Modelling and multi-objective optimisation for simulation of cyanobacterial metabolism

PhD dissertation by:
Maria Siurana Paula

Supervisors:
Dr. Pedro J. Fernández de Córdoba Castellá
Dr. Arnau Montagud Aquino
Dr. Gilberto Reynoso Meza

Departament de Matemàtica Aplicada
Universitat Politècnica de València

October, 2017

This work was supported by a fellowship from the Spanish Ministry of Education, Culture and Sport (FPU2012-5873), and funds from the European Union's Seventh Programme (grant agreement No 308518).

For comments and questions please refer to:
maria.siurana.paula@gmail.com
https://www.researchgate.net/profile/Maria_Siurana
October, 2017

Acknowledgements

En primer lugar, quisiera dar las gracias a aquellos que me brindaron la oportunidad de emprender esta aventura y me han acompañado durante el proceso para que pudiese llevarla a buen término. Gracias a Pedro Fernández de Córdoba y a Javier Urchueguía por abrirme las puertas y permitirme adentrarme en este proyecto. Gracias Pedro por allanar siempre el camino y tener a mano unas palabras de aliento. Gracias Javier por plantear retos y preguntas y estar siempre pendiente de las respuestas. Gràcies a Arnau Montagud, per formar-me i tutelar-me en la meua introducció a la investigació. Gràcies per fomentar sempre el meu esperit crític, per extraure de mi la solució més encertada, i per donar-me el teu consell tant en temes científics i acadèmics, com en alguns altres més personals. Gracias a Gilberto Reynoso, por compartir conmigo su sapiencia y experiencia. Aunque llegaste un poco más tarde, has resultado ser un elemento fundamental para este trabajo, y sin duda también para mi desarrollo científico y personal, gracias por las largas charlas y los buenos consejos. Quiero dar las gracias también a Alberto Conejero y a Daniel Gamermann, que en distintos momentos han aportado su dedicación y su experiencia a este trabajo.

No me puedo olvidar de mi compañero de batallas, David Fuente. Gracias por compartir la carga en los buenos y no tan buenos momentos, y por estar siempre a punto para compartir una duda, un café o lo que se tercie. Tampoco puedo dejar de lado a aquellos que me han acompañado en mi paso por la 210, Joan Vázquez, Andrea Vázquez, Ramón Jaime, Borja Badenes, (y tantos otros, perdonad que no os mencione a todos) con los que he convivido a diario en distintos periodos y que siempre me han ofrecido lo mejor de su compañía, a veces discusiones trascendentales, otras veces divertidas charlas ligeras.

I would like to thank also Röbbe Wünschiers, for welcoming me in his research group (even in your home). Röbbe, I have to thank you for giving me a broader (and so great!) perspective of science world and science geniuses. I thank also Gabriel Kind, who accompanied me in most of my international experiences during this process. Thank you Gabriel for accepting my way of doing things, and for offering me your friendship, it was always comfortable to work and live with you. And I would like to thank Andrew Landels too. "Professor" Landels (I hope my English is good enough) I thank you for sharing your knowledge and personality with me, you were soon a friend, more than a colleague.

Per suposat, també he d'agrair els meus amics, que han aguantat (i aguantaran) les meues "frikades", i sempre han estat a punt per a una cervesa (i el que vinga) quan els he necessitat. Ferran, potser no ens veiem tant com voldriem, però sé que sempre estàs ahí, i que tens paciència infinita amb mi; gràcies. David i Neus, vosaltres també teniu el vostre lloc ací, gràcies per ser més que família, per compartir moments divertits i entranyables, cantant, esquiant o, simplement, estant. David (sí, señor Terrel), a ti también tengo que agradecerte que me permitas robarte un abrazo cuando te pillo desprevenido, y que me consientas que esté siempre mareando la perdiz y complicándoos la vida. Gràcies també a Juan i a Mari, per celebrar cada xicotet triomf. I gràcies als meus companys de la Unió Musical Santa Cecília i Cor Polifònic Santa Cecília de Castellar-Oliveral, per contribuir a complementar la meua vida acadèmica amb la musical.

Per últim, no és casualitat que haja deixat per al final les persones més importants de la meua vida; diuen que les coses bones es fan d'esperar i vosaltres sou el més especial que tinc. Vull agrair als meus pares, Divina i Manel, per haver-me deixat sempre caminar el meu camí, i haver-me ensenyat a fer les meues (incessants, ho sé) preguntes i a respondre-les per mi mateixa. A més vull agrair-vos, encara que la seqüència estiga un poc alterada, haver-me proporcionat la persona que més estime d'entre els meus consanguinis: la meua germana, Divina. Teta (encara que de vegades te diga pel teu nom, eres la meua teta), eres per a mi imprescindible, sempre ho has sigut. M'aportes al mateix temps el teu suport incondicional, (invariable, infinit) i la perspectiva que necessite per a veure que les coses no són només del color que jo les veig. M'encanta que puguem compartir tantes coses i tant diverses, espere que això no canvie mai.

I com no, al pilar de la meua vida, Joan (d'entre tots els noms que tens em quedaré amb este, però vull seguir sent testimoni de tots ells). M'has recolzat i ajudat en totes les decisions i moments clau de la meua vida i, com no, aquesta no ha sigut una excepció. Sé que no ha sigut fàcil lidiar amb el *gremlin* que apareix quan perd el rumb, i tot i això has sigut capaç de suportar-me i d'ajudar-me a tornar al bon camí. N'hi ha tantíssimes coses que t'he d'agrair que potser no tindria prou amb un document de la mateixa extensió que el que ací presente per a fer-ho. Per això tractaré de ser breu: GRÀCIES. Gràcies per completar l'equip.

Abstract

The present thesis is devoted to the development of models and algorithms to improve metabolic simulations of cyanobacterial metabolism. Cyanobacteria are photosynthetic bacteria of great biotechnological interest to the development of sustainable bio-based manufacturing processes. For this purpose, it is fundamental to understand metabolic behaviour of these organisms, and constraint-based metabolic modelling techniques offer a platform for analysis and assessment of cell's metabolic functionality. Reliable simulations are needed to enhance the applicability of the results, and this is the main goal of this thesis.

This dissertation has been structured in three parts. The first part is devoted to introduce needed fundamentals of the disciplines that are combined in this work: metabolic modelling, cyanobacterial metabolism and multi-objective optimisation.

In the second part the reconstruction and update of metabolic models of two cyanobacterial strains is addressed. These models are then used to perform metabolic simulations with the application of the classic *Flux Balance Analysis* (FBA) methodology. The studies conducted in this part are useful to illustrate the uses and applications of metabolic simulations for the analysis of living organisms. And at the same time they serve to identify important limitations of classic simulation techniques based on mono-objective linear optimisation that motivate the search of new strategies.

Finally, in the third part a novel approach is defined based on the application of multi-objective optimisation procedures to metabolic modelling. Main steps in the definition of multi-objective problem and the description of an optimisation algorithm that ensure the applicability of the obtained results, as well as the multi-criteria analysis of the solutions are covered. The resulting tool allows the definition of non-linear objective functions and constraints, as well as the analysis of multiple Pareto-optimal solutions. It avoids some of the main drawbacks of classic methodologies, leading to more flexible simulations and more realistic results.

Overall this thesis contributes to the advance in the study of cyanobacterial metabolism by means of definition of models and strategies that improve plasticity and predictive capacities of metabolic simulations.

Resum

La present tesi està dedicada al desenvolupament de models i algorismes per a millorar les simulacions metabòliques de cianobacteris. Els cianobacteris són bacteris fotosintètics de gran interés biotecnològic per al desenvolupament de bioprocessos productius sostenibles. Per a aquest propòsit, és fonamental entendre el comportament metabòlic d'aquests organismes, i el modelatge metabòlic basat en restriccions ofereix una plataforma per a l'anàlisi i l'avaluació de les funcionalitats metabòliques de les cèl·lules. Es necessiten simulacions fidedignes per a augmentar l'aplicabilitat dels resultats, i aquest és l'objectiu principal d'aquesta tesi.

Aquesta dissertació s'ha estructurat en tres parts. La primera part està dedicada a introduir els fonaments necessaris de les disciplines que es combinen en aquest treball: el modelatge metabòlic, el metabolisme de cianobacteris i l'optimització multiobjectiu.

En la segona part, s'adreça la reconstrucció i l'actualització dels models metabòlics de dos soques de cianobacteris. Aquests models s'empren després per a portar a terme simulacions metabòliques amb l'aplicació de la metodologia clàssica *Flux Balance Analysis* (FBA). Els estudis realitzats en aquesta part són útils per a il·lustrar els usos i aplicacions de les simulacions metabòliques per a l'anàlisi dels organismes vius. I al mateix temps serveixen per a identificar importants limitacions de les tècniques clàssiques de simulació basades en optimització lineal mono-objectiu que motiven la cerca de noves estratègies.

Finalment, en la tercera part, es defineix una nova aproximació basada en l'aplicació al modelatge metabòlic de procediments d'optimització multiobjectiu. Es cobreixen els principals passos en la definició d'un problema multiobjectiu i la descripció d'un algorisme d'optimització que asseguren l'aplicabilitat dels resultats obtinguts, així com l'anàlisi multi-criteri de les solucions. La ferramenta resultant permet la definició de funcions objectiu i restriccions no lineals, així com l'anàlisi de múltiples solucions òptimes en el sentit de Pareto. Aquesta ferramenta evita alguns dels principals inconvenients de les metodologies clàssiques, el que porta a obtenir simulacions més flexibles i resultats més realistes.

En conjunt, aquesta tesi contribueix a l'avanç en l'estudi del metabolisme de cianobacteris per mitjà de la definició de models i estratègies que milloren la plasticitat i les capacitats predictives de les simulacions metabòliques.

Resumen

La presente tesis está dedicada al desarrollo de modelos y algoritmos para mejorar las simulaciones metabólicas de cianobacterias. Las cianobacterias son bacterias fotosintéticas de gran interés biotecnológico para el desarrollo de bioprocesos productivos sostenibles. Para este propósito, es fundamental entender el comportamiento metabólico de estos organismos, y el modelado metabólico basado en restricciones ofrece una plataforma para el análisis y la evaluación de las funcionalidades metabólicas de las células. Se necesitan simulaciones fidedignas para aumentar la aplicabilidad de los resultados, y este es el objetivo principal de esta tesis.

Esta disertación se ha estructurado en tres partes. La primera parte está dedicada a introducir los fundamentos necesarios de las disciplinas que se combinan en este trabajo: el modelado metabólico, el metabolismo de cianobacterias, y la optimización multiobjetivo.

En la segunda parte, se encara la reconstrucción y la actualización de los modelos metabólicos de dos cepas de cianobacterias. Estos modelos se usan después para llevar a cabo simulaciones metabólicas con la aplicación de la metodología clásica *Flux Balance Analysis* (FBA). Los estudios realizados en esta parte son útiles para ilustrar los usos y aplicaciones de las simulaciones metabólicas para el análisis de los organismos vivos. Y al mismo tiempo sirven para identificar importantes limitaciones de las técnicas clásicas de simulación basadas en optimización lineal mono-objetivo que motivan la búsqueda de nuevas estrategias.

Finalmente, en la tercera parte, se define una nueva aproximación basada en la aplicación al modelado metabólico de procedimientos de optimización multiobjetivo. Se cubren los principales pasos en la definición de un problema multiobjetivo y la descripción de un algoritmo de optimización que aseguren la aplicabilidad de los resultados obtenidos, así como el análisis multi-criterio de las soluciones. La herramienta resultante permite la definición de funciones objetivo y restricciones no lineales, así como el análisis de múltiples soluciones en el sentido de Pareto. Esta herramienta evita algunos de los principales inconvenientes de las metodologías clásicas, lo que lleva a obtener simulaciones más flexibles y resultados más realistas.

En conjunto, esta tesis contribuye al avance en el estudio del metabolismo de cianobacterias por medio de la definición de modelos y estrategias que mejoran la plasticidad y las capacidades predictivas de las simulaciones metabólicas.

Contents

Main acronyms

Optimisation techniques, metabolic modelling and multi-objective approaches

DE	Differential Evolution
FBA	Flux Balance Analysis
GFCL	Generate-First Choose-Later
LD	Level Diagrams
LP	Linear Programming
MCDM	Multi-Criteria Decision-Making
MILP	Mixed-Integer Linear Programming
MOEA	Multi-Objective Evolutionary Algorithm
MOMA	Minimisation Of Metabolic Adjustment
MO	Multi-Objective
MOP	Multi-Objective Problem
QP	Quadratic Programming

Metabolites and reactions

2KG	2-ketoglutarate
3PG	3-phosphoglycerate
ACA	acetyl coenzyme A
CIS	citrate synthase
CIT	citrate
CMP	bicarbonate transporter
E4P	erythrose-4-phosphate
ENO	enolase
F6P	fructose-6-phosphate
FBP	fructose-1,6-bisphosphate
FUM	fumarate
FUMH	fumarate hydratase
G6P	glucose-6-phosphate
GAP	glyceraldehyde-3-phosphate
GLC	glucose
GLK	glucokinase
ICD	isocitrate dehydrogenase
ICI	isocitrate
ICL	isocitrate lyase
MAL	malate
OAA	oxaloacetate
PDH	pyruvate dehydrogenase
PEP	phosphoenolpyruvate
PFK	6-phosphofructokinase
PGI	glucose-6-phosphate isomerase
PYR	pyruvate
R5P	ribose-5-phosphate

RBCO	ribulose-bisphosphate carboxylase/oxigenase
RPE	ribulose-phosphate 3-epimerase
RPI	ribose-5-phosphate isomerase
RU5P	ribulose-5-phosphate
RUBP	ribulose-1,5-bisphosphate
SUC	succinate
S7P	sedoheptulose-7-phosphate
TAL	transaldolase
TK	transketolase
X5P	xylulose-5-phosphate

Main symbols

$\boldsymbol{S}$	Stoichiometric matrix representing a metabolic network
$S_{i,j}$	Elements of the stoichiometric matrix
$\boldsymbol{v}$	Vector containing the fluxes of all reactions in a metabolic network
v_j	Flux of reaction j at the vector of fluxes
Z	Objective function
$\boldsymbol{Z}\left(\boldsymbol{v}\right)$	Objective vector
$\boldsymbol{V}_P$	Pareto set
$\boldsymbol{Z}_P$	Pareto front
$\boldsymbol{V}_P^*$	Pareto set approximation
$\boldsymbol{Z}_P^*$	Pareto front approximation
$\boldsymbol{Z}^{ideal}$	Ideal objective vector
$\boldsymbol{Z}^{utopia}$	Utopian objective vector
$\boldsymbol{Z}^{nadir}$	Nadir objective vector

List of Algorithms

List of Figures

List of Tables

Aims, Structure and Contributions of this Thesis

Aims, Structure and Contributions of this thesis

Aims and objectives

Like every applied science, metabolic modelling is situated at the intersection of different disciplines. It combines mathematics, biology and engineering principles to clarify the metabolic behaviour of living organisms with the final goal of designing new biological circuits or alter existing ones that can be used to address new problems. Due to its applied and innovative character, this area offers many open possibilities, but at the same time, its multidisciplinary nature poses a challenge to researchers. The focus of the present dissertation is this intersection, in which mathematical structures and methods are applied to describe and analyse metabolism of living organisms, specifically cyanobacterial metabolism.

In the current energy, environment and socio-economic situation, there exist an increasing trend towards the search of efficient and sustainable manufacturing processes. In this context the use of living organisms as cell factories has gained increasing interest in the past years. In order to develop competitive processes that fulfil the needs of the industry, the natural systems have to be modified to adapt their functionality and enhance their productive capacities. But rational design needs planning: it is fundamental to evaluate potential capacities and maximum yields that can be achieved, as well as to assess how the genetic alterations on the designed cell will affect the whole system. Metabolic models and modelling techniques are a valuable tool for these purposes.

Genome-scale metabolic models are mathematical representations of cellular metabolism that include the description of all biochemical reactions known to occur at a given organism. These models represent the network of substances (metabolites) that are transformed from one to another through the metabolic reactions, which are in most cases catalysed by enzymes. The variable object of study are the metabolic fluxes, that is, the rates of the metabolic reactions, that describe how energy and materials flow through the metabolic network.

Different methods have been used to model metabolism, but one of the most extended approaches is constraint-based modelling, which for many years was the only practical option for large-scale models. Under this approach, a set of constraints are defined to describe the limitations that metabolic systems face in natural environments which restrict the set of allowable phenotypes (represented in the modelling framework by flux distributions). Once the space of feasible solutions is constrained to the set of biologically practical solutions, optimisation can be applied to extract concrete flux distributions that represent optimal phenotypes. The use of optimisation in this context is justified by the assumption that living organisms subject to selective pressure have evolved towards optimal or quasi-optimal performance. This optimisation could involve many phenotypes of the cell (adaptation, efficient use of substrates, *etc*.), but it is almost always considered on biomass growth.

In the present work, these principles are applied to the study of the metabolism of cyanobacteria. Cyanobacteria are photosynthetic bacteria able to perform oxygenic photosynthesis and make use of sunlight and inorganic CO_2 to perform their metabolic functions. These low nutrient requirements have placed them in the spotlight as very promising organisms for the development of sustainable, low-energy-consuming, cell factories. Besides, genetic tools are available for their manipulation to meet commercial needs.

Thus, the main objective of this thesis is:

To contribute with models and strategies to improve plasticity and predictive capacity of simulations of cyanobacterial metabolism.

With this purpose in mind, the following specific aims have been defined:

- Reconstruct and validate up-to-date metabolic networks for two model cyanobacteria *Synechocystis* sp. PCC 6803 and *Synechococcus elongatus* PCC 7942.
- Demonstrate potential applications and uses of these models to simulate metabolic behaviour and design production strategies.
- Identify advantages and limitations of classic simulation methods for constraint-based metabolic modelling.
- Propose a modelling strategy based on multi-objective evolutionary optimisation taking into account the main steps of the process, from constraints and objectives definition, through the optimisation process, to solution analysis and selection.
- Validate the proposed strategy by applying it to the simulation and analysis of metabolic behaviour of the model cyanobacterium *Synechocystis* sp. PCC 6803 under different growth conditions.

Thesis structure

This dissertation is divided in three parts. In the first part, background is given for the two main disciplines participating in this thesis: metabolic modelling and multi-objective optimisation. **Chapter 1** is devoted to explain the basic principles of metabolic modelling and to outline some of the main simulation techniques. It also includes a brief introduction to cyanobacterial metabolism and background on reconstructed metabolic networks of the organisms studied in this thesis. **Chapter 2** is dedicated to give some basic definitions and explain the fundamentals of multi-objective optimisation procedures that are needed for the development of this thesis.

In the second part, contributions to metabolic modelling of cyanobacteria are presented. In **Chapter 3** the update and assembly process of genome-scale metabolic networks of the model cyanobacteria *Synechocystis* sp. PCC 6803 and *Synechococcus elongatus* PCC 7942 is addressed. The resulting metabolic models are used in **Chapter 4** to perform simulations by means of the classic Flux Balance Analysis methodology. Those simulations must serve to illustrate some of the main uses of constraint-based metabolic modelling, and at the same time conclusions are extracted about advantages and limitations of classic simulation strategies.

Finally, the third part is devoted to apply multi-objective optimisation procedures to the problem of constraint-based metabolic simulations. In **Chapter 5** a strategy and an optimisation algorithm are described that apply the principles of multi-objective problem definition and optimisation to perform flux simulations. This tool is applied in **Chapter 6** for its validation through metabolic simulations of the model cyanobacteria *Synechocystis* sp. PCC 6803 under different growth conditions.

Finally, main conclusions of this dissertation are described, as well as some future perspectives for further application of this work.

Contributions

Scientific contributions

Peer-reviewed articles

- Julián Triana, Arnau Montagud, Maria Siurana, David Fuente, Arantxa Urchueguía, Daniel Gamermann, Javier Torres, Jose Tena, Pedro Fernández de Córdoba, Javier Urchueguía. **Generation and Evaluation of a Genome-Scale Metabolic Network Model**

of *Synechococcus elongatus* PCC 7942. *Metabolites* 2014, **4**:680-698
The PhD applicant mainly contributed to the curation process.

- Gabriel Kind, Maria Siurana, Erik Zuchantke, David Fuente, Lenin G. Lemus-Zúñiga, Javier Urchueguía, Röbbe Wünschiers. **CellDesign - An Open-Source, Web-Based Software for Metabolic Modelling and Flux-Balance Analyses.** *Manuscript submitted to BMC Bioinformatics in June 2017.*
  The PhD applicant mainly contributed with the technical guidance about metabolic simulations and supervised the test of use done through application of the tool in academic courses.

- Maria Siurana, Arnau Montagud, Gilberto Reynoso-Meza, J. Alberto Conejero, Javier Sanchis, Lenin G. Lemus-Zúñiga, Javier Urchueguía **Multi-objective evolutionary algorithm allows more accurate genome-scale flux simulations with a small set of experimental values.** *Manuscript in preparation.*

- Maria Siurana, Gilberto Reynoso-Meza, Arnau Montagud, Javier Sanchis **Application and technicalities of multi-objective evolutionary algorithm on genome-scale flux simulations.** *Manuscript in perspective.*

Participation in research project

- **CyanoFactory[1] - Design, construction and demonstration of solar biofuel production using novel (photo)synthetic cell factories.** *Founded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 308518.* December 2012 - November 2015.

[1] Final report available at `cordis.europa.eu/result/rcn/184808_en.html`

Software development

Meta-MODE algorithm described in Chapter 5 will be available at the public MATLAB® repository in the short term after publication of this thesis.

CellDesign[2] web-based metabolic modelling toolbox for non-experts described in Chapter 4 is accessible at `http://celldesign.de`

Conference presentations and posters

Presentations

- Maria Siurana, Arnau Montagud, Gilberto Reynoso-Meza, J. Alberto Conejero, Javier Sanchis, Lenin G. Lemus-Zúñiga, Javier Urchueguía. **Multi-objective evolutionary algorithm for genome-scale flux simulation.** *1st International Solar Fuels Conference.* Uppsala (Sweden), May 2015.

- Marina Pérez-Naveira[3] , Maria Siurana, Javier Urchueguía. **Integration of proteomic and metabolomic data in genome-scale metabolic models and its application to the cyanobacteria *Synechocystis* sp. PCC 6803.** *XIII Symposium on Bioinformatics.* Valencia (Spain), May 2016.

Posters

- Maria Siurana, Arnau Montagud, Javier Urchueguía. ***Synechocystis* metabolic modelling and production strategies assessment.** *1st International Solar Fuels Conference (ISF-1).* Uppsala (Sweden), May 2015.

[2] The PhD applicant mainly contributed with the technical guidance about metabolic simulations and supervised the test of use done through application of the tool in academic courses.

[3] Presented by Marina Pérez-Naveira, supervised by Maria Siurana and Javier Urchueguía during her final Bachelor thesis.

- Gabriel Kind, Arnau Montagud, Erik Zuchantke, Maria Siurana, J. Alberto Conejero, Pedro Fernández de Córdoba, Röbbe Wünschiers, Javier Urchueguía. **CyanoDesign - A web-based tool for the generation and analysis of *in silico* mutants for *Synechocystis* sp. PCC 6803** *1st International Solar Fuels Conference (ISF-1)*. Uppsala (Sweden), May 2015.

- Erik Zuchantke, Maria Siurana, David Fuente, Lenin G. Lemus-Zúñiga, Röbbe Wünschiers, Javier Urchueguía. **Description and application of a pipeline for generating versions of genome-scale metabolic models from sequence information.** *XIII Symposium on Bioinformatics*. Valencia (Spain), May 2016.

- Gabriel Kind, Erik Zuchantke, Maria Siurana, J. Alberto Conejero, Javier Urchueguía, Röbbe Wünschiers. **CyanoFactoryKB - An open-source web-based software program for constructing model organism databases for *Synechocystis* sp. PCC 6803.** *1st International Solar Fuels Conference (ISF-1)*. Uppsala (Sweden), May 2015.

- Raymari Reyes, Maria Siurana, Daniel Gamermann, Arnau Montagud, Julián Triana, Ramón Jaime, Victor M. Nina, David Fuente, Yarlenis Pacheco, Javier Urchueguía, Pedro Fernández de Córdoba. **COPABI: a Computational Platform for Automation on the Genome-Scale Metabolic Models Reconstruction.** *XII Symposium on Bioinformatics*. Sevilla (Spain), September 2014.

Teaching contributions

Teaching Assistant in the following courses

- **Calculus III**. 27 hours. *January - March 2017*

- **Discrete Mathematics**. 38 hours. *September - December 2016*

- **Statistics and Random Signals**. 32 hours. *September - December 2016*
- **Numerical Analysis**. 80 hours. *September 2014 - December 2016*
- **Systems Biology and Systems Metabolic Engineering**. 45 hours. *February 2014 - March 2016*

Bachelor theses supervised

- Marina Pérez-Naveira. **Proteomic and metabolomic data integration in genome-scale metabolic models and its application to a model of cyanobacterium *Synechocystis* sp. PCC 6803**. *Universitat Politècnica de València, 2016.*
- Joan Canet. **Comparative analysis of methodologies to simulate metabolic systems applied to the study of the *E coli Core* model**. *Universitat Politècnica de València, 2017.*

Part I

Introduction

1

Uses of metabolic modelling and aims of its simulations

1.1 Chapter abstract

In this chapter the fundamentals of constraint-based metabolic modelling and its applications to cyanobacteria are presented.

Metabolic models are a valuable tool to reach a system-level comprehension of cells' biochemical functions, which is needed to rationally design new systems for biotechnological applications. In particular, constraint-based models allow the study of genome-wide metabolic networks without the need of extra kinetic or concentration information.

Genome-scale metabolic networks are reconstructed from genome annotation following a four-stage process. Some of those steps can be automatically implemented, but high quality reconstructions always require detailed manual curation and update. Once the topology of the network is described, constraints are imposed that ensure the biological pertinency of the solutions obtained from the simulation. Flux distributions (*i.e.* sets of rates of reactions) satisfying the defined constraints represent allowable metabolic phenotypes. To extract particular solutions from the feasible space optimisation is utilised. It is worth to note that absolute optima are not the goal of these simulations, but realistic solutions that reconcile the biological constraints with the evolutionary objective.

Several algorithms and tools have been described to perform constraint-based optimisation-driven metabolic simulations. Basic principles of all of them are the assumption of basic constraints, *viz.* reaction directionality and stoichiometry, steady-state mass balance, reaction/transport capacity limits and nutrient availability, and the optimisation according to a biological objective (that is usually growth). Differences among methods lie in the inclusion of additional mechanisms to take into consideration further biochemical knowledge that can lead to more comprehensive solutions.

Throughout this PhD dissertation, models and methods will be discussed and developed for metabolic modelling of cyanobacteria. These photosynthetic prokaryotes have revealed of great interest due to their potential to develop low-cost green bio-based production processes.

1.2 Metabolic modelling in systems biology

Systems biology is an area of life sciences that aims at understanding of living organisms at a system level (Kitano, 2001). It aims at explaining how biological systems' characteristics and functions emerge from the behaviour and the interactions of their molecular parts. This system-level understanding involves four key stages (Kitano, 2001, 2002): first, the structures forming the system must be identified, as well as their interactions; then, an analysis of how the system behaves under various conditions has to be performed; once the system's structure and behaviour are deciphered, mechanisms that systematically control the state of the cell can be modulated to avoid malfunctioning or to take advantage of desirable cellular functions; finally, strategies to modify and construct biological systems having desired properties can be devised based on design principles and simulations. Thus, the final goal is not only the comprehension of biological systems, but also the reprogramming of existing systems and design of new ones with applied purposes such as bioengineering or biomedical applications (Ideker et al., 2001; Kitano, 2001, 2002).

Such a holistic approach necessitates systematic data, as it is not possible to investigate a biological system as a whole without them (Chuang et al., 2010). While studying the structure and behaviour of the system (first and second steps above), it must be systematically perturbed (biologically, genetically, or chemically) and its responses must be monitored. All these data have to be gathered and integrated, and ultimately, mathematical models have to be formulated that describe the structure of the system and its response to individual perturbations (Ideker et al., 2001). Thus mathematical models are central in systems biology to organise, understand and exploit as much information as possible from a given biological system.

One area of active research in this field has focused on analysing metabolism (Palsson, 2009; Heinemann and Sauer, 2010; Mardinoglu and Nielsen, 2012; Mardinoglu et al., 2013; Bordbar et al., 2014; Dersch et al., 2016). Metabolic fluxes, which are the rates of metabolic reactions (*i.e.* number of molecules traversing each metabolic reaction per unit time), have an important role in investigating cellular physiology, as they show how the available resources (*e.g.* carbon, reducing equivalents and chemical energy) flow through the metabolism to enable cell function (García Martín et al., 2015). Metabolic phenotypes[4] can be defined in terms of flux distributions through a metabolic network, which can be interpreted and predicted using mathematical modelling and computer simulation (Edwards et al., 2002a). Thus, metabolic models allow researchers to study different genotypes and perturbed environmental conditions, and the resulting phenotypes, constituting a valuable tool

[4] The term “phenotype” refers to the composite of an organism's observable characteristics or traits (Johannsen, 1911). A phenotype results from the expression of an organism's genetic code, its genotype, as well as the influence of environmental factors and the interactions between both.

for quick testing of the consequences of engineering approaches. They have demonstrated being useful for several purposes, such as fundamental understanding of metabolic behaviour (Savinell and Palsson, 1992a,b), discovery and annotation of enzymatic functions (Reed et al., 2006), disease study and drug discovery (Chavali et al., 2012), or bioengineering and design of bacterial strains for biotechnological applications (Yim et al., 2011).

Several techniques have been used for the modelling, simulation and analysis of pathways and networks involved in metabolism (Alon, 2007; Klipp et al., 2016). The different approaches vary in the amount of detail and the broadness of their scope. A simple classification can be posed that separates available techniques between dynamic and static constraint-based modelling approaches (Raman and Chandra, 2009). Dynamic models are kinetic representations of cellular processes that provide detail of dynamic interactions and dependencies among biological entities. However, the vast amount of information (kinetic parameters and concentration of all species, among others) required by those models, and the scarcity of such data in many cases, usually limits the size of the systems described through these methods (Edwards et al., 2002a). A practical alternative to dynamic modelling strategies are static constraint-based models. Under the scope of constraint-based metabolic modelling, governing physicochemical and biological constraints are imposed that narrow the range of achievable functional states that a metabolic system can display (Price et al., 2004). Analysing the resultant allowable flux distributions provides a basis for understanding structure and function of biochemical reaction networks at a system level (Edwards et al., 2002a).

1.2.1 Constraint-based metabolic modelling

As stated before, constraint-based modelling allows the study of components and operation of metabolic networks. The first step in undertaking this task is to address the reconstruction of the network comprising the metabolites, enzymes and biochemical conversions involved in the metabolism of a given cell (Figure 1.1). As system-level comprehension is sought, this network must include all metabolic reactions known to occur in the system under study. A metabolic network is said to be genome-scale when all the enzymes and metabolic functions identified from the genome sequence are considered.

Once the topology of the network has been detailed, the constraint-based paradigm establishes a set of assumptions and constraints that limit the possible metabolic phenotypes. Different methods have been described to determine and analyse the remaining feasible flux distributions that represent the metabolic state of the cell. This thesis uses methods based on optimisation of objective metabolic functions (Figure 1.1). The goal is that the analysis of the obtained flux distributions will give insight into metabolic operation of the system and will serve as a foundation for subsequent modifications and design procedures.

Genome-scale metabolic network reconstruction

A thorough description of the reconstruction process has been reported by Thiele and Palsson (2010). The entire process is divided in four stages. It starts by assembling a draft reconstruction, followed by a refinement of this reconstruction, and its conversion into a mathematical model. The process is completed after validation, when a debugged, iteratively improved network is achieved (Figure 1.2).

The first stage of the reconstruction process consists on the generation of a draft reconstruction from the annotated genome sequence and biochemical databases of the organism under study. Candidate enzymes are identified from the genome annotation, and potential biochemical reactions catalysed by those enzymes are retrieved from biochemical databases such as KEGG (Kanehisa and Goto, 2000; Kanehisa et al., 2017), BRENDA (Schomburg et al., 2000; Placzek et al., 2017) or MetaCyc (Caspi et al., 2016) among others. This step can be carried out manually or by using automated tools (Hamilton and Reed, 2014).

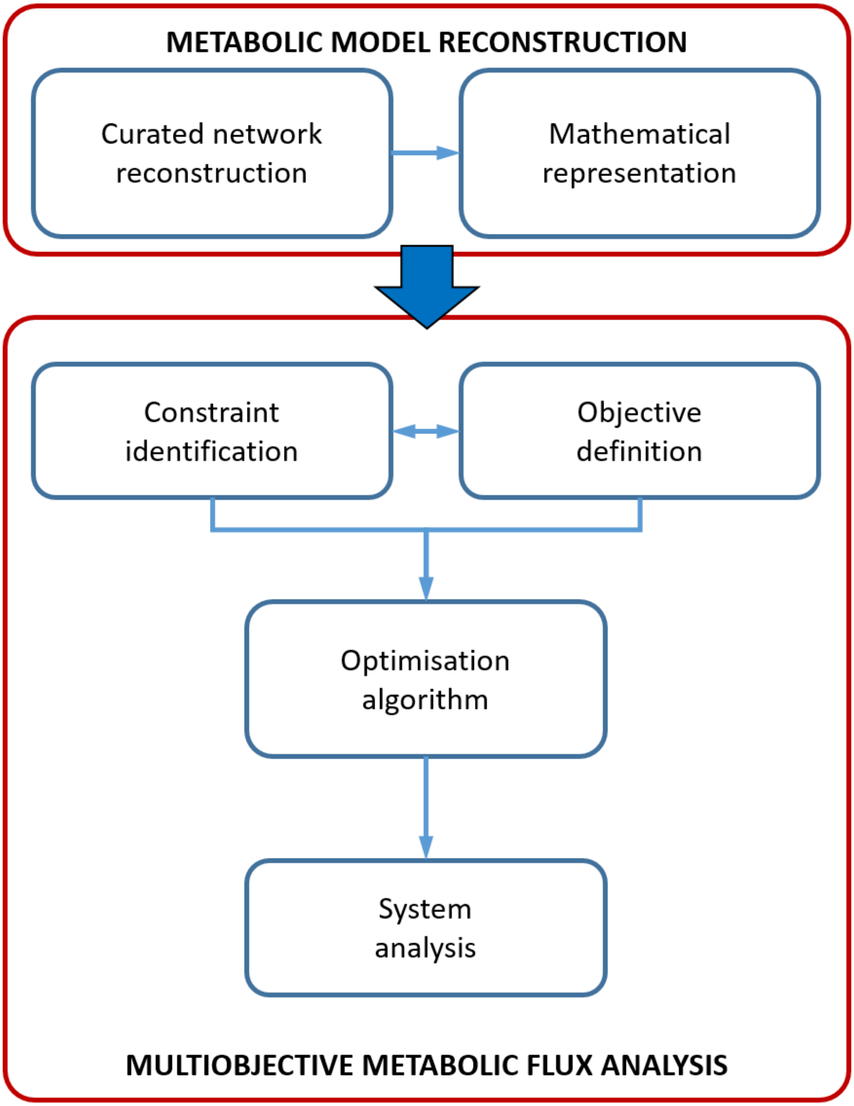

METABOLIC MODEL RECONSTRUCTION
Curated network reconstruction
Mathematical representation
Constraint identification
Objective definition
Optimisation algorithm
System analysis
MULTIOBJECTIVE METABOLIC FLUX ANALYSIS

Figure 1.1: *Major steps in metabolic network reconstruction and constraint-based analysis.*

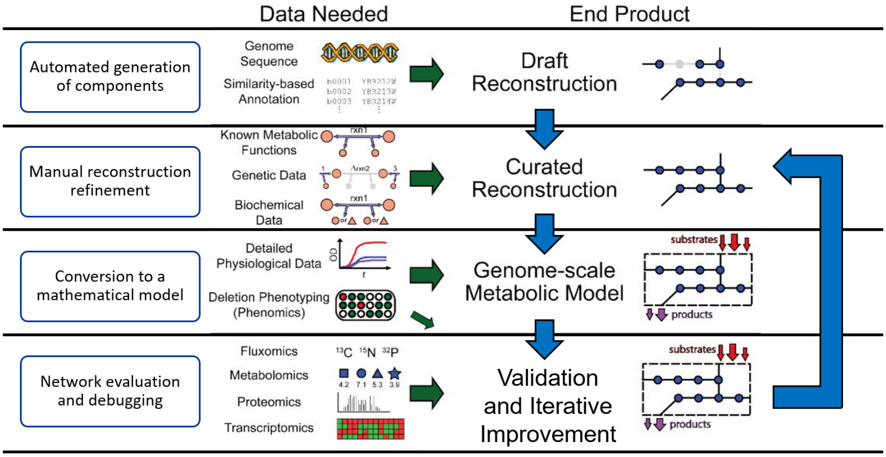

Data Needed
End Product
Automated generation of components
Genome Sequence
Similarity-based Annotation
Draft Reconstruction
Manual reconstruction refinement
Known Metabolic Functions
Genetic Data
Biochemical Data
Curated Reconstruction
Conversion to a mathematical model
Detailed Physiological Data
Deletion Phenotyping (Phenomics)
Genome-scale Metabolic Model
substrates
products
Network evaluation and debugging
Fluxomics
Metabolomics
Proteomics
Transcriptomics
Validation and Iterative Improvement

Figure 1.2: *Steps of the process to reconstruct a genome-scale metabolic network. After Figure 1 from Feist et al. (2009).*

This preliminary reconstruction, however, usually suffers from incompleteness, incorrectness and unspecificities in some reactions (Thiele and Palsson, 2010). Some of the issues associated with the first stage of the reconstruction are due to problems with genome annotations: lack of updated data, incorrect annotations, missing functionalities and non-reported transporter specificity are among the most common. Some others are issues related with databases: unspecific metabolites, reaction imbalances, or lack of detail about reaction directionality and compound protonation states usually affect the resultant draft reconstruction.

In order to obtain a reliable, exhaustive and functional metabolic network manual refinement is needed. During this second stage, high-quality, organism-specific information has to be collected from literature, databases, physiological experiments, and experts' advice, to curate the reconstruction. Although the initial reconstruction step is rapid, especially when automated, the manual curation process is labour-intensive (Feist et al., 2009).

After the refinement stage, a curated reconstruction is obtained, but before it can be used for metabolic simulations, another step is needed:

the reconstructed network has to be converted to a mathematical representation. In this stage, the *stoichiometric matrix* is defined (Definition 1.1).

Definition 1.1 (Stoichiometric matrix $\boldsymbol{S}$)**.** *The stoichiometric matrix $\boldsymbol{S}$ is a matrix of size $m \times n$ containing the stoichiometric coefficients for the reactions that constitute a metabolic network. Rows $i \in \{1, \ldots, m\}$ correspond to metabolites, and columns $j \in \{1, \ldots, n\}$ to reactions. Each entry $S_{i,j}$ is the stoichiometric coefficient of the metabolite i participating in reaction j. The substrates in a reaction are defined to have a negative coefficient, whereas the products have a positive value. A stoichiometric coefficient of zero is used for every metabolite that does not participate in a particular reaction.*

Thus, the stoichiometric matrix represents the topology of the reconstructed network with the stoichiometry of the biochemical reactions involved. $\boldsymbol{S}$ is a sparse matrix because most biochemical reactions involve only a few different metabolites (Orth et al., 2010). Usually, the number of reactions is greater than the number of metabolites ($n > m$) [5].

To complete the mathematical representation of the metabolic system, some parameters must be defined that will serve to delimit actual functional states of the system. This includes reaction directionality, enzymatic capacity, environmental conditions, and/or physiological information (if available), among others. These system parameters are related to the constraints used for flux analysis, and they will be discussed in more detail in the next section.

The fourth stage of the reconstruction process consists of network evaluation and debugging. The metabolic model created at the previous stage is tested for its ability to reproduce known cellular functionalities, such as biomass precursors syntheses, among other things. This evaluation may reveal missing, incomplete or incorrect metabolic func-

[5] Which causes the system of equations that appears when a steady-state mass balance is applied to be under-determined.

tions in the reconstruction, which are corrected by iterating stages 2 and 3. Thus, the reconstruction process is an iterative procedure.

