Correction of the reversibility of reactions and of EC numbers in the model*

The reversibility, along with the direction correction of reactions (in irreversible cases) were done to stop many problems within model. It was corrected manually the reversibility of 71 reactions, and the direction of 33. Both represent about 15 % of the total number of reactions. With a functional model, the *merlin* import from KEGG reactions, along with its automatic tool to correct the reversibility prove to be very useful, because it were only reviewed low percentage of those. The EC number were only changed in about 3 % of the annotated genes, meaning that the annotation method was efficient.

4.2.5 *Transporters prediction*

The number of genes responsible to transport metabolites is 72. Each of those transporters can involve one or more metabolites and even though the model reconstructed in this work does not use all of those, they can be useful when it is wanted to have other metabolites enter of leave the metabolic network, in a simulation. This further demonstrates the plasticity of this metabolic model, when predicting its growth under different environmental conditions.

4.2.6 *Removal of dead-ends*

All of the dead-ends were successfully removed from the model, making it feasible. This was possible by adding reactions to filling the gaps between the dead-end metabolites or by removing problematic ones. The number of reactions imported with no gene association, corresponded to 14 % of the total reactions. The change in EC number, previously mentioned, was responsible for removing dead-ends. Most of the removed reactions were because they were generic, because KEGG represents its pathway maps with generic compounds, unbalancing the model. And most of the added reactions have basis in other databases and in literature, meaning that it is important to consider multiple sources of information to make the model more complete.

4.2.7 *Balance of reactions*

About 10 % of the total reactions had to be reviewed to maintain a stoichiometric model. Along with the generic compounds in KEGG, the balance of required reactions is also affected from this problem. However, most of the reactions were unbalanced by a proton, suggesting that KEGG involved compounds in the oxidized or reduced form in different reactions.

4.2.8 *Biomass precursors*

All the macromolecules average amounts are represented in Table 8, and these are the resultant of the biomass composition, in the metabolic model. This quantitative study was done through multiple sources, described in the previous chapter.

Table 8.: Relative quantity of each macromolecule in the biomass of the model.

Macromolecule	**Quantity (g / g Biomass)**
Protein	0,463
DNA	0,007
RNA	0,040
Carbohydrate	0,352
Lipid	0,099
Inorganic ions	0,010
Cofactor	0,029
Total	1

Qualitatively, there were 66 metabolites inserted into the 7 entities responsible to form the biomass. Most of them were directly retrieved from literature and from *E. coli* iAF1260

GSM model. However the lipid entity was and exception because the compounds in it were deducted, from multiple sources:

This organism lipids can be classified as acetogenic and isoprenoid lipids and have different chemical properties as well as different roles in the cell. The only acetogenic lipids described in literature are hexadecanoate [16:0] and hexadecenoate [16:1] and the only isoprenoid lipids are diploptene e bishomohopanol (Hagen and Goldfine, 1966; Sakata and Seemann, 2008). Acetogenic lipids are known to be incorporated into phospholipids as well as lipopolysaccharides. *E. coli* iAF1260 GSM model considers phospholipids with one type of acetogenic acid at a time, with the same length. Phospholipids with one saturated and one unsaturated acetogenic acid are not considered, though this types of acetogenic acids always come in pairs (for. Hence, eight different phospholipids, four types of phospholipids (Phosphatidylcholine (PC), Phosphatidylethanolamine (PE), Phosphatidylglycerol (PG) and Phosphatidyl-N,N-dimethylethanolamine (PDME)), each of which with two variants (assembled with two saturated or two unsaturated acetogenic lipids), were considered. The amount of each of these phospholipids was inferred from literature. Concerning isoprenoid lipids, bishomopanol is known to be the product of the degradation of Amonibacteriohopanetriol (ABHT), though this reaction was not found in the database or literature. Hence, regarding the synthesis of bishomohopanol, no secondary metabolites were considered in this reaction. The amount of diploptene and bishomohopanol was directly obtained by literature.

Another interesting entity is the lipopolysaccharide one, because although it was not considered, its partitions are still in calculated in the model. This is because there was not found enough evidence to insert these molecules into the model. Beside not finding any information regarding these molecules structure in literature and databases, from the 9 genes that synthesize the most common lipopolysaccharide in bacteria (Lipid A), only one was available in *N. europaea* (Opiyo and Moriyama, 2010).

Finally, one particular and very important network, that is closely related to the biomass, and that was added to the model is the Electron Transport Chain (ETC), for the fact that it is responsible for synthesizing ATP. This pathway was reconstructed based on entries from UniProt and literature. The ETC was found to be responsible for recycling Reduced Nicotinamide Adenine Dinucleotide (NADH) from Nicotinamide Adenine Dinucleotide (NAD) by a process called "reverse electron transfer". The lump of all the reactions responsible of producing energy is described in the following Equation 9:

$$NH_3 + \frac{3}{2}O_2 \rightarrow NO_2^- + H^+ + H_2O \tag{9}$$

The enzyme Hydroxylamine Oxidoreductase (HAO) transfers 2 electrons to the next steps of the ETC, of which 1.65 are used to produce nitrite and 0.35 for recycling NADH (Whittaker et al., 2000).

Because the model does not include electrons, the ratio for ATP synthesis and for NADH recycle was associated in the precursor reaction, performed by AMO.

4.2.9 *Genome-scale metabolic model of N. europaea*

The GSM model reconstructed in this work is available at iPR572, as an SBML file, as well as all the Supplementary material. This model contains 617 reactions, being about 80 % of them inferred from homology. This, and other basic information of the model is presented in Table 9.

Table 9.: General information of the model.

Data	**Number in the model**
Genes	2462
ORF	572
Metabolites	4832
Total Reactions	617
KEGG reactions	7
HOMOLOGY reactions	495
TRANSPORTERS reactions	21
MANUAL reactions	61

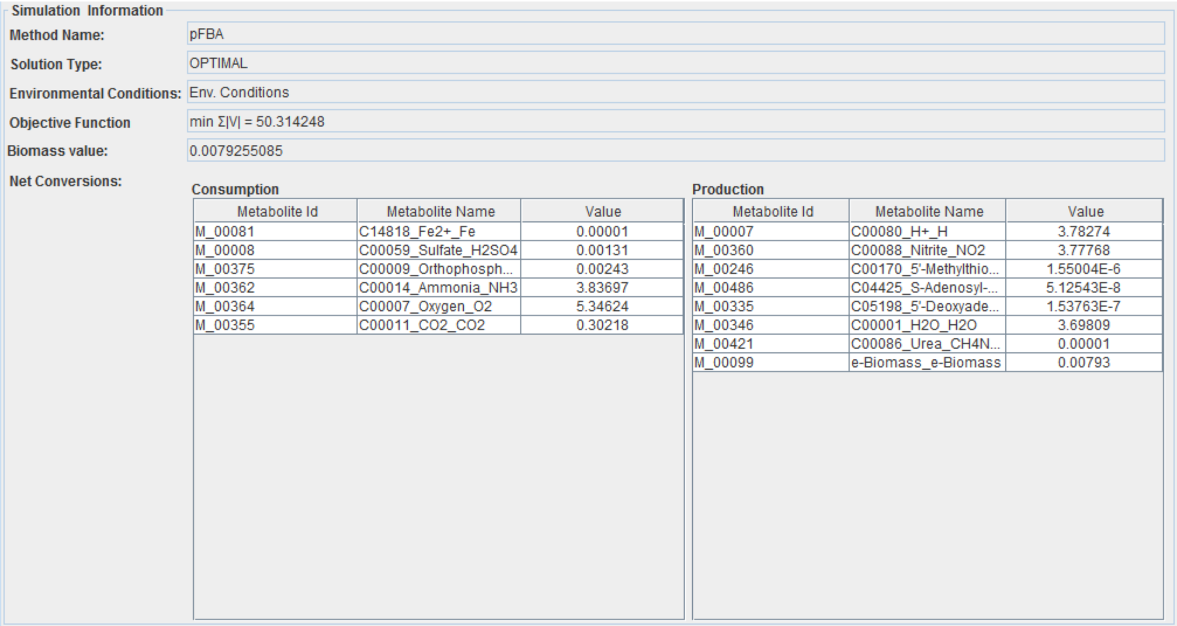
When compared to models of organisms of reference for this work, there are is are not many differences in the number of ORFs or of reactions. The comparison data is presented in Table 10. Because *E. coli* is well studied organism, it has more reactions associated with it. Moreover, *N. europaea* has the shortest genome among the Betaproteobacteria, possibly because they have a limited lifestyle for the fact that all the energy is derived from oxidizing ammonia, leading to a reduction of genes, and reactions (Opiyo and Moriyama, 2010). However the difference of the *N. europaea* and *N. meningitidis* are unnoticeable.

Table 10.: Comparison of models of organisms of reference. iGB555 is the metabolic model of *N. meningitidis*, and iAF120 is of *E. coli*.

Data	**iPR572**	**iGB555**	**iAF1260**
ORF	572	555	1260
Intracellular metabolites	550	471	1039
Total Reactions	617	496	2077
TRANSPORTERS reactions	21	74	690

4.2.10 *Simulation and validation of the genome-scale metabolic model*

It was used a pFBA simulation in a wild-type version of the model, with the drains fluxes described before. The results of the fluxes of consumption and production compounds are presented in Figure 12.



Simulation Information

Method Name: pFBA

Solution Type: OPTIMAL

Environmental Conditions: Env. Conditions

Objective Function: min Σ|V| = 50.314248

Biomass value: 0.0079255085

Net Conversions:

Consumption

Metabolite Id	Metabolite Name	Value
M_00081	C14818_Fe2+_Fe	0.00001
M_00008	C00059_Sulfate_H2SO4	0.00131
M_00375	C00009_Orthophosph...	0.00243
M_00362	C00014_Ammonia_NH3	3.83697
M_00364	C00007_Oxygen_O2	5.34624
M_00355	C00011_CO2_CO2	0.30218

Production

Metabolite Id	Metabolite Name	Value
M_00007	C00080_H+_H	3.78274
M_00360	C00088_Nitrite_NO2	3.77768
M_00246	C00170_5'-Methylthio...	1.55004E-6
M_00486	C04425_S-Adenosyl-...	5.12543E-8
M_00335	C05198_5'-Deoxyade...	1.53763E-7
M_00346	C00001_H2O_H2O	3.69809
M_00421	C00086_Urea_CH4N...	0.00001
M_00099	e-Biomass_e-Biomass	0.00793

Figure 12.: Wild-type simulation of the model, using OptFlux with pFBA. The consumed compounds are only inorganic ones, as expected from a chemolitoautotrophic organism, and the compounds produce include nitrite, protons and biomass.

Note that all the principal consumed and produced compounds are involved in the simulation. Also note that the production of protons justifies the acidification of the medium, in aerobic conditions (Kozlowski et al., 2014). All other compounds produced could not be metabolized so they were secreted by the model.

The *in silico* biomass growth rate was 0.00793 h^{-1}, whereas growth *in vivo* was 0.0078 h^{-1}. This leads to a model accuracy of about 98.36 %.

Another aspect to notice is that the efficiency to produce biomass, with little waste is high. About $99,9903$ % of the carbon consumed is directed to the production of biomass, and only $0,00997$ % is drained through the metabolites secreted by it. The fact that this organism is an chemoautolitotrophic means it spends much of the energy produced in carbon fixation (about 80 %), and for that, it must spend as little resources as possible to thrive in its environment (Baribeau, 2006).

5

CONCLUSIONS AND FUTURE WORK

In this chapter, a brief summary of the applications of the GSM model is presented, in the context of this work, as well as future work to continue to improve this model.

5.1 CONCLUSIONS

A viable GSM model of *N. europaea* was reconstructed in this work, iPR. A robust genome-wide annotation was performed in this work, which allowed providing a reliable representation of the metabolism of this organism in this model. Simulations with iPR allowed predicting *in silico* the behavior of the cell *in vivo*. Moreover, optimizations to improve production of compounds of interest with less effort and resources, for instance, by predicting genes knockouts. Although this model was validated with experiments in aerobic conditions, enzymes to maintain a metabolism in anoxic conditions are also present in the model. However, the model was not validated in this condition because there was no experimental data done in this work to support it.

This model could be used to predict optimizations that would allow using this organism in processes, other than wastewater treatment, to remove N.