The reconstruction can be considered complete, when it has been validated and verified, no severe errors remain, and it has revealed to be sufficient for the simulation purpose for which it was generated (Thiele and Palsson, 2010). Nevertheless, even when a functional reconstruction has been achieved, it is important to regularly revise and update it, including last discoveries and completing (or even adapting) the previous reconstruction for new research challenges.

Constraint-based flux analysis of metabolic networks

Following the procedure described in the previous section, a metabolic network is reconstructed and converted into a metabolic model. This model will describe all the biochemical reactions (metabolites participating, stoichiometry, enzymes involved and genes related to those enzymes) for which there is proof and/or evidence of presence in the target metabolic system. At this point, the model is prepared to be used for flux analysis.

The unknown variable object of this analysis is the flux distribution through the metabolic network, which will give insight into the metabolic states of the cell. The first step to relate topology and stoichiometry of the network with fluxes is to consider the following mass balance:

$$\frac{dX_i}{dt} = S_{i,j} \cdot v_j, \quad \forall i \in \{1, \ldots, m\}, \forall j \in \{1, \ldots, n\} \tag{1.1}$$

Where X_i represents concentration of metabolites ($i \in \{1, \ldots, m\}$), $S_{i,j}$ is the stoichiometric matrix (Definition 1.1), and v_j are the fluxes of all reactions in the network ($j \in \{1, \ldots, n\}$).

The matrix balance presented in Equation (1.1) states that the concentration change of each metabolite i over time is equal to the difference between the rates at which the metabolite is produced and consumed in the various reactions in which it participates.

As explained before, the variable aimed is the flux distribution (which is the flux values for all the reactions) represented by the vector v_j. But genome-scale information about metabolite concentrations is not usually available, therefore some assumption is needed in order to reduce the degrees of freedom of this set of equations. This assumption is that of steady state of the internal metabolites:

$$\frac{dX_i}{dt} = S_{i,j} \cdot v_j = 0 \tag{1.2}$$

Consequently, there is no accumulation or depletion of intracellular metabolites, their concentrations are not allowed to change over time, they are therefore balanced. On the contrary, extracellular metabolites are not balanced (they are outside the system boundaries) and these will be the uptake of substrates and formation of products.

Steady state assumption is widely accepted in systems biology as it can be justified by the fact that metabolic transients are faster than both cellular growth rates and the dynamic changes in the organism's environment. Metabolism typically has transients that are shorter than a few minutes and thus metabolic fluxes are in a quasi-steady state relative to growth and typical process transients (Varma and Palsson, 1994a). Additionally, this assumption allows researchers to avoid the need of detailed dynamic descriptions of metabolism that account for kinetics and regulation of individual enzymes, which has been proven difficult to obtain.

Equation (1.2) describes a static problem simpler than the previous dynamic problem (Equation (1.1)), but it is still under-determined (as number of reactions is greater than number of metabolites), therefore more assumptions need to be made. It is at this point when constraints appear. Biological systems are subject to different types of constraints that limit their cellular functions (Price et al., 2004) and to which all viable phenotypes must comply: physico-chemical constraints, spatial constraints, condition-dependent environmental constraints and regulatory constraints. Thus, identifying those constraints and properly adding them as mathematical restraints to the metabolic model, will restrict the set of feasible solutions to biologically allowable metabolic states.

Mass balance itself is a fundamental physico-chemical constraint. Other important constraints of this type are thermodynamic principles that will define reaction directionality and reversibility. Another type of constraints that play an important role in metabolic modelling are those related to environmental conditions, since they will describe the availability of resources (*e.g.* carbon, inorganic nutrients and salts, reducing molecules, *etc*.). Also maximum enzyme/transporter capacity can be considered, when known, to limit corresponding rates, as well as other physiological information such as experimental measurements of species concentrations or fluxes. The more accurate the definition of constraints is, the better the resultant feasible solutions will represent actual metabolic phenotypes (Raman and Chandra, 2009).

When converting constraints into their mathematical description, generally two groups of restrictions are found: balances and bounds (Price et al., 2004). The conservation of mass is an example of a balance restraint. As it is the case of Equation (1.2), balances result in equality constraints. Sometimes physiological information is available in the form of ratios between fluxes, which would also derive in mathematical balances. Directionality and reversibility of reactions can be expressed in terms of bounds: allowable flux values will belong to interval $(-\infty, +\infty)$ for reversible reactions and $[0, +\infty)$ in the case of irreversible. Constraints related to enzyme capacity or nutrient availability usually lead to bound constraints too, as well as other kinds of physiological evidences such as experimentally measured fluxes.

After adding mathematical constraints to the model, non-feasible phenotypes are excluded from the set of allowable solutions. Some constraints are “hard” constraints that every phenotype must obey (mass balance, reaction reversibility, enzyme capacity), others are condition-dependent constraints that regulate the system's response to particular scenarios. However, even in the cases in which different, highly-informative constraints can be defined (which is the ideal situation, but isn't always possible, depending on the availability of information about the target metabolic system), genome-scale networks typically include large number of reactions (from hundreds to more than one thousand (Erdrich et al., 2015)) and the resulting system of equations will be still grossly under-determined.

Thus, with the addition of constraints it can be analysed what the metabolic network cannot do, although exact flux distributions cannot be determined. From this point on, two main trends have arisen in the field. Some researches have focused on studying the properties of the constraint-defined solution spaces: topology, convex basis, extreme behaviours, or dependencies among fluxes (Papin et al., 2004; Braunstein et al., 2008; Llaneras and Picó, 2010), appear among the most popular techniques within this approach. On the other hand, in some other cases the interest has been put in finding concrete flux distributions to gain insights into particular behaviours and phenotypes. In such cases, optimisation is used to select particular solutions within the feasible space. In this thesis the emphasis will be put in this latter approach.

Optimisation has been at the heart of metabolism simulation as a way to bypass the mathematical hurdle of solving under-determined systems of equations that are found in constraint-based genome-scale models (Stephanopoulos, 1999; Orth et al., 2010). This approach assumes the hypothesis that organisms' evolution under selective pressure tends towards optimality with respect to a metabolic function (Edwards et al., 2002a; Raman and Chandra, 2009; Schuetz et al., 2012; García Martín et al., 2015). In order to obtain significant simulation results, it is fundamental that the objective function mathematically imposed by the researcher is consistent with the evolutionary objective imposed by the environment. Objective functions must, thus, be consistent with known cellular demands and make sense in the light of evolution.

At this point, it is important to stress that optimisation is applied in constraint-based metabolic modelling as an instrument to cope with the high number of degrees of freedom, that cannot be reduced only by imposing constraints. However, it does not mean that “more optimal” simulation solutions are "better" approximations to actual metabolic phenotypes. The great challenge of this discipline is therefore to define meaningful sets of constraints and objective functions to ensure the pertinency of the obtained solutions. In fact, various studies point to the fact that often actual metabolic systems operate close to optimal metabolic performance, but in a quasi-optimal fashion (Ibarra et al., 2002; Fischer and Sauer, 2005; Schuetz et al., 2007, 2012) due to trade-offs among competing objectives, incomplete adaptive evolution under the conditions examined, or observance of other important factors (further objectives) like adaptation to diverse environments.

Together with constraint identification and description, selection of an appropriate objective function (or functions) is a central issue to these modelling techniques. The most commonly used objective function for metabolic simulations is maximisation of biomass yield or growth rate (Feist and Palsson, 2010). Importantly, "biomass formation" or "growth" is not inherently described at a metabolic reconstruction, it is not encoded at the genome in terms of enzymatic functions. Thus, explicit description of what is understood by growth is needed before facing simulations. This is usually a linear function that specifies stoichiometric proportions of biomass components, such as lipids, proteins and nucleic acids (or their precursors), as well as accounts for biosynthetic and maintenance costs. However, biomass composition is not an unvariable trait of living organisms; like all phenotypic features, it results from both genetic and environmental factors. So, mechanisms must be applied to account for such adaptability. In the case of linear contexts, different biomass equations can be defined for different conditions (Montagud et al., 2010).

Apart from biomass formation, other objective functions have been explored (Schuetz et al., 2007) and some of them have proven to be suitable for different analysis and design purposes. Some examples are maximisation of production of ATP (metabolic energy currency used by biological systems), minimisation of overall intracellular flux, maximisation of by-products of interest, or maximisation of metabolic energy efficiency. Often different objectives are of interest for different purposes and/or under different simulation conditions.

Another important issue associated to the application of optimisation to find individual flux distributions concerns the uniqueness of the optimal solution. In large metabolic networks (especially in genome-scale reconstructions), due to the existence of redundant pathways, it is frequent to find alternate optima that are alternative flux distributions with the same optimal value of the objective function (Mahadevan and Schilling, 2003). This fact must be considered when analysing particular solutions since different pathway distributions can be an important factor in some analyses.

Diverse optimisation techniques have been used in constraint-based modelling. The more traditional methods used Linear Programming (LP), but gradually other techniques were introduced to deal with new types of optimisation problems. In the next section, a review of methods using these constraint-based optimisation-focused metabolic modelling approaches is presented.

1.2.2 Simulating constraint-based metabolic models

Numerous constraint-based optimisation-driven methods have been described throughout the last two decades for genome-scale metabolic simulations (Patil et al., 2004; Price et al., 2004; Schellenberger et al., 2011; Zomorrodi et al., 2012; Lewis et al., 2012; King et al., 2015).

Since the appearance of *Flux Balance Analysis* (FBA) (Watson, 1984), a succession of studies and methods have been described that use diverse optimisation techniques and constraints to analyse metabolic phenotypes. Below, some of the most relevant examples are classified according to their fundamentals and purposes.

Flux Balance Analysis

Flux Balance Analysis (FBA) (for a review, see Orth et al. (2010)) is the main representative of a family of methods developed to analyse particular optimal flux distributions that arise after reduction of the mathematical solution space to a set of biologically meaningful flux vectors.

FBA has been satisfactorily applied to predict growth rates and pathway usage of both natural and genetically modified strains of different organisms of interest, as well as to assess the effect of different growth conditions and media compositions (see Famili et al. (2003); Teusink et al. (2009); Lewis et al. (2010); Caspeta et al. (2012) for examples).

According to this method, after imposing steady-state mass balance, and the addition of boundary constraints (derived from reaction directionality, enzyme/transport capacity and specific physiological knowledge), an optimisation problem is solved, using linear programming, to maximise/minimise an objective function which can be any linear combination of fluxes:

$$\max_{\boldsymbol{v}} Z(\boldsymbol{v}) = \boldsymbol{c}^T \cdot \boldsymbol{v} \tag{1.3}$$

subject to:

$$\boldsymbol{S} \cdot \boldsymbol{v} = \boldsymbol{0} \tag{1.4}$$

$$v_{j,rev} \in (-\infty, +\infty) \qquad j \in \{1, \ldots, n\} \tag{1.5}$$

$$v_{j,irr} \in [0, +\infty) \qquad j \in \{1, \ldots, n\} \tag{1.6}$$

$$l_{v_j} \leq v_j \leq u_{v_j} \qquad j \in \{1, \ldots, n\} \tag{1.7}$$

where $\boldsymbol{c}$ is a vector of weights indicating how much each reaction contributes to the objective function.

The solution obtained from FBA is a vector of fluxes leading to an optimal value of the objective function (*e.g.* biomass yield), including individual flux values for each reaction, which is defined as one particular optimal reaction network state. However, as explained before, this vector should be considered as one “potential” pathway distribution since, usually, different flux vectors can lead to the same optimal prediction (alternate optima).

Thus, the necessity of going in depth in the study of flux variations and their interdependencies rises. The following group of methods include techniques used to analyse how particular fluxes can vary within the network, and its derived effects on the objective function value, as well as, on neighbouring reactions.

Flux variability and flux dependencies

Genome-scale metabolic networks often display metabolic redundancies that confer them robustness to genetic and environmental variations. Investigation of how flux perturbations influence optimal objective performance can shed light on the system's plasticity, as well as serve as framework for testing bioengineering approaches based on nutrient modification or gene regulation.

One of the main issues associated to metabolic redundancies and pathway distributions in large-scale metabolic networks, the existence of alternate optima, has already been underlined. *Flux Variability Analysis* (FVA) (Mahadevan and Schilling, 2003) uses an LP-based strategy to calculate the full range of numerical values for each flux in the network leading to same optimal value of a particular objective function.

On the other hand, robustness and sensitivity studies have been performed using FBA to characterise how changes in particular fluxes (nutrient intakes or competing reactions, for example) may affect the optimal performance of the metabolic system. These effects can be analysed by changing one flux at a time, or by constructing the Phenotypic Phase Planes (PhPP) that emerge from simultaneous range-wide variation of two fluxes (Varma et al., 1993; Edwards et al., 2002b). Other parameters, such as biomass stoichiometric composition, can have an impact on flux distribution, and they can also be analysed by using FBA-based approaches (see Schwender and Hay (2012) for an example).

Finally, *Flux Coupling Analysis* (FCA) (Burgard et al., 2004) permits analysing the relationship between fluxes at genome-scale metabolic networks. It allows to identify and classify, qualitatively and quantitatively, dependencies between pairs of fluxes, as well as, to examine further consequences of those dependencies (identification of sets of reactions severely affected by changed intake rates or gene deletions due to flux couplings).

Mutant phenotype evaluation and strain design

As introduced in Section 1.2, the final objective of systems biology is not only to understand how natural systems work, but also to be able to rationally modify them and design new devices and systems with predetermined characteristics. In this context, metabolic modelling has been applied to test and suggest engineering strategies for bioengineering design.

Genes can be deleted from (knocked out) or inserted into (knocked in) the genetic code of a living organism to eliminate or incorporate biological functionalities. Using such genetic manipulations new cells can be created with new capacities, as it is the case of cell factories (*i.e.* bio-based systems for producing chemical commodities), or existing systems can be readjusted to become such a cell factory. If rational design is intended, side effects of modifications have to be previously assessed. Genome-scale metabolic simulations are a valuable tool for this purpose, since system-level metabolic interactions are considered.

FBA has been used to test the effect of genetic variations, yet special considerations must be taken. Although the assumption of optimality for a wild-type (*i.e.* "natural", unmodified version) organism is justifiable, the same argument may not be valid for genetically engineered systems that were not exposed to long-term evolutionary pressure. Specific methods for such cases are available and their main representatives are *Minimisation Of Metabolic Adjustment* (MOMA) (Segrè et al., 2002) and *Regulatory On/Off Minimisation* (ROOM) (Shlomi et al., 2005). MOMA employs Quadratic Programming (QP) to identify a point in flux space, which is closest to the wild-type point, compatible with the constraints imposed by the new condition. ROOM utilises Mixed-Integer Linear Programming (MILP) to find a flux distribution that minimises the total number of significant flux changes with respect to the wild-type flux distribution.

Nevertheless, methods like MOMA or ROOM can be applied to test the effect of genetic engineering strategies, but not to devise those strategies. Specific tools for strain design have been conceived that actively search for strain engineering modifications leading to targeted overproductions. One of the earliest efforts was the *OptKnock* (Burgard et al., 2003), a procedure based on bi-level optimisation that suggests gene knockouts maximising overproduction of a target chemical and growth by coupling biomass formation with chemical production. Later, *OptReg* (Pharkya and Maranas, 2006) extended OptKnock to consider not only knockouts but also finer activation and inhibition of various reactions in the network. In addition, *OptStrain* (Pharkya et al., 2004) allows for knock-ins of non-native functionalities (retrieved from a comprehensive universal database of biochemical reactions) to enable production of desired biochemicals. Another tool, *OptGene* (Patil et al., 2005) makes use of evolutionary programming procedures for the same purpose as OptKnock, but with two distinctive features: evolutionary search speeds up the process (especially when applied to large-scale networks), and it allows definition of non-linear constraints and objectives. Similarly, *Genetic Design through Local Search* (GDLS) (Lun et al., 2009) also reduces computational time required to predict multiple simultaneous gene deletions by employing an approach based on local search. Finally, *OptForce* (Ranganathan et al., 2010), thoroughly predicts minimal sets of fluxes that must actively be forced through genetic manipulations to guarantee a pre-specified overproduction level of a desired biochemical.

Application of additional constraints

The constrains imposed in FBA to form the flux solution space are steady-state mass balance (Equation (1.4)), reaction directionality (Equations (1.5) and (1.6)), and flux bounds stated by enzyme/transporter capacity limits or physiological measurements (Equation (1.7)). Frameworks for imposing other kinds of constraints have been developed.

Energy Balance Analysis (EBA) (Beard et al., 2002) explicitly considers energy balance and thermodynamics of the network reactions to ensure that the predicted flux vector is thermodynamically feasible. As a consequence of the imposition of energy balance constraints biological loops are forced to have no flux.

Dynamic FBA (DFBA) (Varma and Palsson, 1994b; Mahadevan et al., 2002) include additional constraints to account for slow changes in the growth environment. DFBA also allows the incorporation of kinetic expression when the kinetics are well characterised.

Parsimonious enzyme usage FBA (pFBA) (Lewis et al., 2010) employs FBA to optimise the growth rate, followed by minimising the net metabolic flux through all gene-associated reactions in the network. The underlying assumption is that, under growth pressure, there is a selection for strains that can process the growth substrate the most rapidly and efficiently while using the minimum amount of enzyme.

Various methods have approached the question of integrating regulatory constraints into stoichiometric metabolic simulations. First of those methods was *Regulatory FBA* (rFBA) (Covert et al., 2001), that uses Boolean rules (and few dynamic parameters about protein synthesis and degradation) to generate consecutive, time-step separated, optimisation problems in which transcriptional regulation is applied as on/off rules. A variant of the former, *Steady-state Regulatory FBA* (srFBA) (Shlomi et al., 2007) combines Boolean variables for regulation ($r \in \{0, 1\}$) with real variables for fluxes ($v \in \mathbb{R}$) to solve a MILP problem, and employs FVA to explore alternative solutions. *Integrated FBA* (iFBA) (Covert et al., 2008) goes one step further and incorporates an ODE model of cell signalling, while *Integrated Dynamic FBA* (idFBA) (Lee et al., 2008) also differentiates between “fast” reactions under quasi-steady-state conditions and time-delayed “slow” reactions in a similar way to DFBA.

Integration of *omics* datasets

The methods listed in the previous section that try to account for regulation effects present two main drawbacks (Lewis et al., 2012): first, they assume binary responses for all transcription-regulatory interactions, when real biological systems exhibit a range of behaviours, from binary to continuous; second, few organisms have been studied enough to provide adequate regulatory information to describe the required Boolean rules. As a consequence, other methods have explored different approaches based on *omics* data integration to try to elucidate and/or encompass regulation effects over the metabolic network, either explicitly or implicitly.

One of the first attempts to integrate transcriptomic data into the metabolic network topology was made by Patil and Nielsen (2005). They developed an algorithm based on hypothesis-driven data analysis to reveal patterns in the metabolic network that follow a common transcriptional response. This algorithm enables identification of so-called *Reporter Metabolites* (metabolites around which the most significant transcriptional changes occur), and a set of connected genes with significant and coordinated response to genetic or environmental perturbations. In a later development (Oliveira et al., 2008), the authors extend the method for its use with other kinds of bio-molecular networks, to identify further key biological features (*Reporter Features*).

Later, new algorithms appeared that tried to take advantage of transcriptomic data, as well as other kinds of *omics* information, to investigate pathway activation and consequently constrain metabolic networks. Some of those methods use expression data to create context-specific models, thus applying regulation by means of stoichiometric constraints. GIMME (*Gene Inactivity Moderated by Metabolism and Expression*) (Becker and Palsson, 2008) (with its later evolution *Gene Inactivation Moderated by Metabolism, Metabolomics and Expression* (GIM3E) (Schmidt et al., 2013)) and the *Integrative Metabolic Analysis Tool* (iMAT) (Zur et al., 2010) that implements a method previously proposed by Shlomi et al. (2008), use different strategies to classify reactions according to their expression levels and build context-specific models that respectively maximise and minimise the usage of highly and lowly expressed reactions. The *Model-Building Algorithm* (MBA) (Jerby et al., 2010) exploits literature-based knowledge, transcriptomic, proteomic, metabolomic and phenotypic data as evidence to include relevant metabolic functions at the context-specific network, while *Metabolic Context-specificity Assessed by Deterministic Reaction Evaluation* (mCADRE) (Wang et al., 2012) uses a similar approach but adding connectivity-based evidences derived from network topology analyses.

Other methods simultaneously extract information from measurements across multiple conditions looking for a suitable set of flux distributions across all conditions that best match the reported expression levels. Both, *Metabolic Adjustment by Differential Expression* (MADE) (Jensen and Papin, 2011) and *Transcriptionally controlled FBA* (tFBA) (Van Berlo et al., 2011) use the same principles, albeit with a different formulation, to decide about expression patterns by comparing measurements across multiple conditions, that are then considered simultaneously to obtain the final flux distribution.

There exist also algorithms in which the expression information is not used to alter the network topology (or usage), but the bounds stated for the corresponding reactions. The method called *E-Flux* (because it combines flux and expression), directly maps normalised gene expression levels into flux bound constraints. In the first implementation of this method (Colijn et al., 2009), normalisation of the expression level is based on expression differences across all genes; in a later study (Brandes et al., 2012), expression level of the same gene across multiple experiments is considered instead. In the same line, *Probabilistic Regulation Of Metabolism* (PROM) (Chandrasekaran and Price, 2010) goes one step further: given abundant gene expression data measured under multiple conditions, generates a probabilistic model for the gene regulatory network, which is integrated with a constraint-based metabolic model by setting the flux bounds proportional to the associated probabilities calculated. This method provides a way to construct integrated genome-scale regulatory-metabolic networks, nevertheless it requires large amount of experimental data that are not always accessible for every organism.

Consideration of resource limitations

A new trend in metabolic network reconstruction and analysis is moving towards considering not only metabolic function, but also the machinery and resources needed to perform those functions. *Resource Balance Analysis* (RBA) (Goelzer et al., 2011), one of the firsts approaches in this direction, introduces three additional design constraints (metabolic capability constraint, translation capability constraint and density constraint) that account for structural and resource limitations that are found in real cells. The arising problem is equivalent to an LP optimisation problem, and its solution describes more accurately how the available resources in the medium are distributed among the various cellular subsystems.

More recently, a method has been formulated, *Constrained Allocation Flux Balance Analysis* (CAFBA) (Mori et al., 2016), in which a single additional global constraint on fluxes is added that encodes for the tradeoff in the allocation of cellular resources across ribosomal, transport and biosynthetic proteins, conducing to simulation results that account for the biosynthetic costs associated to growth.

However, the most significant step towards consideration of the overall metabolic costs associated to cellular operation was taken by Lerman et al. (2012). In this work, the authors present an integrated model of metabolism and macromolecular expression, *ME-model*, that explicitly accounts for the genotype-phenotype relationship with biochemical representations of transcriptional and translational processes. This new generation of metabolic models will establish a new paradigm in the system-level modelling of cells in the next years, but nevertheless deep knowledge is needed to reconstruct such a complete model and, by now, they are restricted to a few, well-studied model organisms.

1.2.3 Software tools for constraint-based metabolic modelling

There are many open-source software tools available for constraint-based metabolic modelling. Some of them perform only a few functions, but some include a wide range of operations related to reconstruction and analysis of metabolic networks. Some of the most commonly used are listed below.

The *Constraint-Based Reconstruction and Analysis Toolbox* (COBRA Toolbox) (Becker et al., 2007; Schellenberger et al., 2011) is a software package running in the MATLAB® environment that collects numerous tools for simulation and analysis of metabolic phenotypes using genome-scale models. A following development specifically developed for Python, COBRApy (Ebrahim et al., 2013), defines an object-oriented framework that facilitates the representation of the complex biological processes of metabolism and gene expression and includes parallel processing support for computationally intensive processes.

The *BioMet ToolBox* (Cvijovic et al., 2010; Garcia-Albornoz et al., 2014) is a web-based resource that offers a simple graphical user interface and includes two sets of tools for exploiting the capabilities of genome-scale metabolic networks: algorithms for metabolic model analysis and simulation, and algorithms for omics analysis. Some of the tools integrated in the BioMet ToolBox are also available in a downloadable version.

OptFlux (Rocha et al., 2010) is an application, available for multiple platforms, that includes a number of tools to support metabolic computations. Algorithms are included to handle metabolic models in different formats, visualise metabolic networks, perform flux analysis and simulations, and work on strain design.

Pathway Tools (Karp et al., 2002, 2016) is a software system that supports several use cases in bioinformatics and systems biology. One of its components, MetaFlux, allows development of metabolic flux models, FBA-based simulations (including selected and exploratory knockouts), and visualisation of the results in flux maps.

In the present work, *PyNetMet* (Gamermann et al., 2014b) has been used for manipulation and simulation of genome-scale metabolic networks of cyanobacteria. It is a Python library of tools, designed in an object-oriented fashion, to efficiently manage metabolic networks and models, and perform FBA-based metabolic simulations. It was developed in our research group, and the last version can be downloaded from `github.com/CyanoFactory/CyanoFactoryKB`.

PyNetMet describes four classes:

- *Enzyme* objects represent biochemical reactions. Methods and attributes described for this class allow verification and management of properties like stoichiometry, reversibility, substrates and products, *etc.*.
- *Network* class defines a graph (as a set of nodes and edges) and provides classic graph theory methods for its analysis.
- *Metabolism* objects are composed of *enzyme* objects, and represent full metabolic networks. Methods and attributes of this class permit explore and modify all features associated to the metabolic model (stoichiometry, topology, connectivity, constraints, objective function, *etc.*.).
- *FBA* class uses a *metabolism* object as input, and offers tools for flux simulation and analysis. Several FBA-based methods are defined that allow analysis of fluxes, reaction essentiality screening, and robustness and sensitivity analyses among others.

Classes *metabolism* and *FBA* have been extensively used in this work to manage metabolic models and perform FBA-based metabolic simulations.

1.3 Metabolic modelling of cyanobacteria

Cyanobacteria are photoautotrophic[6] prokaryotes able to perform oxygenic photosynthesis. These photosynthetic bacteria become of great importance within the field of scientific research for many aspects. They are thought to be the evolutionary ancestors of chloroplasts under the endosymbiont hypothesis (Douglas, 1998; Raven and Allen, 2003) and they are believed to be the organisms that changed the ancient anoxygenic environment to oxygenic by photosynthesis (Schopf, 2000). In addition, these organisms present a wide ecological distribution and they can be found in a large variety of habitats (oceans, lakes, soils, and even extreme environments) (Herrero and Flores, 2008). Since last century, cyanobacteria have been considered model organisms for the study and the characterisation of a multitude of biological processes, like photosynthesis and its genetic control, atmospheric nitrogen fixation, nitrogen, carbon and hydrogen metabolism, or tolerance to environmental stress (salinity, hight light, nutrient scarcity, *etc.*) (Koksharova and Wolk, 2002).

More recently, thanks to the development and affordability of genetic tools and molecular techniques applied to these organisms (Koksharova and Wolk, 2002; Hess, 2011), it has been revealed the great biotechnological potential of cyanobacteria for many applications (Abed et al., 2009; Ducat et al., 2011; Lau et al., 2015), such as production of biofuels (Angermayr et al., 2009; Dismukes et al., 2008; Parmar et al., 2011; Rodionova et al., 2016), including hydrogen (Tamagnini et al., 2007; Tiwari and Pandey, 2012; Montagud et al., 2015), secondary metabolites of industrial and pharmaceutical interest (Rastogi and Sinha, 2009), and some other products like pigments (Eriksen, 2008) or biopolymers (Li et al., 2001) for example. Moreover, the fact that they are autotrophic organisms with low nutrient requirements, make them even more at-

[6] Photoautotrophic organisms, or photoautotrophs, are organisms that carry out photon capture to acquire energy from light. Generally, these organisms are also carbon fixators: they can use atmospheric carbon to generate molecules used in their metabolism.

tractive as production platforms (Pinto et al., 2015), since they are able to survive and produce using sun energy, atmospheric carbon and a few inorganic nutrients. This way, cyanobacteria have been placed in the spotlight as basis organisms for the development of cell factories for industrial bioprocesses.

To date, more than 120 complete genomes of cyanobacteria have been sequenced (numbers retrieved from NCBI Genome database (NCBI Resource Coordinators, 2016), excluding 'contig' and 'scaffold' levels, on Sept. 29^{th}, 2017) thus enabling the creation of genome-scale reconstructions that may be used to provide insight into intracellular mechanisms and guide the optimisation of producing strains. However, nowadays manually curated genome-scale models have been developed only for a few species (Baroukh et al., 2015).

Among the species for which genome-scale models have been reconstructed, *Synechocystis* sp. PCC 6803 and *Synechococcus elongatus* PCC 7942 will be the focus of this dissertation. Table 1.1 shows the metabolic reconstructions of these two model cyanobacteria reported to date. Differences in the number of genes considered, as well as the resultant number of reactions and metabolites, are shown. The models listed are classified as:

CC Central Carbon metabolic models which include reactions from the core carbon metabolism (*i.e.* glycolisis, citric acid cycle and pentose phosphate pathway), some lumped reactions describing photosynthesis and oxidative phosphorylation and, in some cases, an additional set of reactions producing amino acids; or

GS Genome-Scale metabolic models in which all reactions annotated in the genome are considered following the reconstruction process described in Section 1.2.1. An equation describing biomass assembly from main building blocks (such as amino acids, nucleotides, lipids or carbohydrates) present in cell metabolism must be included.

Additional main differences between the reconstructed networks lie in the degree of detail in the representation of photosynthesis, the definition of subcellular compartments and the degree of detail in pathways of the peripheral metabolism (synthesis of lipids, complex carbohydrates, secondary metabolites, *etc*.).

Synechocystis sp. PCC 6803

Synechocystis sp. PCC 6803 is a unicellular fresh water cyanobacteria, whose complete genome was sequenced, annotated and published in 1996 by Kazusa's laboratory (Kaneko et al., 1995, 1996, 2003). It is one of the best studied cyanobacterial strains, with a large amount of physiological and molecular data available, and it is naturally transformable[7] (Pinto et al., 2012). All of that has contributed to its wide use as a model organism, and has motivated the development of a number of molecular tools for its genetic manipulation (Angermayr et al., 2009; Hess, 2011).

Another important feature that makes it even more attractive for research and biotechnological applications is its adaptability. This cyanobacterium can grow and survive under three different trophic conditions as marked by the utilised energy and carbon sources (Herrero and Flores, 2008), namely (i) *photoautotrophy*, where energy comes from light and carbon from CO_2; (ii) *heterotrophy*, where both energy and carbon source is a saccharide, for instance glucose; and (iii) *mixotrophy*, a combination of the former two, where light is present as well as a combination of two carbon sources: glucose and CO_2.

Synechococcus elongatus PCC 7942

Synechococcus elongatus PCC 7942 is considered a model organism since the early 70s, when successful transformations of exogenous DNA were performed for the first time in a cyanobacterium (Shestakov and Khyen, 1970). Its genome was sequenced, annotated and published in 1980 (van den Hondel et al., 1980; Van der Plas et al., 1992; Chen et al., 2008).

[7] It has the natural ability to take up and incorporate exogenous genetic material.

Table 1.1: *Review of metabolic network reconstructions of the model cyanobacteria Synechocystis sp. PCC 6803 and Synechococcus elongatus PCC 7942.*

References	Type	Genes	Reacs.	Metabs.[a]	Comps.
***Synechocystis* sp. PCC 6803**					
Yang et al. (2002)	CC	NA	20	15	1
Shastri and Morgan (2005)	CC	NA	70	50	1
Hong and Lee (2007)	CC	78	56	63	2
Kun et al. (2008)	GS	383	916	879	1
Navarro et al. (2009)	CC	NA	90	56	1
Fu (2009)	CC[b]	633	831	704	1
Montagud et al. (2010)	GS	669	882	790	2
Knoop et al. (2010)	GS	337	380	291	1
Montagud et al. (2011)	GS	811	976	922	2
Yoshikawa et al. (2011)	GS	393	493	465	2
Nogales et al. (2012)	GS	678	863	795	3
Saha et al. (2012)	GS	731	1156	996	4
Knoop et al. (2013)	GS	677[c]	759	601	4
Erdrich et al. (2014)	GS	-	600	571	1
Maarleveld et al. (2014)	GS	686	904	816	3
Knoop and Steuer (2015)	GS	706	780	601	4
He et al. (2015)	GS	678	865	795	3
Mohammadi et al. (2016)	GS	692	768	557	4
This thesis	GS	842	1059	920	2
***Synechococcus elongatus* PCC 7942**					
Triana et al. (2014)	GS	715	851	838	2
Broddrick et al. (2016)	GS	785	850	768	4

Reacs. = reactions; *Metabs.* = metabolites; *Comps.* = compartments

[a] Non-unique metabolites, species may repeat in different compartments.

[b] This reconstruction includes a genome-wide list of reactions but it lacks a proper genome-scale biomass equation. It uses biomass equation formulation from the central carbon model of Shastri and Morgan (2005), which leads to short-cuts in flux distributions when optimised.

[c] Authors distinguish between a core network and an augmented network including all remaining annotated enzymes with putative metabolic function. Numbers shown here are for the core network. The extended network accounts for 1035 genes.

In particular, it has been used as a paradigm for the study of circadian rhythms in prokaryotes.

Synechococcus elongatus PCC 7942 has a rod-shaped appearance, has the ability to survive in freshwater environments with low nutrients, and is considered an obligate autotroph (Rippka et al., 1979) (*i.e.* it has to use CO_2 as carbon source and light as energy source). These low nutritional requirements make it also a very interesting host organism for the development of biotechnological processes.

2

Fundamentals on multi-objective optimisation procedures

2.1 Chapter abstract

In this chapter the fundamentals about multi-objective optimisation and multi-objective optimisation procedures needed in this work to understand their application to metabolic simulation are presented.

At the beginning of this chapter some important definitions and terminology are introduced. Equation (2.1) presents the general form of a multi-objective optimisation problem. This problem can be solved applying a Generate-First Choose-Later (GFCL) approach, which has the advantage of providing the researcher with more information contained in the Pareto set, although it requires greater understanding of the problem at hand, and often more time.

A complete optimisation procedure based on the GFCL approach involves three stages. In the first stage, the multi-objective problem (MOP) definition, the scope of the problem, as well as the constraints and objectives involved, are analysed, to finally state proper MOP(s) that should lead to sets of solutions pertinent for the situation at hand. In order to generate the Pareto optimal solutions, a multi-objective optimisation process is then launched, that must be conducted by a proper algorithm ensuring the quality of the obtained set of solutions, as well as, the manageability of the problem. Finally, the set of solutions obtained have to be analysed by an expert (the decision-maker) that will choose the more preferable according to his/her criteria. At each of these three stages knowledge is gained about the problem at hand that can be used to improve any of the steps thus enhancing the quality of the final solution(s).

In this work, this approach has been employed for the simulation of metabolic system. At Chapter 5 details about the application of the three steps to the constraint-based metabolic modelling framework are provided. Through this process two pre-existing tools are exploited: sp-MODE algorithm and LD-ToolBox. A brief explanation about these two tools is included in the final section of the present chapter.

2.2 Background on multi-objective optimisation

A multi-objective problem (MOP) can be stated, in a general way, as follows:

$$\min_{\boldsymbol{v}} \boldsymbol{Z}\left(\boldsymbol{v}\right) = \left[Z_1\left(\boldsymbol{v}\right), \ldots, Z_q\left(\boldsymbol{v}\right)\right] \tag{2.1}$$

subject to:

$$\boldsymbol{g}\left(\boldsymbol{v}\right) \leq 0 \tag{2.2}$$

$$\boldsymbol{h}\left(\boldsymbol{v}\right) = 0 \tag{2.3}$$

$$l_{v_j} \leq v_j \leq u_{v_j} \qquad j \in \{1, \ldots, n\} \tag{2.4}$$

where $\boldsymbol{v}$ is the vector of decision variables (or *decision vector*) with n components, $\boldsymbol{Z}\left(\boldsymbol{v}\right)$ is the vector of objective functions (or *objective vector*) with $q \geq 2$ components, $\boldsymbol{g}\left(\boldsymbol{v}\right)$ are the inequality constraints, $\boldsymbol{h}\left(\boldsymbol{v}\right)$

are the equality constraints, and l_{v_j} and u_{v_j} are the lower and upper bounds, respectively, for the variable v_j in the decision space.

For simplicity, minimisation of all the objective functions is assumed. If any of the q objectives is to be maximised, the following transformation can be applied:

$$\max_{\boldsymbol{v}} Z_k(\boldsymbol{v}) = -\min_{\boldsymbol{v}} -Z_k(\boldsymbol{v}) \qquad k \in \{1, 2, \ldots, q\} \tag{2.5}$$

If there would be no conflict between objectives, then a solution could be found where every objective function at the objective vector attains its optimum. However, in general, it is not possible to find a single solution that is optimal for all objectives simultaneously, and therefore several solutions with different trade-off levels may appear. In order to select the "best" solutions some order relation must be defined that allows comparison between objective vectors. To satisfy that necessity the concepts of *dominance* (Definitions 2.1, 2.2 and 2.3) and *Pareto optimality* (Definition 2.4) are introduced. In common words, an objective vector is *Pareto optimal* if none of its components can be improved without worsening at least one of the other components (Miettinen, 1999).

Definition 2.1 (Dominance (Miettinen, 1999))**.** *An objective vector* $\boldsymbol{Z}\left(\boldsymbol{v}^1\right)$ *dominates another objective vector* $\boldsymbol{Z}\left(\boldsymbol{v}^2\right)$*, denoted by* $\boldsymbol{Z}\left(\boldsymbol{v}^1\right) \preceq \boldsymbol{Z}\left(\boldsymbol{v}^2\right)$*, if* $Z_k\left(\boldsymbol{v}^1\right) \leq Z_k\left(\boldsymbol{v}^2\right)$ *for all* $k \in \{1, 2 \ldots, q\}$ *and* $Z_k\left(\boldsymbol{v}^1\right) < Z_k\left(\boldsymbol{v}^2\right)$ *for at least one* $k \in \{1, 2, \ldots, q\}$*.*

Definition 2.2 (Strict Dominance (Miettinen, 1999))**.** *An objective vector* $\boldsymbol{Z}\left(\boldsymbol{v}^1\right)$ *dominates another objective vector* $\boldsymbol{Z}\left(\boldsymbol{v}^2\right)$ *if* $Z_k\left(\boldsymbol{v}^1\right) < Z_k\left(\boldsymbol{v}^2\right) \forall\, k \in \{1, 2, \ldots, q\}$*.*

Definition 2.3 (Weak Dominance (Miettinen, 1999))**.** *An objective vector* $\boldsymbol{Z}\left(\boldsymbol{v}^1\right)$ *weakly dominates another vector* $\boldsymbol{Z}\left(\boldsymbol{v}^2\right)$ *if* $Z_k\left(\boldsymbol{v}^1\right) \leq Z_k\left(\boldsymbol{v}^2\right) \forall\, k \in \{1, 2, \ldots, q\}$*.*

Definition 2.4 (Pareto optimality (Miettinen, 1999))**.** *An objective vector* $\boldsymbol{Z}\left(\boldsymbol{v}^\star\right)$ *is Pareto optimal if there is no other objective vector* $\boldsymbol{Z}\left(\boldsymbol{v}\right)$ *such that* $\boldsymbol{Z}\left(\boldsymbol{v}\right) \preceq \boldsymbol{Z}\left(\boldsymbol{v}^\star\right)$*.*

Thus, Pareto optimal solutions are those for which there is no other solution in the feasible space solution that dominates them. The (infinite) set of solutions that are Pareto optimal is called the *Pareto set* (Definition 2.5). Each solution in the Pareto set defines an objective vector in the *Pareto front* (Definition 2.6).

Definition 2.5 (Pareto set (Miettinen, 1999))**.** *In a multi-objective problem, the Pareto set $\boldsymbol{V_P}$ is the set including all the Pareto optimal solutions.*

Definition 2.6 (Pareto front (Miettinen, 1999))**.** *In a multi-objective problem, the Pareto front $\boldsymbol{Z_P}$ is the set including the objective vectors of all the Pareto optimal solutions in the Pareto set.*

Therefore, Pareto optimal solutions are the "best" possible solutions. However, most of the times the Pareto set is unknown, and then the goal is to find solutions that approximate it. Figure 2.1 shows the solution space of a bi-objective problem ($q = 2$) at the space of objective vectors. The shaded area delimits the feasible space, which is the space within solutions that satisfy the constraints fall on. The bold line represents the (infinite) a priori unknown Pareto front. The solutions represented by empty circles are dominated solutions, since other found solutions (full circles) dominate them. The three solutions represented by full circles are non-dominated solutions, while only the purple one is Pareto optimal, not the other two, since other solutions (not found in this case) dominate them. The set of non-dominated solutions (full circles) build the Pareto front approximation $\boldsymbol{Z_P^*}$, and their corresponding decision vectors build the Pareto set approximation $\boldsymbol{V_P^*}$.