Nevertheless, this model was reconstructed with a focus of helping to stop eutrophication in wastewater treatments. This serious environmental problem, although firstly described in the beginnings of the 20th century, still persists to this very day. By weakening ecological niches worldwide, it sets in motion a varied range of consequences, from economical to health ones. Luckily, with the advances in systems biology and bioinformatics, this issue can be slowed down or even stopped. Therefore it is hoped that this work can contribute to the development of scientific technologies to take a step further to eradicate eutrophication.

5.2 FUTURE WORK

Despite the the good results obtained with this GSM model, there is still room for improvements.

One of them is the compartmentalization of the model, by adding new compartments in which reactions can take place, other than the cytoplasm. A compartment usuallyavailable in Gram negative bacteria is The periplasm, in which of some compounds are synthesized,. Regarding this organisms specifically, The carboxysome is where the fixation of carbon dioxide occurs.

Another improvement would be the validation of the model in anaerobic and anoxic conditions, to fully understand the impressively adaptable metabolism that this organism has.

However, where the true potential of GSM model lies is in the possible simulations *in silico* in different conditions. Hence, gene knock-out, allow increasing the production of compounds of interest, and therefore to increase fluxes in the pathways of the N fixation or in others.

Finally, this organism can be associated in a community of bacteria able to consume nitrite (produced by *N. europaea*), and produce diatomic N - non-reactive N - to stop eutrophication. Therefore, this GSM model could be coupled with metabolic model of those partial denitrifier bacteria, to better understand the community requirements.

BIBLIOGRAPHY

S. K. Agarwal. *Water pollution*. A.P.H. Publishing Corp, New Delhi, 1 edition, 2005.

Rasmus Agren, Liming Liu, Saeed Shoaie, Wanwipa Vongsangnak, Intawat Nookaew, Jens Nielsen, and Costas D Maranas. The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for Penicillium chrysogenum. *PLOS Computational Biology*, 9(3), 2013.

R Apweiler et al. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research*, 39(Database):D214—-D219, jan 2011.

Israel Vogel Arthur. *Vogel's Textbook of Macro and semimicro qualitative inorganic analysis*. Longman, 5 edition, 1979.

Gino Baart and Dirk Martens. Methods in Molecular Biology. In Myron Christodoulides, editor, *Life Sciences*, volume 531, chapter 7, pages 107–125. Humana Press, Hatfield, 1 edition, 2009.

S. Bagchi and T. Nandy. Autotrophic Ammonia Removal Processes: Ecology to Technology. *Critical Reviews in Environmental Science and Technology*, 42(13):1353–1418, 2012.

K. Baker and S. Irvin. The Variation of Nitrifying bacterial Population sizes in a Sequencing Batch Reactor (SBR) treating low, mid, high concentrated synthetic wastewater. *Journal of Environmental Engineering and Science*, 6(March):651–663, 2007.

Hélène Baribeau. Microbiology and Isolation of Nitrifying Bacteria. In Mary Kay Kozyra, editor, *Fundamentals and Control of Nitrification in Chloraminated Drinking Water Distribution Systems*, volume 6, chapter 5, page 270. American Water Works Association, Denver, 1 edition, 2006.

P Benner, R Findeisen, D Flockerzi, U Reichl, and K Sundmacher. *Large-Scale Networks in Engineering and Life Sciences*. Modeling and Simulation in Science, Engineering and Technology. Springer International Publishing, Heidelberg, 1 edition, 2014.

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

Joost Boele, Brett G Olivier, and Bas Teusink. FAME, the Flux Analysis and Modeling Environment. *Systems Biology*, 6(8), 2012.

Bradley J Cardinale, J Emmett Duffy, Andrew Gonzalez, David U Hooper, Charles Perrings, Patrick Venail, Anita Narwani, Georgina M Mace, David Tilman, David A Wardle, Ann P Kinzig, Gretchen C Daily, Michel Loreau, James B Grace, Anne Larigauderie, and Diane S Srivastava. Biodiversity loss and its impact on humanity. *Nature*, 486(7):59–67, 2012.

Ron Caspi and Peter D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 44(1): 471–480, 2016.

Ron Caspi, Richard Billington, Luciana Ferrer, Hartmut Foerster, Carol A. Fulcher, Ingrid M. Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Quang Ong, Suzanne Paley, Pallavi Subhraveti, Daniel S. Weaver, and Peter D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 44(D1):D471–D480, jan 2012.

P. Chain and M. Whittaker. Complete Genome Sequence of the Ammonia-Oxidizing Bacterium and Obligate Chemolithoautotroph Nitrosomonas europaea Complete Genome Sequence of the Ammonia-Oxidizing Bacterium and Obligate Chemolithoautotroph Nitrosomonas europaea . 185(9):2759–2773, 2003.

P C Champe, R A Harvey, and D R Ferrier. *Biochemistry*. Lippincott's illustrated reviews. Lippincott Williams & Wilkins, Philadelphia, 3 edition, 2005.

F. Stuart Chapin, Erika S. Zavaleta, Valerie T. Eviner, Rosamond L. Naylor, Peter M. Vitousek, Heather L. Reynolds, David U. Hooper, Sandra Lavore, Osvaldo E. Sala, Sarah E. Hobbie, Michelle C. Mack, and Sandra Díaz. Consequences of changing biodiversity. *Nature*, 405(11):234–246, 2000.

J. Chase and M. Leibold. Spatial scale dictates the Productivity-Biodiversity relationship. *Nature*, 416(6879):427–430, 2002.

D. Chindler and D. Ilman. Human Alteration of the Global Nitrogen Cycle : Sources and Consequences. *Ecological Applications*, 7(November 1996):737–750, 1997.

Committee on Environmental and Natural Resources. An Assessment of Coastal Hypoxia and Eutrophication in U.S. Waters. pages 1–82, Washington, D. C., 2003.

D. Conley and G. Likens. Controlling Eutrophication: Nitrogen and Phosphorus. *Science*, 323(5917):1014–1015, 2009.

F. Crick. Central Dogma of Molecular Biology. *Nature*, 227(1):561–562, 1970.

O. Dias, M. Rocha, and I. Rocha. iOD907 , the first genome-scale metabolic model for the milk yeast Kluyveromyces lactis. *Biotechnology Journal*, 9:776–790, 2014.

O. Dias, M. Rocha, E. Ferreira, and I. Rocha. Reconstructing genome-scale Metabolic Models with Merlin. *Nucleic Acids Research*, 43(8):3899–3910, 2015.

Oscar Dias and Isabel Rocha. Systems Biology in Fungi. In Dongyou Liu, editor, *Mucormycosis. Food and Water Borne Mycotoxigenic and Mycotic Fungi*, chapter 6, pages 69–92. Taylor & Francis, Boca Raton, 1 edition, 2015.

Oscar Dias and Isabel Rocha. Genome-wide Semi-automated Annotation of Transporter Systems. *Transactions on Computional Biology and Bioinformatics*, 14(1):443–456, 2017.

Quay Dortch, Nancy N Rabalais, R Eugene Turner, and A Naureen. Impacts of Changing Si / N Ratios and Phytoplankton Species Composition. pages 37–48, 2001.

T. Egerton and M. Mulholland. Emergence of Algal Blooms: The Effects of Short-Term Variability in Water Quality on Phytoplankton Abundance, Diversity, and Community Composition in a Tidal Estuary. *Microorganisms*, 2(1):33–57, jan 2014.

S. A. Ensign and D. J. Arp. In vitro activation of ammonia monooxygenase from Nitrosomonas europaea by copper. *Journal of Bacteriology*, 175(7):1971–1980, 1993. ISSN 00219193.

Kazi Fattah. Finding Nutrient-Related Problems in Wastewater Treatment Plants. *International Conference on Environmental, Biomedical and Biotechnology*, 41:181–184, 2012.

Xueyang Feng, You Xu, Yixin Chen, and Yinjie Tang. MicrobesFlux: a web platform for drafting metabolic models from the KEGG database. *Systems Biology*, 6(94), 2012.

Avi Flamholz and Ron Milo. EQuilibrator - The biochemical thermodynamics calculator. *Nucleic Acids Research*, 40(1):770–775, 2012.

J. Galloway. The Global Nitrogen Cycle: Changes and Consequences. *Environmental Pollution*, 102(Suppl. 1):15–24, 1998.

J. Galloway. The Global Nitrogen Cycle: past, present and future. *Science*, 48 Suppl 2(326): 669–678, 2005.

C. P. Grady and G. Daigger. *Biological Wastewater Treatment*. Marcel Dekker, New York, 2 edition, 1999.

P-O. Hagen and H. Goldfine. Phospholipids of Bacteria with Extensive Intracytoplasmic Membranes Tetraethylammonium and Tetrodotoxin : Effects on Cochlear Potentials. *Science*, 151(12):1543–1544, 1966.

Joshua J. Hamilton and Jennifer L. Reed. Software platforms to facilitate reconstructing genome-scale metabolic networks. *Environmental Microbiology*, 16(1):49–59, 2014.

D Harper. *Eutrophication of Freshwaters: Principles, problems and restoration*. Springer Netherlands, Suffolk, 1 edition, 2012.

Roy M. Harrison. *Pollution : causes, effects and control*. Royal Society of Chemistry, Birmingahm, 4 edition, 2001.

Anja Hartmann and Falk Schreiber. Integrative analysis of metabolic models - from structure to dynamics. *Frontiers in bioengineering and biotechnology*, 2:91, 2014.

Christopher Henry, Matthew DeJongh, Aaron A Best, Paul M Frybarger, Ben Linsay, and Rick L Stevens. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, 28(9), 2010.

D Herbert and R E Strange. Chemical Analysis of Microbial Cells. In J R Norris and D W Ribbons, editors, *Methods in Microbiology*, chapter 3, pages 209–344. Elsevier, Porton, 1 edition, 1971.

D. U. Hooper, F. S. Chapin, J. J. Ewel, A. Hector, P. Inchausti, S. Lavorel, J. H. Lawton, D. M. Lodge, M. Loreau, S. Naeem, B. Schmid, H. Setälä, A. J. Symstad, J. Vandermeer, and D. A. Wardle. Effects of biodiversity on ecosystem functioning: A consensus of current knowledge. *Ecological Monographs*, 75(1):3–35, 2005.

M Hucka and J Wang. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics (Oxford, England)*, 19(4):524–31, mar 2003.

Jacques Izard and Ronald J Limberger. Rapid screening method for quantitation of bacterial cell lipids from whole cells. *Journal of Microbiological Methods*, 55:411–418, 2003.

Lu Shi Jing and Hany Alashwal. Database and tools for metabolic network analysis. *Biotechnology and Bioprocess Engineering*, 19(4):568–585, jul 2014.

Nick Juty, Nicolas Le Novère, and Camille Laibe. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic acids research*, 40(Database issue):D580–6, jan 2012.

Lukas Käll and Erik L L Sonnhammer. Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. *Nucleic Acids Research*, 35 (SUPPL.2):429–432, 2007.

M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, jan 2000.

Peter D Karp, Suzanne Paley, and Pedro Romero. The Pathway Tools software. *Bioinformatics*, 18(1):225–232, 2002.

W. Kemp and J. Stevenson. Eutrophication of Chesapeake Bay: Historical trends and Ecological Interactions. *Marine Ecology Progress Series Mar. Ecol. Prog. Ser.*, 303:1–29, 2005.

B N Kholodenko and H V Westerhoff. Metabolic Engineering in the Post Genomic Era. Horizon Bioscience, chapter 11. Horizon Bioscience, Wymondham, 1 edition, 2004.

Jessica A. Kozlowski, Jennifer Price, and Lisa Y. Stein. Revision of N2O-producing pathways in the ammonia-oxidizing bacterium Nitrosomonas europaea ATCC 19718. *Applied and Environmental Microbiology*, 80(16):4930–4935, 2014.

Anders Krogh and Erik L.L Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. *Journal of Molecular Biology*, 305(3):567–580, 2001.

Ram Kulkarni. Metabolic Engineering Biological Art of Producing Useful Chemicals. *Journal of Science Education*, 21(3):233–237, 2016.

H. D. Kumar. *Environmental pollution and waste management*. M D Publications, New Delhi, 1 edition, 1998.

Maren Lang, Michael Stelzer, and Dietmar Schomburg. BKM-react, an integrated biochemical reaction database. *BMC biochemistry*, 12:42, 2011.

S Y Lee and E T Papoutsakis. *Metabolic Engineering*. Biotechnology and Bioprocessing. Taylor & Francis, New York, 1st edition, 1999.

Nathan E Lewis, Kim K Hixson, Tom M Conrad, Joshua A Lerman, Pep Charusanti, Ashoka D Polpitiya, Joshua N Adkins, Gunnar Schramm, Samuel O Purvine, Daniel Lopez-Ferrer, Karl K Weitz, Roland Eils, Rainer König, Richard D Smith, and Bernhard Ø Palsson. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biology*, 6(390), 2010.