Three more vectors appear in Figure 2.1 that are interesting to characterise the ranges of the Pareto front: the *ideal objective vector* (green diamond), the *utopian objective vector* (yellow diamond) and the *nadir objective vector* (red diamond). Following formal definitions are given for those vectors:

Definition 2.7 (Ideal objective vector (Miettinen, 1999))**.** *The ideal objective vector $\boldsymbol{Z^{ideal}}$ is an objective vector whose components Z_k^{ideal} are obtained by minimising each of the objective functions individually subject to the con-*

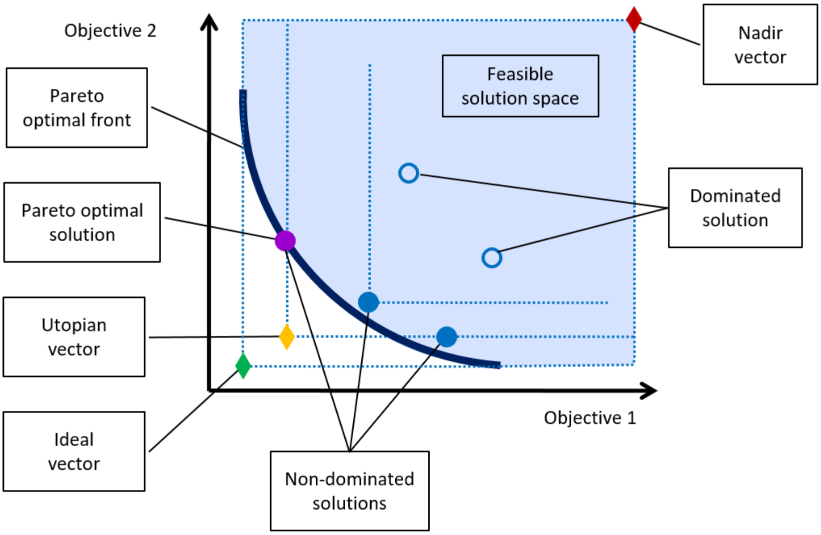


Figure 2.1: *Dominance, Pareto optimality and ranges of the Pareto front.*

straints, that is by solving:

$$\min_{\boldsymbol{v}} Z_k\left(\boldsymbol{v}\right) \qquad \forall\, k \in \{1, 2 \ldots, q\} \tag{2.6}$$

subject to:

$$\boldsymbol{g}\left(\boldsymbol{v}\right) \leq 0 \tag{2.7}$$

$$\boldsymbol{h}\left(\boldsymbol{v}\right) = 0 \tag{2.8}$$

$$l_{v_j} \leq v_j \leq u_{v_j} \qquad j \in \{1, \ldots, n\} \tag{2.9}$$

Thus the *ideal objective vector* is a feasible solution (it satisfies the constraints) that is minimal for all the objectives at the same time. It is obvious that if such a vector would exist it would be the solution of the MOP. However, in general, this is not possible due to the competition between objectives. Nevertheless, the ideal objective vector is an important entity, since it contains all the minimal values for all the objectives, those are the lower bounds of the (unknown) Pareto front. Besides, even when unattainable, it must be kept in mind as a reference point, the ideal goal (Miettinen, 1999).

In practice, since the authentic Pareto front is normally unknown, it is not possible to determine its components. Instead, a *utopian objective vector* is defined as an approximation of the ideal vector. The components of the utopian objective vector can be computed from the lower bounds of the Pareto front approximation.

Definition 2.8 (Utopian objective vector (Miettinen, 1999))**.** *An utopian objective vector* $\boldsymbol{Z}^{utopia}$ *is an infeasible objective vector whose components are formed by*

$$Z_k^{utopian} = Z_k^{ideal} + \epsilon_k \qquad \forall\, k \in \{1, 2 \ldots, q\} \tag{2.10}$$

where $\epsilon_k > 0$ *is a relatively small but computationally significant scalar.*

Definition 2.9 (Nadir objective vector (Miettinen, 1999))**.** *The nadir objective vector* $\boldsymbol{Z}^{nadir}$ *is the vector whose components are the upper bounds of the Pareto front.*

Again, since the Pareto front is usually not known, the components of the nadir objective vector are usually approximated by the upper bounds of the Pareto front approximation, and this approximation is often taken as the true nadir objective vector.

2.3 Stages of multi-objective optimisation procedures

There exist two main approaches to solve a multi-objective problem like the one defined in Equation (2.1) (Mattson and Messac, 2005): the *Aggregate Objective Function* (AOF) approach and the *Generate-First Choose-Later* (GFCL) approach. In the AOF methodology, a single objective function is built by combining the individual objectives (by means of a weighted sum, for example), and then a mono-objective optimisation algorithm is used to optimise this single index. The disadvantage of this methodology is that it produces a single solution, displaying a fixed relation among the objectives, that has been actually imposed when defining the weights. Thus, most of the information that can be retrieved from the Pareto front (dependence and competition between objectives, different trade-off levels) is lost.

On the other hand, in the GFCL approach, the main goal is to generate many potentially desirable Pareto optimal solutions, and then select the most preferable solution(s) among them. Figure 2.2 shows the whole multi-objective optimisation procedure using GFCL methodology (Reynoso-Meza et al., 2017). According to this methodology, the first step involves the description of the variables, constraints and objectives that define the multi-objective problem (MOP). Once the problem has been carefully stated, a multi-objective optimisation process can be performed to generate solutions that approximate the Pareto set. In a subsequent decision-making step, the solutions are analysed by the decision-maker, who selects the best solution(s) according to his/her preferences and the requirements of the problem at hand.

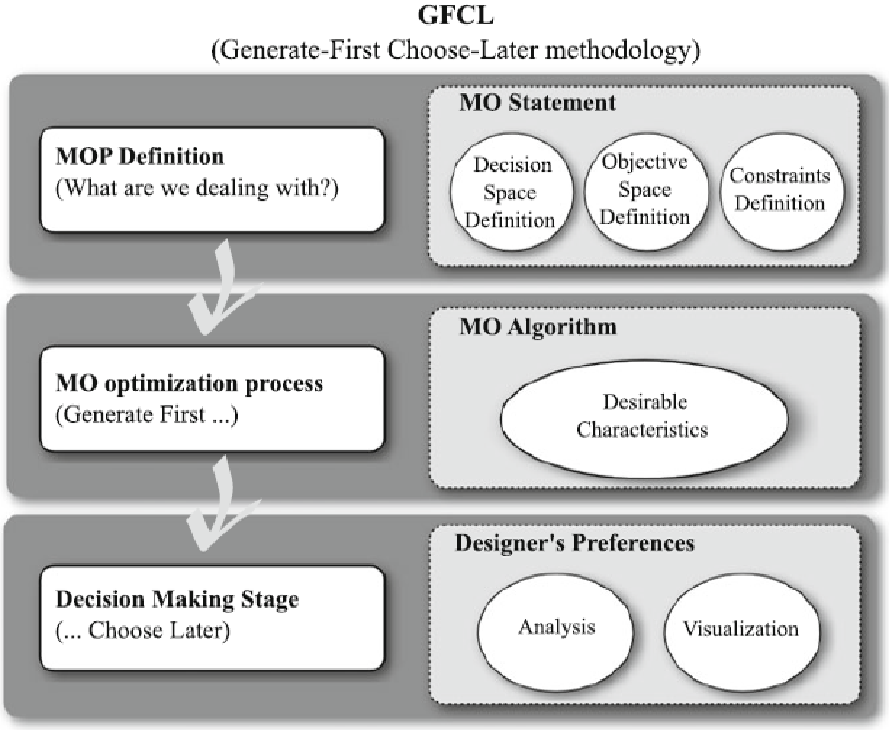


Figure 2.2: *Multi-objective optimisation procedure based on the GFCL methodology. Figure 2.2 from Reynoso-Meza et al. (2017).*

Under the GFCL approach, several solutions are obtained and closely scrutinised through an overall process that constitutes a valuable tool to understand the problem, study the underling relation between objectives, and analyse trade-off levels between them (Mattson and Messac, 2005). Nevertheless, it requires more time and attention from the researcher. At following sections the three steps of a multi-objective optimisation procedure based on the Generate-First Choose-Later approach are explained in further detail.

2.3.1 Multi-objective problem definition

The main goal of this stage is to articulate a multi-objective statement that precisely describes the problem to address and ensures the obtained solutions match the needs of the decision-maker. First, the context must be considered: what are the variables under study, that is the decision variables, and what rules connect them. For this purpose, generally, a parametric model is used that relates the variables among them and describes the standards that govern the system (Mattson and Messac, 2005). This model will set some limitations to the values that the variables can take, as they may influence each other, and will establish the outputs that result when those variables take particular values.

Apart from the general framework and the variables at stake, it is essential to define which is the aim of the procedure, what is pursued with the analysis. This means answering questions such as: what are the results supposed to be used for, what do the solutions represent, what values must be avoided, what features are desirable at the set of solutions. Answering these questions should pave the way to define the objectives to optimise and describe the constraints to impose. Together, the model, the objectives and the constraints strongly influence the solutions that can be found through the process, and thus this is a very important step to ensure the quality of the results (Mattson and Messac, 2005).

In this thesis, multi-objective optimisation is applied to the problem of metabolic simulation. The variables under study are the metabolic fluxes, and the model that encompasses them is the stoichiometric model of the metabolic network. Physico-chemical, biological and environmental constraints that condition the metabolic operation of cells have to be translated into mathematical bounds and balances. Suitable objective functions that account for biologically meaningful objectives, such as optimal resource utilisation, maximal growth yield or minimal energy consumption, must be defined. As stressed in Chapter 1, an accurate definition of constraints and objectives that properly capture the conditions that shape the metabolic response of biological systems is a fundamental requisite to obtain practical simulation results.

2.3.2 Multi-objective optimisation process

The multi-objective optimisation process seeks to approximate the collection of decision variables arrays (that is the Pareto set approximation, $\boldsymbol{V}_P^*$) that give the best Pareto front approximation ($\boldsymbol{Z}_P^*$). For such purpose, an appropriate algorithm must be used that fulfils the desirable characteristics required for the problem at hand. Some of those characteristics are related with the expected quality of the Pareto front approximation, as it is the case of:

- *convergence* (reaching the true, normally unknown, Pareto front),
- *diversity* (getting a useful spreading along the Pareto front approximation), and
- *pertinency* (obtaining useful solutions pertinent for the context).

While others concern how to deal with specific optimisation instances, like

- *constrained* problems (affected by, many times non-linear, inequality or equality restrictions),
- *large-scale* problems (with several decision variables, on the order of hundreds), or

- *multi-modal* problems (where different decision vectors lead to the same objective vector).

Convergence and diversity properties are considered *a must* in multi-objective optimisation (Reynoso-Meza et al., 2014). An additional required characteristic regards to pertinency improvement, which means getting reasonable solutions that fit the needs of the decision-maker. It may happen that several solutions approximated are not practical, due to a strong degradation in some objectives. This is a characteristic to take into account, mainly when performing more than 2 objectives simultaneous optimisation (Ishibuchi et al., 2008).

In the case of metabolic simulation, the pertinency of the solutions imply that they are adequate representations of real metabolic phenotypes. They must present some properties expected from living organisms, and perform as close as possible to known metabolic responses. It has also to be taken into account that when working at genome-scale, the number of decision variables (fluxes of the reactions in the network) is typically around one thousand (Erdrich et al., 2015), and thus the MOP will be a large-scale problem. This large amount of variables is in part due to metabolic redundancies that, at the same time, cause the problem to be multi-modal (alternate optima (Mahadevan and Schilling, 2003)). Besides, restrictions are needed to ensure the realism of the solutions, which combined with the large number of variables, could affect convergence. In the present work, specific mechanisms are proposed to deal with these special features of constraint-based metabolic simulations.

Evolutionary multi-objective optimisation

There exist various classic techniques to solve multi-objective optimisation problems (Marler and Arora, 2004), such as varying weighting vectors, Normal Boundary Intersection (NBI) algorithm (Das and Dennis, 1998), ϵ-constraint (Miettinen, 1999), Physical Programming (Messac and Mattson, 2002), or Normal Constraint (NC) algorithm (Messac et al., 2003). But in recent times the use of evolutionary algorithms to treat multi-objective optimisation problems has become popular (Zhou et al., 2011). Several evolutionary and bio-inspired techniques regularly appear in multi-objective evolutionary algorithms (MOEAs). The most popular include Genetic Algorithms (GA) (Konak et al., 2006), Particle Swarm Optimisation (PSO) (Coello Coello, 2011), and Differential Evolution (DE) (Mezura-Montes et al., 2008; Das and Suganthan, 2010; Das et al., 2016), although nature-inspired techniques like Artificial Bee Colony (ABC), Ant Colony Optimisation (ACO) or Firefly algorithms are becoming common (Yang, 2010).

In this work a multi-objective evolutionary algorithm (MOEA) is used. Evolutionary algorithms are among the so-called *population based* algorithms, because the concept of "population" is applied to the set of candidates solutions (individuals of the population). Algorithm 2.1 shows the general structure of a MOEA (Reynoso-Meza et al., 2017). First of all, an initial population of solutions is generated, normally on the basis of some random distribution (line 1). The individuals at the population are then evaluated for their fitness with respect to the objective functions (line 2). With this information, non-dominated (Definition 2.1) solutions are selected to form the first approximation of the Pareto set (line 3). Then the evolutionary process starts: at each generation a new population of solutions is formed from their predecessors using evolutionary operators (that depend on the particular evolutionary technique) (line 7); the new population is then evaluated for performance (line 8); and the Pareto set approximation is updated with potential preferable solutions coming from the latest population (line 10). The process will stop when the selected solutions are thought to be close enough to the unknown Pareto front.

The particular implementation applied here is an adaptation of the sp-MODE algorithm (Reynoso-Meza et al., 2010) that incorporates specific mechanisms to deal with the particular problem of metabolic simulations. The principles and implementation of sp-MODE are presented in Section 2.4.1.

```
Input: optimisation parameters
Output: Pareto set approximation $\boldsymbol{V}_P^*$
Build initial population $P|_0$ with $N_p$ individuals;
Evaluate $P|_0$;
Build initial Pareto set approximation $\boldsymbol{V}_P^*|_0$;
Set generation counter $G = 0$ ;
while convergence criteria not reached do
    $G = G + 1$;
    Build population $P'|_G$ using $P|_{G-1}$ with an evolutionary
      technique or bio-inspired technique ;
    Evaluate $P'|_G$ ;
    Build Pareto set approximation $\boldsymbol{V}_P^*|_G$ with $\boldsymbol{V}_P^*|_{G-1} \bigcup P'|_G$;
    Update population $P|_G$ with $P'|_G \bigcup P|_{G-1}$
end
return $\boldsymbol{V}_P^*|_G$
```

Algorithm 2.1: Basic MOEA (Reynoso-Meza et al., 2017)

2.3.3 Multi-Criteria Decision-Making

Once the optimisation process is completed, a set of solutions is obtained (the Pareto set approximation) that exhibit different levels of trade-off between the multiple objectives. Although all of them are non-dominated solutions not all of them are necessarily interesting for the researcher (even if specific mechanism to ensure pertinency are applied during the optimisation process). It is he/she who has to judge the obtained solutions and select among them that/those that better fit the needs posed by the particular application. This analysis and selection of the preferable solutions among the set of Pareto optimal solutions is what takes place in the Multi-Criteria Decision-Making (MCDM) stage. This process must be conducted by an expert, the decision-maker, that has to understand the peculiarities of the problem at hand and be able to choose the best solutions according to his/her preferences.

In order to make a reasoned decision as much information as possible from the set of solutions is needed. For this purpose, visualisation tools have been proposed that help the decision-maker to understand objective trade-off (Lotov and Miettinen, 2008). These tools are even more necessary when the dimensions of the problem increase, and they can assist the decision-maker to appreciate how changes in one objective affect all the other indexes and the overall performance. In Lotov and Miettinen (2008) the authors review several techniques that can be used to visualise the Pareto front. In this work, Level Diagrams (Blasco et al., 2008; Reynoso-Meza et al., 2013a) are used to visualise Pareto fronts in four-dimensional objective spaces. The basis and utility of this tool are presented in Section 2.4.2.

Apart from visualisation techniques, sometimes specific metrics are defined to categorise the solutions obtained at the Pareto front approximation and guide the decision-making process (Bonissone et al., 2009). In this work a metric is defined to select among the non-dominated solutions those that better approach realistic metabolic phenotypes, by prioritising those objectives related to known metabolic behaviours.

With the MCDM stage the multi-objective optimisation procedure is completed. However, knowledge gained during this stage can be used to improve the problem definition, and/or to aim for an optimisation algorithm that enhances performance in some aspects. The complete procedure is an organic process in which feedback should be taken from each stage to the others to enrich the obtained results.

2.4 Tools for multi-objective optimisation procedures

In this thesis the multi-objective optimisation procedure described in the previous section has been applied to the problem of constraint-based genome-scale metabolic simulations. Chapter 5 is devoted to deeply discuss the technicalities of this application, addressing the contribution of this work to the three stages of the MO optimisation procedure. Much emphasis is put in this work in the multi-objective problem definition (constraints, objectives, preferable features, *etc*.). Next, a MOEA is proposed to perform metabolic simulations by solving the defined statements. For this purpose an existent MOEA is adapted, by including some specific mechanisms that improve its suitability for this problem. Finally, a previously developed visualisation tool and an *ad hoc* metric are used to analyse the obtained solutions and choose those that best represent the metabolic system under study. The present section is devoted to introduce the two pre-existing tools employed: (i) *sp-MODE algorithm*, for the optimisation process, and (ii) *Level Diagrams Tool* for the visualisation of multi-dimensional Pareto fronts.

2.4.1 Multi-objective Differential Evolution with Spherical Pruning (sp-MODE) algorithm

The algorithm proposed in this work for metabolic simulations through multi-objective optimisation is an adaptation of the sp-MODE[8] algorithm described in Reynoso-Meza et al. (2010). sp-MODE is a MOEA based on the Differential Evolution (DE) algorithm which uses a spherical pruning (sp) mechanism to improve distribution of the solutions along the Pareto front approximation. This algorithm has been already used with success for controller design with several performance objectives and robustness requirements (Reynoso-Meza et al., 2013b).

Next, the main features of sp-MODE, *i.e.* the evolutionary technique and the pruning mechanism, as well as the algorithms generated for its implementation, are explained. In Chapter 5, additional mechanisms are presented which are included to the algorithm to adapt it to the problem of metabolic simulation.

[8] Tool available at `http://www.mathworks.com/matlabcentral/fileexchange/39215`

Evolutionary technique

Differential Evolution (DE) (Storn and Price, 1997; Storn, 2008; Das et al., 2016) was selected as evolutionary mechanism due to its simplicity and proved efficacy in several optimisation instances. The most basic form of DE was applied, which makes use of two operators: mutation (Equation (2.11)) and crossover (Equation (2.12)) to generate the offspring of a given population (Algorithm 2.2).

Mutation: For each target (parent) vector $\boldsymbol{v}^i|_G$, a mutant vector $\boldsymbol{y}^i|_G$ is generated at generation G according to Equation (2.11):

$$\boldsymbol{y}^i|_G = \boldsymbol{v}^{r_1}|_G + F(\boldsymbol{v}^{r_2}|_G - \boldsymbol{v}^{r_3}|_G) \tag{2.11}$$

where $r_1 \neq r_2 \neq r_3 \neq i$ are randomly selected; F is usually known as the scaling factor.

Crossover: For each target vector $\boldsymbol{v}^i|_G$ and its mutant vector $\boldsymbol{y}^i|_G$, a trial (child) vector $\boldsymbol{x}^i|_G = \left[x^i_1|_G, x^i_2|_G, \ldots, x^i_n|_G\right]$ is created as follows:

$$x^i_j|_G = \begin{cases} y^i_j|_G & if \quad rand(0,1) \leq Cr \\ v^i_j|_G & \quad otherwise \end{cases} \tag{2.12}$$

where $j \in \{1, \ldots n\}$ and Cr is named the crossover probability rate.

```
Input: population $P|_G$
Output: offspring $O|_G$ of the population
for i=1:SolutionsInParentPopulation do
    Generate a Mutant Vector $\boldsymbol{y}^i$ (Equation (2.11)) ;
    Generate a Child Vector $\boldsymbol{x}^i$ (Equation (2.12)) ;
end
Offspring $O|_G = X$ ;
return $O|_G$
```

Algorithm 2.2: DE offspring generation mechanism

In basic (mono-objective) Differential Evolution the standard selection mechanism is based on the value of the cost function: *a child is selected* *over its parent (for the next generation) if it has a better cost value*. This selection mechanism is usually known as greedy selection. For multi-objective optimisation the selection criterion in the basic DE is changed from "cost value" to "strict dominance" (Definition 2.2): *a child is selected over its parent if the child strictly dominates its parent*.

```
Input: optimisation parameters
Output: Pareto set approximation $\boldsymbol{V}_P^*$
Build initial population $P|_0$ with $N_p$ individuals;
Evaluate $P|_0$;
Set generation counter $G = 0$;
while stopping criterion unsatisfied do
    $G = G + 1$;
    Build offspring $P'|_G$ using $P|_G$ using DE algorithm operators
      (Algorithm 2.2).;
    Evaluate offspring $P'|_G$;
    Update population $P|_G$ with $P|'_G$ and $P|_{G-1}$ using greedy
      selection mechanism with dominance criteria. ;
end
Build Pareto set approximation $\boldsymbol{V}_P^*|_G = P|_G$ ;
return $\boldsymbol{V}_P^*|_G$
```

Algorithm 2.3: Multi-objective Differential Evolution (MODE) algorithm

A pseudocode for the multi-objective differential evolution (MODE) is presented in Algorithm 2.3. This MODE algorithm is intended to approach the Pareto front (convergence), but it might get close to a single solution, since it lacks any mechanism to spread the solutions along the Pareto front approximation (diversity).

Pruning mechanism

In order to promote diversity properties of the previous algorithm, a pruning mechanism was incorporated. The objective of this pruning is

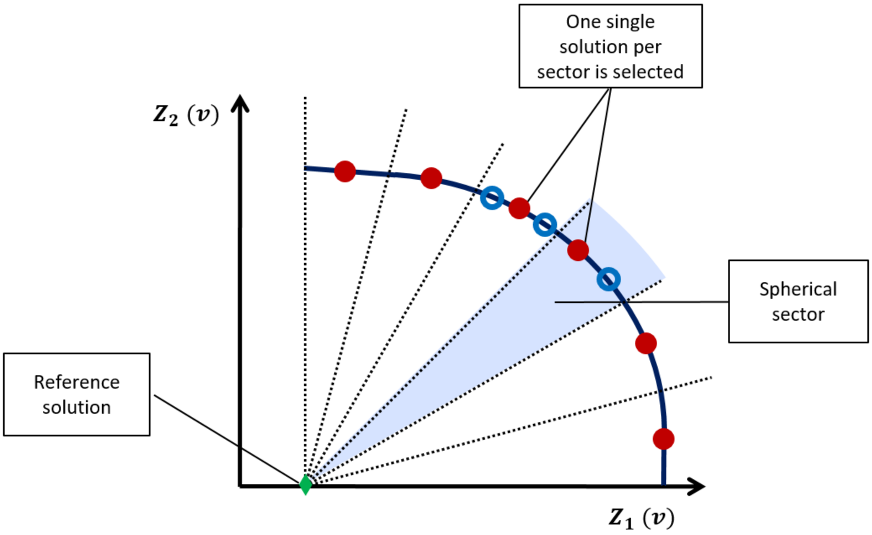


Figure 2.3: Spherical pruning mechanism in a bi-objective space. For each spherical sector, a single solution is selected according to a given metric.

to avoid dense regions in the objective space as well as avoiding missing spots. The implemented pruning mechanism is based on spherical coordinates, normalised with respect to a reference solution. In general terms, what it does is to divide the objective space in spherical sectors, and then select a single solution from each sector according to a given index (Figure 2.3). Before showing the detailed procedure (Algorithm 2.4), some formal definitions are required.

Definition 2.10 (Normalised spherical coordinates (Reynoso-Meza et al., 2010)). *Given a solution $\boldsymbol{v}^1$ and $\boldsymbol{Z}\left(\boldsymbol{v}^1\right)$, let*

$$\boldsymbol{S}(\boldsymbol{Z}\left(\boldsymbol{v}^1\right)) = [\|\boldsymbol{Z}\left(\boldsymbol{v}^1\right)\|_2, \boldsymbol{\beta}(\boldsymbol{Z}\left(\boldsymbol{v}^1\right))] \tag{2.13}$$

be the normalised spherical coordinates from a reference solution $\boldsymbol{Z}^{ref}$ where $\boldsymbol{\beta}(\boldsymbol{Z}\left(\boldsymbol{v}^1\right)) = [\beta_1(\boldsymbol{Z}\left(\boldsymbol{v}^1\right)), \ldots, \beta_{q-1}(\boldsymbol{Z}\left(\boldsymbol{v}^1\right))]$ is the arc vector and $\|\boldsymbol{Z}\left(\boldsymbol{v}^1\right)\|_2$ the Euclidean distance to the reference solution.

The reference solution $\boldsymbol{Z}^{ref}$ must dominate all the solutions existing at that moment. To ensure so, the ideal vector $\boldsymbol{Z}^{ideal}$ (Definition 2.7), or its approximation the utopian vector $\boldsymbol{Z}^{utopia}$ (Definition 2.8), is used as a reference. Note that, since during the evolutionary process the population of solutions is in continuous evolution, the approximation of the Pareto front may vary and so the components of the utopian and nadir vectors have to be assessed at each generation.

Definition 2.11 (Sight range (Reynoso-Meza et al., 2010)). *The sight range from the reference solution $\boldsymbol{Z}^{ref}$ to the Pareto front approximation $\boldsymbol{Z}^*_{\boldsymbol{P}}$ is bounded by $\boldsymbol{\beta}^{\boldsymbol{U}}$ and $\boldsymbol{\beta}^{\boldsymbol{L}}$:*

$$\boldsymbol{\beta}^{\boldsymbol{U}} = \left[\max \beta_1(\boldsymbol{Z}\left(\boldsymbol{v}^i\right), \ldots, \max \beta_{q-1}(\boldsymbol{Z}\left(\boldsymbol{v}^i\right)\right] \quad \forall \boldsymbol{Z}\left(\boldsymbol{v}^i\right) \in \hat{A}|_G \tag{2.14}$$

$$\boldsymbol{\beta}^{\boldsymbol{L}} = \left[\min \beta_1(\boldsymbol{Z}\left(\boldsymbol{v}^i\right), \ldots, \min \beta_{q-1}(\boldsymbol{Z}\left(\boldsymbol{v}^i\right)\right] \quad \forall \boldsymbol{Z}\left(\boldsymbol{v}^i\right) \in \hat{A}|_G \tag{2.15}$$

If $\boldsymbol{Z}^{ref} = \boldsymbol{Z}^{ideal}$, it is straightforward to prove that $\boldsymbol{\beta}^{\boldsymbol{U}} = \left[\frac{\pi}{2}, \ldots, \frac{\pi}{2}\right]$ and $\boldsymbol{\beta}^{\boldsymbol{L}} = [0, \ldots, 0]$.

The set A represents an external archive in which the best set of solutions found so far during the evolutionary process are stored. $\hat{A}|_G$ represents the set of archived solutions at generation G before pruning and $A|_G$ is the same set after pruning.

Definition 2.12 (Spherical grid (Reynoso-Meza et al., 2010)). *Given a set of solutions in the objective space, the spherical grid on the q-dimensional space in arc increments $\boldsymbol{\beta_\epsilon} = [\beta_1^\epsilon, \ldots, \beta_{q-1}^\epsilon]$ is defined as:*

$$\boldsymbol{\Lambda}^{\boldsymbol{Z}^*_P} = \left[\frac{\beta_1^U - \beta_1^L}{\beta_1^\epsilon}, \ldots, \frac{\beta_{q-1}^U - \beta_{q-1}^L}{\beta_{q-1}^\epsilon}\right] \tag{2.16}$$

Definition 2.13 (Spherical sector (Reynoso-Meza et al., 2010)). *The normalised spherical sector of a solution $\boldsymbol{v}^1$ is defined as:*

$$\boldsymbol{\Lambda_\epsilon}(\boldsymbol{v^1}) = \left[\left\lceil \frac{\beta_1(\boldsymbol{Z}\left(\boldsymbol{v}^1\right))}{\Lambda_1^{Z^*_P}} \right\rceil, \ldots, \left\lceil \frac{\beta_{q-1}(\boldsymbol{Z}\left(\boldsymbol{v}^1\right))}{\Lambda_{q-1}^{Z^*_P}} \right\rceil\right] \tag{2.17}$$

Definition 2.14 (Spherical pruning (Reynoso-Meza et al., 2010)). *Given two solutions $\boldsymbol{v}^1$ and $\boldsymbol{v}^2$ from a set, $\boldsymbol{v}^1$ has preference in the spherical sector over $\boldsymbol{v}^2$ iff:*

$$\left[\boldsymbol{\Lambda_\epsilon}(\boldsymbol{v^1}) = \boldsymbol{\Lambda_\epsilon}(\boldsymbol{v^2})\right] \wedge \left[\|\boldsymbol{Z}\left(\boldsymbol{v}^1\right)\|_p < \|\boldsymbol{Z}\left(\boldsymbol{v}^2\right)\|_p\right] \tag{2.18}$$

where $\|\boldsymbol{Z}(\boldsymbol{v})\|_p = \left(\sum\limits_{a=1}^{q} |Z_a(\boldsymbol{v})|^p\right)^{1/p}$ is a suitable p-norm.

```
Input: Archive of best solutions before pruning $\hat{A}|_G$
Output: Archive of best solutions after pruning $A|_G$
Read archive $\hat{A}|_G$;
Read and update extreme values for $Z^{ref}|_G$;
for each member in $\hat{A}|_G$ do
    calculate its normalised spherical coordinates (Definition
      2.10);
end
Build the spherical grid (Definitions 2.11 and 2.12);
for each member in $\hat{A}|_G$ do
    calculate its spherical sector (Definition 2.13);
end
for i=1:SolutionsInArchive do
    Compare with the remainder solutions in $\hat{A}|_G$;
    if no other solution has the same spherical sector then
        it goes to archive $A|_G$;
    end
    if other solutions are in the same spherical sector then
        it goes to archive $A|_G$ if it has the lowest norm (Definition
          2.14);
    end
end
return $A|_G$
```

Algorithm 2.4: Spherical pruning mechanism

The complete implementation combining the DE algorithm for MO (Algorithm 2.3) and the spherical pruning mechanism (Algorithm 2.4) was named as sp-MODE algorithm (see Algorithm 2.5) (Reynoso-Meza et al., 2010). A MATLAB®-based tool is available for its free download at MATLAB® Central[9].

[9] `http://www.mathworks.com/matlabcentral/fileexchange/39215`

```
**Input:** optimisation parameters
**Output:** Pareto set approximation $\boldsymbol{V}_P^*$
Build initial population $P|_0$ with $N_p$ individuals;
Evaluate $P|_0$;
Apply dominance criterion (Definition 2.1) on $P|_0$ to get $\hat{A}|_0$;
Apply pruning mechanism (Algorithm 2.4) to prune $\hat{A}|_0$ to get $A|_0$;
Set generation counter $G = 0$;
**while** *stopping criterion unsatisfied* **do**
    $G = G + 1$;
    Get subpopulation $S|_G$ with solutions in $P|_{G-1}$ and $A|_{G-1}$;
    Generate offspring $O|_G$ with $S|_G$ using DE operators (Algorithm 2.2).;
    Evaluate offspring $O|_G$;
    Update population $P|_G$ with offspring $O|_G$ according to greedy selection mechanism.;
    Apply dominance criterion (Definition 2.1) on $O|_G \bigcup A|_{G-1}$ to get $\hat{A}|_G$;
    Apply pruning mechanism (Algorithm 2.4) to prune $\hat{A}|_G$ to get $A|_G$;
**end**
$\boldsymbol{V}_P^* = A|_G$;
**return** $\boldsymbol{V}_P^*|_G$
```

Algorithm 2.5: sp-MODE

2.4.2 Level Diagrams visualisation tool

In this work Level Diagrams[10] (Blasco et al., 2008; Reynoso-Meza et al., 2013a) are used to visualise four-dimensional Pareto Fronts.

The Level Diagrams visualisation is based on the use of an auxiliary variable that has the same value for all the objectives. This way, a two-dimensional plot can be created for each objective and each decision variable, and a given solution will display the same ordinate at all of them. This allows traceability of the solutions in the different objective sub-spaces.

The auxiliary variable is the distance, according to a preferred norm, to an ideal solution, represented by the utopian objective vector (Definition 2.8). To evaluate this distance, first each objective is normalised with respect to its minimum and maximum values in entire the Pareto front, as shown in Equation (2.19).

$$\hat{Z}_k\left(\boldsymbol{v}\right) = \frac{Z_k\left(\boldsymbol{v}\right) - Z^{k,min}}{Z^{k,max} - Z^{k,min}} \qquad k \in \{1, 2, \ldots, q\} \tag{2.19}$$

The minimum and maximum values for each objective can be extracted from the components of the utopian and the nadir objective vectors (Definitions 2.8 and 2.9) respectively, which leads to Equation (2.20).

$$\hat{Z}_k\left(\boldsymbol{v}\right) = \frac{Z_k\left(\boldsymbol{v}\right) - Z_k^{utopia}}{Z_k^{nadir} - Z_k^{utopia}} \qquad k \in \{1, 2, \ldots, q\} \tag{2.20}$$

Applying this transformation, a normalised objective vector $\hat{\boldsymbol{Z}}\left(\boldsymbol{v}\right) = \left[\hat{Z}_1\left(\boldsymbol{v}\right), \hat{Z}_2\left(\boldsymbol{v}\right), \ldots, \hat{Z}_q\left(\boldsymbol{v}\right)\right]$ is obtained. Then, a p-norm $\|\hat{\boldsymbol{Z}}\left(\boldsymbol{v}\right)\|_p$ is employed to calculate the distance to the utopian objective vector (Definition 2.8).

Norms commonly used are the 1-norm , the 2-norm and the ∞-norm. Given the previous normalisation, distance to the utopian vector can be

[10] Tool available at `https://www.mathworks.com/matlabcentral/fileexchange/62224`

calculated according to these norms as stated in Equations (2.21), (2.22) and (2.23) respectively.

$$\|\hat{\boldsymbol{Z}}(\boldsymbol{v})\|_1 = \sum_{k=1}^{q} \hat{Z}_k(\boldsymbol{v}) \tag{2.21}$$

$$\|\hat{\boldsymbol{Z}}(\boldsymbol{v})\|_2 = \sum_{k=1}^{q} \sqrt{\left(\hat{Z}_k(\boldsymbol{v})\right)^2} \tag{2.22}$$

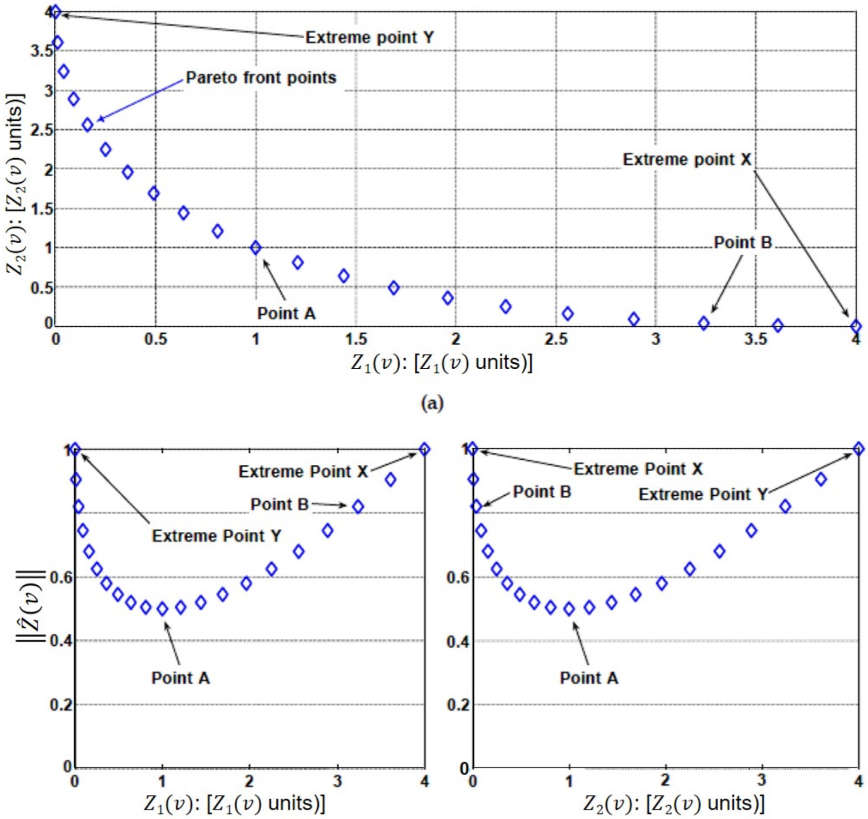
$$\|\hat{\boldsymbol{Z}}(\boldsymbol{v})\|_\infty = \max_{\boldsymbol{v}} |\hat{Z}_k(\boldsymbol{v})| \tag{2.23}$$

Finally, the distance calculated this way is used as the auxiliary variable used to generate two types of graphs: at the *objective sub-graphs* (one for each objective, so q) the ordered pairs $\left(Z_k(\boldsymbol{v}), \|\hat{\boldsymbol{Z}}(\boldsymbol{v})\|_p\right)$ are plotted, and at the *decision variable sub-graphs* (one for each decision variable, so n), the ordered pairs $\left(v_j, \|\hat{\boldsymbol{Z}}(\boldsymbol{v})\|_p\right)$ are shown.

Figure 2.4 shows an example of application of Level Diagrams representation to a bi-objective problem. In this low-dimensional problem is possible to compare the 2-D objective space with the plots generated by means of LD. It can be noticed that, even in this simple case, the distance to the utopian solution provides an added degree of information. This figure also shows how the use of a standard auxiliary variable allows traceability of the solutions along the different objective sub-spaces.

Whatever the selected norm is, the lower the value is, the lower the distance to the (ideal) utopian solution. However, this does not mean that the solution closer to the "ideal" is the best solution from the point of view of the decision-maker (or for the particular problem at hand). That solution could present, for example, an unacceptable degradation in some of the objectives.

Extra metrics that represent additional criteria can be used to colour the solutions presented at the Level Diagrams, which allows the integration of more degrees of information at the same display.



(a)

(b)

Figure 2.4: *Example of representation of the Pareto front for a bi-objective problems using 2-D graph (a) and LD (b). Points at the same level in LD correspond on each graphic. Figure 3 from (Reynoso-Meza et al., 2013a).*

Part II

Works on metabolic modelling of cyanobacteria

3

Metabolic networks of model cyanobacteria

3.1 Chapter abstract

In this chapter work is done in curation and assembly of the metabolic networks of two model cyanobacteria: *Synechococcus elongatus* PCC 7942 and *Synechocystis* sp. PCC 6803.

The metabolic model presented in this chapter for *Synechococcus* was the first genome-scale network assembled for this organism. Following the four-stages process described in Section 1.2.1 (preliminary automated assembly, manual refinement and curation, conversion to a mathematical model, and network evaluation and debugging) a curated genome-scale metabolic model was obtained that allows accurate metabolic simulations of this cyanobacterium.

Additionally, work has been done on updating *Synechocystis*' network from published works: new knowledge has been incorporated and some pathways (such as those involving electron transport and inorganic nutrient assimilation) have been described in greater detail. The implemented changes have led to noticeable simulation improvements

(such as greater plasticity, greater accuracy at electron consuming pathways or more realistic energy needs) that allow more precise metabolic characterisation of the cyanobacterial phenotypes and enhance the predictability of the model.

The metabolic networks of *Synehcocystis* and *Synechococcus* present many common traits. At the end of the chapter a comparison is presented in terms of network features (number of genes, reactions and metabolites), main metabolic pathways present (with special attention to their characteristic electron transport pathways), and network topology and metabolites' connectivity.

Parts of the contents of this chapter are based on the following peer-reviewed article[11]:

- Julián Triana, Arnau Montagud, Maria Siurana, David Fuente, Arantxa Urchueguía, Daniel Gamermann, Javier Torres, Jose Tena, Pedro Fernández de Córdoba, Javier Urchueguía. **Generation and Evaluation of a Genome-Scale Metabolic Network Model of *Synechococcus elongatus* PCC 7942.** *Metabolites* 2014, **4**:680-698

3.2 Introduction

In the current energy, environmental and socio-economic situation, it exists an increasing trend towards the search of efficient and sustainable manufacturing processes. Within this context, the use of biological systems as cell factories for the production of fuels, drugs and other chemicals of industrial interest has been identified as an alternative to traditional processes. However, the use of living organisms for biotechnological processes is not a new topic at all. Bio-manufacturing has been applied since several thousand years ago to obtain traditional products like beer, cheese, wine or bread (Zhang et al., 2016). However, the modern approach is based on rational design procedures: re-

[11] The research leading to the reconstruction of *Synechococcus elongatus* PCC 7942 metabolic network has been done in close collaboration with Julián Triana.

search is driven to gain better understanding of the systems and logically modifying them to obtain the desired outcome.

When facing the development of bioprocesses based on the use of living systems as cell factories, it must be taken into account that biological systems are tuned to optimise their own progress in their natural habitat (Patil et al., 2004). That is why to reach development of competitive processes, regarding cost and performance, genetic modifications are needed to adapt the capacity of the host organism to the productive necessities and to the process conditions, which is known as metabolic engineering (Nielsen, 1997).

However, due to the complexity of cellular systems, in which metabolites, genes, and proteins are interconnected through complex networks (Furusawa et al., 2012), the introduced modifications may produce unexpected effects. Thus, to address the rational design of systems of production based on living organisms, proper understanding of the base cellular metabolism is crucial. This understanding must be reached through a global perspective that allows untangling interaction among the different components (Patil et al., 2004). Genome-scale metabolic models, as introduced in Chapter 1, are valuable tools to achieve such a system-level comprehension.

The four-stage process to reconstruct a genome-scale metabolic network has been reviewed in Chapter 1. This process starts with the (usually automatic) identification of metabolic functions from annotated genome of the organism under study, followed by manual refinement of the draft network, continues with conversion of the curated network into a mathematical model, and can be considered complete after validation, when a functional and verified mathematical representation of all biochemical transformations occurring within the system is achieved (Feist et al., 2009; Thiele and Palsson, 2010). However, this process is an iterative procedure in which continuous curation, maintenance and update of the model is required to extend and enhance its applicability.

Different sources of information can be used for curation and update of a metabolic reconstruction. Biochemical databases containing information about reactions, genes, proteins, enzymes and molecules involved in metabolic processes, are widely used for this purpose. KEGG (Kanehisa and Goto, 2000; Kanehisa et al., 2016, 2017), BRENDA (Schomburg et al., 2000; Placzek et al., 2017), or ChEBI (Hastings et al., 2013) are among the most popular. Information can also be gathered from available literature (reference books and journal articles), physiological experimental datasets or genomic and phylogenetic data.

In this thesis, modelling techniques are applied to the study of the metabolism of two model cyanobacteria. Cyanobacteria are photosynthetic bacteria that can perform oxygenic photosynthesis making use of solar energy and inorganic CO_2 to grow and perform cellular functions. This property causes cyanobacteria to be in the spotlight as very interesting host organisms for the development of solar-fuelled cell factories. In particular, *Synechocystis* sp. PCC 6803 and *Synechococcus elongatus* PCC 7942 are two of the most studied species, they are naturally transformable[12] and several molecular tools have been developed for their genetic manipulation (Hess, 2011; Pinto et al., 2015; Kim et al., 2017). Altogether, these properties have heighten the interest of these species as potential cell factories and have motivated the necessity of studying their metabolic functionalities.

This chapter is dedicated to address the update of *Synechocystis*' and first assembly of *Synechococcus*' metabolic models. The resulting networks are then presented and, at the end of the chapter, they are compared between them.

[12] They have the natural ability to take up and incorporate exogenous genetic material.

3.3 *i*Syn842: updated metabolic network of *Synechocystis* sp. PCC 6803

3.3.1 Previous version: *i*Syn811

In 2010, our research group published the first version of the *i*Syn model, *i*Syn669 (Montagud et al., 2010), that was later updated to *i*Syn811 (Montagud et al., 2011).

One of the most important contributions of these models was the definition of a genome-scale biomass equation accounting for 39 biomass components and precursors (Table 3.1). As explained in Chapter 1, a proper genome-scale biomass equation must account for all cell building blocks, as well as for the energy consumed during the assembly process, and describe, as accurately as possible, the proportions among them (Feist and Palsson, 2010). This equation is needed in almost all constraint-based methods to represent cell growth. It is usually employed as the biological objective function when performing mono-objective simulations. If a different objective is defined, generally, the biomass function has to be used as a constraint, in order to consider energy and mass associated drains. The biomass equation of reference (Montagud et al., 2011) has been conserved in the current version of the model.

More than five years have passed since the publication of *i*Syn811, and some improvements have been done on the model in order to keep it updated. New annotation information, new physiological discoveries, and the application of the model to new studies have motivated this continuous update. In the following section, some of the main changes are reported.

3.3.2 Update process

For the construction of a detailed, up-to-date and accurate genome-scale metabolic model of *Synechocystis* sp. PCC 6803 the departing point

Table 3.1: *Biomass formulation in iSyn811 (Montagud et al., 2011) and iSyn842. All units are mmole per gram of dry cell weight* ($mmol/g_{DC}$)
.

Metabolites	Coefficients	Metabolites	Coefficients
Aminoacids			
Alanine	0.499149	Leucine	0.437778
Arginine	0.28742	Lysine	0.333448
Aspartate	0.234232	Methionine	0.149336
Asparagine	0.234232	Phenylalanine	0.180021
Cysteine	0.088988	Proline	0.214798
Glutamine	0.255712	Serine	0.209684
Glutamate	0.255712	Threonine	0.246506
Glycine	0.595297	Tryptophan	0.055234
Histidine	0.092056	Tyrosine	0.133993
Isoleucine	0.282306	Valine	0.411184
Deoxyribonucleotides		**Ribonucleotides**	
dATP	0.0241506	AMP	0.14038929
dTTP	0.0241506	UMP	0.14038929
dGTP	0.02172983	GMP	0.12374585
dCTP	0.02172983	CMP	0.12374585
Lipids			
16C-lipid	0.20683718	(9Z,12Z)18C-lipid	0.03568367
(9Z)16C-lipid	0.01573412	(9Z,12Z,15Z)18C-lipid	0.01797109
18C-lipid	0.00351776	(6Z,9Z,12Z)18C-lipid	0.05031906
(9Z)18C-lipid	0.03188596	(6Z,9Z,12Z,15Z)18C-lipid	0.01448179
Carbohydrates		**Antenna chromophores**	
Glycogen	1.171.827	Chlorophyll a	0.02728183
		Carotenoids	0.00820225

was the model *i*Syn811 (Montagud et al., 2011) on which various modifications were implemented leading to noticeable simulation improvements.

Electron transport chain

Cyanobacterial energy metabolism is one of their most characteristic traits, since they form complex electron transport chains that combine photosynthesis and oxidative phosphorylation in the same compartment, the thylakoid (Vermaas, 2001). While photosynthesis takes place only in thylakoid membranes, oxydative phosphorylation appears both in thylakoid and cytoplasmic membranes. Thus, in the thylakoid both electron transport chains are present and share elements, leading to connections between photosynthetic and respiratory electron flows (Figure 3.1).

Besides, *Synechocystis* has a hydrogenase which contributes as a transient electron sink during light-regime changes (Khanna and Lindblad, 2015). This fact has also strongly attracted the interest over this organism, since it is an appealing candidate for hydrogen production mediated by water splitting in photosynthesis.

The electron transport chain described in the previous version of the model was already a detailed set of reactions (in contrast with lumped reactions used in central carbon metabolic models) including the main complexes depicted in Figure 3.1. However, new information became available since *i*Syn811 model was published. For instance, new reactions were incorporated, such as those corresponding to the use of cytochrome cM as a soluble electron transporter (from plastocyanin to terminal cytochrome oxidases) and the Mehler reaction (oxygen photoreduction directly from NADPH). Also soluble quinol transporter was changed from ubiquinol to plastoquinol, whose synthesis was added to the set of reactions. Finally, proton pumping stoichiometry was reviewed and updated. The energetic needs, especially under autotrophic conditions, are now more precise, and the model accounts for more detailed reactions at the electron transport chain.

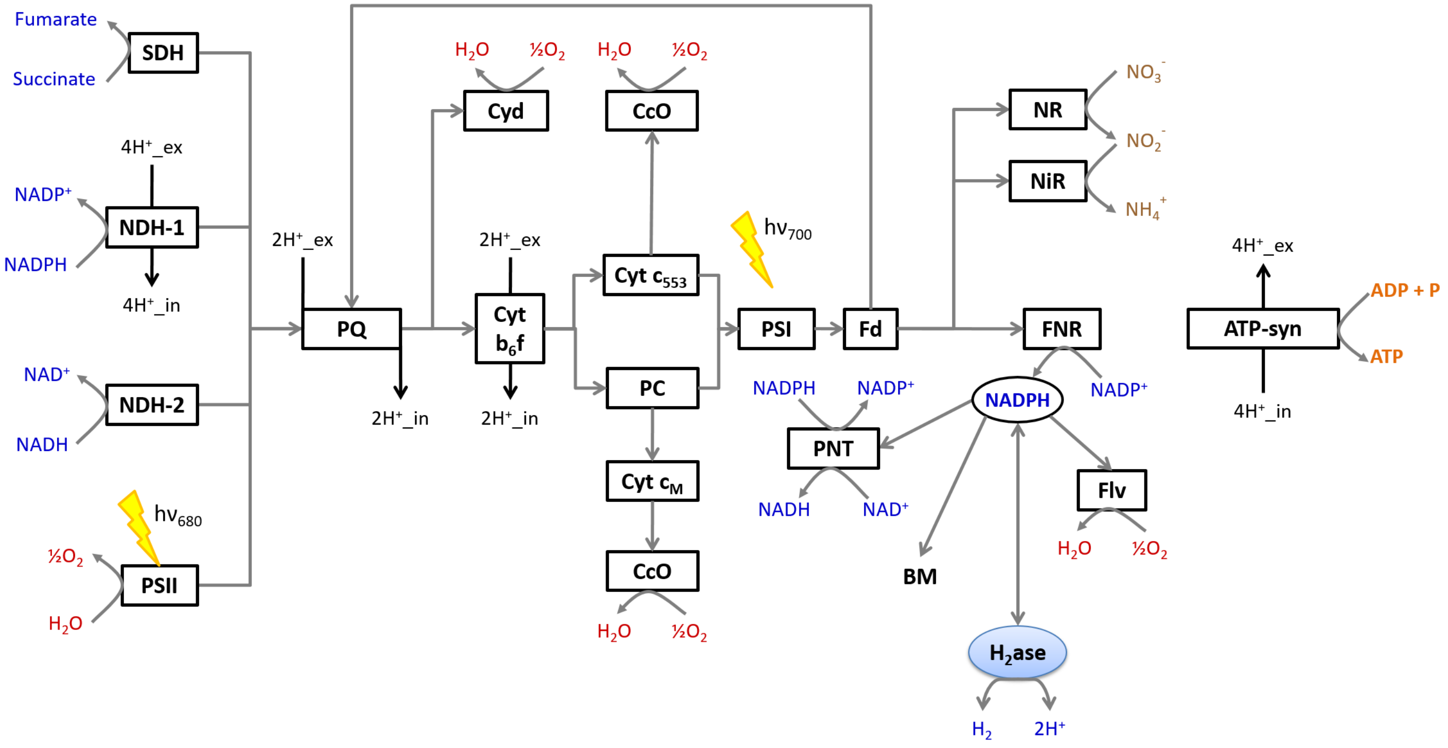


Figure 3.1: *Outline of electron transport chain in Synechocystis' thylakoid membranes. Abbreviations used - SDH: succinate deshydrogenase; NDH-1: type I NADPH deshydrogenase; NDH-2: type II NADH deshydrogenase; PSII: photosystem II; PQ: plastoquinone; Cyd: cytochrome bd-oxidase; Cyt b6f: cytochrome b6f; Cyt c553: cytochrome c553; PC: plastocyanin; PSI: photosystem I; Fd: ferredoxin; FNR: ferredoxin-NADP+ reductase; NR: nitrate-reductase; NiR: nitrite-reductase; Cit cM: cytochrome cM; CcO: cytochrome c-oxidase; Flv: flavoprotein; PNT: pyridin-nucleotide transhydrogenase; BM: biomass; H2ase: hydrogenase; ATP-syn: ATP synthase;*

Electron transport carriers

Other electron carriers, such as NAD(P)H, ferredoxin and thioredoxin, are also crucial when studying energy metabolism and biochemical production. In this version of the model, the reactions in which they are implied have been improved by reviewing stoichiometric inconsistencies inherited from public databases and adding new reactions that allow electron exchanges between them. In this process more than 120 reactions have been modified or added to the model. After these changes, several constraints, previously needed to avoid futile cycles, were found to be of no use and were eliminated leading to a more flexible model able to better adapt to different environmental conditions, as natural organisms do.

Inorganic nutrient assimilation pathways

Various authors have shown also interest in inorganic nutrient assimilation pathways, especially nitrogen and sulphur (Burrows et al., 2009; Takahashi et al., 2011), which imply electron consumption and thus compete with other pathways, like hydrogen or secondary metabolites production. In these pathways of inorganic elements incorporation, the electron stoichiometry described in the network was reviewed and corrected and new spontaneous reactions were added. Also transport reactions that allow alternative nutrient sources were included. In Chapter 4, these alternatives are investigated as interesting nutrient sources in terms of hydrogen production (Section 4.4.3).

Global *i*Syn842 inspection

Apart from particular pathways and metabolic hubs, the network was globally examined to assess some properties, like network connectivity; and, lastly, annotation updates were considered.

In order to detect reactions fully disconnected from the network and metabolic dead-ends (*i.e.* metabolites that only appear once in the whole network), a set of functions from *PyNetMet* (Gamermann et al., 2014b) (see Section 1.2.3) specifically designed for this purpose were used. The importance of detecting such reactions is that they are blocked under steady state condition: there will never be flux through them and they do not contribute to the flux map. The appearance of these disconnected reactions can be due to a lack of information at the time of the reconstruction, or to differences in nomenclature. Using this update process, about a hundred disconnected reactions were detected and were updated, corrected or deleted depending on each case.

Information from public databases of biochemical knowledge, like for example KEGG (Kanehisa and Goto, 2000; Kanehisa et al., 2016, 2017), was used as well to update names of enzymes that had changed, new gene-enzyme-reaction assignments, or stoichiometric differences that had been unravelled during the last years. This way, around one hundred enzymes and reactions were updated and included if necessary.

Finally, experimental data obtained from collaborator' research groups were used to check for physiological evidence of many of the reported reactions. For instance, a wide proteomics analysis from the *Chemical Engineering at the Life Science Interface* from The University of Sheffield was used to identify enzymes detected and their EC numbers and reactions associated (Andrew Landels, personal communication). This information has been used to identify blind-spots and contradictory records in the model by double-checking the list of enzymes detected during the proteomic analyses and present in the metabolic model.

3.3.3 Resulting *i*Syn842 network

The result of this tuning process is a model that consists of 1059 metabolic reactions and 920 metabolites (see Table 3.2 for a comparative with previous versions). It accounts for 540 enzymes performing 804 different reactions. It also includes 53 transport reactions allowing the intake of inorganic nutrients like nitrogen, sulphur, phosphate and some metals, and carbon substrates like sugars, aminoacids or inorganic carbon compounds, the transport of gases, like oxygen or hydrogen, or the

Table 3.2: Comparison of the network features of several versions of iSyn metabolic model of Synechocystis sp. PCC 6803: iSyn669 (Montagud et al., 2010), iSyn811 (Montagud et al., 2011) and iSyn842 (present work).

Network features	*i*Syn669	*i*Syn811	*i*Syn842
genes	669	811	842
reactions	882	956	1059
enzymatic	591	769	804
transport	20	46	53
electron transport chain	21	21	25
metabolites	790	911	920

input of photons. It further incorporates several spontaneous chemical conversions and a few artificial reactions needed for biomass simulations (such as secretion of biomass components). It is equipped with a detailed biomass equation (Table 3.1) which takes into account amino acids, nucleic acids, lipids, carbohydrates, ribonucleotides, deoxyribonucleotides, and antenna chromophores. Additional file 3.1 (see page 249) contains a complete description of the resulting *i*Syn842 model.

As reported above, compared with the previous version of the model, important improvements in the simulations have been achieved as a consequence of all the implemented changes: greater plasticity, greater accuracy at electron consuming pathways, and more realistic energy needs. These improvements allow for more precise metabolic characterisation of the cyanobacterial phenotypes and improve the predictability of the model. Using this model different simulations of cellular growth and production of different substances, can be performed. Different mutants can be tested by easily adding or removing selected reactions, and several environmental conditions can be assessed by changing simulation parameters and media-related constraints. Thus, it can be used for fluxomic and metabolic characterisation of different strains and conditions (an example can be seen in Section 4.4.3).

3.4 *i*Syf715: first metabolic network assembled for *Synechococcus elongatus* PCC 7942

3.4.1 Assembly process

The process followed to reconstruct *Synechococcus*' metabolic model was conducted following the four-stage process described in the previous part of the present dissertation (see Figure 1.2).

The genome sequence and annotation were gathered from NCBI Entrez Gene database (NCBI Resource Coordinators, 2016). For the first automated assembly two software were applied, Pathway Tools (Karp et al., 2002, 2016) and COPABI (Reyes et al., 2012), which allowed double checking the resulting automated reconstructions. Specific functions for gap filling and duplicate check from COPABI were applied. After this first stage of automated generation of components, a draft reconstruction was obtained that accounted for 540 enzymes encoded by 672 genes, and included 898 reactions. Using the gap filling function, incomplete pathways were completed based on probabilistic criteria of unicity and completeness (for details of this process please check reference Reyes et al. (2012)).

Next the draft reconstruction was examined and curated using the available sources of information (databases, literature, experiments, *etc*..). It was at this stage of the process of manual refinement where the PhD. applicant mostly contributed. First, the list of reactions added automatically to fill the gaps in incomplete pathways was manually checked, distinguishing between indispensable reactions and those that were not. Those reactions that were essential were maintained (or added in some cases), and the most appropriate enzyme and stoichiometry was chosen for them based on phylogenetic proximity and sequence comparison with cognate genes performing the same function. Non-enzymatic (spontaneous) reactions reported in *Synechococcus*' metabolism were added as well. Also, EC numbers and stoichiometry assigned to the whole set of reactions were verified with databases such as KEGG

(Kanehisa and Goto, 2000; Kanehisa et al., 2016, 2017), BRENDA (Schomburg et al., 2000; Placzek et al., 2017) and MetaCyc (Caspi et al., 2016) and reviewing state-of-the-art literature. Reaction reversibility was also verified, and when no conclusive irreversibility evidence was reported, reactions were set to be reversible. Finally, unspecific metabolite names obtained from the databases (such as "alcohol") were converted to the corresponding organism-specific metabolites (Thiele and Palsson, 2010).

Next stage of the reconstruction process was to convert the network to a mathematical model, by adding the biomass equation and system parameters. The biomass equation considered amino acid, carbohydrates, chromophores, nucleic acids, and lipids as building blocks for biomass assembly. Table 3.3 shows the biomass composition considered in *i*Syf715. As parameters, flux bounds were added to transport reactions carrying nutrients such as phosphate, water, sulphate, nitrate, ammonia, as well as carbon monoxide and hydrogen peroxide.

The final stage of genome-scale network reconstruction consists in network evaluation and model validation. During this process, biomass formation was verified and obtained flux distributions were verified to check if they used the expected set of metabolic pathways. In this validation process some reactions were corrected. Also, some of the reversible reactions involving NADH and NADPH were constrained to be irreversible so that spurious trans-hydrogenation was controlled, and internal loops thermodynamically unfeasible were removed (Thiele and Palsson, 2010).

3.4.2 Resulting *i*Syf715 network

The resulting network, *i*Syf715, accounted for 715 genes, which encoded 530 enzymes and 23 transporter proteins. It included 838 metabolites and 902 reactions, among which 710 were enzymatic conversions (Table 3.4). Additionally, a set of reactions with no cognate genes was added on the basis of biochemical evidence or physiological considerations: 13 non-enzymatic conversions, 16 passive transport reactions, and 76 non-annotated reactions.

Table 3.3: *Biomass formulation in iSyf715. All units are mmole per gram of dry cell weight* ($mmol/g_{DC}$)

Metabolites	Coefficients	Metabolites	Coefficients
Amino acids			
Alanine	0.897	Leucine	1.128
Arginine	0.526	Lysine	0.417
Aspartate	0.518	Methionine	0.194
Asparagine	0.374	Phenylalanine	0.406
Cysteine	0.102	Proline	0.512
Glutamine	0.576	Serine	0.548
Glutamate	0.614	Threonine	0.580
Glycine	0.702	Tryptophan	0.149
Histidine	0.197	Tyrosine	0.294
Isoleucine	0.628	Valine	0.638
Deoxyribonucleotides		**Ribonucleotides**	
dATP	0.0201156	AMP	0.140389293
dTTP	0.0201156	UMP	0.140389293
dGTP	0.02538445	GMP	0.123745851
dCTP	0.02538445	CMP	0.123745851
Lipids		**Antenna chromophores**	
14C-lipid	0.028	Zeaxanthin	0.00079
16C-lipid	0.0042	Beta-carotene	0.000875
18C-lipid	0.00448	Trans-lycopene	0.00820225
(9Z)16C-lipid	0.0066	Chlorophyll a	0.0057
(9Z)18C-lipid	0.00625		
Carbohydrates			
Glycogen	1.171.827		

Table 3.4: *Network features of iSyf715 metabolic model of Synechococcus elongatus PCC 7942.*

Network features	***i*Syf715**
genes	715
reactions	902
enzymatic	710
transport	41
electron transport chain	21
metabolites	838

The final model included central metabolic pathways such as the glycolysis/gluconeogenesis pathway, the Calvin-Benson-Bassham cycle, the pentose phosphate pathway, incomplete reactions within the citric acid cycle, as well as anabolic pathways involved in the biosynthesis of amino acids, nucleotides, lipids, chlorophyll, glycogen, vitamins, cofactors, and other secondary metabolites. Photosynthetic and respiratory electron transport chains were represented by a set of 21 reactions, including light captured by photosystem II and photosystem I, intermediate electron transporter complexes, cyclic electron transfer, final oxidases, and a bidirectional hydrogenase. Additional file 3.2 (see page 249) contains a complete description of the resulting *i*Syf715 model.

3.5 *Synechocystis*' *vs.* *Synechococcus*' metabolic networks comparison

In previous sections, the process of update and reaction assembly have been described, leading to the construction of *i*Syn842 and *i*Syf715, metabolic models of *Synechocystis* sp. PCC 6803 and *Synechococcus elongatus* PCC 7942, respectively. In the present section, qualitative properties of these networks, regarding network components and topology are assessed and compared. In the following chapter flux analyses are conducted that help further characterise, quantitatively, main traits of *Synechocystis*' and *Synechococcus*' metabolic networks.

Table 3.5: *Comparison of iSyn842 and iSyf715 network features.*

Network features	***i*Syn842**	***i*Syf715**
genes	842	715
reactions	1059	902
enzymatic	804	710
transport	53	41
electron transport chain	25	21
metabolites	920	838

The main features of the *i*Syn842 and *i*Syf715 networks are collected in Table 3.5. *Synechocystis*' reconstructed network includes a greater number of genes, reactions and metabolites than *Synechococcus*', but before any consideration on difference on the reconstruction are made, it must be taken into account that those organisms have quite different numbers of annotated genes: *Synechocystis* sp. PCC 6803 has 3564 while for *Synechococcus elongatus* PCC 7942 has 2661 genes leading to proteins (numbers retrieved from KEGG database, Kanehisa and Goto (2000), on May 19th, 2017).

Synechocystis sp. PCC 6803 and *Synechococcus elongatus* PCC 7942 are two cyanobacteria closely phylogenetically related (Gamermann et al., 2014a) that have many traits in common. The core of their metabolic networks, as it happens with most cyanobacteria (Baroukh et al., 2015), is similar. In general terms, their metabolic networks can be schematically decomposed as:

- photosynthesis, to produce energy from light;
- Calvin-Benson-Bassham cycle, to assimilate inorganic carbon;
- glycolysis, to produce energy from glucose and generate precursor metabolites;
- the citric acid cycle, to produce other precursor metabolites;

- the pentose phosphate pathway, to produce reductive power and precursor metabolites;
- oxidative phosphorylation (also termed cellular respiration), to produce energy from reduced species;
- inorganic nitrogen and sulphur assimilation;
- carbohydrate and lipid synthesis, to build cell components and store carbon; and
- synthesis (and degradation) of proteins, DNA, RNA, cofactors, chlorophyll and other secondary metabolites from inorganic nutrients and precursor metabolites.

Among those pathways, their energy metabolism must be highlighted since it is distinctive of cyanobacteria. As it has been mentioned, both *Synechocystis* sp. PCC 6803 and *Synechococcus elongatus* PCC 7942 combine elements from the photosynthetic and respiratory electron transport chains at the thylakoid membranes, which gives place to many cross-talks among pathways (Nogales et al., 2012) that are used by the cell to regulate its energetic state. Among these alternative pathways, both organisms have a bidirectional hydrogenase which make them very interesting candidate organisms for the design of photosynthetic cell factories for sustainable production of hydrogen.

However, as it was already noted in Section 1.3, there is an important difference between these two organisms: while *Synechocystis* sp. PCC 6803 is able to grow both under autotrophic and heterotrophic conditions (as well as under the intermediate mixotrophic conditions), *Synechococcus elongatus* PCC 7942 is an obligate autotroph that depends on light and carbon dioxide to survive. This difference is also reflected on the metabolic models obtained in this chapter, being the main reason the absence of sugar transporters in *Synechococcus*' network, which agrees with experimental observations (McEwen et al., 2013).

Regarding the topology of the networks, both *i*Syn842 and *i*Syf715 exhibit the same pattern ubiquitously present in cellular networks, in

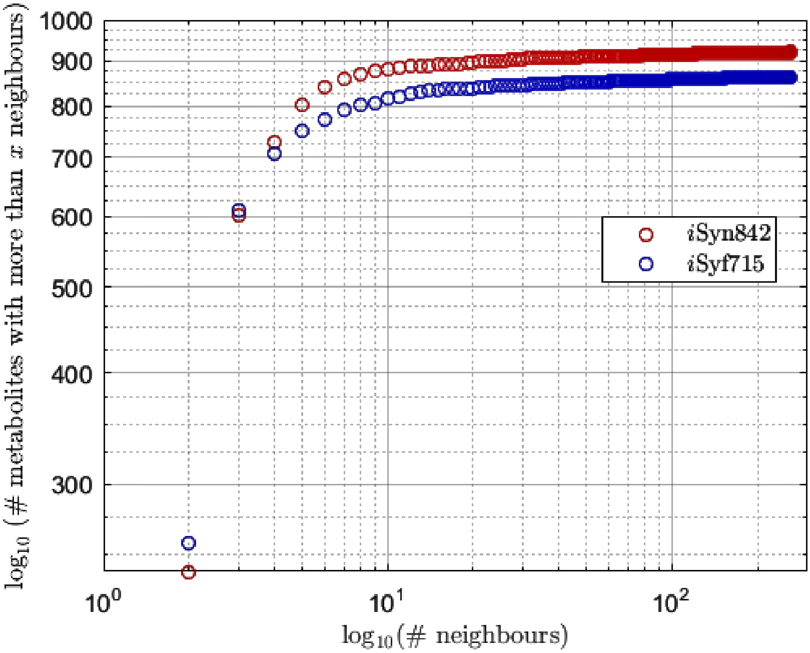


Figure 3.2: *Connectivity of metabolites in iSyn842 and iSyf715 networks.*

which few highly connected nodes (termed hubs in graph theory) dominate the topology by linking the rest of the less connected nodes to the system (Jeong et al., 2000; Barabasi and Oltvai, 2004). Figure 3.2 shows the cumulative distribution of the number of metabolites with more than a given number of connections. In both cases, many metabolites have few connections while few metabolites have large number of connections. Besides, both networks present a very similar distribution, with the only difference that *Synechocystis*' includes more metabolites.

Finally, in Table 3.6 the most connected species of *i*Syn842 and *i*Syf715 are shown. Metabolic hubs are common to both networks (as they are also in other microorganisms (Montagud et al., 2010)), and comprise mainly cofactors (H_2O, H^+, O_2, ATP/ADP/AMP, PO_4^{3-}, $P_2O_7^{4-}$, $NAD(P)^+$/NAD(P)H) that play an important role connecting reactions and transporting energy and reductive power from one pathway to another.

Table 3.6: Most connected metabolites in iSyn842 and iSyf715 networks.

Metabolite	**Number of neighbours**	
	***i*Syn842**	***i*Syf715**
H_2O	261	230
H^+	183	153
ATP	162	153
PO_4^{3-}	124	114
ADP	120	109
$P_2O_7^{4-}$	95	97
$NADP^+$	80	61
NADPH	79	60
CO_2	77	71
NAD^+	66	55
NADH	62	51
O_2	53	38
L-glutamate	47	44
AMP	39	41

Altogether, their main features (Table 3.5), their central metabolic pathways, and their topology (Figure 3.2 and Table 3.6) show that the reconstructed networks for *Synechocystis* sp. PCC 6803 and *Synechococcus elongatus* PCC 7942 are similar in many aspects, with main differences being their size and availability to grow on sugar substrates. In next chapter, quantitative analyses are conducted that serve to further compare these two models, which leads to finding further likenesses and spot other differences between them.

4

Flux Balance Analysis on cyanobacteria metabolic models

4.1 Chapter abstract

In this chapter the metabolic networks of *Synechococcus elongatus* PCC 7942 and *Synechocystis* sp. PCC 6803 presented in the previous chapter are used to perform flux simulations using Flux Balance Analysis methodology. This methodology has been used to analyse and compare the robustness of these metabolic networks through an analysis of essential reactions. Next, a comparison of the flux landscapes obtained with both networks under autotrophic conditions has been performed that revealed significant similarities between them. And finally, an analysis was performed with *Synechocystis*' metabolic network to test the effect of perturbed culture conditions over the production of a metabolite of interest.

These simulations serve to illustrate the applications of constraint-based flux simulations in different studies, and to identify important features and opportunities for the development of new simulation algorithms.

Some limitations of the classic FBA methodology (such as impossibility of accounting for regulation, necessity of hard constraints to yield realistic energy costs, and impossibility of optimising several objectives simultaneously) are identified in this chapter, and this knowledge will be applied in Part III of this thesis dissertation to develop new algorithms to extend and enrich the possibilities of the simulations.

At the final part of this chapter, a contribution to the development of a web-based software tool, termed CellDesign, is presented. This tool aims at facilitating the access and application of metabolic simulations by non-experts.

Some of the results from this chapter appear in the following journal article:

- Gabriel Kind, Maria Siurana, Erik Zuchantke, David Fuente, Lenin G. Lemus-Zúñiga, Javier Urchueguía, Röbbe Wünschiers. **CellDesign - An Open-Source, Web-Based Software for Metabolic Modelling and Flux-Balance Analyses.** *Manuscript submitted to BMC Bioinformatics in June 2017.*

4.2 Introduction

As it has been discussed in Chapter 1, Systems Biology is a discipline that aims at the study of biological systems with the aim of understanding the behaviour of living organisms from a system-level point of view. Systems Biology also addresses the modification of existent systems and the design of new devices to perform newly desired functions such as disease treatment, chemical production, *etc.*(Ideker et al., 2001; Kitano, 2002). Analysis of metabolic fluxes contributes to the study of biochemical functions of the cells, to the analysis of their behaviour and to the assessment of the effect of selected modifications over the whole system. This understanding is fundamental to address the generation of cell factories: rationally designed living organisms that behave like production systems (Patil et al., 2004).

The object of study of metabolic modelling are the metabolic fluxes: the rates of metabolic reactions, which are the number of molecules being transformed at each metabolic reaction per unit time. These fluxes have an important role in cellular physiology, since they describe how materials and energy flow through the metabolic network of reactions (García Martín et al., 2015). From all the different types of modelling techniques that have been applied to the metabolism (for a good review, please refer to reference Klipp et al. (2016)), the present thesis is focused in constraint-based metabolic modelling.

The starting point to perform constraint-based metabolic simulations is to carefully reconstruct the metabolic network describing all biochemical transformations occurring in the system, which was addressed in the previous chapter.

Once an accurate and complete metabolic network has been obtained, constraint-based modelling methodology imposes a steady state on the metabolism. This allows to mathematically solve an underdetermined problem, as the number of reactions are greater than the number of metabolites. The rationale behind this steady-state imposition is the fact metabolic conversions are much faster than both cellular growth rates and the dynamic changes in the organism's environment (Varma and Palsson, 1994a). Consequently, a steady-state mass balance is applied to the system that reduces the space of possible flux distributions. Additionally, further constraints are defined that capture the physicochemical, environmental and biological limitations that real cell's metabolism faces in natural environments (Price et al., 2004). After applying these constraints, the remaining set of biologically feasible solutions describes the diversity of metabolic phenotypes that are physically possible.

However, with the addition of constraints and the analysis of the resulting feasible phenotypes it can be analysed what the metabolic model cannot do, but concrete flux distributions cannot be determined yet (Edwards et al., 2002b). One approach applied in constraint-based modelling to extract particular flux distributions from the set of feasible solutions consists on the optimisation of an objective function. This approach is based on the assumption that evolution under selective pressure has caused organisms to work in an optimal or quasi-optimal metabolic regime. In this sense, an objective function has to be chosen that describes appropriately the processes and different goals that characterise the metabolic behaviour of real organisms (Raman and Chandra, 2009).

The main representative of the methodologies used to calculate optimal flux distributions from biologically constrained spaces of solutions is Flux Balance Analysis (FBA) (see Section 1.2.2 in Chapter 1 and Section 4.3.1 bellow). This methodology can be applied to study flux distributions in natural conditions, as well as to analyse the effect that changes in environmental and genetic conditions have over the metabolic behaviour of the organism under study.

As part of the design strategy to obtain metabolic systems with desired properties genes can be deleted (knocked out) or inserted (knocked in) into the genetic code of an organism. Using such genetic manipulations, cells with new capacities can be created, as in cell factories, or existing systems can be readjusted to become such a cell factory. Additionally to genetic interventions, different environmental conditions can be tested that increase or decrease some fluxes of reactions of pathways that produce a metabolite of interest. Constraint-based modelling can be used in this context to compare the metabolic flux distributions of wild-type (non-mutated) and other mutated organism.

In this chapter, examples are presented that illustrate applications of Flux Balance Analysis methodology on cyanobacterial metabolism. In particular, FBA is applied to the models of *Synechococcus elongatus* PCC 7942 and *Synechocystis* sp. PCC 6803 presented in the previous chapter.

4.3 Materials and Methods

4.3.1 Flux Balance Analysis

Flux Balance Analysis methodology was applied to simulate flux distributions. Briefly (see Section 1.2.2 for details) and according to this methodology, a metabolic network is represented by its stoichiometric matrix $\boldsymbol{S}$ (as explained in Definition 1.1), and a steady-state mass balance is then applied to calculate the metabolic fluxes (gathered in vector $\boldsymbol{v}$). Constraints are imposed to the system that limit the range of allowable fluxes by considering reaction directionality, enzyme/transport capacity and specific physiological knowledge. The following linear optimisation problem is then stated to maximise/minimise an objective function which can be any linear combination of fluxes:

$$\max_{\boldsymbol{v}} Z(\boldsymbol{v}) = \boldsymbol{c}^T \cdot \boldsymbol{v} \tag{4.1}$$

subject to:

$$\boldsymbol{S}{\cdot}\boldsymbol{v} = \boldsymbol{0} \tag{4.2}$$

$$v_{j,rev} \in (-\infty, +\infty) \qquad j \in \{1, \ldots, n\} \tag{4.3}$$

$$v_{j,irr} \in [0, +\infty) \qquad j \in \{1, \ldots, n\} \tag{4.4}$$

$$l_{v_j} \leq v_j \leq u_{v_j} \qquad j \in \{1, \ldots, n\} \tag{4.5}$$

where $\boldsymbol{c}$ is a vector of weights indicating how much each reaction contributes to the objective function, $v_{j,rev}$ and $v_{j,irr}$ are the fluxes of the reversible and irreversible reactions respectively, and l_{v_j} and u_{v_j} are the lower and upper flux bounds for reaction j respectively.

Flux bounds defined for FBA simulations have two origins:

- directionality bounds (Equations (4.3) and (4.4)), that establish that reversible reactions can take negative values (meaning flux in the direction opposite to the described by the stoichiometric equation), while irreversible reactions can only have positive values.

- capacity and availability bounds (Equation (4.5)), that are imposed only on certain reactions, generally exchange reactions, and account for maximum enzyme/transport capacities and environmental nutrient availability.

The most commonly used objective function for Flux Balance Analysis is maximisation of biomass yield or growth rate (Feist and Palsson, 2010). As explained in Chapter 1, a biomass equation has to be defined during the reconstruction process. This biomass equation must account for all cell building-blocks, such as lipids, proteins and nucleic acids (or their precursors), as well as for the energy consumed during the assembly process and cell maintenance. It must describe as accurately as possible the proportions among the considered biomass components. This equation is used in FBA to represent cell growth and is usually employed as biological objective function. In the case that a different objective is defined, generally, the biomass function has to be used as a constraint, to consider energy and mass associated drains. In this Chapter, the biomass equations described in Tables 3.1 and 3.3 (Chapter 3) for *Synechocystis* sp. PCC 6803 and *Synechococcus elongatus* PCC 7942 respectively are used as objectives to represent growth.

The solution obtained from FBA is a vector of fluxes with individual flux values for each reaction.

In this thesis, the software tool *PyNetMet*[13] (Gamermann et al., 2014b) was used to perform FBA simulations. This Python-based toolbox (that was briefly described in Section 1.2.3) is designed to manipulate metabolic networks and perform flux simulation and analysis.

4.3.2 Metabolic models and simulation conditions

In this chapter, the metabolic networks of *Synechococcus elongatus* PCC 7942 and *Synechocystis* sp. PCC 6803 presented in Chapter 3 are used to perform metabolic simulations.

[13] Tool available at `github.com/CyanoFactory/CyanoFactoryKB`

As explained in Section 1.3, *Synechocystis* sp. PCC 6803 is able to grow under three trophic conditions that differ in the chosen energy and carbon sources (Montagud et al., 2010). These growth modes are:

(i) *photoautotrophy*, where energy comes from light and carbon from CO_2,

(ii) *heterotrophy*, where a sugar, often glucose, is the source of both energy and carbon, and

(iii) *mixotrophy*, a combination of the former two, where all three elements (light, CO_2 and glucose) are combined.

A part from these, *Synechocystis* also requires a variety of trace inorganic elements such as nitrogen, sulphur, phosphorus, iron, molybdenum, magnesium and manganese.

Thus, *Synechocystis* sp. PCC 6803 is what is known as a *facultative* autotroph, that is, it can grow in autotrophic conditions, but it is also able to obtain energy and carbon from reduced organic species, like glucose. On the other hand, as it was already mentioned in Section 1.3, *Synechococcus elongatus* PCC 7942 is an obligate autotroph (Rippka et al., 1979), which means that it can only survive under autotrophy, extracting energy and reducing equivalents from light and carbon from CO_2. Thus, it is unable to use reduced organic species to survive.

In the present chapter simulations are performed to compare both metabolic networks, and other simulations are conducted to explore their productive capacities. In all cases, the autotrophic growth mode has been chosen as it is the only trophic mode that they have in common. Table 4.1 shows the set of constraints applied to simulate autotrophic growth in both *i*Syn842 and *i*Syf715 models. In the case of simulations conducted to assess productive capacities, as it was already highlighted, the ability of cyanobacteria to perform photosynthesis and grow on sunlight, atmospheric CO_2 and some inorganic nutrients, is one of the main reasons for their interest as organisms for the development of cell factories. Thus, the goal is to design productive processes based on these organisms growing and producing under photoautotrophic conditions.