Yu Chieh Liao, Ming Hsin Tsai, Feng Chi Chen, and Chao A. Hsiung. GEMSiRV: A software platform for GEnome-scale metabolic model simulation, reconstruction and visualization. *Bioinformatics*, 28(13):1752–1758, 2012.

H Lieth and R H Whittaker. *Primary Productivity of the Biosphere*. Ecological Studies. Springer Berlin Heidelberg, Berlin, 1 edition, 1975.

Se Lohrenz, Gl Fahnenstiel, Dg Redalje, Ga Lang, X Chen, and Mj Dagg. Variations in primary production of northern Gulf of Mexico continental shelf waters linked to nutrient inputs from the Mississippi River. *Marine Ecology Progress Series*, 155:45–54, 1997.

S Malcolm and T H J Goodship. *Genotype to Phenotype*. Human molecular genetics series. BIOS Scientific, San Diego, 1 edition, 2001.

Karimi Mandana and Arezoo Tahmurespour. the Effect of Phenol on Heterotrophic Ammonia-Oxidizing Bacteria in Soils and Wastewaters. *Journal of Research in ...*, 8(1): 13–21, 2012.

C D Maranas and A R Zomorrodi. *Optimization Methods in Metabolic Networks*. John Wiley & Sons, New Jersey, 1 edition, 2016.

Olivier Martin and Marco Pagni. MetaNetX / MNXref reconciliation of metabolites and biochemical reactions to bring together genome-scale S ebastien. *Nucleic Acids Research*, 44:523–526, 2016.

Marjan De Mey and Erick Vandamme. Comparison of DNA and RNA quanti W cation methods suitable for parameter estimation in metabolic modeling of microorganisms. *Analytical Biochemistry*, 353:198–203, 2006.

Peter B. Moyle and Robert A. Leidy. Loss of Biodiversity in Aquatic Ecosystems: Evidence from Fish Faunas. In Peter B Moyle and Robert A Leidy, editors, *Conservation Biology*, chapter 6, pages 127–169. Springer US, Boston, MA, 1992.

National Academy of Sciences. *Eutrophication: Causes, Consequences, Correctives; Proceedings of a Symposium*. National Academy of Sciences, Washington, D. C., 1969.

National Research Council. *Tropospheric Transport of Pollutants and Other Substances to the Oceans (1978)*. National Academy of Sciences, Washington, 1 edition, 1978.

J Nielsen. *Metabolic Engineering*. Advances in Biochemical Engineering/Biotechnology. Springer Berlin Heidelberg, Berlin, 1st edition, 2003.

J Nielsen and S Arnold. *Biotechnology for the Future*. Number vol. 100 in Advances in Biochemical Engineering / Biotechnology Series. Springer, Berlin, 1 edition, 2005.

Jens Nielsen and Michael C Jewett. Impact of systems biology on metabolic engineering of Saccharomyces cerevisiae. 2007.

Stephen O Opiyo and Etsuko N Moriyama. Evolution of the Kdo 2 -lipid A biosynthesis in bacteria. *Biomedcentral*, 10(362), 2010.

Rashidi Othman, Nurul Azlen, Bt Hanifah, Razanah Ramya, Ayuni Bt, Mohd Hatta, Wan Syibrah, Hanisah Bt, Wan Sulaiman, Bt Yaman, Zainul Mukrim, and Bin Baharuddin. Aquatic plants as ecological indicator for urban lakes eutrophication status and indices. *International Journal of Sustainable Energy and Environmental Research*, 3(34):178–184, 2014.

Stephan Pabinger, Robert Rader, Rasmus Agren, Jens Nielsen, and Zlatko Trajanoski. MEMOSys: Bioinformatics platform for genome-scale metabolic models. *Systems Biology*, 5(20), 2011.

Yongzhen Peng and Guibing Zhu. Biological nitrogen removal with nitrification and denitrification via nitrite pathway. *Applied Microbiology and Biotechnology*, 73(1):15–26, 2006.

Octavio Perez-garcia and Naresh Singhal. ScienceDirect Clarifying the regulation of NO/N2O production in Nitrosomonas europaea during anoxic-oxic transition via flux balance analysis of a metabolic network model. *Water Research*, 60:267–277, 2014.

James Pfafflin and Edward Ziegler. *Encyclopedia of Environmental Science and Engineering*. CRC Press, Boca Raton, 5th editio edition, 2006.

Esa Pitkä nen, Paula Jouhten, Jian Hou, Muhammad Fahad Syed, Peter Blomberg, Jana Kludas, Merja Oja, Liisa Holm, Merja Penttilä, Juho Rousu, Mikko Arvas, and Jens Nielsen. Comparative Genome-Scale Reconstruction of Gapless Metabolic Networks for Present and Ancestral Species. *PLoS Comput Biol*, 10(2), 2014.

Nancy N. Rabalais, R. Eugene Turner, and William J. Wiseman. Gulf of Mexico Hypoxia, A.K.A. The Dead Zone. *Annual Review of Ecology and Systematics*, 33(1):235–63, nov 2002.

Isabel Rocha, Jochen Forster, and Jens Nielsen. Rocha, Isabel Jochen Förster, and Jens Nielsen. In *Microbial Gene Essentiality: Protocols and Bioinformatics*, Methods in Molecular Biology. Humana Press, New Delhi, 1 edition, 2008.

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

M. H. Saier, Can V Tran, and Ravi D Barabote. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Research*, 34(90001):D181–D186, jan 2006.

Susumu Sakata and Myriam Seemann. Organic Geochemistry Stable carbon-isotopic compositions of lipids isolated from the ammonia-oxidizing chemoautotroph Nitrosomonas europaea. *Organic Geochemistry*, 39(12):1725–1734, 2008.

Ilan Samish and Rafael J Najmanovich. Achievements and challenges in structural bioinformatics and computational biophysics. *Bioinformatics*, 31(1):146–150, 2015.

Joanne M. Savinell and Bernhard O. Palsson. Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. *Journal of Theoretical Biology*, 154(4):421–454, 1992.

E. W. Sayers et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 41(D1):D8—-D20, jan 2013.

D. Schindler and J. Vallentyne. The Algal Bowl: Overfertilization of the World's Freshwaters and Estuaries. *The Canadian Field-Naturalist*, 123:188, 2008.

I. Schmidt and M. Jetten. Aerobic and Anaerobic Ammonia Oxidizing Bacteria - competitors or natural partners? *FEMS Microbiology Ecology*, 39(3):175–181, 2002.

Ida Schomburg, Antje Chang, and Dietmar Schomburg. BRENDA, enzyme data and metabolic information. *Nucleic acids research*, 30(1):47–9, jan 2002.

Daniel Segrè, Dennis Vitkup, and George M Church. Analysis of optimality in natural and perturbed metabolic networks. *PNAS*, 99(23), 2002.

Tomer Shlomi, Omer Berkman, and Eytan Ruppin. Regulatory onoff minimization of metabolic flux changes after genetic perturbations. *PNAS*, 102(21), 2005.

Niranjan Kumar Shrestha, Shigeru Hadano, Toshiaki Kamachi, Ichiro Okura, Niranjan Kumar Shrestha, Shigeru Hadano, Toshiaki Kamachi, and Ichiro Okura. Conversion of Ammonia to Dinitrogen in Wastewater by Nitrosomonas europaea. *Applied Biochemistry and Biotechnology*, 90(3):221–232, 2001.

T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

V. Smith. Eutrophication of freshwater and coastal marine Ecosystems - A Global Problem. *Environmental Science and Pollution Research*, 10(2):126–139, 2003.

G J F Smolders, J Van Der Meij, M C M Van Loosdrecht, and J J Heijnen. Model of the Anaerobic Metabolism of the Biological Phosphorus Removal Process: Stoichiometry and pH Influence. *Biotechnology and Bioengineering*, 43:461–470, 1994.

C Smolke. *The Metabolic Pathway Engineering Handbook: Fundamentals*. The Metabolic Pathway Engineering Handbook. CRC Press, Boca Raton, 1st edition, 2009.

J. Sprent. *The Ecology of the Nitrogen Cycle*. Cambridge University Press, New York, 1 edition, 1987.

G. Stephanopoulos, A. Aristidou, and J. Nielsen. *Metabolic Engineering: Principles and Methodologies*. Academic Press, San Diego, 1 edition, 1998.

Neil Swainston, Kieran Smallbone, Pedro Mendes, Douglas B Kell, and Norman W Paton. The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks. *Journal of Integrative Bioinformatics*, 8(2):186, 2011.

C. Tamm. Introduction: Geochemical Occurrence of Nitrogen. Natural Nitrogen Cycling and Anthropogenic Nitrogen Emissions. In C Tamm, editor, *Nitrogen in Terrestrial Ecosystems: Questions of Productivity, Vegetational Changes, and Ecosystem Stability*, chapter 1, pages 1–6. Springer Berlin Heidelberg, Berlin, 1 edition, 1991.

M Terzer and J Stelling. Genome-scale metabolic networks. *Wiley interdisciplinary reviews.Systems biology and medicine*, 1(3):285–297, 2009.

Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93–121, jan 2010.

N W Tschoegl. *Fundamentals of Equilibrium and Steady-State Thermodynamics*. Elsevier Science, Amsterdam, 1 edition, 2000.

Y Tuan and C Dove. Predicting integrated protein nutritional quality Part 1: Amino Acid Availability Corrected Amino Acid Score and nitrogen balance data fitted to linear and non-linear models for test proteins. *Nutrition Research*, 19(12):1791–1805, 1999.

Neeraja Vajrala and Daniel J Arp. Role of a Fur homolog in iron metabolism in Nitrosomonas europaea. *BMC Microbiology*, 11(1):37, 2011.

Cornelis Verduyn and Johannes P. Van Dijken. Energetics of Saccharomyces cerevisiae in anaerobic glucose-limited chemostat cultures. *Journal of General Microbiology*, 136(1990): 405–412, 1990.

Vladimir Ivanovic Vernadsky. *The biosphere*. Springer Science, New York, 1 edition, 1998.

T. N. Veziroglu. *The Biosphere, Problems and Solutions*. Elsevier, Miami Beach, 1 edition, 1984.

John Villadsen. Principles of Metabolic Engineering. In Horst Doelle, J Stefan Rokem, and Marin Berovic, editors, *Biotechnology - Volume III: Fundamentals in Biotechnology*, chapter 12, pages 226–256. EOLSS Publications, Singapore, 1 edition, 2009.

Tilmann Weber and Hyun Uk Kim. The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. *Synthetic and Systems Biotechnology*, 69(79), 2016.

Mark Whittaker, David Bergmann, David Arciero, and Alan B Hooper. Electron transfer during the oxidation of ammonia by the chemolithotrophic bacterium Nitrosomonas europaea. *Biochemistry and Biophysics*, 1459:346–355, 2000.

C Wittmann and S Y Lee. Genome-scale network modeling. In C Wittmann and S Y Lee, editors, *Systems Metabolic Engineering*, chapter 1. Springer Netherlands, Dordrecht, 1 edition, 2012.

Xiao-E Yang, Xiang Wu, Hu-Lin Hao, and Zhen-Li He. Mechanisms and assessment of water eutrophication. *J Zhejiang Univ Sci B*, 9(3):197–209, 2008.

Ran Yu, Xiaohua Fang, Ponisseril Somasundaran, and Kartik Chandran. Chemosphere Short-term effects of TiO 2 , CeO 2 , and ZnO nanoparticles on metabolic activities and gene expression of Nitrosomonas europaea. *Chemosphere*, 128:207–215, 2015.