Table 4.1: *Main constraints for autotrophic growth simulations in Synechocystis' and Synechococcus' metabolic models. All units are mmole per gram of dry cell weight per hour,* $mmol/(g_{DC} \cdot h)$.

Input fluxes	**Lower bound**	**Upper bound**
Light in PSI[a]	0	23.5
Light in PSII[a]	0	23.5
CO_2	0	1.7
HCO_3^-	0	1.7
NO_3^-	0	100
SO_4^2-	0	100
PO_4^3-	0	100
glucose[b]	0	0

[a] Carbon-limited conditions. Values of light constraints corresponding to *Synechocystis*' model, obtained through a two-steps optimisation process (see Montagud et al. (2010) for details).

[b] Glucose intake is disabled in *Synechocystis*, no glucose transporter described in *Synechococcus*.

4.3.3 Essential reactions

One trait that can be used to characterise and compare metabolic networks and to evaluate network robustness is reaction essentiality.

The analysis consists on systematically deleting, one at a time, all the reactions in the network and use FBA to simulate the resulting model for growth functionality, that is, for its ability to produce biomass. To simulate the deletion of the reaction, which at the laboratory would be done by deleting the corresponding gene, there are two options that are equivalent: (i) delete the corresponding column in the stoichiometric matrix, or (ii) set the corresponding upper and lower flux bounds equal to zero. To simulate cell growth the biomass equation is used as objective function.

According to the growth rate obtained the reactions can be classified as:

- reactions whose deletion does not affect growth rate (resulting growth is the same as in the wild type $\pm 5\%$),
- reactions whose deletion reduces the growth rate with respect to the wild type simulation (reduction greater than 5% of wild type growth), and
- essential reactions, whose deletion totally disrupts biomass formation.

This analysis was applied on both *Synechocystis*' and *Synechococcus*' metabolic networks in order to compare their network robustness.

4.3.4 Comparison of flux landscapes

In the present chapter the flux landscapes obtained from autotrophic growth simulations of *Synechocystis* sp. PCC 6803 and *Synechococcus elongatus* PCC 7942 are compared. For both simulations to be comparable the same carbon intake constraint had to be set (see Table 4.1).

In order to compare the flux distributions, first the sets of reactions of the two networks were analysed and a subset of common reactions was identified and their values were plotted using a scatterplot. In this plot, similar fluxes will be close to the $y = x$ line, and distance from this line can be interpreted as a measure of dissimilarity between these two reactions in the two networks. This way, differences can be visually identified and analysed.

4.4 Flux Balance Analysis applied to *Synechocystis*' and *Synechococcus*' metabolic models

4.4.1 Essential reactions in *Synechocystis*' and *Synechococcus*' metabolic models

Synechocystis' and *Synechococcus*' metabolic networks obtained at Chapter 3 have been analysed in terms of reaction essentiality. In the case of *Synechocystis*, essential reactions under autotrophic, heterotrophic and mixotrophic conditions were evaluated. In the case of *Synechococcus*, only autotrophic conditions were analysed.

Under autotrophic conditions, both organisms exhibit a very similar behaviour (Figure 4.1 and Table 4.2), which was expectable, given that they are closely related phylogenetically (Gamermann et al., 2014a). *Synechocystis* (A) shows slightly higher robustness than *Synechococcus* (B), given the lower percentage of reactions that cause decrease or total disruption of growth when deleted (yellow and red codes in Figure 4.1 and Table 4.2).

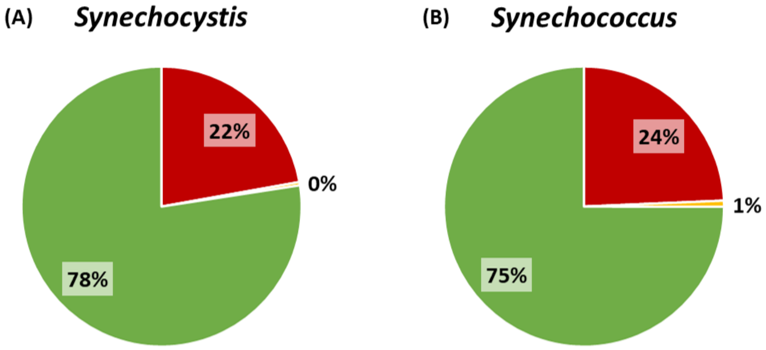


Figure 4.1: *Essential reactions in Synechocystis' and Synechococcus' metabolic networks under autotrophic conditions. Red corresponds to no growth, yellow to reduced growth and green to wild type growth ±5%.*

Table 4.2: Essential reactions in Synechocystis' and Synechococcus' metabolic networks under autotrophic conditions.

Effect on growth	*Synechocystis*		*Synechococcus*	
🟥 No growth	235	22.19%	207	24.32%
🟨 Reduced growth	4	0.38%	6	0.71%
🟩 Wild type growth ±5%	820	77.43%	638	74.97%

In the case of *Synechocystis*, essential reactions under heterotrophic and mixotrophic conditions were also analysed (Figure 4.2 and Table 4.3). Under autotrophic conditions (A) the proportion of reactions that cause organism's death is slightly higher, while no reactions are found to cause constrained growth, which altogether indicates that under these conditions the metabolic network is less flexible to cope with impaired reactions. If heterotrophic (B) and mixotrophic (C) modes are compared, the proportion of reactions that can be deleted without causing changes in growth greater than 5% of wild-type growth (green sectors) is slightly higher in heterotrophy. These modes present similar amount of essential reactions, and under mixotrophic conditions the proportion of reactions that cause constrained growth is slightly higher. Furthermore, independence of distribution was assessed using Pearson's Chi-squared test that proved non-significant for the distributions across trophic conditions. Meaning that independence could not be rejected.

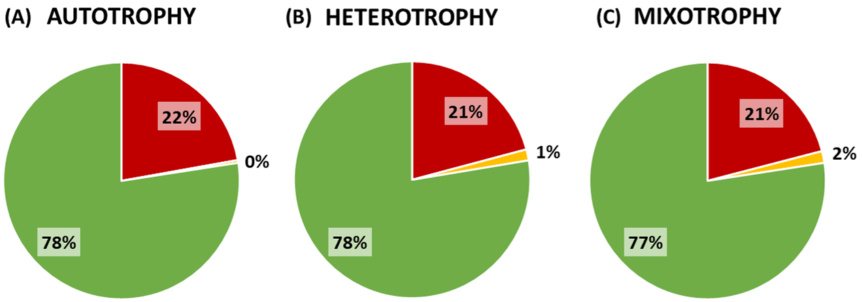


Figure 4.2: Essential reactions in Synechocystis' metabolic network under three trophic conditions: (A) autotrophy, (B) heterotrophy, and (C) mixotrophy. Red corresponds to no growth, yellow to reduced growth and green to wild type growth ±5%.

Table 4.3: *Essential reactions in Synechocystis' metabolic network under three trophic conditions.*

Effect on growth	Autotrophy		Heterotrophy		Mixotrophy	
🟥 No growth	235	22.19%	221	20.87%	222	20.96%
🟧 Reduced growth	4	0.38%	16	1.51%	17	1.61%
🟩 Wild type growth ±5%	820	77.43%	822	77.62%	820	77.43%

However, it must be taken into account that by means of these simulations what is evaluated is the direct effect of the presence (or absence) of a given reaction in the metabolic network. Indirect effects such as its gene's or enzyme's regulation are not being considered in this kind of analyses. Thus, these results may vary if regulatory information is considered. In the following part of this dissertation, an algorithm is proposed to improve the realism of the obtained results by including experimental data.

4.4.2 Comparison of flux landscapes of *Synechocystis* and *Synechococcus*

As seen in Chapter 3 *Synechocystis* sp. PCC 6803 and *Synechococcus elongatus* PCC 7942 metabolic networks present the same main pathways, *viz.* : (i) pathways integrating the central carbon metabolism (glycolisis, citric acid cycle and pentose phosphate pathway), (ii) those related to energy metabolism (photosynthesis and oxidative phosphorylation), (iii) pathways for the assimilation of inorganic nutrients (like nitrogen and sulphur), and (iv) anabolic pathways for the synthesis of cell components, biomass precursors and secondary metabolites.

Along these pathways, certain reactions are found that are common and have the same stoichiometry. In the present section, fluxes of 451 common reactions are compared under autotrophic growth conditions. Figure 4.3 shows a scatter plot of the flux values of these reactions in *i*Syn842 (x-axis) and *i*Syf715 (y-axis). Most of those fluxes have similar values in both systems, and so they appear close to the $y = x$ line. Only one reaction is separated from this line more than 1 unit (dashed lines), which indicates strong similarity between the flux landscapes obtained with *i*Syn842 and *i*Syf715 under photoautotrophic conditions. In fact, Spearman correlation of these two subsets has a rho of 0.6925 with a p-value inferior to 2.2e-16.

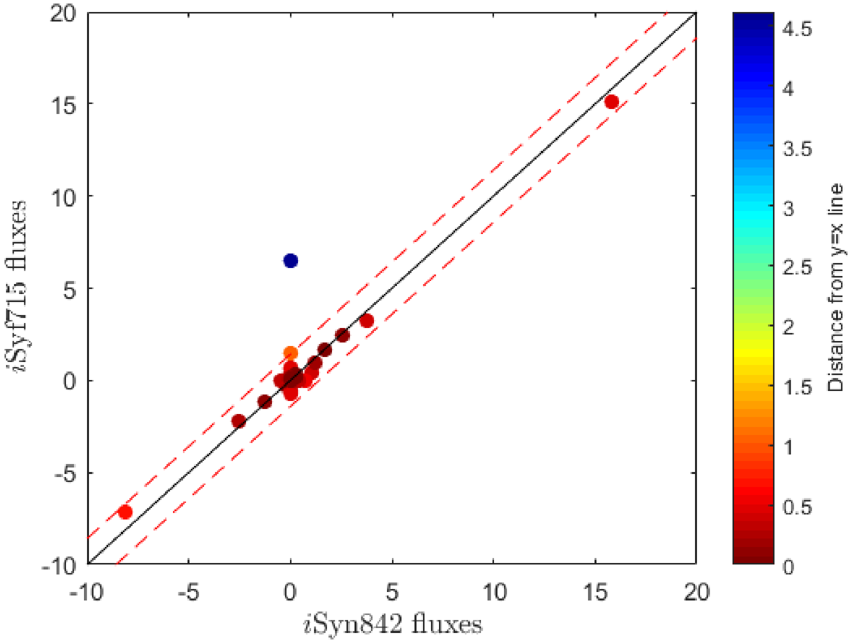


Figure 4.3: *Correspondence between fluxes in 451 common reactions of iSyn842 and iSyf715 obtained under autotrohic growth conditions. Flux units are mmole per gram of dry cell weight per hour,* $mmol/(g_{DC} \cdot h)$*. Dashed line marks the distance of one unit from the diagonal.*

In order to check if these subsets were statistically different from their parent sets of data we used Kruskal-Wallis rank sum test. This test resulted in non-significance when comparing the 451 values of common fluxes of *i*Syn842 to the 1059 values of this model. Same non-significance was obtained when comparing the 451 values of common fluxes of *i*Syf715 to the 902 values of this model. Thus, these subsets have the same distributions than their parent sets of data. Figure 4.4 shows histograms of the whole sets of reaction fluxes in *i*Syn842 (A) and *i*Syf715 (B).

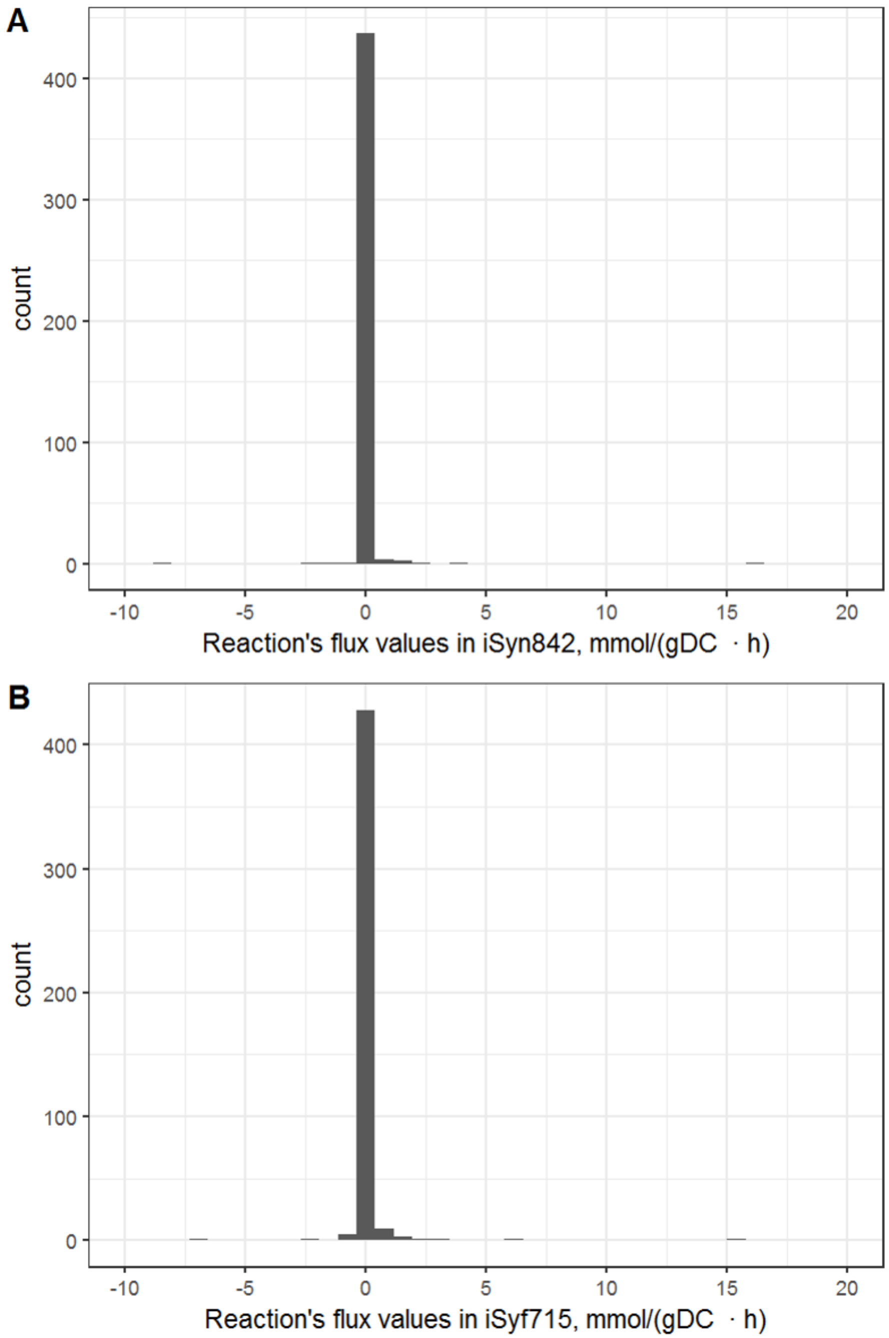


Figure 4.4: *Histogram of flux values distributions of 451 common reactions in (A) iSyn842 and (B) iSyf715.*

Additional file 4.1 (see page 249) contains the flux distributions resulting from the simulations of *i*Syn842 and *i*Syf715 models under autotrophic conditions (see Table 4.1 for constraints), and the list of common reactions between them.

The single dark blue point located over the $y = x$ line in the plot in Figure 4.3 represents the reaction with the highest difference of flux value between *i*Syn842 and *i*Syf715 of the common 451 reactions. This reaction is trans-hydrogenation from NADPH to NADH:

$$1.6.1.2 : \mathrm{NAD}^{+} + \mathrm{NADPH} \quad \rightleftharpoons \quad \mathrm{NADH} + \mathrm{NADP}^{+}$$

This reaction has no flux in the case of *i*Syn842 (flux = 0) while it has a considerable flux in the case of *i*Syf715 (flux = 6.53, Figure 4.3). This difference might mean that *Synechococcus* has preference for NADH while *Synechocystis* makes more use of NADPH. However, even though *i*Syf715 was validated to the best of our knowledge, it could be also the case of a spurious trans-hydrogenation, which would require some more constraints in *i*Syf715 to avoid such loops when applying FBA. In fact, this is one of the drawbacks of this methodology, that many times hard constraints are needed in order to avoid unpractical simulation results as well as to account for realistic energy costs.

The only other reaction placed outside the threshold of 1 unit distance from the diagonal is the point with flux values *i*Syn842 flux = 0 and *i*Syf715 flux = 1.51. The corresponding reaction is:

$$3.6.1.1 : \mathrm{H_2O} + \mathrm{diphosphate} \quad \longrightarrow \quad 2\ \mathrm{PO}_4^{3}-$$

which is hydrolysis of diphosphate to form two molecules of phosphate.

Interestingly, Figure 4.3 also shows that most of the reactions are between the ranges of -5 and 5 $mmol/(g_{DC} \cdot h)$, while only a few show higher values, but never higher than 20 $mmol/(g_{DC} \cdot h)$. Among the highest fluxes, the point at the (15,15) crossing of the plot represents ATP synthesis at the ATP-ase (E.C. 3.6.3.14), and the point at the (-8,-7) of the plot corresponds (considering the direction of the flux) to electron transfer from thioredoxin to NADPH catalysed by enzyme 1.8.1.9.

4.4.3 Testing physiological perturbed conditions in *Synechocystis* sp. PCC 6803

It has been already mentioned that most of the work done in this thesis was motivated by the participation in a European project of the Seventh Framework Programme, CyanoFactory, which aimed at the design of cell factories based on cyanobacteria to produce biofuels, specifically hydrogen. Within this framework the metabolic model of *Synechocystis* sp. PCC 6803 described in Chapter 3 was used to test environmental and genetic conditions to enhance H_2 production.

Assessment of nitrogen and sulphur substrates to enhance H_2 production

As it was seen in Chapter 3 (Figure 3.1), *Synechocystis* sp. PCC 6803 has a complex electron transport chain that combines elements from both photosynthesis and oxidative phosphorylation (Vermaas, 2001). These crosstalks cause that electrons flowing through this part of the network can reach many alternative pathways (Nogales et al., 2012): they can stay within the electron transport network to increase the ATP/NADPH ratio through different cyclic routes; or can be directed to other pathways, including CO_2 fixation (through the Calvin-Benson-Bassham cycle) and inorganic nitrogen and sulphur reduction, to be used to synthesise biomass components; or they can end in final electrons sinks, such as reduction of oxygen into water or protons into hydrogen, that are used by the cell to adjust to high or changing light conditions. Thus, hydrogen production competes for electrons with many other pathways.

Nitrogen and sulphur assimilation pathways are among the pathways that consume electrons from the electron transport chain. Since they are related to nutrient assimilation, changes in the growth media can be done in order to increase the amount of electrons available to produce hydrogen. In this section, a substrate study is described whose goal was to determine the best sources of inorganic nitrogen and sulphur that would increase the hydrogen production.

Previous studies had pointed that inhibiting (Gutthann et al., 2007) or disrupting (Baebprasert et al., 2011) the nitrate assimilation pathway and feeding ammonia to *Synechocystis*, instead of nitrate or nitrite, increased hydrogen production was achieved (Figure 4.5 (A) and (B)). Apart from regulatory effects that may occur (especially in the case of gene disruption), one of the main reasons for this increase in hydrogen evolution is the electrons saved from being used in nitrate reduction to nitrite and ammonia that can be then redirected to other electron sinks, like hydrogen production.

Figure 4.6 represents the main steps of nitrogen (red) and sulphur (blue) assimilation pathways. Both pathways receive their electrons directly from ferredoxin, that obtains them from the electron transport chain. Considering these observations, the effect of changing nitrogen and sulphur sources over maximum H_2 yield was investigated through FBA simulations of *i*Syn842.

The standard mineral BG-11 medium (Rippka et al., 1979) contains nitrate (NO_3^-) and sulphate (SO_4^{2-}) as nitrogen and sulphur sources respectively. The species considered in this study were the framed ones in Figure 4.6: nitrate (NO_3^-), nitrite (NO_2^-) and ammonia (NH_4^+) as nitrogen sources; and sulphate (SO_4^{2-}), sulphite (SO_3^{2-}), thiosulphate ($S_2O_3^{2-}$), sulphide (S^{2-}), cysteine (Cys), methionine (Met) and glutathione as sulphur sources. Previous studies from Gutthann et al. (2007) and Baebprasert et al. (2011) pointed to the use of different sources of nitrogen. In the case of sulphur, no previous study was found in literature in which different species were tested for hydrogen production.

The simulations where performed solving a sequence of optimisation stages in which different objective functions were considered:

- *Biomass formation*: In this study the biomass formulation defined in *i*Syn842 model (Table 3.1) was applied, as it is standard in metabolic FBA simulations (see Section 1.2.1 and Feist and Palsson (2010)).

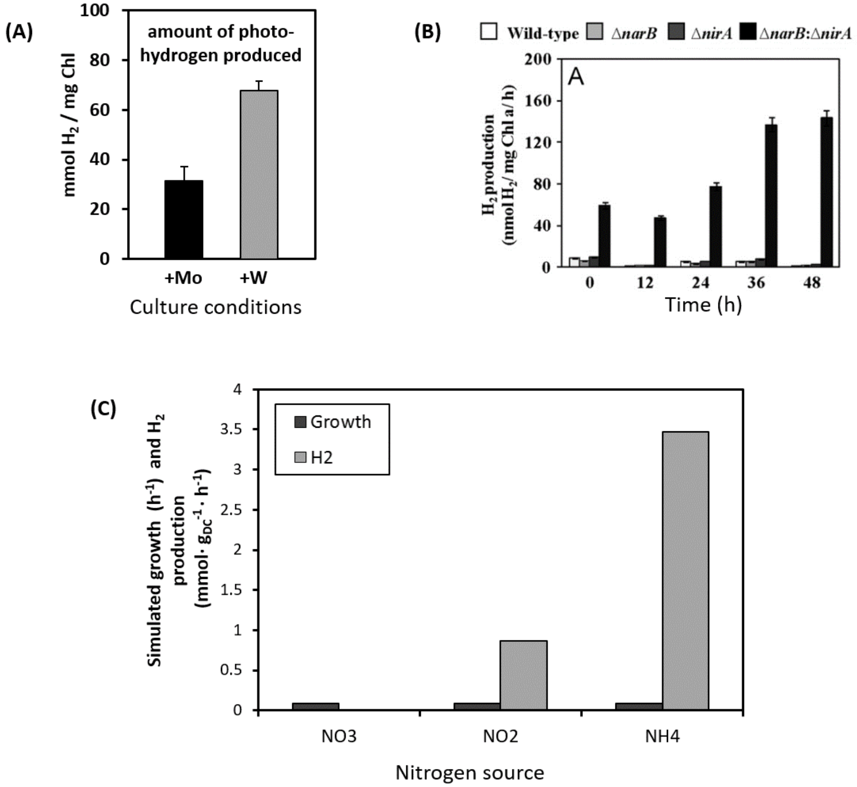


Figure 4.5: *Experimental vs. simulated data of hydrogen production in Synechocystis sp. PCC 6803 with different nitrogen sources. (A) Ammount of photo-hydrogen produced in normal Synechocystis cultures with normal (+Mo) and inhibited (+W) nitrate reductase enzyme. Adapted from Figure 6 (Gutthann et al., 2007), improvement of hydrogen production when the enzyme nitrate reductase is inhibited by the presence of tungstate (+W) and ammonia is present as sole nitrogen source. (B) From Figure 6 (Baebprasert et al., 2011), improvement of hydrogen production in mutants with the nitrate assimilation pathway disrupted. (C) Simulated hydrogen production at a given growth rate with different sources of inorganic nitrogen. Simulations were carried out under photoautotrophic conditions, with fixed amount of photons and* CO_2*, at a fixed growth ratio (maximum growth for this carbon intake) using Algorithm 4.1, and allowing the input of different nitrogen sources: nitrate (NO3), nitrite (NO2) and ammonia (NH4).*

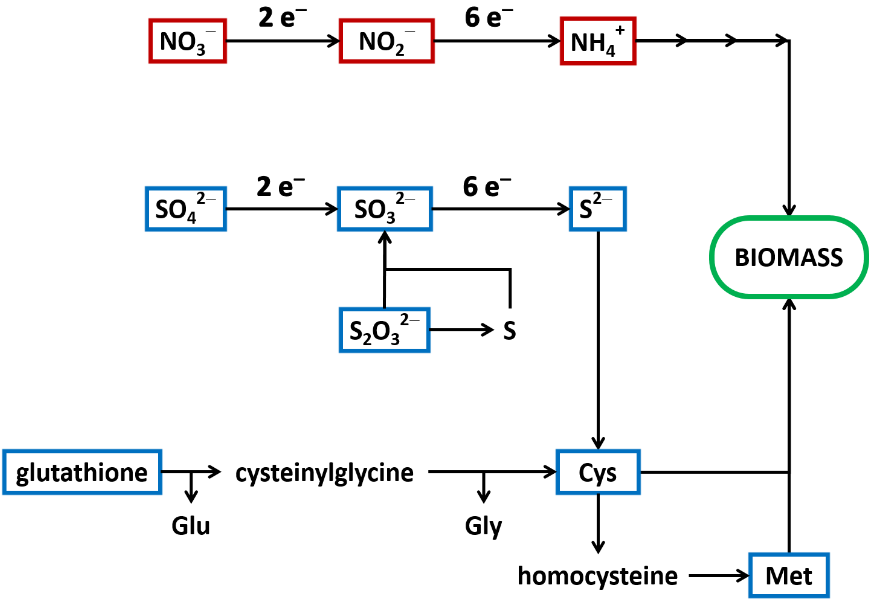


Figure 4.6: Outline of nitrogen (red) and sulphur (blue) assimilation pathways in Synechocystis sp. PCC 6803. Abbreviated species: nitrate (NO_3^-), nitrite (NO_2^-), ammonia (NH_4^+), sulphate (SO_4^{2-}), sulphite (SO_3^{2-}), thiosulphate ($S_2O_3^{2-}$), sulphide (S^{2-}), cysteine (Cys), methionine (Met), glutamate (Glu) and glycine (Gly).

- *Intake of nitrogen source* n: defines the reaction flux that describes transport of the given nitrogen source

$$n \in \mathscr{N} := \{NO_3^-; NO_2^-; NH_4^+\}$$

- *Intake of sulphur source* s: defines the reaction flux that describes transport of the given sulphur source

$$s \in \mathscr{S} := \{SO_4^{2-}; SO_3^{2-}; S_2O_3^{2-}; S^{2-}; Cys; Met; glutathione\}$$

- *Hydrogen production*: defines the positive flux of the bi-directional hydrogenase reaction, *i.e.* the hydrogen production

$$r_{H_2} : \mathrm{NADPH} + 2\,\mathrm{H}^+ \rightleftharpoons \mathrm{H}_2$$

Algorithm 4.1 shows the sequence of steps followed to obtain the maximum growth and hydrogen yields achievable if the cells are grown on a culture media containing each of the nitrogen and sulphur sources considered. First (lines 1-4), a simulation is run maximising biomass production under autotrophic conditions with standard BG-11 nitrogen and sulphur sources:

$$0 \leq v_{NO_3^-} \leq 1000, \quad v_n = 0, \ \forall n \in \mathscr{N} \mid n \neq NO_3^- \tag{4.6}$$

$$0 \leq v_{SO_4^{2-}} \leq 1000, \quad v_s = 0, \ \forall s \in \mathscr{S} \mid s \neq SO_4^{2-} \tag{4.7}$$

From this simulation the value of normal growth under autotrophic conditions in standard BG-11 media is obtained (line 5).

Next, for each alternative substrate considered (t), if it is a nitrogen source (lines 7-9), nitrogen intake is restricted to t and sulphur source is set to be SO_4^{2-}:

$$0 \leq v_t \leq 1000, \quad v_n = 0, \ \forall n \in \mathscr{N} \mid n \neq t \tag{4.8}$$

$$0 \leq v_{SO_4^{2-}} \leq 1000, \quad v_s = 0, \ \forall s \in \mathscr{S} \mid s \neq SO_4^{2-} \tag{4.9}$$

And if alternate sulphur sources are considered (lines 10-12), nitrogen is limited to be NO_3^- and sulphur intake is restricted to t :

$$0 \leq v_{NO_3^-} \leq 1000, \quad v_n = 0, \ \forall n \in \mathscr{N} \mid n \neq NO_3^- \tag{4.10}$$

$$0 \leq v_t \leq 1000, \quad v_s = 0, \ \forall s \in \mathscr{S} \mid s \neq t \tag{4.11}$$

Then the biomass formation is constrained to the value obtained from the first optimisation (line 13):

$$v_{biomass} = \mu \tag{4.12}$$

and a new simulation is run minimising the intake of the given nitrogen or sulphur species to obtain the minimum flux v_{in}^t needed to achieve normal growth using this nitrogen or sulphur source (lines 14-16).

Finally, a last simulation is run (lines 17-20) maximising hydrogen production while biomass formation is fixed to be equal to μ and the maximum intake of the substrate under study is set to v_{in}^t:

$$0 \leq v_t \leq v_{in}^t \tag{4.13}$$

```
Input: metabolic network and constraints
Output: maximum growth and hydrogen yields with each
        considered nitrogen and sulphur source
Set constraints to autotrophic growth (Table 4.1) ;
Set nitrogen and sulphur sources to BG-11 species (Equations
  (4.6) and (4.7)) ;
Set objective function to maximise biomass formation ;
Calculate flux distribution $\boldsymbol{v}|_1$ using FBA methodology ;
Extract from $\boldsymbol{v}|_1$ the flux of biomass formation $\mu$ ;
for each substrate $t$ considered do
    if $t \in \mathscr{N}$ then
        Restrict nitrogen sources to $t$ (Equation (4.8)) ;
        Restrict sulphur sources to $SO_4^{2-}$ (Equation (4.9)) ;
    else if $t \in \mathscr{S}$ then
        Restrict nitrogen sources to $NO_3^-$ (Equation (4.10)) ;
        Restrict sulphur sources to $t$ (Equation (4.11)) ;
    Constrain flux of biomass formation to be equal to $\mu$
      (Equation (4.12)) ;
    Set objective function to minimise intake of the given
      substrate ;
    Calculate flux distribution $\boldsymbol{v}|_2$ using FBA methodology ;
    Extract from $\boldsymbol{v}|_2$ the flux of substrate intake $v_{in}^t$ ;
    Set upper bound of flux of substrate intake to $v_{in}^t$ (Equation
      (4.13)) ;
    Constrain flux of biomass formation to be equal to $\mu$
      (Equation (4.12));
    Set objective function to maximise hydrogen production ;
    Calculate flux distribution $\boldsymbol{v}|_3$ using FBA methodology ;
    Extract from $\boldsymbol{v}|_3$ the flux of hydrogen production $v_{H_2}$ ;
    Store the pair $\mu$ and $v_{H_2}$ at row $t$ of the matrix of results $R$
end
return $R$
```

Algorithm 4.1: Sequence of steps to evaluate maximum hydrogen yield in presence of different nitrogen and sulphur sources.

This complex process is needed to approach the most optimal combination of substrate intake, growth and hydrogen production by using the mono-objective linear optimisation-based methodology employed in Flux Balance Analysis. It will be seen later in this thesis that multi-objective optimisation paradigm can be applied to metabolic simulations to permit the consideration of several simultaneous objectives.

The results obtained at the simulations of alternative nitrogen sources (Figure 4.5, page 106) qualitatively match experimental observations reported by previous works (Gutthann et al., 2007; Baebprasert et al., 2011). Both experimental and simulated data show that the amount of hydrogen produced can be increased by using more reduced nitrogen sources. This is due to the fact that the form in which *Synechocystis* is able to assimilate nitrogen is as ammonia, and so all the electrons that are not used to reduce the nitrogen to this form (8 mmol of electrons per mmol of nitrogen) are available to produce hydrogen.

In Figure 4.7, the effect of using alternate nitrogen and sulphur sources can be seen separately (A and B) and combined (C). In the case of sulphur, an increase is observed from $\mathrm{SO_4^{2-}}$ to other inorganic sources (like $\mathrm{SO_3^{2-}}$, $\mathrm{S_2}O_3^{2-}$ or $\mathrm{S^{2-}}$) which is due to the savings in electrons that do are not used to reduce the inorganic sulphur (up to 8 mmol of electron per mmol of sulphur). When using cysteine or glutathione a higher increase is possible, because these molecules are also sources of carbon and electrons. However, with methionine as source of sulphur no increase is observed. This may be due to the fact that electrons available in methionine molecules cannot be extracted through catabolic reactions, and thus cannot be used to produce hydrogen. The effect of changing sulphur sources is less dramatic than using alternate nitrogen sources. This is because the amount of sulphur needed for growth is significantly smaller, and thus fewer electrons are consumed in total to reduce the oxidized sources to produce sulphide.

Additional file 4.2 (see page 249) contains the flux distributions resulting from all the simulations performed in this study.

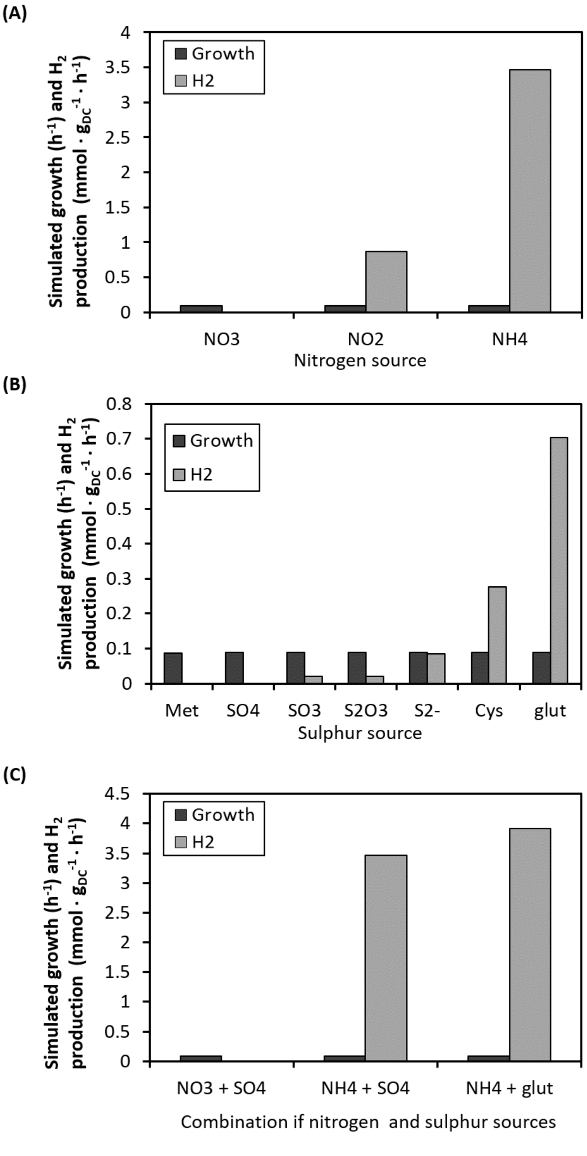


Figure 4.7: *Results of the simulation of hydrogen production at a given growth rate with different sources of nitrogen and sulphur. Simulations were carried out under photoautotrophic conditions, with fixed amount of photons and CO_2, at a fixed growth ratio (maximum growth for this carbon intake), and allowing the input of different (A) nitrogen sources: nitrate (NO3), nitrite (NO2) and ammonia (NH4); (B) sulphur sources: methionine (Met), sulphate (SO4), sulphite (SO3), thiosulphate (S2O3), sulphide (S2-), cysteine (Cys) and glutathione (glut); and (C) combination of nitrogen and sulphur sources leading at best maximum optimal hydrogen production.*

The conclusions of this study show that by choosing the appropriate source of these inorganic substrates, maximum optimal H_2 production can be increased substantially (Figure 4.7 (C)). Besides, numeric results of the simulations can be used to assist experimental design of growth media by calculating the minimum amount of each nitrogen or sulphur source needed at a given growth rate.

4.5 Development of software tools for FBA simulations

Metabolic models and Flux Balance Analysis are valuable tools to analyse metabolic behaviour of cells and to guide design of new strains by assessing how different modifications in media or genomes can affect growth or productivity of desired metabolites. In order to easy the access of different scientific communities to perform this kind of metabolic simulations using metabolic models, it is important to provide user-friendly open-source tools. In this sense, and as a part of the mentioned European research project, the PhD applicant has collaborated with the research group led by Dr. Röbbe Wünschiers from the University of Applied Sciences Mittweida in the creation of CellDesign[14].

CellDesign is a web-based environment that allows non-specialist researchers to simulate and evaluate environmental and genetic perturbations, and design and integrate parts and devices into the cell wild type represented by a metabolic model.

For the design of this tool two key aspects were taken into consideration. First the system should be easy to use. To accomplish this goal, a deep study of similar existing tools was performed, that allowed to realise that many of these tools are oriented to experts (*i.e.* COBRA Toolbox (Becker et al., 2007; Schellenberger et al., 2011) or COBRApy (Ebrahim et al., 2013) come to mind). The second aspect to be considered was the information architecture: an appropriate structure should

[14] Home page `http://celldesign.de`

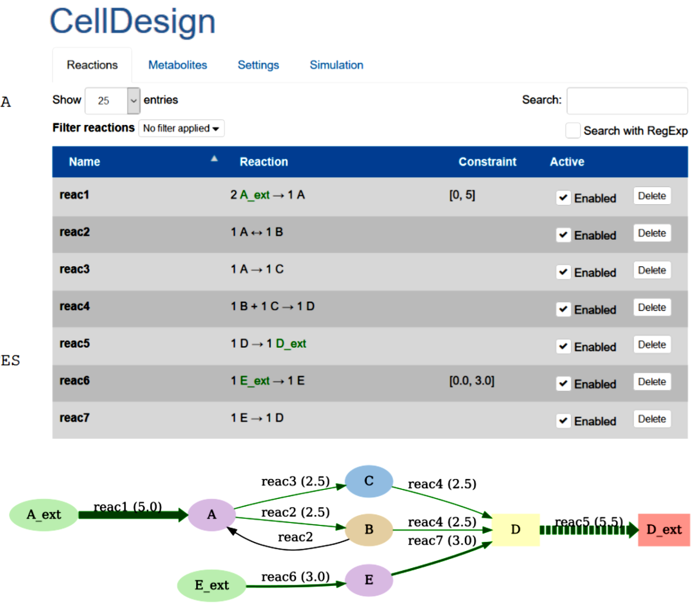
allow the user to organize, select and show all the information in a straightforward way. They were identified the tool inputs (main settings, experiment customization) and outputs (results summary and visualization), so that the relationship between both remains coherent and modular, ensuring that the system is flexible and scalable.

The PhD applicant was part of the team that designed and built this tool. Her role focused on the technical guidance about metabolic simulations, use of PyNetMet for these, and promote and supervise the use of this tool in a Bachelor's level metabolic engineering and modelling course throughout three years as Teaching Assistant.

The alpha version was available at the end of 2014, when it was presented to the researchers integrating the CyanoFactory research project (most of them with an experimental background). It was also tested as an educational tool in a curse on metabolic engineering and modelling from 2014 until 2017. With the feedback obtained from the consortium partners and students, some new desirable features were identified and implemented, which contributed to improve the user-interface doing it friendlier for non-experts.

In the current version of CellDesign users can load a pre-built model or create their own from a template, and also some pre-defined models (including the *i*Syn842 model presented in Chapter 3) are available. Reactions and metabolites can be added, modified or deleted, and changes can be saved in the same model or as a new one. The layout is organized by tabs: the first tab is devoted to reactions and their features (metabolites involved, stoichiometry, reversibility, constraints), the second tab presents metabolites and their features (reactions in which the metabolite is produced or consumed, external/internal), at the third tab the settings of the simulation can be adjusted (objective, type of simulation, formats of the results), and at the fourth tab the results are presented after running the simulations. CellDesign uses an updated version of PyNetMet (Gamermann et al., 2014b) to calculate metabolic fluxes through Flux Balance Analysis, that are shown using different graphical resources (bar plots, flux maps and tables), with some addi-

```
-REACTIONS

reac1 : 2 A_ext -> A
reac2 : A <-> B
reac3 : A -> C
reac4 : B + C -> D
reac5 : D -> D_ext
reac6 : E_ext -> E
reac7 : E -> D

-CONSTRAINTS

reac1 [0.0, 5.0]
reac6 [0.0, 3.0]

-EXTERNAL METABOLITES

A_ext
D_ext
E_ext

-OBJ

reac5 1 1

-DESIGNOBJ

reac5 1 1
```



CellDesign

Reactions | Metabolites | Settings | Simulation

Show 25 entries | Search:

Filter reactions: No filter applied ▾ | □ Search with RegExp

Name	Reaction	Constraint	Active	
reac1	2 A_ext → 1 A	[0, 5]	✓ Enabled	Delete
reac2	1 A ↔ 1 B		✓ Enabled	Delete
reac3	1 A → 1 C		✓ Enabled	Delete
reac4	1 B + 1 C → 1 D		✓ Enabled	Delete
reac5	1 D → 1 D_ext		✓ Enabled	Delete
reac6	1 E_ext → 1 E	[0.0, 3.0]	✓ Enabled	Delete
reac7	1 E → 1 D		✓ Enabled	Delete

Figure 4.8: *The CellDesign workflow. A metabolic model (left), here an example Toy model, is imported to CellDesign and displayed inside the website. All reactions and their corresponding constraints are shown and external metabolites are highlighted in green (top right). The model is fully customisable and the modeller applies required adjustments to the model. After simulation a graph representation of the model is displayed and annotated with the calculated fluxes. The objective function is highlighted through a dashed line (bottom right).*

tional information about the simulation, and can also be exported in different formats.

4.6 Conclusions of this chapter

In this chapter uses of Flux Balance Analysis have been illustrated using the metabolic networks of two model cyanobacteria described in Chapter 3. Different analyses have been performed to assess the effect of gene knock-outs on growth rate, to quantitatively compare the flux landscapes of the two cyanobacteria, and to test the effect of modifications in growth media on the production of a chemical of interest, namely hydrogen. These examples emphasize the convenience of using simulation techniques to assess design strategies, that require reduced time and materials investment, before trying them in the laboratory.

However, the variety of analyses that can be performed with FBA, is limited. Some limitations have been pointed out in the present chapter such as impossibility of accounting for regulation effects, necessity of hard constraints to avoid futile cycles and to match realistic energy costs, and impossibility of optimising several simultaneous objectives. Limitations that lead us to require the use of other algorithms to extend and enrich the possibilities of solutions. In the following Part of this dissertation, an algorithm is presented that is based on multi-objective evolutionary optimisation and aims at addressing these limitations, without losing the simplicity of constraint-based metabolic simulations that is one of their greatest strength.

Finally, a web-based software tool has been presented that aims at facilitating the access of non-specialists to these modelling techniques and providing an open platform for analysis and assessment of design strategies.

Part III

Metabolic modelling by means of multi-objective optimisation techniques

5

Meta-MODE: a multi-objective evolutionary algorithm for constraint-based flux simulations

5.1 Chapter abstract

In this chapter a new tool for genome-scale metabolic simulations based on multi-objective optimisation is described. The presented algorithm, referred to as Meta-MODE, is used in order to solve a multi-objective optimisation process based on differential evolution applied to metabolism simulation. It incorporates specific mechanisms to improve pertinency of the obtained solutions, as well as to prepare the algorithm to deal with the large-scale, strongly constrained, and multi-modal optimisation problem that arises when performing genome-scale constraint-based flux simulations. It also includes the mathematical formulation

of a set of constraints and objectives relevant to ensure the biological meaningfulness of the solutions.

The resultant algorithm aims to avoid some of the main drawbacks of classic methods for constraint-based metabolic analysis, like the need of defining a fixed equation accounting for biomass assembly, the strong dependency of the solutions on the selected objective or the limitations encountered to define informative constraints or objectives that exceed the mathematical capabilities of linear and/or mono-objective optimisation techniques.

A comprehensive description of the mathematical formulation of the algorithms and processes involved in this simulation tool is presented in the present chapter. In the next chapter, an extensive example of application of this tool is discussed.

5.2 Introduction

As it has been discussed in Chapter 1, Systems Biology aims at the study of living organisms as a whole, trying to understand how a cell's behaviour emerges from the interaction of its molecular parts (Ideker et al., 2001; Kitano, 2002). In particular, analysis of metabolic fluxes can assist researchers in determining physiological states of the cells. Different phenotypes arising from different genotypes and environmental conditions can be studied to investigate metabolic functions and aid in rational design of biological systems performing desired roles. Metabolic models and, specifically, constraint-based metabolic models, provide a functional tool for this purpose.

The starting point to perform constraint-based metabolic simulations is to carefully reconstruct the metabolic network describing all biochemical transformations occurring in the system. The variable object of study are the metabolic fluxes, that describe how materials and energy are converted and flow through this network. In order to delimit the set of practical flux distributions that characterise feasible phenotypes, a combination of constraints must be imposed that specify physico-chemical, environmental and biological limitations affecting real metabolic systems. Flux distributions remaining in the biologically feasible solution space will represent allowable metabolic phenotypes. Thus, a precise and complete description of the governing constraints is a key issue in this modelling technique. Meaningful, often complex, rules are needed to properly account for real metabolic restrains.

In order to extract particular flux distributions from the set of all the allowable solutions, optimisation is applied based on the assumption that evolution under selective pressure has made organisms perform in an optimal (or quasi-optimal) metabolic regime. In this regard, it is crucial that the mathematical objective function appropriately represents the evolutionary objective. Thus, selection of proper biological objectives is another essential point to obtain meaningful results.

However, it is difficult to account for all mechanisms and processes that can favour evolutionary advantage of an organism as a single objective. Often, living systems face complex situations (adaptation to shifting conditions, multiple functions, external threats, *etc.*) in which they are forced to find a trade-off between different, seldomly opposed objectives (Schuetz et al., 2012; Metallo and Vander Heiden, 2013). Classic optimisation algorithms described for constraint-based metabolic modelling (see Section 1.2.2 for a review) are mainly based on mono-objective optimisation techniques, which don't allow consideration of competing metabolic goals. Only a few research works have applied multi-objective procedures to perform flux simulations (Nagrath et al., 2007, 2010; Schuetz et al., 2012).

Apart from considering solely single objectives, some of the classical tools suffer from other important limitations, like (i) depending on the definition of a biomass equation, (ii) producing results strongly dependent on objective functions, (iii) being restricted to linear constraints and/or linear objective functions, or (iv) yielding too optimally unrealistic flux distributions.

As it was mentioned in Section 1.2.1, many methods rely on the definition of a function describing biomass assembly . Usually, this function is a fixed stoichiometric combination of biomass components and precursors. This approach, has satisfactorily been used to solve simple FBA-based optimisation problems[15], but it has significant shortcomings. First, like all phenotype traits, biomass composition is a result of both genetic and environmental factors, and thus it may vary under different conditions. A way to circumvent this drawback is to define different biomass equations for the distinct combinations of conditions, but this way the final trait (biomass composition) is imposed beforehand, instead of emerging as a consequence of the studied conditions.

A second implication of using single (even when assorted) biomass functions to account for growth objective is that the resulting flux distribution will strongly depend on the elements present in this equation and their proportions. Accurate determination of biomass components and their corresponding coefficients is a hard task (see Feist and Palsson (2010) for a detailed description of the process), and thus it is convenient to attenuate the effect that errors in these results may have on the final flux distributions.

Another common feature among FBA-based methods is that they turn to Linear Programming optimisation techniques, which forces all the relations stated among constraints and conditions to be linear. Even methods that employ non-linear optimisation techniques to consider non-linear objectives usually limit their constraints to linear functions. Non-linear rules and relationships often appear to control metabolic behaviour (Metallo and Vander Heiden, 2013) and thus, it is important to provide means to consider such relationships when performing metabolic simulations.

Nevertheless, even if the selected optimisation algorithms allow for the inclusion of non-linear rules, the way to define metabolic regulation is not straightforward (for example Covert et al. (2001) and Shlomi et al. (2007) add gene regulation to the optimisation process, but more

[15] See section **Flux Balance Analysis** at page 27 for a description of this method.

equations and parameters are needed). Deciphering all the underlying regulation mechanisms and expressing them in terms of mathematical rules requires great amounts of information and a deep knowledge of the system which is not always at hand. However, nowadays with the advent of high-throughput technologies, the field of systems biology has amassed plentiful omics data, which supposes a great opportunity for metabolic modelling. Since the results displayed by experimental measurements are a consequence of the internal regulation of the system, combining those data with metabolic models can aid to implicitly consider their effects. Furthermore, sets of data obtained under different genetic and/or environmental conditions may show different behaviour. Thus, available data can be included within metabolic simulations to heighten their predictive capabilities and their plasticity when dealing with perturbed conditions.

In this work, a tool is proposed that uses a multi-objective evolutionary optimisation algorithm to perform constraint-based steady-state flux simulations of genome-scale metabolic networks. This tool aims to offset some of the limitations found in classic algorithms, with the final goal of increasing predictability and improving accuracy of simulations.

Due to the multi-objective character of the proposed tool, it allows for considering multiple, potentially competing objectives, which aligns better with the situation occurring with adaptive evolution. In addition, simultaneous optimisation of various metabolic functions allows for the analyses of their interdependences. As a result of the optimisation process, instead of obtaining a single optimal flux distribution, as it is the case of FBA, a set of solutions is achieved, each of them describing different trade-offs between objectives. These solutions, in general, will appear to be suboptimal for single objectives considered separately, but they might match with more realistic flux distributions.

On the other hand, this multi-objective optimisation algorithm is based on differential evolution, and so it is flexible to admit non-linear constraints and objectives. The way constraints are handled during the optimisation process eases the definition of restrictions, that can be described in a simple way without needing intricate or cryptic mathematical definitions.

Furthermore, this tool gives a framework to include experimental information that can be used throughout the simulation process to tune the results. Using this tool, as it will be described in following sections, measured metabolic fluxes are included in a way that they do not hardly restrain the solution space, but they favour the appearance of solutions that quantitatively match the experimental measurements.

The following sections of this chapter are devoted to thoroughly describe the applied strategy, and to explain in detail the implementation of the algorithm. A broad example of application of this simulation tool is later addressed in Chapter 6.

5.3 Multi-objective metabolic flux analysis

As it was explained in Chapter 2, in order to implement a successful multi-objective approach, three fundamental steps are required: the multi-objective problem (MOP) definition, the multi-objective optimisation process, and the multi-criteria decision-making (MCDM) stage. In the context of constraint-based flux simulations, these three steps will extend the phases of metabolic flux analysis that were shown in Figure 1.1 (Chapter 1, page 18). Starting from the stoichiometry of the metabolic network, previously carefully reconstructed and mathematically represented, the multi-objective metabolic flux analysis stage comprises the following steps (Figure 5.1):

1st) During the **MOP definition**, a set of constraints must be identified that ensure the solutions remaining at the feasible space appropriately symbolise practical metabolic phenotypes. Besides, suitable objective functions, that represent biologically relevant objectives, must be defined. Together, constraints and objectives must guarantee the pertinency (in this context, meaning the validity of the solutions from a metabolic point of view) of the obtained solutions.

2nd) In the **multi-objective optimisation process** an appropriate algorithm is used to find the set of solutions that form the best Pareto front approximation, given the defined MOP. This algorithm must include the necessary mechanisms to ensure convergence, diversity and pertinency of the solutions of the obtained Pareto front approximation, and to deal with the specific properties of the optimisation problem (constraints, large scale and multi-modality[16]). In the present work a multi-objective evolutionary algorithm (MOEA) is used. It is based on the sp-MODE[17] algorithm defined by Reynoso-Meza et al. (2010) (see Section 2.4.1 for a description).

3rd) In the **MCDM** stage the solutions obtained from the optimisation algorithm are visualised and examined. Among them, those that are thought to best represent the actual metabolic behaviour of the system, are selected. These selected solutions can then be analysed to contribute to further comprehension of the metabolic functioning of the system under study. Interestingly, multiple, slightly different, solutions that represent equally Pareto-optimal metabolic phenotypes can be selected in this process, thus enriching the results of these simulations.

5.3.1 Addressing special features of flux simulations

When applying a multi-objective evolutionary algorithm to determine flux distributions at a genome scale the optimisation problem that arises has some special characteristics that must be considered. First, it will be a large-scale problem, since the decision variables set is the set of all reaction rates, typically around 1000 (in some cases even more than

[16] A multi-modal optimisation problem is that in which different decision vectors, that is vectors of variables, lead to the same objective vector

[17] Tool available at `http://www.mathworks.com/matlabcentral/fileexchange/39215`

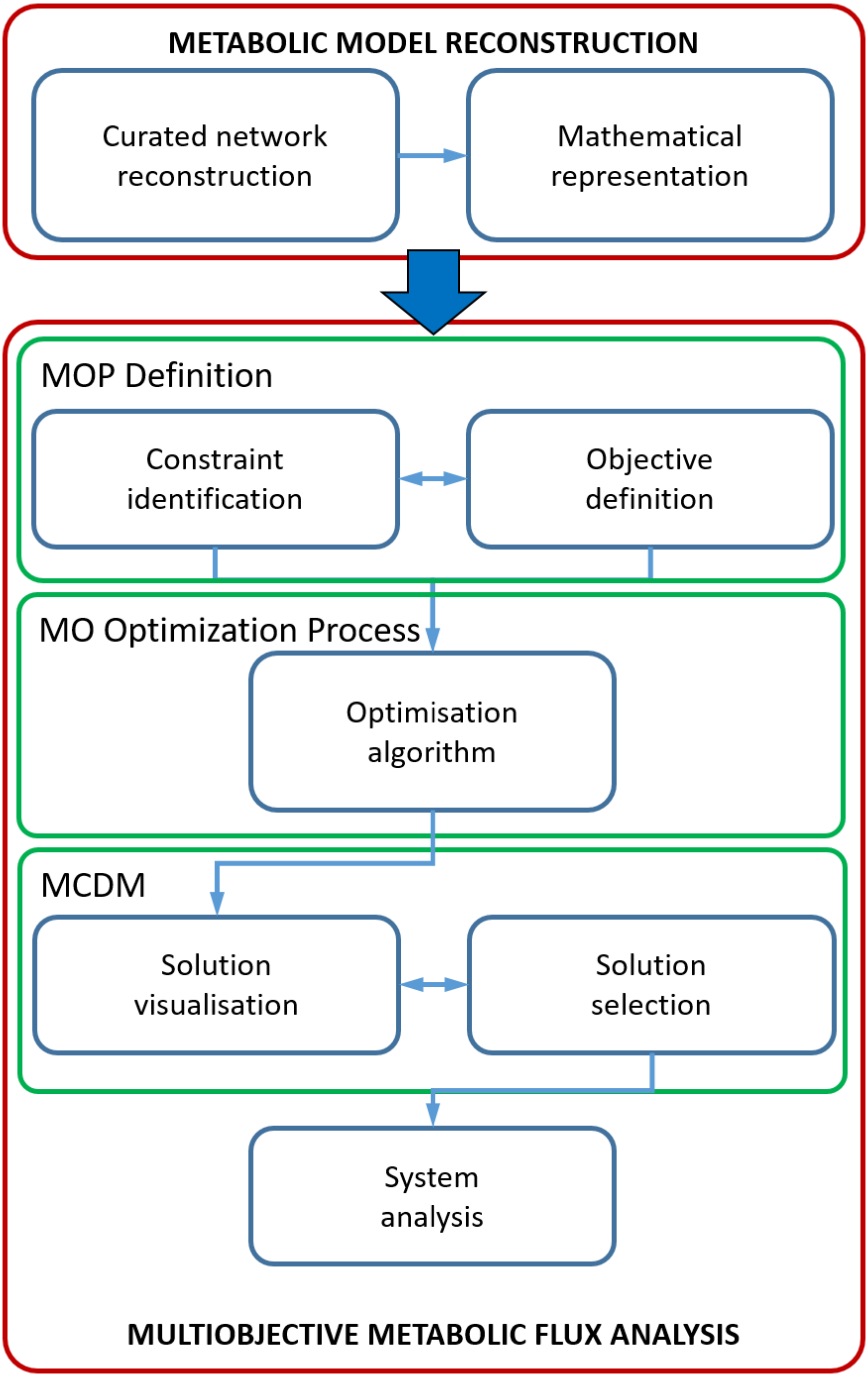

METABOLIC MODEL RECONSTRUCTION
Curated network reconstruction
Mathematical representation
MOP Definition
Constraint identification
Objective definition
MO Optimization Process
Optimisation algorithm
MCDM
Solution visualisation
Solution selection
System analysis
MULTIOBJECTIVE METABOLIC FLUX ANALYSIS

Figure 5.1: *Extension of metabolic network constraint-based analysis by means of multi-objective optimisation.*

2000) when working at genome scale (Erdrich et al., 2015). It will be also a strongly constrained problem, even if the amount of boundary constraints is minimised, as the steady state condition sets out one equality constraint per metabolite, usually from several hundreds to more than one thousand. Besides, due to the diversity of internal (often redundant) pathways, different combinations of internal fluxes are possible for an identical output, even when an absolute optimum is considered (alternate optima (Mahadevan and Schilling, 2003; Schuetz et al., 2007)), which makes the problem also multi-modal. But not every combination is valid from the point of view of cellular metabolism: several solutions can describe a mathematically feasible and optimal distribution which is not biologically meaningful. Therefore, it is also important to consider the pertinency of the resultant flux distributions.

In order to address these particular features, specific mechanisms have to be considered during the MOP definition and the MO optimisation process to ensure both that the optimisation algorithm can deal properly with the problem at hand, and that the results it generates are practical for the purpose at hand.

Parsimony

It is common that living organisms exhibit several redundant pathways in their metabolic networks, which gives them plasticity to adapt to and survive under different environmental conditions (Güell et al., 2014). When using these metabolic networks for optimisation-based simulations this redundancy leads to alternate solutions with different internal flux distributions, even when boundary restrictions are applied to input and output reactions. This effect is magnified when stochastic procedures are used to generate candidate solutions. In particular, flux landscapes may appear with excellent nutrient consumption and product yield features that include unrealistically high fluxes at internal cyclic pathways. In order to avoid such undesirable solutions and to facilitate the appearance of biologically meaningful flux distributions, the principle of parsimony is applied in this work: the optimisation process is driven so that flux vectors with lower overall metabolic activity are favoured.

In terms of biological meaning, strains that require lower overall flux through the metabolic network represent a parsimonious enzyme usage, which, in the end, means reduced expenses for enzyme synthesis (Schuetz et al., 2007; Lewis et al., 2010).

The inclusion of a parsimonious criterion during the optimisation helps both to ensure the obtained solutions are pertinent, and to meliorate internal distribution of the fluxes through pathways, avoiding futile cycles and unproductive extremely high fluxes.

Closeness to experimental fluxes

Nowadays, with high-throughput technologies, the field of systems biology has gained access to sets of physiological data that were not achievable in former times. ^{13}C-based metabolic flux analysis studies allow for determining *in vivo* sets of metabolically steady-state reaction rates that can be combined with metabolic models to heighten the predictive capabilities of computational simulations and their plasticity when dealing with perturbed conditions. The algorithm described here enables the inclusion of experimental fluxes within the optimisation process in such a way that the calculated fluxes are close to the experimental ones.

But the amount of metabolic fluxes that can be obtained from *in vivo* experiments is limited, and much smaller than the number of variables to determine. The Meta-MODE algorithm, that will be described in Section 5.5.5, uses a few experimental values to adjust the whole set of fluxes. This mechanism provides a way to deal with multi-modality and improves the pertinency of the solutions, as the flux landscapes obtained must be coherent with the biological reality described by the experimental data. Therefore, the resultant flux vectors will better approximate the real pathway distribution.

Pairwise flux ratios

Organisms and cells have evolved a variety of mechanisms to modulate reaction rates through metabolic pathways. Among them, it is well known that cell control systems often detect and adjust the relative flux proportions around certain metabolic hubs (Metallo and Vander Heiden, 2013). Important examples of ratios that can be used to characterise the physiologic state of a cell are the P/O (Phosphate/Oxygen) ratio, the ratio between NADH and ATP synthesis, or the ratio between linear and cyclic electron flow in photosynthetic organisms (Dong and Wei, 2004; Kwon et al., 2013). This kind of information is sometimes obtainable through specific experimental setups and it would be very valuable to regulate the flux response of a metabolic network.

However, a ratio between two fluxes is not a linear relationship and so there is no straightforward way to include this information when using linear optimisation. Researcher could use iterations of constraint-based simulations in order to force a given ratio between two reactions, but such strategies imply a methodological artifice that frequently reduces the plasticity of the model, highly limiting its predictive value. In contrast, the use of evolutionary optimisation techniques facilitates the definition of non-linear constraints and objectives, thus allowing the incorporation of flux ratios to improve the metabolic response of the simulations. In Meta-MODE, the ratio between the flux of a pair of reactions can be defined, and the solutions will be selected such that the proportion between these fluxes respects the given value, without enforcing individual values for each reaction.

Anchor points

During an evolutionary optimisation process successive populations of solutions are generated from their predecessors, evaluated, and selected so that they optimise the defined objectives. To start this process, an initial population is needed. Typically, the initial population is generated stochastically, selecting random individuals from the decision space. However, in the case of flux simulations, given the large amount of variables, and the strong constraint imposed by the steady state, the proportion of feasible solutions is quite low, so it is not to be expected that randomly selected flux vectors satisfy this condition. To deal with this situation, anchor points (Definition 5.1) were added to the initial set of solutions.

Definition 5.1 (Anchor points). *Given the multi-objective optimisation problem stated at Equation* (2.1), *anchor points* ($\boldsymbol{v}^{*k}$) *are points that correspond to the optimal value of the individual objectives in the feasible space. That is, those that are solution of the problem*

$$\min_{\boldsymbol{v}} Z_k\left(\boldsymbol{v}\right), \quad k \in \{1, 2, \ldots, q\}$$

subject to

$$\begin{aligned} g\left(\boldsymbol{v}\right) &\leq 0 \\ h\left(\boldsymbol{v}\right) &= 0 \\ l_{v_j} \leq v_j &\leq u_{v_j} \qquad j \in \{1, \ldots, n\} \end{aligned}$$

Thus, anchor points are solutions of the problem (since they are inside the feasible space) that are optimal for individual objectives. As they are solutions they satisfy all the constraints, so they form a starting set of feasible flux vectors; and they are optimal for individual objectives, so they will be at the ending parts of the Pareto front. The addition of anchor points among the initial population, together with other random vectors, improves the convergence to feasible solutions and decreases the computational cost of the algorithm.

5.4 MOP definition for metabolic flux simulation

In this thesis, an approach is proposed based on multi-objective optimisation to perform constraint-based metabolic simulations. Taking into account what was introduced in Section 2.2 about MOPs, and what was explained in Section 1.2.1 about constraint-based flux analysis, the MOPs stated for constraint-based metabolic simulations will be of the form:

$$\min_{\boldsymbol{v}} \boldsymbol{Z}\left(\boldsymbol{v}\right) = \left[Z_1\left(\boldsymbol{v}\right), Z_2\left(\boldsymbol{v}\right), \ldots, Z_q\left(\boldsymbol{v}\right)\right] \tag{5.1}$$

subject to:

$$\boldsymbol{S}\cdot\boldsymbol{v} = \boldsymbol{0} \tag{5.2}$$

$$v_{j,rev} \in (-\infty, +\infty) \qquad j \in \{1, \ldots, n\} \tag{5.3}$$

$$v_{j,irr} \in [0, +\infty) \qquad j \in \{1, \ldots, n\} \tag{5.4}$$

$$l_{v_j} \leq v_j \leq u_{v_j} \qquad j \in \{1, \ldots, n\} \tag{5.5}$$

$$\boldsymbol{g}\left(\boldsymbol{v}\right) \leq 0 \tag{5.6}$$

$$\boldsymbol{h}\left(\boldsymbol{v}\right) = 0 \tag{5.7}$$

where:

$\boldsymbol{Z}\left(\boldsymbol{v}\right)$ is the vector of objective functions,
$\boldsymbol{S}$ is the stoichiometric matrix,
$\boldsymbol{v}$ is the vector of fluxes,
$v_{j,rev}$ are the fluxes of the reversible reactions,
$v_{j,irr}$ are the fluxes of the irreversible reactions,
l_{v_j} is the lower flux bound for reaction j,
u_{v_j} is the upper flux bound for reaction j,
$\boldsymbol{g}\left(\boldsymbol{v}\right)$ are other inequality constraints[18], and
$\boldsymbol{h}\left(\boldsymbol{v}\right)$ are other equality constraints[19].

As it has been stressed in this dissertation, the definition of constraints and objectives that properly describe the limitations and challenges that cells must face during evolution under selective pressure is crucial to ensure the biological meaningfulness of the solutions that will

[18] Like for example closeness or ratio constraints defined later in this chapter.

[19] Like for example in the case of fixing biomass production rate to experimental value of growth rate.

be obtained later at the optimisation process. In this section, the constraints and objectives proposed in this work to state appropriate multi-objective problems for constraint-based metabolic simulations are described.

5.4.1 Definition of metabolic constraints

As most of the optimisation problems, flux simulations of genome-scale metabolic networks require considering constraints. The most common kind of constraints at any optimisation problem are the bound constraints, which define the boundary of the decision space. In the case of metabolic steady-state flux simulations, apart from lower and upper bounds for each reaction flux, it is essential to ensure each metabolite's mass balance . Besides, other constraints can be convenient to improve pertinency of the obtained solutions. Below, the different kinds of constraints pre-defined in Meta-MODE are described.

Flux bounds

Flux bounds that appear in constraint-based metabolic flux analysis mainly comprise two types:

- directionality bounds (Equations (5.3) and (5.4)), that establish that reversible reactions can take negative values (meaning flux in the opposite direction to the one described by the stoichiometric equation), while irreversible reactions can only have positive values. In practice, in numerical simulations, numbers much higher than normal flux values (usually 100 or 1000) are used to represent ∞ (see Section 4.4.2 in Chapter 4 for normal ranges of metabolic fluxes).

- capacity and availability bounds (Equation (5.5)), that are imposed only at certain reactions, generally exchange reactions, and account for maximum enzyme/transport capacities and environmental nutrient availability.

Mass balance under steady state

Equation (5.2) establishes a steady-state mass balance around every internal metabolite in the network. These steady-state balances become equations that set out one equality constraint per metabolite.

There exist some metabolites considered *external*, which means they are out of the system boundaries. Here the word "system" refers to the mathematical representation of the cell, not the physical cell. This way, *external metabolites* don't have to be necessarily secreted (or incorporated from) outside the cell membrane. They are metabolites that are excluded from the mass balance, and thus they can accumulate or be consumed. Consideration of such especial metabolites is fundamental to describe inputs, outputs, accumulation of substances (such as reserve substances) and net production of components, like biomass.

Closeness to experimental fluxes as constraints

Previously, in Section 5.3.1, it has been pointed how inclusion of *in vivo* measured sets of fluxes can help to improve the quality of the obtained solutions. One way to include this information is as an optimisation constraint. In order to ensure the calculated fluxes are close to the experimental ones, different ways to measure the divergence between them have been defined in the algorithm.

One option is to apply the concept of deviation. First, individual deviations between each experimental value and the corresponding calculated value at a given flux vector are evaluated as relative deviations with respect to the experimental value:

$$\delta_e = \frac{\left|\boldsymbol{v}_e^{exp} - \boldsymbol{v}_e^{calc}\right|}{\left|\boldsymbol{v}_e^{exp}\right|}, \qquad e \in E \tag{5.8}$$

where E is the set of reactions with experimental flux measurements.

To represent these individual deviations as a single value, several options are available among the simulation settings: norm of the vector

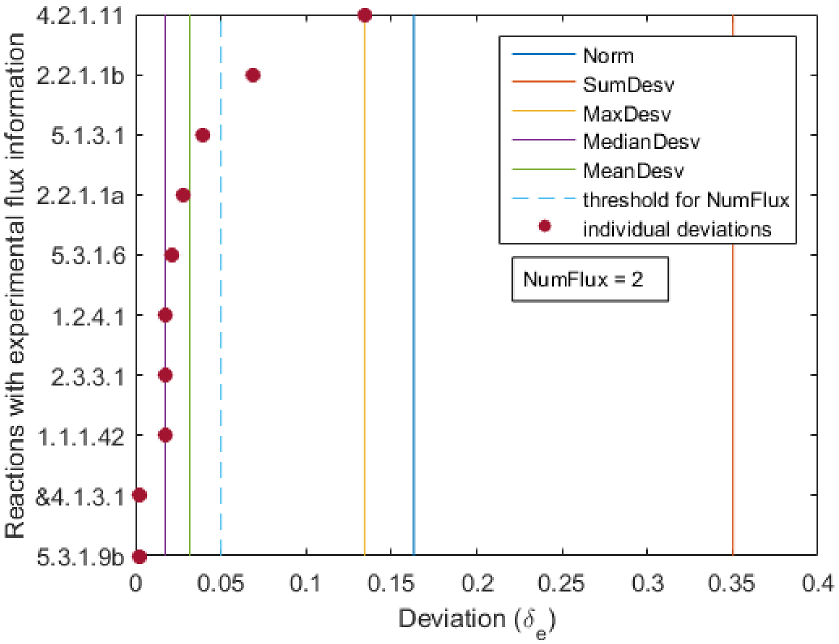


Figure 5.2: *Behaviour of the criteria for the selection of a value of deviation from individual deviations with respect to the experimental fluxes. The options available in Meta-MODE are: Norm: norm of the vector of individual deviations; SumDesv: cumulative sum of the individual deviations; MaxDesv: maximum of the individual deviations, MedianDesv: median of the individual deviations; MeanDesv: mean of the individual deviations; NumFlux; number of fluxes with deviation greater than a given threshold.*

δ, cumulative sum of the relative deviations, or mean, median or maximum of the relative deviations, as well as the number of fluxes with deviation greater than a given threshold. .An example of the behaviour of the different criteria is shown in Figure 5.2.

The other option available in the algorithm to calculate the closeness between experimental and calculated flux values uses a similarity coefficient based on the standard fuzzy metric. Prior to define the proposed metric, it is necessary to introduce some previous definitions to explain what is the standard fuzzy metric.

Definition 5.2 (Fuzzy metric space in the sense of George and Veeramani (George and Veeramani, 1994)). *The 3-tuple $(X, M, *)$ is said to be a fuzzy metric space if X is a non-empty set, $*$ is a continuous t-norm and M is a fuzzy set on $X \times X \times]0, +\infty[$ satisfying the following conditions:*

$$M(x, y, t) > 0 \tag{5.9}$$

$$M(x, y, t) = 1 \text{ if and only if } x = y \tag{5.10}$$

$$M(x, y, t) = M(y, x, t) \tag{5.11}$$

$$M(x, y, t) * M(y, z, s) \leq M(x, z, t + s) \tag{5.12}$$

$$M(x, y, \cdot) :]0, +\infty[\to [0, 1] \text{ is left continuous} \tag{5.13}$$

$x, y, z \in X$ and $t, s > 0$

Definition 5.3 (Fuzzy metric). *(George and Veeramani, 1994) If $(X, M, *)$ is a fuzzy metric space, then $(M, *)$ is said to be a fuzzy metric in X.*

Definition 5.4 (Standard fuzzy metric). *(George and Veeramani, 1994) Let (X, d) be a metric space. Define $a * b = ab$ for any $a, b \in [0, 1]$. Let M_d be the function defined in the set $X \times X \times]0, +\infty[$ as follows:*

$$M_d(x, y, t) = \frac{t}{t + d(x, y)} \tag{5.14}$$

Then, the function M_d is a fuzzy metric induced by the metric d. M_d is called the standard fuzzy metric.

As the authors remark in George and Veeramani (1994), a fuzzy metric $M(x, y, t)$ defined according to Definitions 5.2 and 5.3 can be thought of as the degree of nearness between x and y with respect to t. It is identified $M(x, y, t) = 1$ with $x = y$ (identical points), and $M(x, y, t) = 0$ with ∞ (infinitely separated points). Thus, a fuzzy metric constitutes a good measure of the closeness between two points, in this case, two fluxes, with respect to a certain reference value t.

In this work a metric is defined based on the standard fuzzy metric to measure closeness between experimental and calculated fluxes taking as d the Euclidean distance between the fluxes, and as t the experimental flux (Equation (5.15)).

$$M(\boldsymbol{v}_e^{calc}, \boldsymbol{v}_e^{exp}, |\boldsymbol{v}_e^{exp}|) = \frac{|\boldsymbol{v}_e^{exp}|}{|\boldsymbol{v}_e^{exp}| + |\boldsymbol{v}_e^{exp} - \boldsymbol{v}_e^{calc}|} \tag{5.15}$$

Given equations (5.8) and (5.15) the relation between δ and M is:

$$M = \frac{1}{1+\delta} \qquad \text{and} \qquad \delta = \frac{1-M}{M} \tag{5.16}$$

It is important to take into account that M accounts for similarity while δ accounts for deviation, so they must be treated oppositely: in the case of M the aimed values will be high (high similarity), while when considering δ low values will be pursued (low deviation).

Other user-defined distances or metrics can be easily included to the algorithm to describe closeness between experimental and calculated fluxes. The selection of one particular method depends on the purpose and characteristics of the given simulation.

Closeness to experimental fluxes can be used both as a constraint, establishing an upper limit for the admissible deviation (lower limit for the desired closeness), or as an objective, aiming at the minimum possible deviation (maximum possible closeness).

When closeness to experimental fluxes is used as a constraint, to obtain a single value that represents the set of deviation/closeness values, it is preferable to use criteria that focus on individual values, such as maximum, minimum or number of fluxes trespassing a threshold, rather than norm, sum, median or mean which refer to the set as a whole.

Pairwise flux ratios as constraints

As discussed in Section 5.3.1, including previous knowledge about reactions that are known to keep a proportion between their relative rates in nature can also enhance pertinency of the resulting flux distributions. Those proportions (or ratios) can be incorporated into the simulation as constraints. To have a measurement of how similar are the calculated ratios to the stated ones, again the concepts of relative deviation (Equation (5.18)) or similarity (Equation(5.19)) can be applied.

$$\rho^{calc}_{r1,r2} = \frac{\boldsymbol{v}^{calc}_{r_1}}{\boldsymbol{v}^{calc}_{r_2}} \qquad r_1, r_2 \in P \tag{5.17}$$

$$\delta_{r_1,r_2} = \frac{\left|\rho^{exp}_{r1,r2} - \rho^{calc}_{r1,r2}\right|}{\left|\rho^{exp}_{r1,r2}\right|}, \qquad r_1, r_2 \in P \tag{5.18}$$

$$M_{r_1,r_2} = \frac{\left|\rho^{exp}_{r1,r2}\right|}{\left|\rho^{exp}_{r1,r2}\right| + \left|\rho^{exp}_{r1,r2} - \rho^{calc}_{r1,r2}\right|} \qquad r_1, r_2 \in P \tag{5.19}$$

where P is the set of reactions with information about ratios.

Like in the case of experimental fluxes, researchers can define other metrics to account for divergence or nearness between calculated and experimental ratios. Also, the information about ratios can be applied both as constraint and as objective.

5.4.2 Definition of metabolic objectives

There exist different objective functions that have been frequently used in the field of systems biology to perform constraint-based optimisations of genome-scale metabolic networks. Among the most popular are the maximisation of biomass (or other extracellular products) yield, the minimisation of the overall intracellular flux, the minimisation of nutrient consumption and the minimisation of ATP production (Schuetz et al., 2007). When applied independently to a mono-objective (linear or non-linear) optimisation problem, most of these objectives need additional constraints (growth rate limits, maintenance energy drains, bounds for some cofactor use ...) in order to obtain solutions that properly account for the behaviour of actual metabolic systems (Price et al., 2004; Schuetz et al., 2007). The algorithm described here allows the researchers to simultaneously optimise multiple competing objectives, thereby mimicking how organisms operate. Using this strategy, the amount of further constraints can be reduced which gives more plasticity to the simulations and avoids overfitting.

Various kinds of objectives are available at Meta-MODE that are hereby described in detail.

Maximisation/minimisation of single reaction flux

This is the simplest type of objective that can be set in Meta-MODE. Examples of application are the minimisation of ATP production (aiming at an efficient use of energy (Savinell and Palsson, 1992a; Schuetz et al., 2007)) or the maximisation of a desired product secretion (when analysing producing capacities). In this case the cost associated to the objective is the flux of the corresponding reaction, with the appropriate sign: positive for minimisation and negative for maximisation.

$$Z_{react_t}(\boldsymbol{v}) = \boldsymbol{v}_t \ , \quad t \in \{1, \ldots, n\} \tag{5.20}$$

Product of two reactions

Even when a multi-objective scheme is proposed, sometimes it can be interesting to consider competing objectives simultaneously, for example if they are required to be coupled. Such is the case of the coupled yield of two reactions (typically biomass and some product of interest) (Patil et al., 2005; Montagud et al., 2011). In this case the product of the two reactions of interest is taken as the corresponding cost (again negative sign is established for maximisation and positive for minimisation).

$$Z_{prod_{j_1,j_2}}(\boldsymbol{v}) = \boldsymbol{v}_{j_1} \times \boldsymbol{v}_{j_2} \ , \quad j_1, j_2 \in \{1, \ldots, n\} \tag{5.21}$$

Growth

When linear optimisation is used to perform metabolic simulations, in order to account for growth, it is necessary to describe a biomass formulation based on a fixed linear combination of metabolites. Even when multiple biomass equations can be defined to try to consider different physiologic situations, at the moment of the simulation only one can be used as either objective or constraint. This fact makes the simulation to be tightly constrained by the chosen biomass composition.

In this context, one of the great advantages of using an evolutionary optimisation algorithm is that non-linear objectives can be considered, just like that of non-rigid biomass formation.

At Meta-MODE growth is considered as a pool of biomass components, with no pre-defined proportion among them. In order to maximise the growth yield, several drains of individual biomass precursors are defined as separated reactions, and the unconstrained sum of fluxes of those individual reactions, with negative sign, is established as the cost associated to this objective:

$$Z_{growth}\left(\boldsymbol{v}\right) = -\sum_{j} v_j \ , \quad j \in B \tag{5.22}$$

where B is the set of reactions describing individual biomass precursor drains.

This way, the algorithm allows the simulation to evolve biomass components in an unrestricted way. When experimental fluxes are used during the simulation process, either as objective or as a constraint, these values will influence the production fluxes of biomass elements. Therefore, the algorithm allows the study of the proportion of those biomass components under different growth conditions, as they are results of the simulation, instead of pre-established constraints.

Parsimony as objective

The technical and biological importance of including a parsimonious criterion during the optimisation process was already justified in Section 5.3.1. It helps to avoid solutions containing extremely high fluxes or mathematical artefacts with no biological meaning, thus improving the pertinency of the obtained solutions.

The parsimony criterion was introduced as a design objective. This objective aims at a low overall metabolic activity, that is, low values of the sum of reaction rates. To do so, the total sum of the absolute value of fluxes was taken as the associated cost. Since some of the reactions are described as reversible, they can have negative fluxes, which is why the absolute value of the fluxes was considered and not their sign, since the sign only designates the direction of the reaction and the reaction rate is related only to the module.

$$Z_{pars}\left(\boldsymbol{v}\right) = \sum_{j=1}^{n} |v_j| \tag{5.23}$$

When the algorithm tries to minimise this cost, it is trying to minimise the overall sum of reaction rates.

Closeness to experimental fluxes as objective

As mentioned in the previous section, closeness of the calculated flux values to the experimental ones can be used as a constraint or as a design objective (or both). When using it as an objective, the corresponding cost will consider the deviation δ (or the similarity M) between both sets and greater costs will be assigned to greater deviations (lower similarities).

$$Z_{dev}\left(\boldsymbol{v}\right) = D \tag{5.24}$$

$$Z_{clos}\left(\boldsymbol{v}\right) = -C \tag{5.25}$$

where D and C are the unidimensional values chosen to represent, respectively, the deviation or closeness between two sets of experimental and calculated fluxes. As explained before, different options are available at the algorithm to represent all the individual deviations as a single value: norm of the vector, cumulative sum of coordinates, maximum, minimum, mean, or median of coordinates, as well as the number of fluxes trespassing a given threshold (Figure 5.2). When closeness to experimental fluxes is used as an objective, it is preferable to use criteria that better capture the nature of the entire set of values, such as norm, sum, median or mean, rather than max, min or number of fluxes beyond a given threshold, that focus more on individual values.