A

ANNEX

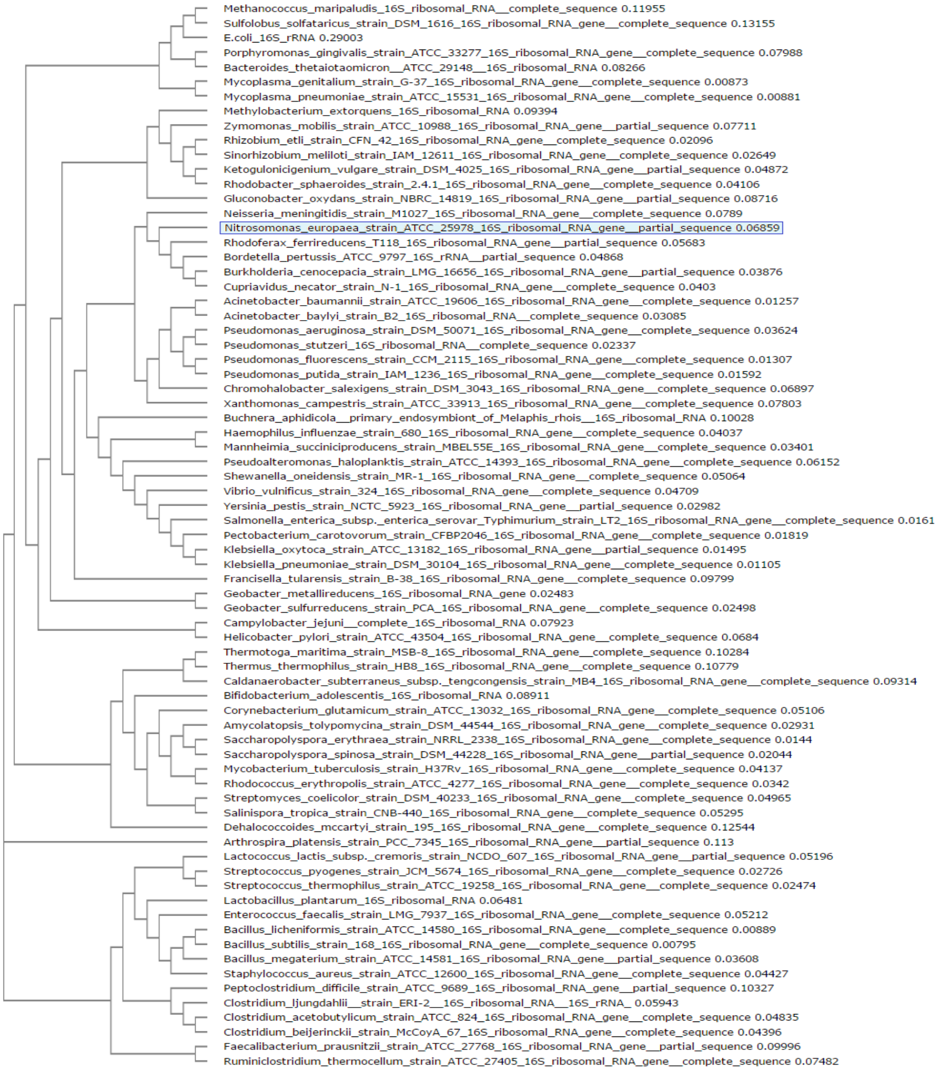

Methanococcus_maripaludis_16S_ribosomal_RNA__complete_sequence 0.11955
Sulfolobus_solfataricus_strain_DSM_1616_16S_ribosomal_RNA_gene__complete_sequence 0.13155
E.coli_16S_rRNA 0.29003
Porphyromonas_gingivalis_strain_ATCC_33277_16S_ribosomal_RNA_gene__complete_sequence 0.07988
Bacteroides_thetaiotaomicron__ATCC_29148__16S_ribosomal_RNA 0.08266
Mycoplasma_genitalium_strain_G-37_16S_ribosomal_RNA_gene__complete_sequence 0.00873
Mycoplasma_pneumoniae_strain_ATCC_15531_16S_ribosomal_RNA_gene__complete_sequence 0.00881
Methylobacterium_extorquens_16S_ribosomal_RNA 0.09394
Zymomonas_mobilis_strain_ATCC_10988_16S_ribosomal_RNA_gene__partial_sequence 0.07711
Rhizobium_etli_strain_CFN_42_16S_ribosomal_RNA_gene__complete_sequence 0.02096
Sinorhizobium_meliloti_strain_IAM_12611_16S_ribosomal_RNA_gene__complete_sequence 0.02649
Ketogulonicigenium_vulgare_strain_DSM_4025_16S_ribosomal_RNA_gene__partial_sequence 0.04872
Rhodobacter_sphaeroides_strain_2.4.1_16S_ribosomal_RNA_gene__complete_sequence 0.04106
Gluconobacter_oxydans_strain_NBRC_14819_16S_ribosomal_RNA_gene__partial_sequence 0.08716
Neisseria_meningitidis_strain_M1027_16S_ribosomal_RNA_gene__complete_sequence 0.0789
Nitrosomonas_europaea_strain_ATCC_25978_16S_ribosomal_RNA_gene__partial_sequence 0.06859
Rhodoferax_ferrireducens_T118_16S_ribosomal_RNA_gene__partial_sequence 0.05683
Bordetella_pertussis_ATCC_9797_16S_rRNA__partial_sequence 0.04868
Burkholderia_cenocepacia_strain_LMG_16656_16S_ribosomal_RNA_gene__partial_sequence 0.03876
Cupriavidus_necator_strain_N-1_16S_ribosomal_RNA_gene__complete_sequence 0.0403
Acinetobacter_baumannii_strain_ATCC_19606_16S_ribosomal_RNA_gene__complete_sequence 0.01257
Acinetobacter_baylyi_strain_B2_16S_ribosomal_RNA__complete_sequence 0.03085
Pseudomonas_aeruginosa_strain_DSM_50071_16S_ribosomal_RNA_gene__complete_sequence 0.03624
Pseudomonas_stutzeri_16S_ribosomal_RNA__complete_sequence 0.02337
Pseudomonas_fluorescens_strain_CCM_2115_16S_ribosomal_RNA_gene__complete_sequence 0.01307
Pseudomonas_putida_strain_IAM_1236_16S_ribosomal_RNA_gene__complete_sequence 0.01592
Chromohalobacter_salexigens_strain_DSM_3043_16S_ribosomal_RNA_gene__complete_sequence 0.06897
Xanthomonas_campestris_strain_ATCC_33913_16S_ribosomal_RNA_gene__complete_sequence 0.07803
Buchnera_aphidicola__primary_endosymbiont_of_Melaphis_rhois__16S_ribosomal_RNA 0.10028
Haemophilus_influenzae_strain_680_16S_ribosomal_RNA_gene__complete_sequence 0.04037
Mannheimia_succiniciproducens_strain_MBEL55E_16S_ribosomal_RNA_gene__complete_sequence 0.03401
Pseudoalteromonas_haloplanktis_strain_ATCC_14393_16S_ribosomal_RNA_gene__complete_sequence 0.06152
Shewanella_oneidensis_strain_MR-1_16S_ribosomal_RNA_gene__complete_sequence 0.05064
Vibrio_vulnificus_strain_324_16S_ribosomal_RNA_gene__complete_sequence 0.04709
Yersinia_pestis_strain_NCTC_5923_16S_ribosomal_RNA_gene__partial_sequence 0.02982
Salmonella_enterica_subsp._enterica_serovar_Typhimurium_strain_LT2_16S_ribosomal_RNA_gene__complete_sequence 0.0161
Pectobacterium_carotovorum_strain_CFBP2046_16S_ribosomal_RNA_gene__complete_sequence 0.01819
Klebsiella_oxytoca_strain_ATCC_13182_16S_ribosomal_RNA_gene__partial_sequence 0.01495
Klebsiella_pneumoniae_strain_DSM_30104_16S_ribosomal_RNA_gene__complete_sequence 0.01105
Francisella_tularensis_strain_B-38_16S_ribosomal_RNA_gene__complete_sequence 0.09799
Geobacter_metallireducens_16S_ribosomal_RNA_gene 0.02483
Geobacter_sulfurreducens_strain_PCA_16S_ribosomal_RNA_gene__complete_sequence 0.02498
Campylobacter_jejuni__complete_16S_ribosomal_RNA 0.07923
Helicobacter_pylori_strain_ATCC_43504_16S_ribosomal_RNA_gene__complete_sequence 0.0684
Thermotoga_maritima_strain_MSB-8_16S_ribosomal_RNA_gene__complete_sequence 0.10284
Thermus_thermophilus_strain_HB8_16S_ribosomal_RNA_gene__complete_sequence 0.10779
Caldanaerobacter_subterraneus_subsp._tengcongensis_strain_MB4_16S_ribosomal_RNA_gene__complete_sequence 0.09314
Bifidobacterium_adolescentis_16S_ribosomal_RNA 0.08911
Corynebacterium_glutamicum_strain_ATCC_13032_16S_ribosomal_RNA_gene__complete_sequence 0.05106
Amycolatopsis_tolypomycina_strain_DSM_44544_16S_ribosomal_RNA_gene__complete_sequence 0.02931
Saccharopolyspora_erythraea_strain_NRRL_2338_16S_ribosomal_RNA_gene__complete_sequence 0.0144
Saccharopolyspora_spinosa_strain_DSM_44228_16S_ribosomal_RNA_gene__partial_sequence 0.02044
Mycobacterium_tuberculosis_strain_H37Rv_16S_ribosomal_RNA_gene__complete_sequence 0.04137
Rhodococcus_erythropolis_strain_ATCC_4277_16S_ribosomal_RNA_gene__complete_sequence 0.0342
Streptomyces_coelicolor_strain_DSM_40233_16S_ribosomal_RNA_gene__complete_sequence 0.04965
Salinispora_tropica_strain_CNB-440_16S_ribosomal_RNA_gene__complete_sequence 0.05295
Dehalococcoides_mccartyi_strain_195_16S_ribosomal_RNA_gene__complete_sequence 0.12544
Arthrospira_platensis_strain_PCC_7345_16S_ribosomal_RNA_gene__partial_sequence 0.113
Lactococcus_lactis_subsp._cremoris_strain_NCDO_607_16S_ribosomal_RNA_gene__partial_sequence 0.05196
Streptococcus_pyogenes_strain_JCM_5674_16S_ribosomal_RNA_gene__complete_sequence 0.02726
Streptococcus_thermophilus_strain_ATCC_19258_16S_ribosomal_RNA_gene__complete_sequence 0.02474
Lactobacillus_plantarum_16S_ribosomal_RNA 0.06481
Enterococcus_faecalis_strain_LMG_7937_16S_ribosomal_RNA_gene__complete_sequence 0.05212
Bacillus_licheniformis_strain_ATCC_14580_16S_ribosomal_RNA_gene__complete_sequence 0.00889
Bacillus_subtilis_strain_168_16S_ribosomal_RNA_gene__complete_sequence 0.00795
Bacillus_megaterium_strain_ATCC_14581_16S_ribosomal_RNA_gene__partial_sequence 0.03608
Staphylococcus_aureus_strain_ATCC_12600_16S_ribosomal_RNA_gene__complete_sequence 0.04427
Peptoclostridium_difficile_strain_ATCC_9689_16S_ribosomal_RNA_gene__partial_sequence 0.10327
Clostridium_ljungdahlii__strain_ERI-2__16S_ribosomal_RNA__16S_rRNA_ 0.05943
Clostridium_acetobutylicum_strain_ATCC_824_16S_ribosomal_RNA_gene__complete_sequence 0.04835
Clostridium_beijerinckii_strain_McCoyA_67_16S_ribosomal_RNA_gene__complete_sequence 0.04396
Faecalibacterium_prausnitzii_strain_ATCC_27768_16S_ribosomal_RNA_gene__partial_sequence 0.09996
Ruminiclostridium_thermocellum_strain_ATCC_27405_16S_ribosomal_RNA_gene__complete_sequence 0.07482

Figure 13.: Phylogenetic tree of the organisms with a GSMM, found in three databases. The distance of each organism is presented in front of its name. Tree constructed by MUSCLE, using 16S rRNA as base of reference.

Table 11.: Manual annotation of 50 random genes of *N. europaea*. Each gene is identified by its Gene ID, and has it own score calculated by *merlin*. The genes are grouped by a score, and each group ranges by 0.1 score. The EC number and Function reflects the annotation itself and is evaluated by a Classification Level (CL), which tells how reliable the annotation is.