Pairwise flux ratios as objectives

Like in the case of sets of measured fluxes, information about *in vivo* flux ratios can help tuning the response of the model to better reproduce the real metabolic behaviour of the organism under study. Flux ratios can be also included as optimisation constraint, as design objective, or both as constraint and objective. If applied as objective, either deviation (δ) or similarity (M) between the calculated and the experimentally measured ratios can be considered:

$$Z_{rat\_dev}\left(\boldsymbol{v}\right) = \delta_{r_1,r_2} \tag{5.26}$$

$$Z_{rat\_sim}\left(\boldsymbol{v}\right) = -M_{r_1,r_2} \tag{5.27}$$

where δ_{r_1,r_2} is the relative deviation calculated according to Equation (5.18) and M_{r_1,r_2} is the similarity calculated by means of the standard fuzzy metrics as stated in Equation (5.19). Thus, greater discrepancies are allowed, but with higher costs.

5.5 Multi-objective optimisation process: the Meta-MODE algorithm

In the previous section (5.3), the required constraints and pursued objectives applicable to constraint-based flux simulations where described in detail. Thus, the multi-objective problem was formulated, the next step is to describe how the optimisation process is conducted.

In this work a Multi-Objective Evolutionary Algorithm (MOEA) is applied to solve the MOPs defined according to the previous section. This algorithm is based on the sp-MODE[20] algorithm (Reynoso-Meza et al., 2010) that was introduced in Section 2.4.1. However, sp-MODE algorithm by itself can't deal properly with the particular problem of metabolic simulations. As commented before, this problem requires an algorithm with the following properties:

- *Convergence* to the (unknown) Pareto front.
- *Diversity* of solutions along the Pareto front approximation.
- Metabolic *pertinency* of the solutions at the Pareto front approximation.
- Mechanisms to deal with a *constrained* problem.
- Mechanisms to deal with a *large-scale* problem.
- Mechanisms to deal with a *multi-modal* problem.

The tuning parameters in Tables 5.1 and 5.2 are used in order to improve the convergence of the algorithm. Furthermore, as explained in Section 5.3.1, anchor points (Definition 5.1) are incorporated within the initial population to improve convergence. The parameters presented, together with the inclusion of anchor points within the initial population also allow that the algorithm addresses the large-scale problem. Diversity is addressed with the spherical pruning mechanism

[20] Tool available at `http://www.mathworks.com/matlabcentral/fileexchange/39215`

explained in Section 2.4.1: the space covered by the Pareto front approximation is divided in spherical sectors and only one solution is selected from each sector (Figure 2.3, page 59).

Table 5.1: *Tuning guidelines for DE's parameters.*

Parameter	Value	Comments
	DE algorithm	
F (Scaling factor)	0.5	Recognized as good initial choice . according to Storn and Price (1997)
Cr (Crossover rate)	1.0	Value recognized for highly non-separable problems according to Reynoso-Meza et al. (2011) and Das and Suganthan (2010).
N_p (Population size)	50	50 individuals is proposed by default (Reynoso-Meza et al., 2010).

Table 5.2: *Tuning guidelines for the pruning mechanism as described in Reynoso-Meza et al. (2017).*

Parameter	Value	Comments
	Spherical pruning mechanism	
$\boldsymbol{\beta_\epsilon}$ (Arcs)	100	It has been proposed for bi-objective problems, to bound the approximated Pareto front to 100 design alternatives.
	$[10, 10]$	It has been proposed for 3-objective problems, to bound the approximated Pareto front to $10^2 = 100$ design alternatives.
	$\overbrace{[m, \ldots, m]}^{m-1}$	It has been proposed for m-objective problems, to bound the approximated Pareto front to m^{m-1} design alternatives.
p (p-norm)	1	It has been proposed as default value.

Even when, given the characteristics of the problem at hand, convergence and diversity can be achieved by adjusting the parameters of the algorithm, that is insufficient for the other points. Metabolic pertinency of solutions is improved by stating a multi-objective constrained optimization problem, applying the constraints and objectives commented above. For this purpose, a penalty scheme is used, with the guidelines stated in Reynoso-Meza et al. (2012). Particular constraints and objectives proposed in this work to improve metabolic pertinency are closeness to experimental fluxes, pairwise flux ratios and parsimony. These mechanisms also help to deal with the multi-modal problem, as explained in Section 5.3.1.

The current section is devoted to describe these mechanisms introduced in this work to deal with the specific problem of genome-scale constraint-based metabolic flux analysis.

5.5.1 Mechanisms to improve metabolic pertinency

Parsimony

Two approaches are included in the proposed algorithm to improve parsimony of the solutions: on the one hand, parsimony is introduced as objective, as described in previous section (Equation (5.23)); on the other hand, a routine has been incorporated in the algorithm that filter out spurious flux loops (Algorithm 5.1). In order to cut down computational efforts, a list of previously identified reactions that may lead to flux loops (reactions with opposite sense and same metabolites) is given to the algorithm as part of the metabolic network description.

Closeness to experimental fluxes

As explained in the previous section, similarity between experimental and calculated fluxes is considered by the algorithm either in terms of deviation or closeness. In both cases, individual deviations (Equation

```
Input: simulation parameters and vector of calculated fluxes $\boldsymbol{v}$
Output: vector of calculated fluxes with simplified/cleaned
        loops $\boldsymbol{v}'$
Read list of potential loops ;          /* from parameters */
for each potential loop do
    Read list of reactions involved in the loop;
    Check reactions' direction and reversibility;
    Extract from $\boldsymbol{v}$ reactions' flux values;
    Calculate net flux though the loop $\phi$;
    if $\phi = 0$ then
        Assign null flux to all reactions involved in the loop;
    else
        Identify reaction $t$ of the loop with higher activity;
        Assign flux $\phi$ to reaction $t$;
        Assign flux $\phi$ to reactions up and downstream from
          reaction $t$ needed to satisfy mass balance;
        Assign null flux to the rest of reactions involved in the
          loop;
end
Update $\boldsymbol{v}$ with the new flux distribution to create $\boldsymbol{v}'$;
return $\boldsymbol{v}'$
```

Algorithm 5.1: Look for loops function

(5.8)) or closeness coefficients (Equation (5.15)) are calculated from a set of measured fluxes and their corresponding values in the flux vector. Different criteria are available in the algorithm to choose a single value that represents the vector of individual coefficients. Algorithms 5.2 and 5.3 show the process for deviation and closeness indexes respectively.

Pairwise flux ratios

Pairwise constraints and objectives are treated by the algorithm in a way similar to the closeness to experimental fluxes. In the case of ratios, individual single values of deviation or similarity are calculated according to Equations (5.18) and (5.19). In this case, those values are used directly to apply constraints and objectives based on flux ratios (see Sections 5.5.2 and 5.5.4 bellow).

5.5.2 Handling constraints

In the proposed algorithm, bound constraints and optimisation constraints are treated in a different way. Bound constraints directly define the limits of the space within which random solutions are generated by the evolutionary process. To deal with optimisation constraints the algorithm uses penalty functions that add additional costs, proportional to the constraint violation, to the objective function. The description of the specific processes and penalty functions defined in the algorithm to handle bounds and optimisation constraints can be found below.

Flux bounds constraints

Bounds for flux values accounting for reaction directionality and enzyme/transport capacity limits (see Section 5.4.1) , are provided as part of the simulation parameters. At every moment the algorithm ensures that the generated random solution vectors are within the space defined by these limits: individuals of the initial population are directly

```
Input: simulation parameters and vector of calculated fluxes $\boldsymbol{v}$
Output: deviation from experimental fluxes $D$
Read matrix containing the indexes of reactions with
  experimental measurements ($e \in E$) and the corresponding
  measured values;
for all reactions with experimental measurement ($e \in E$) do
    Calculate relative deviations (Equation (5.8));
end
switch deviation criterion do
    case Norm do                          /* norm of the vector */
        $D = \|\delta\|$;
    case Sum do                             /* cumulative sum */
        $D = \sum_{e \in E} \delta_e$
    case Max do                              /* maximum value */
        $D = \max_{e \in E}\{\delta\,(e)\}$;
    case Median do                      /* median of the values */
        $D = \operatorname{median}_{e \in E}\{\delta\,(e)\}$;
    case Mean do                          /* mean of the values */
        $D = \operatorname{mean}_{e \in E}\{\delta\,(e)\}$;
    case NumFlux do            /* # fluxes over threshold */
        Read deviation threshold $DevThr$;
        $S = \{e : \delta\,(e) > DevThr\,,\, e \in E\}$;
        $D = |S|$
end
return $D$
```

Algorithm 5.2: Deviation from experimental fluxes

```
Input: simulation parameters and vector of calculated fluxes $\boldsymbol{v}$
Output: closeness to experimental fluxes $C$

Read matrix containing the indexes of reactions with
  experimental measurements ($e \in E$) and the corresponding
  measured values;
for all reactions with experimental measurement ($e \in E$) do
    Calculate fuzzy closeness (Equation (5.15));
end
switch closeness criterion do
    case Norm do                         /* norm of the vector */
        $C = \|M\|$;
    case Sum do                              /* cumulative sum */
        $C = \sum_{e \in E} M_e$
    case Min do                               /* minimum value */
        $C = \min_{e \in E}\{M(e)\}$;
    case Median do                     /* median of the values */
        $C = \mathrm{median}_{e \in E}\{M(e)\}$;
    case Mean do                         /* mean of the values */
        $C = \mathrm{mean}_{e \in E}\{M(e)\}$;
    case NumFlux do          /* # fluxes under threshold */
        Read closeness threshold $ClosThr$;
        $S = \{e : M(e) < ClosThr\,,\ e \in E\}$;
        $C = |S|$
end
return $C$
```

Algorithm 5.3: Closeness to experimental fluxes

assembled inside this space, and at each generation the offspring individuals are checked and migrated within the bounds if needed.

Mass balance constraint

The steady-state mass balance stated in Equation (5.2) defines a set of equality constraints, one per each internal metabolite, that must be fulfilled. Algorithm 5.4 describes the penalty function used in MetaMODE to penalise mass balance inconsistencies.

```
  Input: simulation parameters and vector of calculated fluxes $\boldsymbol{v}$
  Output: mass balance constraint violation penalty $PenMassCon$
1 $PenMassCon = 0$;
2 Calculate the balance vector $\beta = S * \boldsymbol{v}$;
3 for each element in $\beta$ do
4     Compare the value of the balance to a given admissible error
        $\epsilon$;
5     if $|\beta_a| > \epsilon$ then
6         $PenMassCon = PenMassCon + |\beta_a|$
7 end
8 return $PenMassCon$
```

Algorithm 5.4: Penalty function for mass balance constraint violation

First, a null value is assigned to the constraint violation penalty (line 1), and then the vector containing the resulting fluxes for the individual balances is calculated (line 2). For each element of this vector, the balance is checked comparing the obtained value to a (small) admissible error (line 4). In the case that the mass balance is not fulfilled the absolute value of the discrepancy is added to the constraint violation penalty (line 6). The final value of the constraint violation penalty, which contains the sum of all discrepancies, is returned when the loop is completed (line 8).

Closeness constraints

Closeness constraints can be included in the simulation, by fixing a limit for the maximum admissible deviation or the minimum desired closeness. Algorithm 5.5 shows the penalty function to account for this kind of constraints.

```
Input: simulation parameters and vector of calculated fluxes $\boldsymbol{v}$
Output: closeness to experimental fluxes constraint violation
        penalty $PenClosenessCon$
1  if closeness constraints are applied then
2      switch selected metric do
3          case relative deviation (Equation (5.8)) do
4              Calculate deviation from experimental fluxes $D$
                 (Algorithm 5.2);
5              Compare the value with the admissible limit $MaxDev$;
6              if $D > MaxDev$ then
7                  $PenClosenessCon = D - MaxDev$;
8          case fuzzy closeness (Equation (5.15)) do
9              Calculate closeness to experimental fluxes $C$
                 (Algorithm 5.3);
10             Compare the value with the admissible limit $MinClos$;
11             if $C < MinClos$ then
12                 $PenClosenessCon = MinClos - C$;
13     end
14 else
15     $PenClosenessCon = 0$
16 end
17 return $PenClosenessCon$
```

Algorithm 5.5: Penalty function for closeness to experimental fluxes constraint violation

In the case that closeness constraints are considered, the first step is to choose the preferred metric to account for divergencies (line 2): pre-programmed metrics include relative deviation (Equation (5.8)) and fuzzy closeness (Equation (5.15)), but researchers can use their own metrics if needed. In any case, the value of the parameter (deviation or closeness) is calculated and compared with the admissible limit (lines 4 to 5 and 9 to 10). If the parameter exceeds the limit, the difference between them is taken as constraint violation penalty (lines 7 and 12). As expected, greater penalties are imposed to greater deviations (lower closeness).

If the criterion selected at Algorithm 5.2 (or 5.3) is *NumFlux* it will return the number of fluxes beyond a given threshold, and so the value for the $MaxDev$ ($MinClos$) parameter must be the maximum (minimum) number of fluxes accepted to violate the threshold.

As mentioned before, when closeness to experimental fluxes is used as a constraint, it is preferable to use deviation/closeness criteria that focus on individual values, such as *Max*, *Min* or *NumFlux*, rather than *Norm*, *Sum*, *Median* or *Mean* which refer to the set as a whole.

Ratio constraints

Ratio constraints are managed in a very similar way to closeness constraints, as it is shown in Algorithm 5.6.

5.5.3 Initial population

The first step of a multi-objective evolutionary optimisation process is always to assemble an initial population. As it was explained in Section 5.3.1, in Meta-MODE anchor points are inserted in the initial set of individuals to improve convergence and reduce computational cost. Anchor points (Definition 5.1) are solutions of the problem that are optimal for individual objectives. However, sometimes the objectives considered are quite complex, and algorithms can last too long to find the absolute optimum of each individual objective. In such cases, faster methods will be considered to search for very advantageous (even when not optimal) solutions instead. These solutions can be surpassed during

```
Input: simulation parameters and vector of calculated fluxes $\boldsymbol{v}$
Output: ratio constraint violation penalty $PenRatioCon$
if ratio constraints are applied then
    Compute ratio between calculated fluxes (Equation (5.17));
    switch selected metric do
        case relative deviation do
            Calculate deviation $\delta_{r_1,r_2}$ from experimental ratio (Equation (5.18));
            Compare the value with the admissible limit $MaxRatDev$;
            if $\delta_{r_1,r_2} > MaxRatDev$ then
                $PenRatioCon = \delta_{r_1,r_2} - MaxRatDev$;
        case fuzzy closeness do
            Calculate closeness $M_{r_1,r_2}$ to experimental ratio (Equation (5.19));
            Compare the value with the admissible limit $MinRatClos$;
            if $M_{r_1,r_2} < MinRatClos$ then
                $PenRatioCon = MinRatClos - M_{r_1,r_2}$;
    end
else
    $PenRatioCon = 0$
end
return $PenRatioCon$
```

Algorithm 5.6: Penalty function for ratio constraint violation

the optimisation process if better combinations are found, so not taking the absolute optima from the beginning does not exclude those optima from being present at the final Pareto set.

Algorithm 5.7 describes the general directions used in Meta-MODE to calculate anchor points: first the considered constraints are read from the set of simulation parameters (line 1). Then, for each objective, the algorithm checks its mathematical formulation and selects the appropriate algorithm to calculate a flux distribution optimal for the given objective (lines 3 and 4). For all the objectives, if additional non-linear constraints (closeness to experimental fluxes or pairwise flux ratios) are considered, the optimal flux distribution calculated initially must be recalculated so that the obtained flux vector fulfils all the imposed restrictions (lines 5 to 8).

```
   Input: simulation parameters
   Output: flux vectors of anchor points $\boldsymbol{v}^{*k}$, $\forall k \in \{1, 2, \ldots, q\}$
 1 Read constraints applied;
 2 for each objective $k$ do
 3     Check mathematical formulation of objective $k$;
 4     Calculate optimal flux distribution using the appropriate
         algorithm (see below) ;
 5     if closeness constraints are applied then
 6         Recalculate optimal flux distribution to fulfil closeness
             constraints (Algorithm 5.8, Section 5.5.3);
 7     if ratio constraints are applied then
 8         Recalculate optimal flux distribution to fulfil ratio
             constraints (Algorithm 5.9, Section 5.5.3);
 9 end
10 return $\boldsymbol{v}^{*k}$, $\forall k \in \{1, 2, \ldots, q\}$
```

Algorithm 5.7: Anchor points calculation scheme in Meta-MODE

The calculation is performed following this sequence, instead of ensuring constraint fulfilment from the beginning, because it results in a faster algorithm: preceding solutions are used as initial points for subsequent optimisations which speeds up the process. Next, the steps taken to calculate the initial optimal flux distribution depending on the objectives considered, and to recalculate the solutions to meet the applied non-linear constraints, are detailed.

Initial calculation of the optimal flux distribution

Since anchor points must be optimal for individual objectives, a mono-objective optimisation problem has to be solved for every objective. Depending on the nature of the objective in question an appropriate algorithm will be chosen.

Maximisation/minimisation of single reaction flux Once more, the simplest option is when the objective function is the maximisation or minimisation of a single reaction flux. In this case, linear programming is used to optimise the selected reaction, subject to bound constraints and mass balance constraints:

$$\min_{\boldsymbol{v}} \boldsymbol{Z}(\boldsymbol{v}) = \boldsymbol{v}_t \qquad t \in \{1, \ldots, n\} \tag{5.28}$$

$$\max_{\boldsymbol{v}} \boldsymbol{Z}(\boldsymbol{v}) = -\min_{\boldsymbol{v}} \boldsymbol{Z}(\boldsymbol{v}) = -\boldsymbol{v}_t \qquad t \in \{1, \ldots, n\} \tag{5.29}$$

subject to:

$$\boldsymbol{S} \cdot \boldsymbol{v} = \mathbf{0} \tag{5.30}$$

$$l_{v_j} \leq v_j \leq u_{v_j} \qquad j \in \{1, \ldots, n\} \tag{5.31}$$

Thus, the problem is equivalent to solve a FBA problem.

Growth When growth objective is considered, the aim is to maximise all the individual reactions that are part of the biomass components. Although non-fixed proportions among them are pursued, to calculate the anchor points linear programming is applied, maximising them all in a ratio of one to one.

$$\max_{\boldsymbol{v}} \boldsymbol{Z}(\boldsymbol{v}) = c^T \cdot \boldsymbol{v} \tag{5.32}$$

$$\boldsymbol{S} \cdot \boldsymbol{v} = \boldsymbol{0} \tag{5.33}$$

$$l_{v_j} \leq v_j \leq u_{v_j} \qquad j \in \{1, \ldots, n\} \tag{5.34}$$

with $c_t = 1 \ \forall t \in B$ and $c_t = 0 \ \forall t \notin B$, where B is the set of reactions describing individual biomass precursor drains.

Which is equivalent to solve an FBA problem, with objective function determined by the vector of weights c. This method is applied, instead of selecting some non-linear optimisation algorithm, because it accelerates the computation and the results obtained are good enough to serve as seeds for Pareto optimal growth solutions.

Parsimony The aim of the parsimony objective is to minimise the sum of fluxes through the metabolic network. If there were no flux bounds, the obvious solution would be the null vector, that is, no flux in any reaction. However, when bound constraints are applied, some intake fluxes may appear that have positive minima, so a flux vector must be found that has minimal activity but not zero. In this case, again a strategy based on linear programming is preferred due to its faster response: all non-reversible reactions are minimised with equal coefficients.

$$\min_{\boldsymbol{v}} \boldsymbol{Z}(\boldsymbol{v}) = c^T \cdot \boldsymbol{v} \tag{5.35}$$

subject to:

$$\boldsymbol{S} \cdot \boldsymbol{v} = \boldsymbol{0} \tag{5.36}$$

$$l_{v_j} \leq v_j \leq u_{v_j} \qquad j \in \{1, \ldots, n\} \tag{5.37}$$

with $c_t = 1 \ \forall t \notin R$ and $c_t = 0 \ \forall t \in R$, where R is the set of reversible reactions.

Only non-reversible reactions are minimised because minimising reversible reactions would result in maximising their negative flux (absolute value is not a linear function unless its domain is restricted to non-negative or non-positive numbers and thus it cannot be used as an objective function for linear programming).

Closeness to experimental fluxes The aim of closeness objective is to minimise the difference between the experimental and the calculated fluxes. The ideal solution would be whose fluxes would be identical to the experimental ones. But some problems can appear that make impossible to precisely fit all the values: numerical and computational precision errors, small experimental errors at the *in vivo* measured fluxes, and inconsistencies between the metabolic network used for ^{13}C-based flux determination and the genome-scale network used for simulation are the most common sources of disagreement. This fact makes it necessary to consider some small percentage error.

Taking this into account, the strategy used to calculate the anchor points for closeness objective is as follows: upper and lower bounds of the reactions with experimental fluxes are modified so that they take into account the experimental values ± the error; linear programming is used to optimise some function (in this case it is programmed to use the same objective function as for the Parsimony objective because this objective must always stay); if no solution is found that satisfy the constraints, the error is increased and the process repeated until a feasible solution is found. This way, an iterative process is used to find a flux distribution coherent with the experimental values, with the smallest bound error.

$$\min_{\boldsymbol{v}} \boldsymbol{Z}\left(\boldsymbol{v}\right) = c^T \cdot \boldsymbol{v} \tag{5.38}$$

subject to:

$$\boldsymbol{S} \cdot \boldsymbol{v} = \boldsymbol{0} \tag{5.39}$$

$$l_{v_j} \leq v_j \leq u_{v_j} \qquad j \in \{1, \ldots, n\} \tag{5.40}$$

$$\boldsymbol{v}_e^{exp} \times (1 - error) \leq \boldsymbol{v}_e \leq \boldsymbol{v}_e^{exp} \times (1 + error) \qquad e \in E \tag{5.41}$$

with $c_t = 1 \; \forall t \notin R$ and $c_t = 0 \; \forall t \in R$, where R is the set of reversible reactions. And where E is the set of reactions with experimental flux measurements.

Pairwise flux ratios When ratios are used as objective, the aim is also to minimise discrepancy between calculated values and a reference one. But in this case there is an important difference: no individual values must be forced, only their proportion. Thus, the strategy used for closeness is not suitable now, as it involved setting a fixed value for each flux. In this case, a non-linear optimisation was preferred to maintain the freedom of the individual values. MATLAB® *fmincon* interior point algorithm is used to solve the optimisation problem:

$$\min_{\boldsymbol{v}} \boldsymbol{Z}(\boldsymbol{v}) = Z_{rat} \tag{5.42}$$

subject to:

$$\boldsymbol{S} \cdot \boldsymbol{v} = \boldsymbol{0} \tag{5.43}$$

$$l_{v_j} \leq v_j \leq u_{v_j} \qquad j \in \{1, \ldots, n\} \tag{5.44}$$

with Z_{rat} calculated as stated in Equation (5.26) or (5.27) (page 141).

Recalculation to meet non-linear constraints

Anchor points calculated following the instructions detailed in the previous section satisfy both bound and mass balance constraints, which are the constraints that must always be considered in steady-state flux simulations. But, as it was justified previously, other non-linear constraints such as closeness and ratio constraints contribute to improve realism of the obtained solutions. Then, in cases where these constraints are applied the previous anchor point flux distributions must be recalculated to satisfy them.

Closeness constraints ^{13}C-based metabolic flux analyses allow determining *in vivo* fluxes of some internal reactions with respect to a carbon substrate intake reaction which is used as a reference. Thus, all the values of the fluxes obtained are expressed as a proportion to the reference. This fact is leveraged in Meta-MODE to describe a routine (Algorithm 5.8) that modifies the flux bounds of the reactions with experimental information taking into account the experimental value and the admissible discrepancy limit (maximum deviation or minimum closeness) fixed in the simulation parameters set. The new bounds will be calculated as described in Equations (5.45) and (5.46).

$$\boldsymbol{v}_e^{exp} \times (1 - MaxDev) \leq \quad \boldsymbol{v}_e \ \leq \boldsymbol{v}_e^{exp} \times (1 + MaxDev) \qquad e \in E \tag{5.45}$$

$$\boldsymbol{v}_e^{exp} \times \left(2 - \frac{1}{MinClos}\right) \leq \quad \boldsymbol{v}_e \ \leq \boldsymbol{v}_e^{exp} \times \left(\frac{1}{MinClos}\right) \qquad e \in E \tag{5.46}$$

where E is the set of reactions with experimental flux measurements.

Once the bounds have been updated, a new optimisation problem is solved to obtain a new flux distribution that is optimal for the corresponding objective and satisfies closeness constraints.

```
Input: simulation parameters and preliminary flux distribution
       of an anchor point $\boldsymbol{v}^*|_0$
Output: flux distribution of an anchor point $\boldsymbol{v}^*$ fulfilling
        closeness constraints

Extract from $\boldsymbol{v}^*|_0$ the value of the reference flux $\boldsymbol{v}^*_{ref}$;
for each flux with experimental information $e$ do
    Calculate the value of the flux $\boldsymbol{v}^*_e$ according to $\boldsymbol{v}^*_{ref}$ ;
    Calculate lower and upper bounds for flux $e$ based on the
      value $\boldsymbol{v}^*_e$ and the admissible discrepancy limit (Equation
      (5.45) or (5.46));
    Calculate $\boldsymbol{v}^*$ by solving a new optimisation problem with the
      new bounds and an appropriate objective function
      (Equations from (5.28) to (5.44));
end
return $\boldsymbol{v}^*$
```

Algorithm 5.8: Recalculation of anchor points to satisfy closeness constraints

Ratio constraints In the case of establishing ratios between reaction rates, there is no reference reaction that can be used to readjust the flux bounds. Instead, a new optimisation problem is formulated including non-linear constraints based on deviation/closeness with respect to the expected ratio (Equations (5.47) and (5.48)), and an appropriate objective function, depending on the objective (Algorithm 5.9).

$$\delta_{r_1,r_2} - MaxRatDev \leq 0 \tag{5.47}$$

$$MinRatClos - M_{r_1,r_2} \leq 0 \tag{5.48}$$

where δ_{r_1,r_2} is deviation calculated as stated in Equation (5.18), M_{r_1,r_2} is closeness calculated as stated in Equation (5.19), and $MaxRatDev$ and $MinRatClos$ are the maximum admissible deviation and minimum admissible closeness, respectively.

In Meta-MODE the MATLAB® *fmincon* interior point algorithm is used to solve the resultant constrained non-linear multi-variable optimisation problem.

```
Input: simulation parameters and preliminary flux distribution
       of an anchor point $v^*|_0$
Output: flux distribution of an anchor point $v^*$ fulfilling
        closeness constraints
Define non-linear constraints (Equation (5.47) or (5.48));
Calculate $v^*$ by solving a new optimisation problem with an
  appropriate objective function (Equations from (5.28) to (5.44));
return $v^*$
```

Algorithm 5.9: Recalculation of anchor points to satisfy ratio constraints

When closeness and ratio constraints are simultaneously considered, the re-calculation sequence will ensure that the final anchor points satisfy them both, because the modified flux bounds defined to account for closeness constraints (Algorithm 5.8) will also apply to solve the non-linear optimisation problem defined for ratio constraints (Algorithm 5.9).

This way, (almost) optimal solutions for individual objectives are found that satisfy all the constraints imposed to the problem. These anchor points will be included in the initial population to improve convergence of the algorithm and reduce computational cost, as previously explained.

5.5.4 Evaluation of the population

To evaluate the performance of the individuals that constitute the population of flux vectors at a given generation, a cost function has to be defined. This cost function will serve to compute costs associate to the different design objectives and, if needed, will account for penalty costs for constraint violation. Later, dominance criteria will be applied based on those costs.

Algorithm 5.10 describes the cost function built for Meta-MODE. First, for each individual (flux vector) in the population (set of flux vectors) the various costs associated to the different design objectives are calculated using the corresponding equations described at Section 5.4.2 *Definition of metabolic objectives* (lines 1 to 16). Then, each individual is tested for constraints observance: mass balance, closeness and ratio constraints are checked and total penalty for constraint violation is obtained as the sum of the individual penalties (lines 18 to 21). If the total penalty is not zero, then the cost vector of the given individual is modified: for each objective the worst cost value of all the population for that objective is taken and the total penalty is summed (lines 22 to 25). This way, the flux vectors that transgress constraints will have cost values worse than all the individuals in the Pareto set, thus ensuring they will be dominated. The addition of the constraint violation penalty helps to order the transgressor individuals so that if some of them need to be incorporated to the population the least bad are selected.

```
Input: simulation parameters and set of flux vectors
    $X = \{\boldsymbol{v}^1, \ldots, \boldsymbol{v}^{N_P}\}$
Output: set of cost vectors $\{\boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^{N_P}\}$
for each individual $\boldsymbol{v}$ do
    for each objective $k$ do
        switch type of objective do
            case Max/min single reaction flux do
                Calculate cost $Z_k(\boldsymbol{v})$ using Equation (5.20);
            case Product of two reactions do
                Calculate cost $Z_k(\boldsymbol{v})$ using Equation (5.21);
            case Growth do
                Calculate cost $Z_k(\boldsymbol{v})$ using Equation (5.22);
            case Parsimony do
                Calculate cost $Z_k(\boldsymbol{v})$ using Equation (5.23);
            case Closeness to experimental fluxes do
                Calculate cost $Z_k(\boldsymbol{v})$ using Equation (5.24) or (5.25);
            case Pairwise flux ratios do
                Calculate cost $Z_k(\boldsymbol{v})$ using Equation (5.26) or (5.27);
        end
    end
end
for each individual $\boldsymbol{v}$ do
    Calculate penalty for mass balance constraint violation
      $PenMassCon$ (Algorithm 5.4);
    Calculate penalty for closeness constraint violation
      $PenClosenessCon$ (Algorithm 5.5);
    Calculate penalty for ratio constraint violation $PenRatioCon$
      (Algorithm 5.6);
    Calculate total penalty for constraint violation
      $PenCon = PenMassCon + PenClosenessCon + PenRatioCon$;
    if $PenCon > 0$ then
        for each objective $k$ do
            $Z_k = \max_{\boldsymbol{v} \in X} Z_k + PenCon$ ;        /* worst cost of
             objective $k$ for the whole set of flux
             vectors plus total penalty */
        end
end
return $\{\boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^{N_P}\}$
```

Algorithm 5.10: Cost function of Meta-MODE

5.5.5 Algorithm proposed for multi-objective metabolism optimisation: Meta-MODE

In previous sections equations and procedures have been described that are needed to build a multi-objective optimisation algorithm suitable for steady-state flux simulations. Taking them into account it is possible to rewrite Algorithm 2.5 (sp-MODE) to adapt it to the problem under study. Consequently, Meta-MODE algorithm has been created, and a complete description of the operations is shown at Algorithm 5.11.

The input data for the algorithm is the complete set of simulation parameters: the metabolic network of the organism under study, the list of selected objectives, vectors describing flux bounds, the list of chosen constraints and the related information needed to apply them (experimental fluxes or known flux ratios, selected metrics and admissible thresholds), as well as other optimisation parameters inherited from sp-MODE, such as population size, number of arcs for the spherical pruning or stopping criteria.

After loading the complete list of input parameters, an initial population, including anchor points, is generated and evaluated, and non-dominated individuals are selected. Then successive generations of flux vectors are created from their predecessors, using differential evolution operators, and the new individuals are once more evaluated and subjected to dominance filters. At every generation, the complete set of selected individuals is analysed to look for the best values for individual objectives, which leads to anchor points updating if the previous ones have been surpassed. Then, the spherical pruning mechanism is applied to obtain a new approximation of the Pareto set and front. When one of the stopping criteria (usually maximum number of function evaluations, or maximum number of generations) is satisfied, the algorithm terminates and the Pareto set and Pareto front approximations are returned as a result together with a list of parameters including all the input values and some report variables of the optimisation

```
Input: simulation parameters
Output: Pareto set approximation $\boldsymbol{V}_P^*$
Read simulation parameters;
Calculate anchor points $\boldsymbol{v}^{*k},\ \forall k \in \{1, 2, \ldots, q\}$ (Algorithm 5.7);
Build initial population $P|_0$ with $N_p$ individuals;
Add anchor points to $P|_0$ ;
Evaluate $P|_0$ using cost function (Algorithm 5.10);
Apply dominance criterion (Definition 2.1) on $P|_0$ to get $\hat{A}|_0$;
Apply pruning mechanism (Algorithm 2.4) to prune $\hat{A}|_0$ to get
  $A|_0$;
Update anchor points $\boldsymbol{v}^{*k} = \inf_{\boldsymbol{v} \in P|_0} Z_k,\ \forall k \in \{1, 2, \ldots, q\}$;
Set generation counter $G = 0$ ;
while stopping criterion unsatisfied do
    $G = G + 1$;
    Get subpopulation $S|_G$ with solutions in $P|_{G-1}$ and $A|_{G-1}$;
    Add anchor points to $S|_G$;
    Generate offspring $O|_G$ with $S|_G$ using DE operators
      (Algorithm 2.2);
    Evaluate offspring $O|_G$ using cost function (Algorithm 5.10);
    Update population $P|_G$ with offspring $O|_G$ according to
      greedy selection mechanism;
    Apply dominance criterion (Definition 2.1) on $O|_G \bigcup A|_{G-1}$ to
      get $\hat{A}|_G$;
    Update anchor points $\boldsymbol{v}^{*k} = \inf_{\boldsymbol{v} \in \hat{A}|_G} Z_k,\ \forall k \in \{1, 2, \ldots, q\}$;
    Apply pruning mechanism (Algorithm 2.4) to prune $\hat{A}|_G$ to
      get $A|_G$;
end
$\boldsymbol{V}_P^* = A|_G$;
return $\boldsymbol{V}_P^*|_G$
```

Algorithm 5.11: Meta-MODE

process (starting and finishing time, number of generations and functional evaluation achieved, and Pareto front extremes).

5.6 Conclusions of this chapter

In this chapter, a new tool to perform steady-state metabolic flux simulations by means of multi-objective optimisation has been presented. The optimisation kernel of this tool is based on a previous evolutionary algorithm (Reynoso-Meza et al., 2010), but several specific mechanisms have been added to improve pertinency of the obtained solutions, as well as to prepare the algorithm to deal with the large-scale, strongly constrained and multi-modal optimisation problem that arises when performing genome-scale constraint-based flux simulations. The resultant algorithm, Meta-MODE, aims to avoid some of the main drawbacks of classic methods for constraint-based metabolic analysis, like the need of defining a fixed equation accounting for biomass assembly, the strong dependency of the solutions on the selected objective or the limitations encountered to define informative constraints or objectives that exceed the mathematical capabilities of linear and/or mono-objective optimisation techniques. This tool is used in the next chapter to simulate flux landscapes using the metabolic network of *Synechocystis* sp. PCC 6803 under different trophic conditions, and it is benchmarked against the classic Flux Balance Analysis algorithm.

6

Multi-objective optimisation procedure applied to metabolic simulations of *Synechocystis* sp. PCC 6803

6.1 Chapter abstract

In this chapter, the multi-objective optimisation tool for metabolic simulations described in the previous chapter (Meta-MODE) is applied to the study of the metabolic network of *Synechocystis* sp. PCC 6803. It is used to simulate five different growth conditions, that range from pure autotrophy to pure heterotrophy through different combinations of both regimes, in order to demonstrate the plasticity that this formulation confer to the simulation results. Besides, the results obtained with Meta-MODE are compared with classic mono-objective FBA simulations (with and without consideration of experimental internal fluxes) in order to benchmark the proposed tool.

The results of this study show that using the multi-objective algorithm proposed in this work, metabolic simulations can be performed in which

the inclusion of a few internal fluxes obtained from measurements allows to readjust the whole flux distribution, thus avoiding the necessity of imposing hard restrictions. Besides, the inclusion of this experimental information and the mathematical formulation of the objectives within this tool also allow to avoid the necessity of a biomass equation to account for biomass formation; instead, under this scheme, the biomass composition arises as a consequence of the studied conditions, instead of being imposed beforehand.

The solutions obtained from simulations with Meta-MODE describe a quasi-optimal state in which a balance is found between metabolic performance and metabolic pertinency. It is shown in this chapter that these solutions approximate the experimental data closer than other classic methods that have already proved their applicability. Finally, the results show that the algorithm is flexible to tally experimental observations under different environmental conditions adapting the flux behaviour to the specific situation.

Contents of this chapter appear in the following journal article:

- Maria Siurana, Arnau Montagud, Gilberto Reynoso-Meza, J. Alberto Conejero, Javier Sanchis, Lenin G. Lemus-Zúñiga, Javier Urchueguía **Multi-objective evolutionary algorithm allows more accurate genome-scale flux simulations with a small set of experimental values.** *Manuscript in preparation.*

6.2 Introduction

As introduced in Chapter 1, the study of intracellular flux landscapes of microorganisms is widely used in biotechnology in order to gain knowledge on the metabolic potentialities of a targeted organism. Reliable, reproducible and realistic simulations are needed to foster industrial uses of production platforms. Models have also to cope with the plasticity of their metabolic behaviour when responding to environmental and genetic perturbations, let them be caused by adaptation or human intervention.

Optimisation has been at the heart of metabolism simulation as a way to bypass the mathematical hurdle of solving under-determined systems of equations that are found in constraint-based genome-scale metabolic modelling and as an approximation to study biological growth and metabolic behaviour (Stephanopoulos et al., 1999; Orth et al., 2010). For some time, researchers have been using mono-objective optimisation algorithms to solve such problems. These linear-programming-based mono-objective algorithms solved appropriately some biotechnological problems such as Flux Balance Analysis, but, unfortunately, present some limitations like being restricted to single objectives, making difficult the definition of non-linear constraints or objectives, or needing hard constraints like biomass equation or energy expenses to reflect observed metabolic responses.

In order to be able to optimise for more than one objective function in mono-objective algorithms some mathematical tweaks or constructions had to be done: either performing serial optimisations where the result of the former was fed as constraint in the latter or building an objective function as a weighted drain of different variables and optimising for it, like in the case of biomass formation in metabolic models.

Alternatively, multi-objective optimisation allows having a set of objective functions that are simultaneously optimised and, more interestingly, allows spotting dependencies upon them and define strategies to fine tune their behaviour. Pareto fronts are described as a decision-making tool that gathers all the points that represent an optimal trade-off between different objectives. These Pareto points will be optimal for the set of objectives established by researchers and allow them to browse among these objectives and to choose one, or several, that suit their needs. Pareto fronts and sets have been used widely in science (Kung et al., 1975), economics (Greenwald and Stiglitz, 1986) and technology (Martínez-Iranzo et al., 2009) as a tool that clarifies dependencies and trade-offs among different optimisation objectives. Furthermore, this kind of analysis would represent a leap in systems biology, allowing researchers to optimise for different functions or non-linear objective functions using non-linear constraints, of great importance for biological fluxes' study.

Also, accurate flux balance analysis optimisation requires the compilation of a set of boundary limits, such as drain of substrates and production of by-products. These boundary limits, are typically acquired from experiments and literature, and are usually tiresome to gather and even sometimes cryptic to understand.

In Chapter 5 of this dissertation, a multi-objective evolutionary algorithm was presented that allows the simulation of metabolism minimising the use of these boundary limits, with the inclusion of a few experimental flux values, and retrieving metabolic flux landscapes that are much closer to the real ones. Additionally, Meta-MODE algorithm, due to its evolutionary nature, allows the use of non-linear constraints, and the maximisation of non-linear objectives.

In this chapter this algorithm is applied to the genome-scale metabolic model of *Synechocystis* sp. PCC 6803 described in Chapter 3 to simulate its metabolic response under different trophic conditions.

6.3 Materials and Methods

6.3.1 Metabolic model and simulation conditions

The present study has been conducted using the metabolic network of *Synechocystis* sp. PCC 6803 presented in Chapter 3. As explained in Section 1.3 this cyanobacterium can grow under three trophic conditions that differ in the chosen energy and carbon sources. These growth modes are:

(i) *photoautotrophy*, where energy comes from light and carbon from CO_2,

(ii) *heterotrophy*, where a sugar, often glucose, is the source of both energy and carbon, and

(iii) *mixotrophy*, a combination of the former two, where all three elements (light, CO_2 and glucose) are combined.

In the case of heterotrophy, some authors, like Vermaas (1996), consider different variants:

(a) *dark heterotrophy*, where the cyanobacteria is grown in the darkness[21],

(b) *light-activated heterotrophy*, where the cyanobacteria is grown in darkness after a short period of light exposure, and

(c) *photoheterotrophy* (or *light heterotrophy*), where the cyanobacteria is grown in presence of light but this light is not used as energy source, , *i.e.* the photosynthesis is not done completely (not to be confused with mixotrophy, where the photosynthesis is completely used).