Group Score	Gene ID	Score	EC number	Function	CL
0.0	NE0330	0.05	-	Glycosyl transferase, family 2	F
	NE0396	0.02	-	Uncharacterized protein	-
	NE0458	0.01	-	Uncharacterized protein	-
	NE0476	0.05	-	Uncharacterized protein	-
	NE0902	0.05	-	Periplasmic component of the Tol biopolymer transport system	G
0.1	NE0017	0.15	1.1.1.144	Anaerobic nitric oxide reductase transcription regulator NorR	D
	NE0522	0.16	1.8.-.-	Glutathione S-transferase, C-terminal domain	E
	NE0020	0.19	5.1.99.6	Bifunctional NAD(P)H-hydrate repair enzyme Nnr	D
	NE0025	0.13	3.1.-.-	Ribonuclease TTHA0252	E
	NE0029	0.18	1.-.-.-	Uncharacterized oxidoreductase YciK	D
0.2	NE0035	0.2	3.4.21.-	Putative Lon protease homolog	D
	NE0048	0.27	4.1.2.13	Aldolase	F
	NE0063	0.28	3.6.3.31	Spermidine/putrescine import ATP-binding protein PotA	C

Continued on next page

Table 11 – continued from previous page

Group Score	Gene ID	Score	EC number	Function	CL
	NE0080	0.24	3.1.4.52	Diguanylate cyclase/phosphodiesterase domain 2 (EAL)	D
	NE0085	0.26	2.1.1.176	Ribosomal RNA small subunit methyl--transferase B	D
0.3	NE0005	0.39	5.4.99.21	23S rRNA pseudouridine(2604) synthase	D
	NE1738	0.33	2.7.13.3	Histidine kinase	F
	NE0034	0.37	2.6.1.-	Serine–pyruvate aminotransferase	E
	NE0040	0.38	2.8.1.1	Rhodanese/cdc25 fold	F
	NE0107	0.3	3.6.3.-	Methionine import ATP-binding protein MetN	E
0.4	NE0015	0.4	2.7.13.3	Sensor histidine kinase GlrK	D
	NE1652	0.45	2.4.1.83	Dolichyl-phosphate beta-D-mannosyl--transferase	E
	NE0071	0.44	1.13.12.-	Ammonia monooxygenase subunit C	F
	NE0142	0.4	2.6.1.11	Acetylornithine aminotransferase	A
	NE0190	0.4	2.7.7.-	Nucleotidyl transferase	G
0.5	NE0010	0.56	3.1.-.-	Single-stranded-DNA -specific exonuclease RecJ	D
	NE0056	0.5	3.2.2.-	Adenine DNA glycosylase	D
	NE0076	0.53	3.4.24.84	CAAX prenyl protease 1 homolog	E

Continued on next page

Table 11 – continued from previous page

Group Score	Gene ID	Score	EC number	Function	CL
	NE0208	0.51	2.3.1.157 2.7.7.23	Bifunctional protein GlmU	A
	NE0362	0.5	3.5.4.9 1.5.1.5	Bifunctional protein FolD	A
0.6	NE0023	0.65	3.1.11.2	Exodeoxyribonuclease III:Exodeoxyribonuclease III xth	D
	NE0026	0.63	2.6.1.11	Adenosylmethionine-8--amino-7-oxononanoate aminotransferase	D
	NE0037	0.62	2.1.1.77	Protein-L-isoaspartate O-methyltransferase	E
	NE0053	0.61	3.4.16.4	D-alanyl-D-alanine carboxypeptidase	E
	NE0054	0.68	3.6.3.-	Lipid A export ATP-binding/permease protein MsbA	D
0.7	NE0014	0.7	4.1.3.27	Anthranilate synthase component 2	E
	NE0049	0.72	2.7.13.3	Sensory transduction histidine kinases	D
	NE0093	0.74	3.6.4.12	ATP-dependent DNA helicase RecQ	D
	NE0141	0.77	2.7.7.7	DNA polymerase III subunit epsilon	D
	NE0110	0.77	3.1.-.-	CRISPR-associated endoribonuclease Cas2 1	A
0.8	NE0111	0.81	3.1.-.-	CRISPR-associated endonuclease Cas1 1	E
	NE0112	0.83	3.1.-.-	CRISPR-associated endoribonuclease Cas2 2	A
	NE0204	0.87	3.6.3.14	ATP synthase subunit alpha	A
	NE0206	0.86	3.6.3.14	ATP synthase subunit beta	A

Continued on next page

Table 11 – continued from previous page

Group Score	Gene ID	Score	EC number	Function	CL
	NE0213	0.88	3.6.4.12	Holliday junction ATP-dependent DNA helicase RuvB	A
0.9	NE0002	0.94	2.7.7.7	DNA polymerase III subunit beta	E
	NE0003	0.94	5.99.1.3	DNA gyrase subunit B	E
	NE0012	0.95	4.1.1.48	Indole-3-glycerol phosphate synthase	E
	NE0013	0.94	2.4.2.18	Anthranilate phosphoribosyltransferase	A
	NE0019	0.95	6.3.5.3	Phosphoribosylformyl- -glycinamidine synthase	D

Table 12.: Manual annotation of the 385 genes of *N. europaea*, being 348 of them the pruposed ones by the New metric. Each gene is identified by its Gene ID, and has it own score calculated by *merlin*. The genes are grouped by a score, and each group ranges by 0.1 score. The EC number and Function reflects the annotation itself and is evaluated by a Classification Level (CL), which tells how reliable the annotation is.

Gene ID	EC number	Function	CL
NE0005	5.4.99.21	23S rRNA pseudouridine(2604) synthase	D
NE0010	3.1.-.-	Single-stranded-DNA-specific exonuclease RecJ	D
NE0015	2.7.13.3	Sensory transduction histidine kinases	D
NE0016	-	Uncharacterized protein	-
NE0026	2.6.1.82	Putrescine aminotransferase	D
NE0034	2.6.1.-	Aminotransferase class-V	E
NE0037	2.1.1.77	Protein-L-isoaspartate O-methyltransferase	E
NE0040	2.8.1.1	Rhodanese	F
NE0048	-	Class II Aldolase and Adducin N-terminal domain	F
NE0053	3.4.16.4	D-alanyl-D-alanine carboxypeptidase	E
NE0056	3.2.2.-	Adenine DNA glycosylase	D
NE0060	2.7.1.69	PTS sugar transporter subunit IIA	F

Continued on next page

Table 12 – continued from previous page

Gene ID	EC number	Function	CL
NE0063	3.6.3.31	Spermidine/putrescine import ATP-binding protein PotA	C
NE0071	2.7.1.113	Deoxynucleoside kinase	F
NE0076	3.4.24.84	CAAX prenyl protease 1 homolog	E
NE0080	3.1.4.52	Diguanylate cyclase/phosphodiesterase domain 2	D
NE0085	2.1.1.176	Ribosomal RNA small subunit methyltransferase B	D
NE0091	2.7.-.-	Probable protein kinase UbiB	D
NE0097	-	HAD-superfamily subfamily IB hydrolase, TIGR01490	F
NE0103	-	Uncharacterized protein YyaL	E
NE0107	3.6.3.-	Methionine import ATP-binding protein MetN	D
NE0142	2.6.1.11	Acetylornithine aminotransferase	A
NE0159	3.6.3.-	Lipid A export ATP-binding/permease protein MsbA	E
NE0174	2.8.1.7	Cysteine desulfurase IscS	D
NE0184	3.6.1.-	NUDIX hydrolase	D
NE0190	2.7.7.-	Nucleotidyl transferase	F
NE0202	-	ATP synthase subunit b	A
NE0205	-	ATP synthase subunit gamma chain	A
NE0208	2.7.7.23 2.3.1.157	Bifunctional protein GlmU	A
NE0214	3.1.2.-	4-hydroxybenzoyl-CoA thioesterase family active site	E
NE0217	-	Proline-rich region	I
NE0224	2.3.1.n3	Glycerol-3-phosphate acyltransferase	A
NE0233	-	Toprim domain	I
NE0278	1.3.99.-	Oxygen-independent coproporphyrinogen-III oxidase-like protein YggW	D
NE0292	3.1.-.-	Toxin YhaV	D
NE0310	3.2.1.-	Peptidoglycan hydrolase FlgJ	D

Continued on next page

Table 12 – continued from previous page

Gene ID	EC number	Function	CL
NE0320	-	GCN5-related N-acetyltransferase	G
NE0334	1.1.1.95	D-3-phosphoglycerate dehydrogenase	E
NE0335	4.2.1.51	Prephenate dehydratase (PDT):Chorismate mutase:ACT domain	D
NE0337	1.3.1.12	Prephenate dehydrogenase	E
NE0343	2.7.13.3	Sensor protein QseC	D
NE0347	2.5.1.16	Polyamine aminopropyltransferase	A
NE0355	3.1.26.12	Ribonuclease E	D
NE0362	1.5.1.5 3.5.4.9	Bifunctional protein FolD	A
NE0368	2.7.6.5	GTP pyrophosphokinase	E
NE0370	1.1.5.3	Aerobic glycerol-3-phosphate dehydrogenase	D
NE0371	3.1.4.46	Glycerophosphoryl diester phosphodiesterase	F
NE0376	3.6.3.-	Lipid A export ATP-binding/permease protein MsbA	D
NE0377	-	Sensor signal transduction histidine kinases	F
NE0378	2.7.8.31	UDP-glucose:undecaprenyl-phosphate glucose-1-phosphate transferase	D
NE0379	3.4.21.-	Serine proteases, trypsin family	G
NE0380	3.1.17	L-serine dehydratase 2 4	D
NE0382	3.1.21.3	Putative type I restriction enzyme HindVIIP R protein	E
NE0384	3.1.21.3	Restriction modification system, type I	I
NE0385	2.1.1.72	Putative type I restriction enzyme HindVIIP M protein	E
NE0397	1.1.1.44	6-phosphogluconate dehydrogenase, decarboxylating	D
NE0438	1.1.1.38	NAD-dependent malic enzyme, mitochondrial	D
NE0439	3.1.3.3	Phosphoserine phosphatase	E

Continued on next page

Table 12 – continued from previous page

Gene ID	EC number	Function	CL
NE0441	3.4.11.10 3.4.11.1	Probable cytosol aminopeptidase	A
NE0442	2.7.7.7	DNA polymerase III subunit chi	F
NE0456	-	Alpha/beta hydrolase fold protein	G
NE0466	2.4.1.1	Glycogen phosphorylase	E
NE0483	3.6.3.40	Teichoic acids export ATP-binding protein TagH	E
NE0499	-	Glycosyl transferase, family 2	F
NE0500	5.1.3.2	putative UDP-glucose 4-epimerase	E
NE0502	4.2.1.115	Polysaccharide biosynthesis protein CapD	E
NE0515	2.7.13.3	Signal transduction histidine-protein kinase BaeS	D
NE0525	1.1.1.205	CBS domain	H
NE0532	2.1.1.107 1.3.1.76 4.99.1.4	Siroheme synthase	A
NE0553	3.1.-.-	Ribonuclease VapC	I
NE0569	3.4.11.10	Probable cytosol aminopeptidase	A
NE0585		Uncharacterized protein	-
NE0591	4.2.1.75	Uroporphyrinogen-III synthase	D
NE0592	2.1.1.107	Possible uroporphyrin-III C-methyltransferase	G
NE0594	1.18.1.3	putative reductase oxidoreductase protein	E
NE0604	2.1.1.163 2.1.1.201	Possible ubiE ubiquinone/menaquinone biosynthesis methyltransferase	F
NE0611	2.7.1.-	Uncharacterized sugar kinase MJ0406	E
NE0612	2.3.1.51	1-acyl-sn-glycerol-3-phosphate acyltransferase	E
NE0618	3.6.3.8	Calcium-transporting ATPase	E
NE0620	4 1.1.1.1	Alcohol dehydrogenase	E
NE0626	3.4.11.2	Aminopeptidase N	D
NE0640	3.-.-.-	HIT (Histidine triad) family	E

Continued on next page

Table 12 – continued from previous page

Gene ID	EC number	Function	CL
NE0650	2.8.1.10 1.5.3.-	Bifunctional protein ThiO/ThiG	E
NE0652	-	Biotin carboxyl carrier protein of acetyl-CoA carboxylase	E
NE0653	6.3.4.14	Acetyl-/propionyl-coenzyme A carboxylase alpha chain	E
NE0674	1.1.3.15	Glycolate oxidase subunit GlcE	D
NE0675	-	Glycolate oxidase subunit GlcD	D
NE0678	5.1.3.13	dTDP-4-dehydrorhamnose 3,5-epimerase	D
NE0679	5.1.3.2	UDP-glucose 4-epimerase	E
NE0683	1.9.3.1	Cytochrome c oxidase, subunit I	F
NE0684	1.9.3.1	Cytochrome c oxidase, subunit I	F
NE0696	6.3.2.17 6.3.2.12	Bifunctional protein FolC	D
NE0713	3.1.-.-	Toxin YoeB	D
NE0723	3.4.21.-	Extracellular serine protease	E
NE0728	2.7.13.3	Histidine kinase	G
NE0741	1.17.99.1	4-cresol dehydrogenase [hydroxylating] flavoprotein subunit	E
NE0742	1.1.1.1	Alcohol dehydrogenase 4	E
NE0757	6.6.1.1	Magnesium-chelatase subunit ChlH, chloroplastic	E
NE0772	1.11.1.15	Putative peroxiredoxin bcp	E
NE0774	1.8.1.4	Dihydrolipoyl dehydrogenase	E
NE0775	2.1.1.-	Uncharacterized Transfer Ribonucleic Acid (tRNA)/rRNA methyltransferase slr1673	E
NE0777	3.1.1.3	Esterase/lipase/thioesterase family active site	G
NE0781	3.1.3.25	Inositol-1-monophosphatase	E
NE0782	1.-.-.-	Multicopper oxidase com	E
NE0793	1.1.1.193	Riboflavin biosynthesis protein RibD	D
NE0794	-	Glycosyl transferase, group 1	F
NE0795	-	Glycosyl transferase, group 1	F

Continued on next page

Table 12 – continued from previous page

Gene ID	EC number	Function	CL
NE0796	-	Putative CapK protein	F
NE0800	-	Uncharacterized protein	-
NE0811	-	Cytochrome c1	E
NE0820	1.1.1.1	Zinc-containing alcohol dehydrogenase superfamily	E
NE0825	3.6.3.-	Macrolide export ATP-binding/permease protein	A
NE0826	3.6.3.-	Macrolide export ATP-binding/permease protein	A
NE0827	3.6.3.-	Macrolide export ATP-binding/permease protein	A
NE0832	3.1.2.28	1,4-dihydroxy-2-naphthoyl-CoA hydrolase	D
NE0833	3.6.4.13	HrpA-like helicases	D
NE0848	-	Phosphoglycerate mutase family	G
NE0850	2 3.1.2.-	Acyl-protein thioesterase	E
NE0855	1.8.4.8	Phosphoadenosine phosphosulfate reductase	E
NE0859	1.6.1.2	NAD(P) transhydrogenase subunit alpha	D
NE0860	1.6.1.2	NAD(P) transhydrogenase subunit alpha part 2	E
NE0863	1.16.3.1	Bacterioferritin	E
NE0873	1.14.13.-	2-octaprenyl-6-methoxyphenol hydroxylase	D
NE0876	2.1.2.3 3.5.4.10	Bifunctional purine biosynthesis protein PurH	D
NE0880	3.6.4.12	probable ATP-dependent DNA helicase-related protein	D
NE0882	5.2.1.8	PpiC-type peptidyl-prolyl cis-trans isomerase	A
NE0899	1.1.1.28	D-lactate dehydrogenase	D
NE0903	1.1.99.3	probable cytochrome c	E
NE0909	4.1.1.17	Ornithine decarboxylase	E

Continued on next page

Table 12 – continued from previous page

Gene ID	EC number	Function	CL
NE0917	2.7.7.6	RNA polymerase factor sigma-70	F
NE0919	3.6.3.-	Lipid A export ATP-binding/permease protein MsbA	E
NE0922	6.3.2.30	Cyanophycin synthetase	E
NE0923	6.3.2.3	Glutathione biosynthesis bifunctional protein GshAB	E
NE0924	1.7.2.1	Copper-containing nitrite reductase	E
NE0932	-	Putative isomerase	G
NE0938	-	Uncharacterized protein	-
NE0940	-	Putative DNA transport competence protein, ComEA	F
NE0943	1.13.12.-	Ammonia monooxygenase	A
NE0944	1.13.12.-	Ammonia monooxygenase, acetylene-binding	A
NE0945	1.13.12.-	Ammonia monooxygenase subunit C	F
NE0947	3.1.3.-	hydrolase family	E
NE0969	2.1.1.171	Ribosomal RNA small subunit methyltransferase D	E
NE0970	-	Uncharacterized zinc protease-like protein y4wB	E
NE0974	-	PemK-like protein	G
NE0981	-	HhH-GPD	G
NE0985	2.4.1.129	Peptidoglycan synthase FtsI	E
NE1003	5.1.3.15	Putative glucose-6-phosphate 1-epimerase	D
NE1004	-	Uncharacterized protein	-
NE1009	1.1.2.4	D-lactate dehydrogenase [cytochrome], mitochondrial	E
NE1013	1.9.3.1	Cytochrome c oxidase subunit 3	E
NE1019	3.6.3.54	Copper-exporting P-type ATPase A	D
NE1024	3.4.21.-	Putative signal peptide peptidase SppA	D
NE1031	3.6.3.31	Spermidine/putrescine import ATP-binding protein PotA	A

Continued on next page

Table 12 – continued from previous page

Gene ID	EC number	Function	CL
NE1033	4.2.2.n	Membrane-bound lytic murein transglycosylase B	D
NE1046	1.3.5.1	Succinate dehydrogenase, cytochrome b subunit	F
NE1047	1.3.5.1	Succinate dehydrogenase subunit D	F
NE1067	3.1.-.-	Putative plasmid stability-like protein	I
NE1123	3.5.1.97	Acyl-homoserine lactone acylase QuiP	E
NE1125	6.2.1.-	Probable crotonobetaine/carnitine-CoA ligase	E
NE1126	4.1.1.20	Diaminopimelate decarboxylase	E
NE1127	6.3.5.4	Asparagine synthetase [glutamine-hydrolyzing] 1	E
NE1137	2.7.7.7	DNA polymerase III subunit delta	D
NE1160	2.5.1.10	Farnesyl diphosphate synthase	D
NE1165	1.-.-.-	Short-chain dehydrogenase/reductase (SDR) superfamily	E
NE1168	5.4.99.17	Probable squalene–hopene cyclase	E
NE1170	1.1.1.219	Putative dihydroflavonol 4-reductase	E
NE1174	2.1.1.264	Ribosomal RNA large subunit methyltransferase K/L	E
NE1184	2.3.1.51	Phospholipid/glycerol acyltransferase	F
NE1212	2.7.1.4	PfkB family of carbohydrate kinase	F
NE1213	2 2.4.1.14	Probable sucrose-phosphate synthase	E
NE1216	3.6.3.54	Copper-exporting P-type ATPase A	D
NE1227	-	Uncharacterized protein RC0076	E
NE1232	-	Uncharacterized protein y4iL	E
NE1237	1.1.-.-	Glucose-methanol-choline (GMC) oxidoreductase	E
NE1239	1.13.11.34	Arachidonate 5-lipoxygenase	E
NE1241	1.14.18.1	Tyrosinase	E
NE1247	5.4.3.-	DUF160	E
NE1250	2.7.13.3	Chemotaxis protein CheA	D
NE1273		Uncharacterized protein YhiN	D
NE1288	2.7.13.3	Phosphate regulon sensor protein PhoR	D

Continued on next page

Table 12 – continued from previous page

Gene ID	EC number	Function	CL
NE1294	6.3.2.2	Putative glutamate–cysteine ligase	F
NE1299	4.6.1.2	Guanylate cyclase	G
NE1306	-	Uncharacterized protein	-
NE1314	2.7.13.3	Sensor histidine kinase RegB	E
NE1321	2.7.8.8	CDP-diacylglycerol–serine O-phosphatidyltransferase	E
NE1324	2.2.1.6	Acetolactate synthase isozyme 3 small subunit	D
NE1327	3.5.-.-	Hydrolase sll0601	E
NE1331	3.4.19.13	Gamma-glutamyltranspeptidase	D
NE1332	-	Possible capsular polysaccharide biosynthesis/export transmembrane	G
NE1333	-	Short-chain dehydrogenase/reductase (SDR) superfamily	F
NE1334	2.4.1.-	Glycosyl transferase, family 2	E
NE1336	-	Glycosyl transferase, group 1	F
NE1343	1.1.1.22	UDP-glucose 6-dehydrogenase	D
NE1370	-	Glycosyl transferase, family 2	E
NE1373	-	Uncharacterized protein	-
NE1388	2.3.1.47	8-amino-7-oxononanoate synthase	A
NE1389	2.3.1.-	putative type I polyketide synthase WcbR	E
NE1398	2.7.7.7	Putative DNA polymerase-related protein, bacteriophage-type	F
NE1399	2.3.1.128	GCN5-related N-acetyltransferase	F
NE1403	-	Uncharacterized protein YeaO	D
NE1404	3.6.3.-	Macrolide export ATP-binding/permease protein MacB	D
NE1407	-	Uncharacterized protein	-
NE1408	3.1.3.-	Sensor protein PhoQ	E
NE1411	1.13.12.-	Ammonia monooxygenase subunit C	F
NE1414	3.6.3.-	Methionine import ATP-binding protein MetN	D
NE1416	3.4.24.-	Insulinase family (Peptidase family M16)	E

Continued on next page

Table 12 – continued from previous page

Gene ID	EC number	Function	CL
NE1420	1.18.1.-	Nitric oxide reductase FlRd-NAD(+) reductase	D
NE1425	2.7.4.7	Hydroxymethylpyrimidine/phospho-methylpyrimidine kinase	D
NE1430	3.4.24.-	Peptidase family M23/M37	E
NE1454	3.6.3.-	Lipoprotein-releasing system ATP-binding protein LolD	A
NE1455	3.1.-.-	GDSL lipolytic enzyme	E
NE1463	6.3.2.5	Coenzyme A biosynthesis bifunctional protein CoaBC	E
NE1467	-	Fatty acid desaturase, type 2:Fatty acid desaturase, type 1	F
NE1485	3.4.16.4	D-alanyl-D-alanine carboxypeptidase DacC	E
NE1486	2.6.1.21	D-alanine aminotransferase	E
NE1496	3.6.1.25	Inorganic triphosphatase	A
NE1498	2.5.1.44	Homospermidine synthase	E
NE1503	-	Rieske iron-sulfur protein 2Fe-2S subunit	F
NE1508	3.4.21.107	Periplasmic pH-dependent serine endoprotease DegQ	D
NE1510	4.2.2.n	Membrane-bound lytic murein transglycosylase A	E
NE1512	1.14.13.-	2-octaprenylphenol hydroxylase	D
NE1514	-	tRNA-modifying protein YgfZ	D
NE1516	3.1.-.-	Uncharacterized protein family	D
NE1517	4.2.3.12	6-pyruvoyl tetrahydropterin synthase	G
NE1528	1.1.1.35	Probable 3-hydroxyacyl-CoA dehydrogenase	E
NE1543	1.-.-.-	Hephaestin	E
NE1548	1.3.99.-	Acyl-coenzyme A dehydrogenase	D
NE1549	6.2.1.3	Putative long-chain-fatty-acid–CoA ligase	E

Continued on next page

Table 12 – continued from previous page

Gene ID	EC number	Function	CL
NE1567	-	Short-chain dehydrogenase/reductase (SDR) superfamily	F
NE1569	2.7.7.82	CMP-N,N′-diacetyllegionaminic acid synthase	E
NE1570	2.5.1.97	Pseudaminic acid synthase	E
NE1589	-	Uncharacterized protein	-
NE1613	3.6.3.-	Lipid A export ATP-binding/permease protein MsbA	D
NE1614	3.1.3.- 2.7.7.72 3.1.4.-	Multifunctional CCA protein	A
NE1615	4.2.2.n	Soluble lytic murein transglycosylase	E
NE1625	3.1.13.1	Ribonuclease II domain	F
NE1635	-	Uncharacterized protein	-
NE1651	4.2.2.-	Endolytic murein transglycosylase	D
NE1652	2.4.1.83	Dolichol-phosphate mannosyltransferase	E
NE1655	2.3.1.47	8-amino-7-oxononanoate synthase	E
NE1658	6.2.1.-	AMP-dependent synthetase and ligase	E
NE1666	2.4.2.9	Bifunctional protein PyrR	E
NE1678	1.20.4.1	Arsenate reductase	F
NE1687	4.1.2.52	4-hydroxy-2-oxo-heptane-1,7-dioate aldolase	E
NE1688	1.1.1.95	D-3-phosphoglycerate dehydrogenase	E
NE1689	-	Possible epimerase	F
NE1691	3.1.1.31	6-phosphogluconolactonase	E
NE1697	2.5.1.48	Cystathionine gamma-synthase	D
NE1700	3.1.4.52	Diguanylate cyclase/phosphodiesterase domain 2 (EAL)	D
NE1701	1.8.4.11	Peptide methionine sulfoxide reductase MsrA	E
NE1726	3.4.24.-	Putative integral membrane transmembrane protein	E
NE1733	3.6.1.3	Chaperone protein ClpB	E

Continued on next page

Table 12 – continued from previous page

Gene ID	EC number	Function	CL
NE1738	2.7.13.3	Sensory transduction histidine kinases	F
NE1745	3.6.1.40	Guanosine-5′-triphosphate,3′-diphosphate pyrophosphatase	D
NE1763	1.1.1.-	Putative membrane-bound dehydrogenase oxidoreductase protein	E
NE1765	1.6.5.-	NAD(P)H-quinone oxidoreductase chain 4 1	E
NE1766	1.6.5.-	NAD(P)H-quinone oxidoreductase subunit 5, chloroplastic	E
NE1768	1.6.5.11	NADH-quinone oxidoreductase subunit J	E
NE1772	1.6.5.11	NADH-quinone oxidoreductase subunit F	D
NE1773	1.6.5.11	NADH-quinone oxidoreductase subunit E	E
NE1782	3.4.21.-	Probable CtpA-like serine protease	E
NE1783	2.8.1.11	NAD binding site:UBA/THIF-type NAD/FAD binding fold	E
NE1784	3.6.3.-	Lipid A export ATP-binding/permease protein MsbA	D
NE1785	3.1.3.48	Putative low molecular weight protein-tyrosine-phosphatase slr0328	E
NE1795	6.3.5.4	Asparagine synthetase [glutamine-hydrolyzing] 1	E
NE1796	-	Glycosyl transferase group 1	F
NE1803	-	Putative capsular polysaccharide biosynthetic protein-like protein	F
NE1804	-	Lipopolysaccharide biosynthesis	G
NE1806	3.6.3.-	ATPase components of ABC transporters with duplicated ATPase domains	E
NE1807	3.6.4.13	putative ATP-dependent RNA helicase protein	D
NE1809	3.1.2.30	probable beta subunit of citrate lyase	E
NE1851	3.5.99.10	YER057c/YjgF/UK114 family	E

Continued on next page

Table 12 – continued from previous page

Gene ID	EC number	Function	CL
NE1857	2.1.1.-	Uncharacterized RNA methyltransferase	A
NE1866	2.7.13.3	Chemotaxis protein CheA	D
NE1895	5.4.2.10	Probable phosphoglucosamine mutase	E
NE1897	3.4.-.-	Beta-barrel assembly-enhancing protease	E
NE1899	3.6.3.-	Macrolide export ATP-binding/permease protein MacB	A
NE1900	3.6.3.-	Macrolide export ATP-binding/permease protein MacB	A
NE1901	1.6.5.3	NADH-ubiquinone oxidoreductase chain	E
NE1908	2.5.1.18	Possible glutathione S-transferase family protein	F
NE1909	3.1.4.52	Diguanylate cyclase/phosphodiesterase domain 2 (EAL)	D
NE1915	2.5.1.-	Prenyl transferase	E
NE1919	-	Protein CbbQ	E
NE1952	-	Putative glutathione S-transferase protein	E
NE1954	3.1.3.25	Inositol-1-monophosphatase	E
NE1958	2.5.1.21	Squalene and phytoene synthases	F
NE1974	2.7.13.3	Nitrogen regulation protein NtrY	E
NE1979	3.4.-.-	Uncharacterized protease YegQ	D
NE1982	3.1.5.1	Deoxyguanosinetriphosphate triphosphohydrolase-like protein	A
NE1983	-	Uncharacterized protein	-
NE1999	3.1.-.-	Toxin YhaV	E
NE2000	1.2.1.8	NAD/NADP-dependent betaine aldehyde dehydrogenase	E
NE2003	-	Nitric oxide reductase subunit C	E
NE2004	1.7.2.5	Nitric oxide reductase subunit B	E
NE2005	-	Protein NorQ	E
NE2014	-	Uncharacterized protein	-
NE2015	-	Possible capK protein, putative	F

Continued on next page

Table 12 – continued from previous page

Gene ID	EC number	Function	CL
NE2031	-	Glycosyl hydrolase family 57	F
NE2032	3.2.1.1	Alpha-amylase	E
NE2044	1.7.2.6	Hydroxylamine oxidoreductase	A
NE2053	3.6.5.3	Elongation factor G	A
NE2064	1.13.12.-	Ammonia monooxygenase subunit C	F
NE2067	2.4.1.129	Penicillin-binding protein PbpB	E
NE2075	3.1.4.52	Diguanylate cyclase/phosphodiesterase domain 2 (EAL)	D
NE2082	2.7.13.3	Signal transduction histidine-protein kinase AtoS	D
NE2086	3.6.3.14	Flagellum-specific ATP synthase	D
NE2110	-	Bacterial regulatory proteins, AsnC family	G
NE2112	-	PIN (PilT N terminus) domain	G
NE2113	-	Uncharacterized protein	-
NE2119	-	conserved hypothetical protein	E
NE2123	1.4.1.13	Ferredoxin-dependent glutamate synthase	E
NE2126	1.1.-.-	L-lactate dehydrogenase	D
NE2144	1.13.11.24	Putative quercetin 2,3-dioxygenase PA2418	E
NE2147	3.4.11.9	Xaa-Pro aminopeptidase	D
NE2166	3.6.4.12	DNA helicase	I
NE2167	2.6.1.87	UDP-4-amino-4-deoxy-L-arabinose--oxoglutarate aminotransferase	D
NE2171	2.-.-.-	Formyl transferase N-terminus	I
NE2173	2.4.2.53	Undecaprenyl-phosphate 4-deoxy-4-formamido-L--arabinose transferase	D
NE2178	-	Uncharacterized protein	-
NE2182	3.5.1.28	1,6-anhydro-N-acetylmuramyl-L-alanine amidase AmpD	D
NE2183	-	Uncharacterized protein	-

Continued on next page

Table 12 – continued from previous page

Gene ID	EC number	Function	CL
NE2187	-	SAM (And some other nucleotide) binding motif	I
NE2188	-	Uncharacterized protein	-
NE2192	4.2.99.18	DNA-(apurinic or apyrimidinic site) lyase	E
NE2199	1.14.13.8	Flavin-containing monooxygenase (FMO)	E
NE2212	-	Putative transmembrane protein	G
NE2213	2.3.1.1 2.3.1.35	Arginine biosynthesis bifunctional protein ArgJ	A
NE2215	-	NUDIX hydrolase	I
NE2216	1.6.99.3	FAD-dependent pyridine nucleotide-disulphide oxidoreductase	F
NE2226	4.-.-.-	SLT domain	E
NE2235	6.2.1.3	AMP-dependent synthetase and ligase	E
NE2237	1.17.99.1	4-cresol dehydrogenase [hydroxylating] flavoprotein subunit	E
NE2244	3.6.3.-	Lipoprotein-releasing system ATP-binding protein LolD	A
NE2249	5.1.3.14	UDP-N-acetylglucosamine 2-epimerase	D
NE2250	2.7.7.13	Mannose-1-phosphate guanylyltransferase	D
NE2252	6.1.1.23	Aspartate–tRNA(Asp/Asn) ligase	A
NE2253	3.6.1.-	NUDIX hydrolase	D
NE2259	4.2.1.130	Glutathione-independent glyoxalase HSP31	E
NE2262	2.3.1.9	Acetyl-CoA acetyltransferase	D
NE2267	-	Glycosyl transferases group 1	F
NE2276	1.1.1.336	UDP-N-acetyl-D-mannosamine dehydrogenase	E
NE2277	5.1.3.6	NAD dependent epimerase/dehydratase family	E
NE2278	3.1.3.48	Probable low molecular weight protein-tyrosine-phosphatase EpsP	D

Continued on next page

Table 12 – continued from previous page

Gene ID	EC number	Function	CL
NE2280	2.7.10.-	Tyrosine-protein kinase etk	D
NE2311	-	Possible helicase (Snf2/Rad54 family)	I
NE2317	3.4.-.-	Penicillin-binding protein 1A	D
NE2329	3.4.21.107	Probable periplasmic serine endoprotease DegP-like	D
NE2344	3.2.1.-	Possible unsaturated glucuronyl hydrolase	G
NE2348	1.3.99.-	Acyl-CoA dehydrogenase	E
NE2349	6.2.1.-	AMP-dependent synthetase and ligase	D
NE2368	2.6.1.1	Probable aspartate aminotransferase	E
NE2379	3.6.3.-	Nod factor export ATP-binding protein I	E
NE2384	3.6.3.25	Sulfate/thiosulfate import ATP-binding protein CysA	A
NE2398	-	CBS domain	H
NE2400	4.4.1.8	Protein MalY	D
NE2416	3.5.1.54	Urea amidolyase	E
NE2417	-	Uncharacterized protein	-
NE2418	-	Uncharacterized protein	-
NE2420	6.4.1.2	Biotin carboxylase	D
NE2421	4.2.2.n	Membrane-bound lytic murein transglycosylase D	E
NE2431	-	PIN (PilT N terminus) domain	G
NE2456	1.1.1.81	Putative hydroxypyruvate reductase	E
NE2460	6.3.4.15	Bifunctional ligase/repressor BirA	D
NE2463	2.6.1.83	LL-diaminopimelate aminotransferase	E
NE2465	1.11.1.15	Alkyl hydroperoxide reductase subunit C	D
NE2480	3.6.3.-	Methionine import ATP-binding protein MetN	D
NE2496	3.1.21.3	Restriction modification system, type I	G
NE2497	2.1.1.72	Type I restriction enzyme EcoEI M protein	E
NE2499	3.1.21.3	Type I restriction enzyme EcoKI R protein	D
NE2501	1.1.1.-	L-sorbosone dehydrogenase	E
NE2505	-	Uncharacterized protein	-
NE2510	1.6.99.1	NADH:flavin oxidoreductase/NADH oxidase	E

Continued on next page

Table 12 – continued from previous page

Gene ID	EC number	Function	CL
NE2520	3.6.4.12	ATP-dependent DNA helicase RecQ	D
NE2522	2.1.1.72	Putative type I restriction enzyme MjaXP M protein	E
NE2524	2.1.1.72	Uncharacterized adenine-specific methylase MJ1220	E
NE2526	3.1.21.3	Restriction modification system, type I	H
NE2527	3.1.21.3	Putative type-1 restriction enzyme MjaXP R protein	E
NE2528	3.1.21.-	AAA ATPase superfamily	D
NE2546	3.1.3.18	Phosphoglycolate phosphatase 2	E
NE2547	2.1.1.64 2.1.1.222	Ubiquinone biosynthesis O-methyltransferase	A
NE2555	2.5.1.9	Riboflavin synthase	E
NE2561	3.5.1.42	CinA-like protein	D
NE2564	3.6.4.12	ATP-dependent DNA helicase RecQ	D
NE2567	1.8.5.-	Sulfide:quinone oxidoreductase, mitochondrial	E
NE2571	3.-.-.-	Beta-lactamase hydrolase-like protein	E

Table 13.: Accuracy confusion matrix, resultant from the comparison of the EC number between manual and automatic annotation, of a random 50 gene sample of *N. europaea*. This gene classification was done in 11 instances of different alpha values and in 9 cases of different thresholds. The red, 2 x 2, table below represents where each genes classification is organized in the 99 examples (11 instances x 9 cases).

		Alpha Value																					
		0.0		0.1		0.2		0.3		0.4		0.5		0.6		0.7		0.8		0.9		1.0	
Threshold Value	0.1	32	0	33	0	33	0	33	0	33	0	32	0	32	0	31	1	30	5	30	7	27	13
		13	5	12	5	12	5	12	5	12	5	13	5	13	5	13	5	10	5	8	5	5	5
	0.2	32	0	33	0	33	0	32	1	32	3	30	5	30	6	29	9	27	14	23	21	21	23
		13	5	12	5	12	5	12	5	10	5	10	5	9	5	7	5	4	5	1	5	1	5
	0.3	31	2	32	3	31	5	31	6	30	9	29	10	28	12	26	18	22	22	20	24	19	25
		12	5	10	5	9	5	8	5	6	5	6	5	5	5	1	5	1	5	1	5	1	5
	0.4	30	6	30	8	30	9	30	10	29	12	27	15	24	20	21	23	19	25	17	27	17	28
		9	5	7	5	6	5	5	5	4	5	3	5	1	5	1	5	1	5	1	5	0	5
	0.5	29	9	30	10	29	12	29	12	26	18	24	20	19	25	18	26	17	28	16	29	15	30
		7	5	5	5	4	5	4	5	1	5	1	5	1	5	1	5	0	5	0	5	0	5
	0.6	28	11	28	13	26	17	25	19	20	24	19	25	17	28	16	29	15	30	14	31	14	31
		6	5	4	5	2	5	1	5	1	5	1	5	0	5	0	5	0	5	0	5	0	5
	0.7	26	16	25	19	22	22	18	26	16	29	15	30	14	31	14	31	13	32	13	32	13	32
		3	5	1	5	1	5	1	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5
	0.8	21	23	16	28	12	33	10	35	10	35	10	35	10	35	10	35	11	34	12	33	12	33
		1	5	1	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5
	0.9	16	28	8	37	5	40	5	40	5	40	5	40	5	40	5	40	5	40	5	40	5	40
		1	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5

VP	FN
FP	VN

Table 14.: Lower threshold confusion matrix, resultant from the comparison of the EC number between manual and automatic annotation, of a random 50 gene sample of *N. europaea*. This gene classification was done in 11 instances of different alpha values and in 9 cases of different thresholds. The red, 2 x 2, table below represents where each genes classification is organized in the 99 examples (11 instances x 9 cases).

		Alpha Value																					
		0.0		0.1		0.2		0.3		0.4		0.5		0.6		0.7		0.8		0.9		1.0	
Threshold Value	0.1	33	0	34	0	34	0	34	0	34	0	33	0	33	0	32	1	31	2	31	2	27	7
		15	2	11	5	11	5	11	5	11	5	12	5	12	5	12	5	9	8	7	10	5	11
	0.2	33	0	34	0	34	0	33	1	33	1	31	2	31	2	29	4	27	6	23	10	21	13
		12	5	11	5	11	5	11	5	9	7	9	8	8	9	7	10	4	13	1	16	1	15
	0.3	32	1	33	1	32	2	32	2	30	4	29	4	28	5	26	7	22	11	20	13	19	15
		11	6	9	7	8	8	7	9	6	10	6	11	5	12	1	16	1	16	1	16	1	15
	0.4	31	2	30	4	30	4	30	4	29	5	27	6	24	9	21	12	19	14	17	16	17	17
		8	9	7	9	6	10	5	11	4	12	3	14	1	16	1	16	1	16	1	16	0	16
	0.5	29	4	30	4	29	5	29	5	26	8	24	9	19	14	18	15	17	16	16	17	15	19
		7	10	5	11	4	12	4	12	1	15	1	16	1	16	1	16	0	17	0	17	0	16
	0.6	28	5	28	6	26	8	25	9	20	14	19	14	17	16	16	17	15	18	14	19	14	20
		6	11	4	12	2	14	1	15	1	15	1	16	0	17	0	17	0	17	0	17	0	16
	0.7	26	7	25	9	22	12	18	16	16	18	15	18	14	19	14	19	13	20	13	20	13	21
		3	14	1	15	1	15	1	15	0	16	0	17	0	17	0	17	0	17	0	17	0	16
	0.8	21	12	16	18	12	22	10	24	10	24	10	23	10	23	10	23	11	22	12	21	12	22
		1	16	1	15	0	16	0	16	0	16	0	17	0	17	0	17	0	17	0	17	0	16
	0.9	16	17	8	26	5	29	5	29	5	29	5	28	5	28	5	28	5	28	5	28	5	29
		1	16	0	16	0	16	0	16	0	16	0	17	0	17	0	17	0	17	0	17	0	16

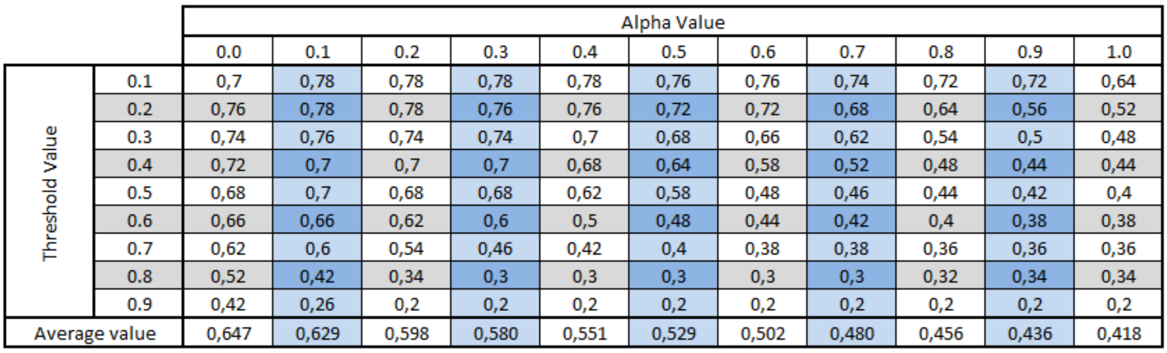
VP	FN
FP	VN

Table 15.: Upper threshold confusion matrix, resultant from the comparison of the EC number between manual and automatic annotation, of a random 50 gene sample of *N. europaea*. This gene classification was done in 11 instances of different alpha values and in 9 cases of different thresholds. The red, 2 x 2, table below represents where each genes classification is organized in the 99 examples (11 instances x 9 cases).

		Alpha Value																					
		0.0		0.1		0.2		0.3		0.4		0.5		0.6		0.7		0.8		0.9		1.0	
Threshold Value	0.1	45	0	45	0	45	0	45	0	45	0	45	0	45	0	44	1	40	5	38	7	32	13
		3	2	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5
	0.2	45	0	45	0	45	0	44	1	42	3	40	5	39	6	36	9	31	14	24	21	22	23
		0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5
	0.3	43	2	42	3	40	5	39	6	36	9	35	10	33	12	27	18	23	22	21	24	20	25
		0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5
	0.4	39	6	37	8	36	9	35	10	33	12	30	15	25	20	22	23	20	25	18	27	17	28
		0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5
	0.5	36	9	35	10	33	12	33	12	27	18	25	20	20	25	19	26	17	28	16	29	15	30
		0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5
	0.6	34	11	32	13	28	17	26	19	21	24	20	25	17	28	16	29	15	30	14	31	14	31
		0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5
	0.7	29	16	26	19	23	22	19	26	16	29	15	30	14	31	14	31	13	32	13	32	13	32
		0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5
	0.8	22	23	17	28	12	33	10	35	10	35	10	35	10	35	10	35	11	34	12	33	12	33
		0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5
	0.9	17	28	8	37	5	40	5	40	5	40	5	40	5	40	5	40	5	40	5	40	5	40
		0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5	0	5

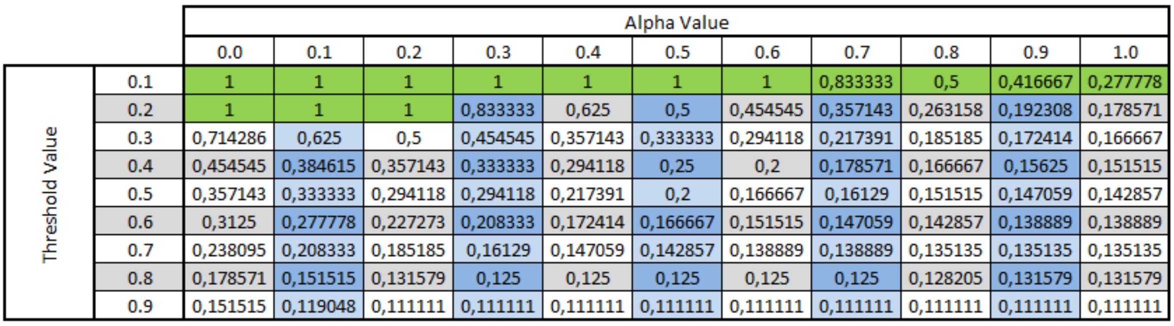
VP	FN
FP	VN

Table 16.: Accuracy matrix, resultant of the calculation of Accuracy of its confusion matrix, with the average value in each instance.



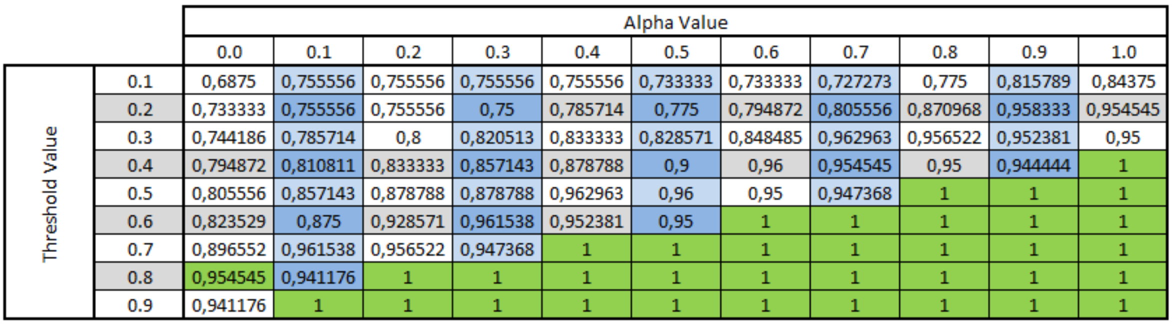
		Alpha Value										
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Threshold Value	0.1	0,7	0,78	0,78	0,78	0,78	0,76	0,76	0,74	0,72	0,72	0,64
	0.2	0,76	0,78	0,78	0,76	0,76	0,72	0,72	0,68	0,64	0,56	0,52
	0.3	0,74	0,76	0,74	0,74	0,7	0,68	0,66	0,62	0,54	0,5	0,48
	0.4	0,72	0,7	0,7	0,7	0,68	0,64	0,58	0,52	0,48	0,44	0,44
	0.5	0,68	0,7	0,68	0,68	0,62	0,58	0,48	0,46	0,44	0,42	0,4
	0.6	0,66	0,66	0,62	0,6	0,5	0,48	0,44	0,42	0,4	0,38	0,38
	0.7	0,62	0,6	0,54	0,46	0,42	0,4	0,38	0,38	0,36	0,36	0,36
	0.8	0,52	0,42	0,34	0,3	0,3	0,3	0,3	0,3	0,32	0,34	0,34
	0.9	0,42	0,26	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2
Average value		0,647	0,629	0,598	0,580	0,551	0,529	0,502	0,480	0,456	0,436	0,418

Table 17.: Negative Predictive Value matrix, resultant of the calculation based of the Lower Threshold matrix. The green cells represent the best value in each instance.



		Alpha Value										
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Threshold Value	0.1	1	1	1	1	1	1	1	0,833333	0,5	0,416667	0,277778
	0.2	1	1	1	0,833333	0,625	0,5	0,454545	0,357143	0,263158	0,192308	0,178571
	0.3	0,714286	0,625	0,5	0,454545	0,357143	0,333333	0,294118	0,217391	0,185185	0,172414	0,166667
	0.4	0,454545	0,384615	0,357143	0,333333	0,294118	0,25	0,2	0,178571	0,166667	0,15625	0,151515
	0.5	0,357143	0,333333	0,294118	0,294118	0,217391	0,2	0,166667	0,16129	0,151515	0,147059	0,142857
	0.6	0,3125	0,277778	0,227273	0,208333	0,172414	0,166667	0,151515	0,147059	0,142857	0,138889	0,138889
	0.7	0,238095	0,208333	0,185185	0,16129	0,147059	0,142857	0,138889	0,138889	0,135135	0,135135	0,135135
	0.8	0,178571	0,151515	0,131579	0,125	0,125	0,125	0,125	0,125	0,128205	0,131579	0,131579
	0.9	0,151515	0,119048	0,111111	0,111111	0,111111	0,111111	0,111111	0,111111	0,111111	0,111111	0,111111

Table 18.: Precision matrix, resultant of the calculation based of the Upper Threshold matrix. The green cells represent the best value in each instance.



		Alpha Value										
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Threshold Value	0.1	0,6875	0,755556	0,755556	0,755556	0,755556	0,733333	0,733333	0,727273	0,775	0,815789	0,84375
	0.2	0,733333	0,755556	0,755556	0,75	0,785714	0,775	0,794872	0,805556	0,870968	0,958333	0,954545
	0.3	0,744186	0,785714	0,8	0,820513	0,833333	0,828571	0,848485	0,962963	0,956522	0,952381	0,95
	0.4	0,794872	0,810811	0,833333	0,857143	0,878788	0,9	0,96	0,954545	0,95	0,944444	1
	0.5	0,805556	0,857143	0,878788	0,878788	0,962963	0,96	0,95	0,947368	1	1	1
	0.6	0,823529	0,875	0,928571	0,961538	0,952381	0,95	1	1	1	1	1
	0.7	0,896552	0,961538	0,956522	0,947368	1	1	1	1	1	1	1
	0.8	0,954545	0,941176	1	1	1	1	1	1	1	1	1
	0.9	0,941176	1	1	1	1	1	1	1	1	1	1