The case of *photoheterotrophy* is a growth mode that can be induced in the laboratory. Under this mode, the cyanobacteria is grown in presence of light, which can lead to the activation and operation of some cellular processes, but it cannot make use of this light as energy source because of the inhibition of the photosynthetic function by means of the addition of some chemical or by genetic modification.

It is of note that this organism requires other substances in order to sustain growth such as nitrogen, sulphur, phosphorus, chlorine, and metals like magnesium, molybdenum, sodium or iron among others.

In this study four of these growth modes are simulated (Table 6.1), *viz.* light-activated heterotrophy (hereafter called heterotrophy), photoheterotrophy, mixotrophy and photoautotrophy). In order to determine the specific constraints that characterise each mode, some experimental information is needed. In this study this information has been obtained from the journal articles specified in Table 6.1. The measurements and

[21] Some authors have pointed to the inability of *Synechocystis* sp. PCC 6803 to grow under complete darkness unless previously exposed to a pulse of light (minutes) (Anderson and Mcintosh, 1991), which is the growth mode called *light-activated heterotrophy*.

growth conditions described in the articles have been translated to constraints (see next section) with the aim of reproducing *in silico* the conditions of the *in vivo* experiments, which allows for comparison.

Table 6.1: *Growth modes of Synechocystis sp. PCC 6803 simulated in this study.*

Mode	Carbon source	Energy source	Light exposure	Photosynthesis	Data from
Photoautotrophy	CO_2	light	continuous	active	[1]
Mixotrophy	CO_2 + glucose	light + glucose	continuous	active	[2,3]
Photoheterotrophy	glucose	glucose	continuous	inhibited	[2]
Heterotrophy†	glucose	glucose	pulse	absent	[3]

† Light-activated heterotrophy (Anderson and Mcintosh, 1991).
[1] Young et al. (2011); [2] Nakajima et al. (2014); [3] Yang et al. (2002).

Biomass-glucose yield

A good index to measure metabolic efficiency in glucose-consuming microbial cultures is biomass-glucose yield ($Y_{X/S}$) (Stephanopoulos et al., 1999). In this work, the mass-mass yield is used, which is defined as grams of biomass produced per gram of glucose consumed. Since carbon obtained from glucose is incorporated into biomass precursor molecules, this yield is expected to be less than one when glucose is the only carbon source. If other sources are added, like in the case of mixotrophic conditions where CO_2 is also a carbon source, this yield can achieve greater values.

In this work biomass-glucose yield is used to compare solutions obtained by different simulation methodologies, and under different environmental conditions. This yield was calculated from the different simulation results as growth rate divided by glucose consumption rate (in grams). The experimental value was calculated in the same way (from experimental rates extracted from references indicated in table 6.1) and verified with values provided by the corresponding journal articles.

As shown in Table 6.1, data for the simulation of mixotrophic conditions were extracted from two different bibliographic sources. These works describe two variants of mixotrophic cultivation and can be classified according to their biomass-glucose yields:

(i) Experiment from Nakajima et al. (2014), presents higher $Y_{X/S}$, which corresponds to higher contribution of CO_2 and light. This condition is tagged as *mixotrophic (H)*.

(ii) Experiment from Yang et al. (2002) results in lower $Y_{X/S}$, which corresponds to lower contribution of CO_2 and light. This condition is tagged as *mixotrophic (L)*.

6.3.2 Optimisation statements

Two approaches have been applied in this chapter to simulate metabolic flux landscapes of the cyanobacterium *Synechocystis* sp. PCC 6803: mono-objective optimisation using the FBA methodology, and multi-objective optimisation by means of the algorithm described in Chapter 5. Next, the particular optimisation problems stated for each case are explained.

Mono-objective optimisation

To benchmark the solutions obtained with the MO method proposed in this work against classic methods used in constraint-based metabolic modelling, Flux Balance Analysis (FBA) has been applied as representative of the classic methods. Briefly (see Section 1.2.2 for details) and according to this methodology, a metabolic network is represented by its stoichiometric matrix $\boldsymbol{S}$ (as explained in Definition 1.1), and a steady-state mass balance is then applied to calculate the metabolic fluxes through the network (gathered in vector $\boldsymbol{v}$). Constraints are imposed to the system that limit the range of allowable fluxes by considering reaction directionality, enzyme/transport capacity and specific physiological knowledge. The following linear optimisation problem is then stated to maximise/minimise an objective function which can be any linear combination of fluxes:

$$\max_{\boldsymbol{v}} Z\left(\boldsymbol{v}\right) = \boldsymbol{c}^{T} \cdot \boldsymbol{v} \tag{6.1}$$

subject to:

$$\boldsymbol{S}{\cdot}\boldsymbol{v} = \boldsymbol{0} \tag{6.2}$$

$$v_{j,rev} \in (-\infty, +\infty) \qquad j \in \{1, \ldots, n\} \tag{6.3}$$

$$v_{j,irr} \in [0, +\infty) \qquad j \in \{1, \ldots, n\} \tag{6.4}$$

$$l_{v_j} \leq v_j \leq u_{v_j} \qquad j \in \{1, \ldots, n\} \tag{6.5}$$

where $\boldsymbol{c}$ is a vector of weights indicating how much each reaction contributes to the objective function, $v_{j,rev}$ and $v_{j,irr}$ are the fluxes of the reversible and irreversible reactions respectively, and l_{v_j} and u_{v_j} are the lower and upper flux bounds for reaction j respectively.

In the present work, the considered objective Z is the maximisation of growth yield, the most common function used in constraint-based metabolic simulations (Feist and Palsson, 2010). Formulation of the equation accounting for biomass assembly (the biomass equation) of *i*Syn842 model was presented in Chapter 3 (Table 3.1, page 74).

Different optimisation statements of this form are defined by substituting the corresponding flux bounds for the different trophic conditions. Two different approaches have been applied for the definition of these bounds:

(i) Simulations labelled as "**FBA**" use the classic approach according to which flux bounds are defined for some intake reactions on the basis of experimental measurements and growth conditions. In this case those values are obtained from literature (see Table 6.1 and Additional file 6.1).

(ii) Simulations labelled as "**FBA_exp**" include information about internal flux measurements retrieved from literature (see Table 6.1). A set of 10 internal fluxes (common to all simulation conditions) was selected (Table 6.2 and Figures from 6.1 to 6.5). Bounds are imposed to those fluxes with experimental data allowing an error of $\pm 25\%$ of the experimental value. Minimal additional bounds related with experimental set-up (light, glucose, CO_2 and some other nutrients availability) are extracted from the description of the experiments included in Table 6.1 (see Additional file 6.2).

In no case the experimental measurements of growth were applied to tune or constrain the simulations, with the aim of using these values as a benchmark.

Multi-objective optimisation

The main goal of this chapter is to illustrate the functionality of the multi-objective optimisation tool for constraint-based flux simulations presented in Chapter 5. Here, this tool is applied to optimise MOP statements of the form (see Section 5.4):

$$\min_{\boldsymbol{v}} \boldsymbol{Z}(\boldsymbol{v}) = [Z_1(\boldsymbol{v}), Z_2(\boldsymbol{v}), \ldots, Z_q(\boldsymbol{v})] \tag{6.6}$$

subject to:

$$\boldsymbol{S} \cdot \boldsymbol{v} = \boldsymbol{0} \tag{6.7}$$

$$v_{j,rev} \in [-100, 100] \qquad j \in \{1, \ldots, n\} \tag{6.8}$$

$$v_{j,irr} \in [0, 100] \qquad j \in \{1, \ldots, n\} \tag{6.9}$$

$$l_{v_j} \leq v_j \leq u_{v_j} \qquad j \in \{1, \ldots, n\} \tag{6.10}$$

$$\max\{\delta_e(v_{j,exp})\} \leq 0.25 \qquad j \in \{1, \ldots, n\} \tag{6.11}$$

where:

$\boldsymbol{S}$ is the stoichiometric matrix,

$\boldsymbol{v}$ is the flux vector,

$v_{j,rev}$ are the fluxes of the reversible reactions,

$v_{j,irr}$ are the fluxes of the irreversible reactions,

l_{v_j} is the lower flux bound for reaction j,

u_{v_j} is the upper flux bound for reaction j,

$v_{j,exp}$ are the fluxes of those reactions with experimental information, and

Table 6.2: *Experimental fluxes considered for the simulations with Meta-MODE and FBA_exp. First two rows contain the fluxes at the reference reactions in* $^{mmol}/_{g_{DC}} \cdot h$*. For all trophic modes with glucose intake (heterotrophy, photoheterotrophy and mixotrophy) the reference reaction is the phosphorylation of glucose to glucose-6-phosphate (first step of glycolysis), while in the case of autotrophy the reference reaction is* H_2CO_3 *intake. Flux values of all remaining reactions are expressed as a percentage of flux at the reference reaction. See page xx for the list of acronyms.*

Reaction	**Name in *i*Syn842**	**Autotrophy**	**Mixotrophy (H)**	**Mixotrophy (L)**	**Photoheterotrophy**	**Heterotrophy**
CMP	H2CO3_in	3.70	-	-	-	-
GLK	2.7.1.2a	-	0.53	0.38	0.58	0.85
PGI	5.3.1.9b	-19.00	5.66	86.10	-144.83	5.70
ENO	4.2.1.11	23.60	215.09	211.00	103.45	142.20
RPE	5.1.3.1	-75.90	-186.79	-146.90	117.24	56.50
RPI	5.3.1.6	35.60	83.02	64.50	-63.79	-33.70
TK2	2.2.1.1b	-38.50	-98.11	-76.40	56.90	26.40
TK1	2.2.1.1a	-37.30	-90.57	-70.50	60.34	30.10
PDH	1.2.4.1	11.80	98.11	161.80	46.55	117.50
CIS	2.3.3.1	3.20	20.75	56.80	8.62	42.50
ICD	1.1.1.42	3.00	20.75	13.90	8.62	9.30
ICL	&4.1.3.1	0.20	0.00	42.90	0.00	33.20

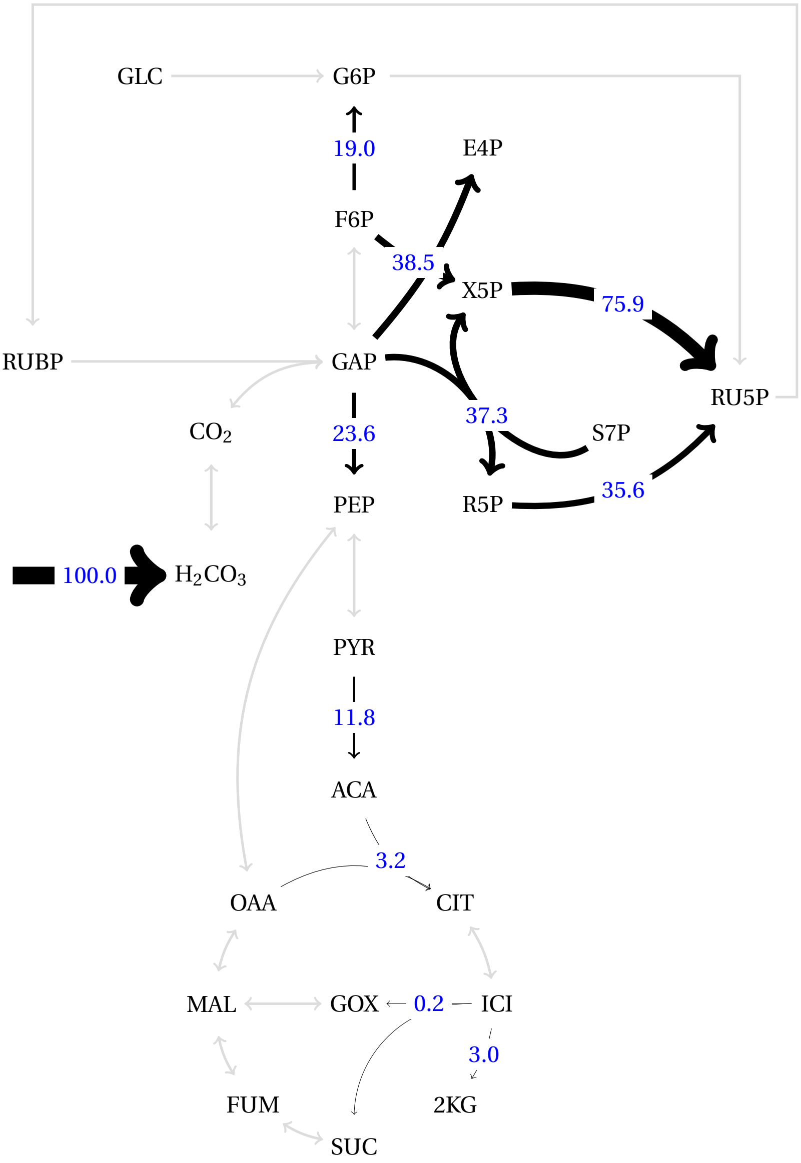


Figure 6.1: *Experimental fluxes considered for the simulations with Meta-MODE and FBA_exp under autotrophic conditions (Young et al., 2011). Reactions whose flux was not used at the simulations are represented by grey arrows. Reactions whose flux was used at the simulations are represented by black arrows. In these latter reactions, arrow thickness and label show flux values expressed as a percentage of flux at the reference reaction (H_2CO_3 intake).See page xx for the list of acronyms.*

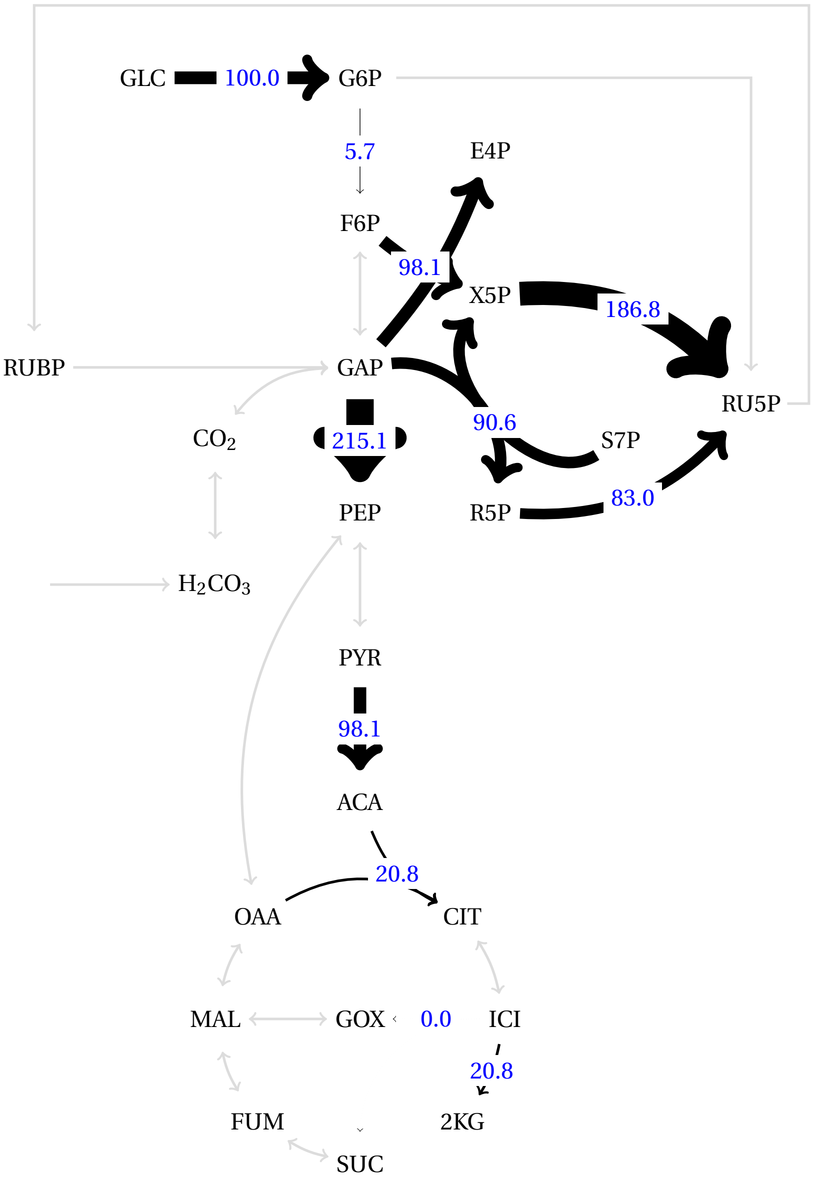


Figure 6.2: *Experimental fluxes considered for the simulations with Meta-MODE and FBA_exp under mixotrophic (H) conditions (Nakajima et al., 2014). Reactions whose flux was not used at the simulations are represented by grey arrows. Reactions whose flux was used at the simulations are represented by black arrows. In these latter reactions, arrow thickness and label show flux values expressed as a percentage of flux at the reference reaction (phosphorylation of glucose to glucose-6-phosphate). See page xx for the list of acronyms.*

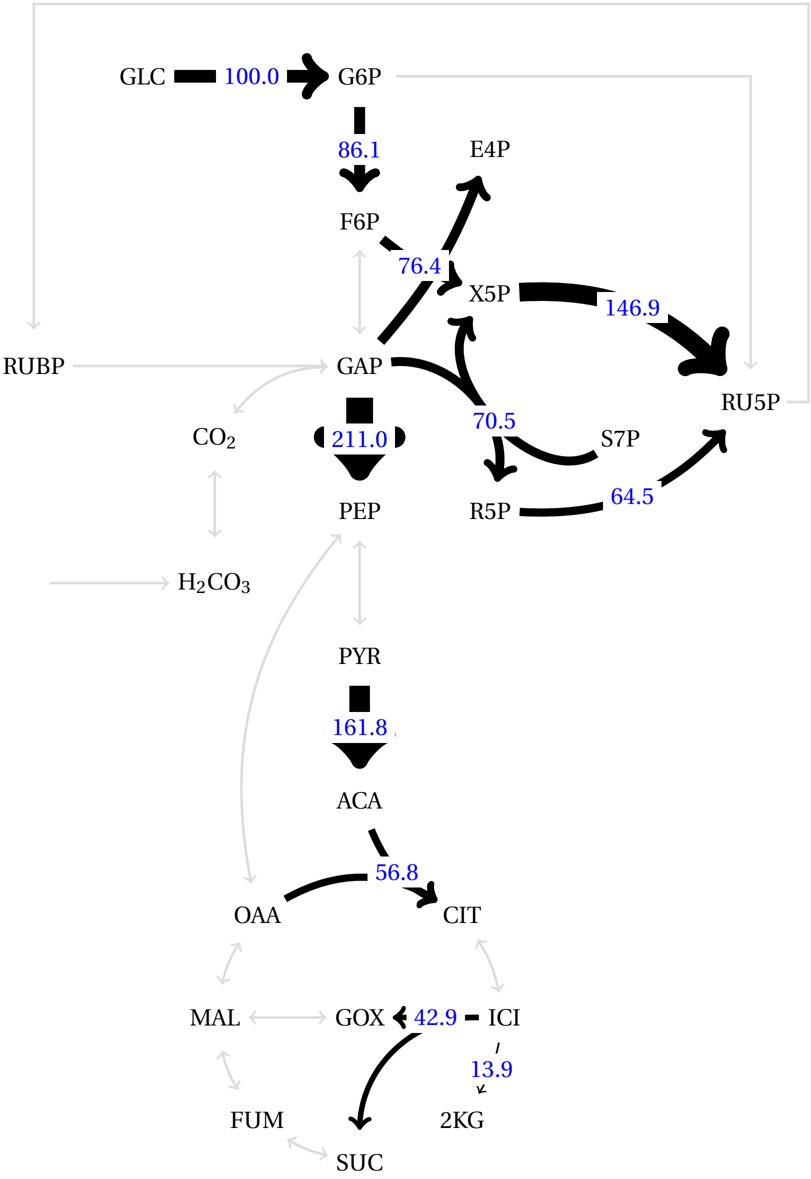


Figure 6.3: *Experimental fluxes considered for the simulations with Meta-MODE and FBA_exp under mixotrophic (L) conditions (Yang et al., 2002). Reactions whose flux was not used at the simulations are represented by grey arrows. Reactions whose flux was used at the simulations are represented by black arrows. In these latter reactions, arrow thickness and label show flux values expressed as a percentage of flux at the reference reaction (phosphorylation of glucose to glucose-6-phosphate). See page xx for the list of acronyms.*

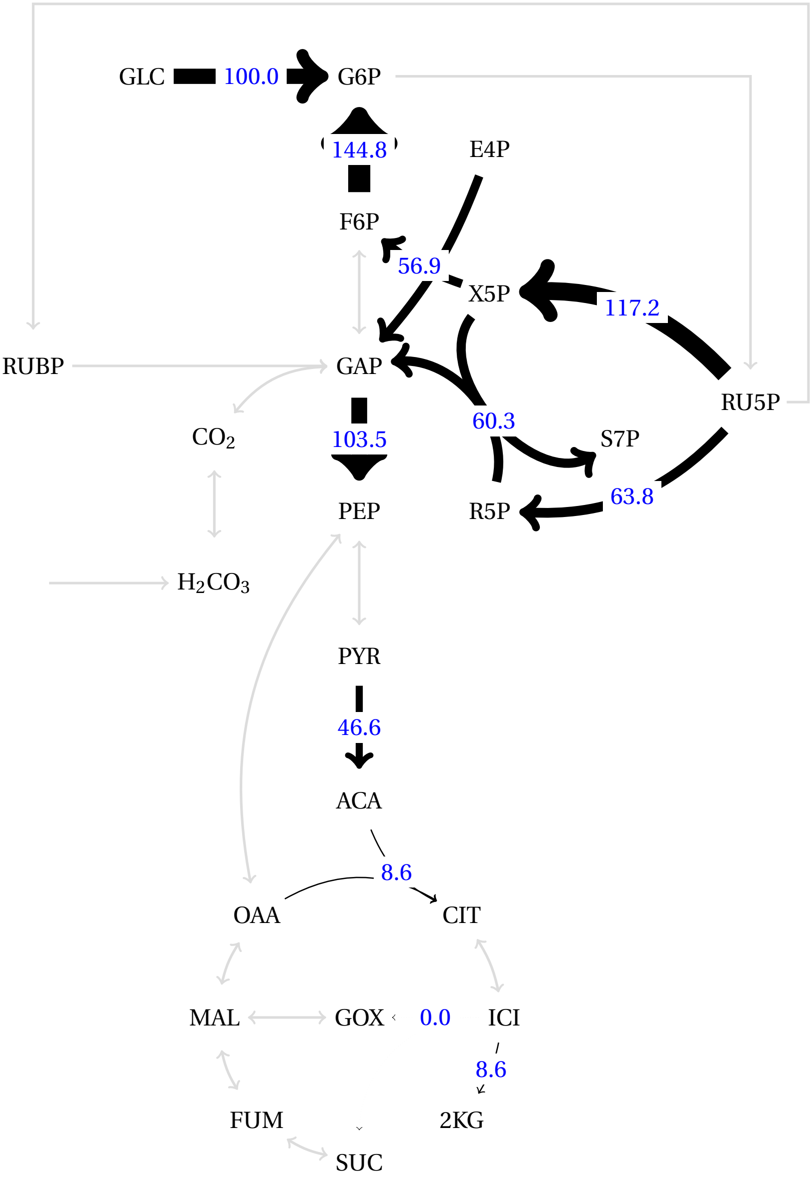


Figure 6.4: *Experimental fluxes considered for the simulations with Meta-MODE and FBA_exp under photoheterotrophic conditions (Nakajima et al., 2014). Reactions whose flux was not used at the simulations are represented by grey arrows. Reactions whose flux was used at the simulations are represented by black arrows. In these latter reactions, arrow thickness and label show flux values expressed as a percentage of flux at the reference reaction (phosphorylation of glucose to glucose-6-phosphate). See page xx for the list of acronyms.*

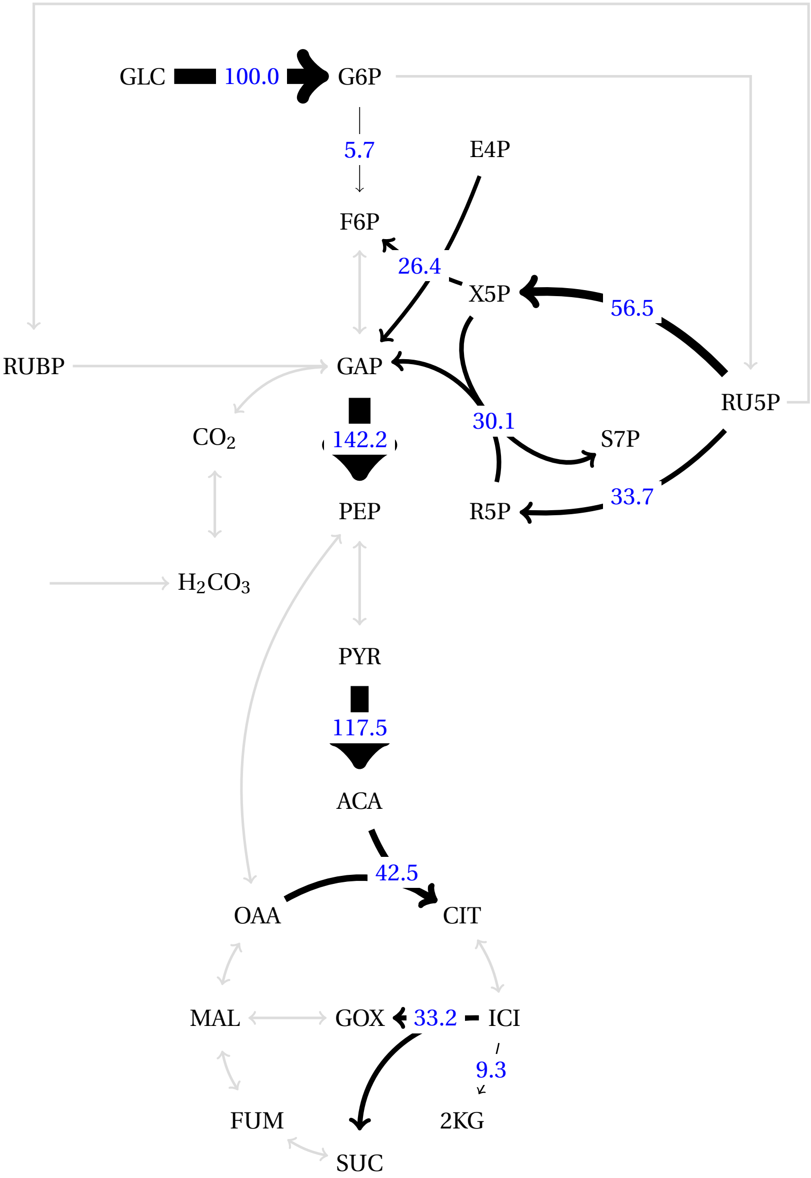


Figure 6.5: *Experimental fluxes considered for the simulations with Meta-MODE and FBA_exp under heterotrophic conditions (Yang et al., 2002). Reactions whose flux was not used at the simulations are represented by grey arrows. Reactions whose flux was used at the simulations are represented by black arrows. In these latter reactions, arrow thickness and label show flux values expressed as a percentage of flux at the reference reaction (phosphorylation of glucose to glucose-6-phosphate). See page xx for the list of acronyms.*

$\delta_e\left(v_{j,exp}\right)$ is the vector deviation from experimental fluxes (Equation (5.8)),

and $\boldsymbol{Z}\left(\boldsymbol{v}\right)$ is the objective vector, within which the following objectives are defined:

$Z_1\left(\boldsymbol{v}\right)$: Maximise growth (Equation (5.22))

$Z_2\left(\boldsymbol{v}\right)$: Minimise ATP synthesis (Equation (5.20))

$Z_3\left(\boldsymbol{v}\right)$: Minimise median deviation from experimental fluxes (closeness) (Equation (5.24))

$Z_4\left(\boldsymbol{v}\right)$: Minimise total sum of fluxes (parsimony) (Equation (5.23))

The selection of these four objectives implies the combination of two objectives accounting for high metabolic performance (hight production of biomass components, $Z_1\left(\boldsymbol{v}\right)$, with low energy expenses, $Z_2\left(\boldsymbol{v}\right)$) and two objectives aiming at the metabolic pertinency of the solutions (resembling real measurements, $Z_3\left(\boldsymbol{v}\right)$, and avoiding unrealistic high fluxes, $Z_4\left(\boldsymbol{v}\right)$). As explained before, maximisation of growth is the most common objective function applied in constraint-based metabolic simulations, and it represents the evolutionary advantage of fast-growing phenotypes (Savinell and Palsson, 1992a; Feist and Palsson, 2010). Minimisation of ATP synthesis implies efficient use of energy and, together with maximisation of growth, results in the cell aiming to grow while using the minimum amount of energy necessary, thereby conserving ATP (Savinell and Palsson, 1992a; Knorr et al., 2007). Inclusion of the closeness objective aims to minimise the deviation between calculated and experimentally measured flux values, which helps to shape the whole flux landscape based on the information available for a small set of fluxes. And finally, minimisation of the total sum of fluxes helps to improve metabolic pertinency since it helps to avoid solutions containing extremely high fluxes or futile cycles with no biological meaning, and biologically it represents parsimonious usage of enzymes which means reduced expenses for enzyme synthesis (Schuetz et al., 2007; Lewis et al., 2010).

The set of constraints included in the above statement are:

- Steady-state mass balance (Equation (6.7)).
- Reaction reversibility (Equations (6.8) and (6.9)).
- Flux bounds (Equation (6.10)).
- Closeness constraints (Equation (6.11)).

Closeness constraints are handled along the optimisation process by means of the penalty function described in Algorithm 5.5. The sets of internal flux measurements have been extracted from the journal articles listed in Table 6.1, and are the same 10 internal fluxes used for the FBA_exp simulations (Table 6.2 and Figures from 6.1 to 6.5). The maximum deviation allowed was 25%.

This set of 10 internal fluxes (common to all simulation conditions, see Table 6.2) has been used for the closeness objective and constraints. Remarkably, none of them were the growth value (or any of the biomass components). Again, growth or biomass components from the experimental datasets were not used as inputs or constraints in any way[22] which allows using the experimental growth values as benchmark.

Minimal additional flux bounds related with experimental set-up (light, glucose, CO_2 and some other nutrients availability) were imposed. The corresponding values, extracted from the description of the experiments included in the selected journal articles, were the same as the ones used for the FBA_exp simulations (see Table 6.1 and Additional file 6.3).

Optimisation tools

To solve the mono-objective optimisation problems stated above, the software tool *PyNetMet*[23] Gamermann et al. (2014b) was used. This Python-based toolbox (briefly described in Section 1.2.3) is designed to manipulate metabolic networks and perform flux simulation and anal-

[22] The only constraints imposed to the reactions yielding biomass components are the reversibility bounds, imposed as stated in Equations (6.8) and (6.9)

[23] Tool available at `github.com/CyanoFactory/CyanoFactoryKB`

ysis. In this chapter it has been applied to solve the FBA and FBA_exp instances.

For the multi-objective optimisation approach, a MATLAB®-based implementation of the Meta-MODE algorithm (Algorithm 5.11) presented in Chapter 5 has been applied.

6.3.3 Solution visualisation and selection (MCDM stage)

Transformation of growth objective units

As it was explained in Chapter 5 (Section 5.4.2), in Meta-MODE the growth objective is defined as the sum of all fluxes yielding biomass components (with negative sign for maximisation). The units of these fluxes in *i*Syn842 are defined to be $^{mmol}/_{g_{DC}} \cdot h$, that is, millimole per gram of dry cell (biomass dry weight) per hour. Consequently, those are also the units of the growth objective function. However, usually the growth rate is measured at the laboratory in h^{-1} (which comes from $^{g}/_{g} \cdot h$), and thus the units must be converted to address comparisons. To do so, for each individual biomass precursor millimoles are converted to grams and then all weights of biomass components are summed:

$$\mu = \sum_j v_j \cdot M^{bm_j} \cdot 10^{-3} \ , \quad j \in B \tag{6.12}$$

where B is the set of reactions describing individual biomass precursor drains, and M^{bm_j} is molar weight of the biomass precursor drained in reaction j.

The value of the growth objective is shown in all figures using these units.

Visualisation plots

Level Diagrams visualisation was used to visualise all four components of the resulting Pareto fronts, by means of the LD-ToolBox[24] (Reynoso-Meza et al., 2013a) developed for MATLAB®.

Also 3-D plots showing three of the four objectives were depicted for comparison purposes. The selected objectives for these visualisations were at the *x*-axis $Z_1(\boldsymbol{v})$ (growth); at the *y*-axis $Z_4(\boldsymbol{v})$ (parsimony); and at the *z*-axis $Z_3(\boldsymbol{v})$ (closeness). The reason of choosing this combination is because parsimony and closeness are critical objectives that aim at improving biological meaningfulness of the solutions, while growth objective is an objective aiming at best metabolic performance which was considered critical due to its relation with the FBA's "biomass" objective.

Parsimony-closeness trade-off score

For the selection of solutions among the Pareto front, an index has been defined based on the trade-off between the two pertinency-based objectives, that is between $Z_4(\boldsymbol{v})$ (minimisation of total sum of fluxes - *parsimony*), and $Z_3(\boldsymbol{v})$ (minimisation of median deviation from experimental fluxes - *closeness*).

First, the values of these two objectives are normalised with respect to their minimum and maximum values in the entire Pareto front:

$$\hat{Z}_k(\boldsymbol{v}) = \frac{Z_k(\boldsymbol{v}) - Z_k^{utopia}}{Z_k^{nadir} - Z_k^{utopia}} \qquad k = [3, 4] \tag{6.13}$$

Then, a weighted distance is defined that prioritises the closeness objective over the parsimony objective by assigning a greater weight to deviation:

[24] Tool available at `https://www.mathworks.com/matlabcentral/fileexchange/62224`

$$\omega = \sqrt{\left(\hat{Z}_{pars}\left(\boldsymbol{v}\right)\right)^{2} + \left(2 \times \hat{Z}_{dev}\left(\boldsymbol{v}\right)\right)^{2}} \tag{6.14}$$

The reason for giving more importance to closeness than to parsimony is because closeness objective is related to actual measured values, thus it is a highly trustworthy index, while parsimony is based on the intuition that living systems will tend to minimise enzyme usage, but there is no reference value of the "normal" total sum of fluxes.

This index has been used in the representation of the Pareto front approximations to colour the dots that represent the solutions at each Pareto set. As it is defined by a (weighted) distance, better solutions in terms of metabolic pertinency will have lower values. With this idea on mind the solutions are ranked and the chosen (five) solutions are the ones with lower values of this index.

Similarity measure and clustering dendrograms of fluxes

In order to evaluate similarity between flux vectors, a fuzzy metric based on the standard fuzzy metric (Definition 5.4, Chapter 5, page 135) was used. This metric can be defined to calculate similarity between any two flux values, given a reference:

$$M(\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_{ref}) = \frac{|\boldsymbol{v}_{ref}|}{|\boldsymbol{v}_{ref}| + |\boldsymbol{v}_1 - \boldsymbol{v}_2|} \tag{6.15}$$

In this work, experimental values are always used as reference values. Thus, this metric compares the difference between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ with respect to the value of the flux measured for this same reaction. If the difference between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ is big compared with the experimental value, the similarity coefficient will be close to 0; if the difference between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ is small compared with the experimental value, the similarity coefficient will be close to 1.

This metric was used in this study to compare the solutions from different simulation methods. The flux values of the subset of reactions with experimental measurements were extracted from the different simulation solutions (columns in the table below). These vectors of fluxes were pairwise compared between them and with the experimental values using the metric above. For each pair of vectors, the median of the similarities was taken. This way, for each pair of flux vectors a single similarity value is obtained.

	Flux vectors			
Reactions with measurements	method 1	method 2	$\cdots$	Experimental
r_1	-	-	$\cdots$	-
r_2	-	-	$\cdots$	-
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r_n	-	-	$\cdots$	-

These median similarities where later used to depict dendrograms in which "more similar" solutions will appear clustered together. Two families of dendrograms were plotted: (i) in one case, only the subset of 10 internal fluxes, common to all conditions (Table 6.2), used to define closeness constraints and objective, was considered; and (ii) in the other case, extended sets of fluxes including all the measurements available for each condition were used (33, 23, 19, 21 and 19 fluxes for the autotrophic, mixotrophic (h) and (l), photoheterotrophic and heterotrophic conditions respectively, among which the subset of 10 was always included).

This way, it can be checked if the obtained fluxes resemble the experimental values used to tune the simulations and if, by adjusting this small subset, other fluxes are also close to the real values.

6.3.4 Convergence and stochastic behaviour

Since Meta-MODE is an algorithm based on differential evolution that uses stochastic techniques both to build the initial population and to derive offspring individuals from their parents, it is important to verify the stability of the obtained solutions. In Figure 6.6 an approximation of the front hypervolume (Branke et al., 2008) against generations is shown. Dots are drawn for the solution with median hypervolume at a given generation while lines represent first and third quartiles. Notably, the hypervolume is used here to monitor the evolution of the front size and shape, but not as a "quality" measurement[25]. That is, it will be used in order to gain some insight about convergence rate of the algorithm and therefore, being able to suggest a reasonable amount of generations (or function evaluations).

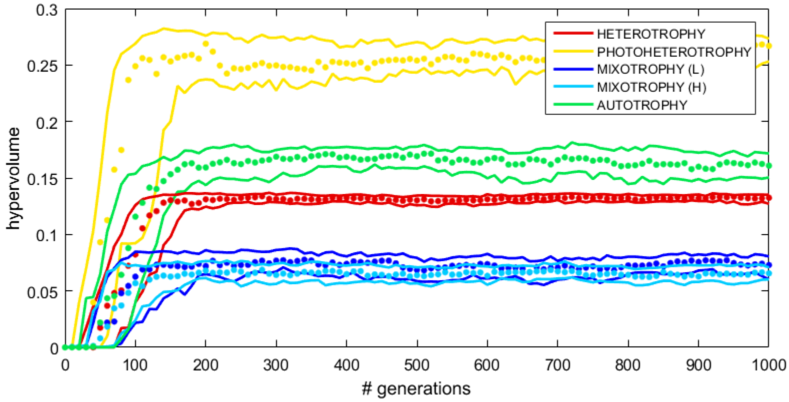


Figure 6.6: *Convergence and stochastic behaviour of the Meta-MODE simulations under all five simulation conditions.*

To perform this analysis, for each condition 51 repeats of the simulation were run, and results were gathered every 10 generations until 1000 generations were reached. As it can be seen in Figure 6.6, the algorithm yields stable Pareto front approximations from around 200 generations for all conditions studied. All results shown in this chapter correspond to generation 400, in particular to the solution showing median hypervolume. Therefore, this value is suggested altogether with the parameters of Tables 5.1 and 5.2 as a set of suitable parameters in order to use this algorithm with other organisms.

[25] As explained in Chapter 2 when using optimisation techniques to simulate metabolism emphasis is put in the realism of the solutions, more than in the mathematical optimality.

6.4 Results and discussion

Additional file 6.4 (see page 250) contains the results of all simulations performed in this chapter, as well as the complete lists of experimental fluxes available from literature for each growth mode.

6.4.1 Pareto front approximations obtained with multi-objective optimisation algorithm

Figures from 6.7 to 6.11 show the obtained Pareto front approximations for all five conditions simulated, *i.e.* (i) photoautotrophy (Figure 6.7), (ii) mixotrophy with higher CO_2 component (H) (Figure 6.8), (iii) mixotrophy with lower CO_2 component (L) (Figure 6.9), (iv) photoheterotrophy (Figure 6.10), and (v) heterotrophy (Figure 6.11).

In these figures the Pareto fronts are depicted using three dimensional plots that exclude one of the objectives (ATP production), as well as using Level Diagrams (LD). The great advantage of Level Diagrams representation is that it allows to visualise all components of the Pareto front when more than three objectives are considered. Besides, it eases the analysis of independent objectives, as well as to spot their interdependences.

An interesting trait that can be observed in the Level Diagrams plots, is that the solutions that show lower normalised distance to the utopian solution (vertical axis in LD visualisation, see Section 2.4.2) are not exactly the same that show better parsimony-closeness trade-off, although they are always close by. This fact illustrates that metabolic reality often implies quasi-optimal flux distributions, which highlights the importance of using multi-objective approaches in order to analyse further Pareto optimal states (instead of a single optimal solution). The biggest differences are seen in the case of heterotrophic growth, which will be discussed later in detail (Section 6.4.4).

If objectives are analysed individually from the LD plots in Figures 6.7 to 6.11, some important observations arise. Looking at objective

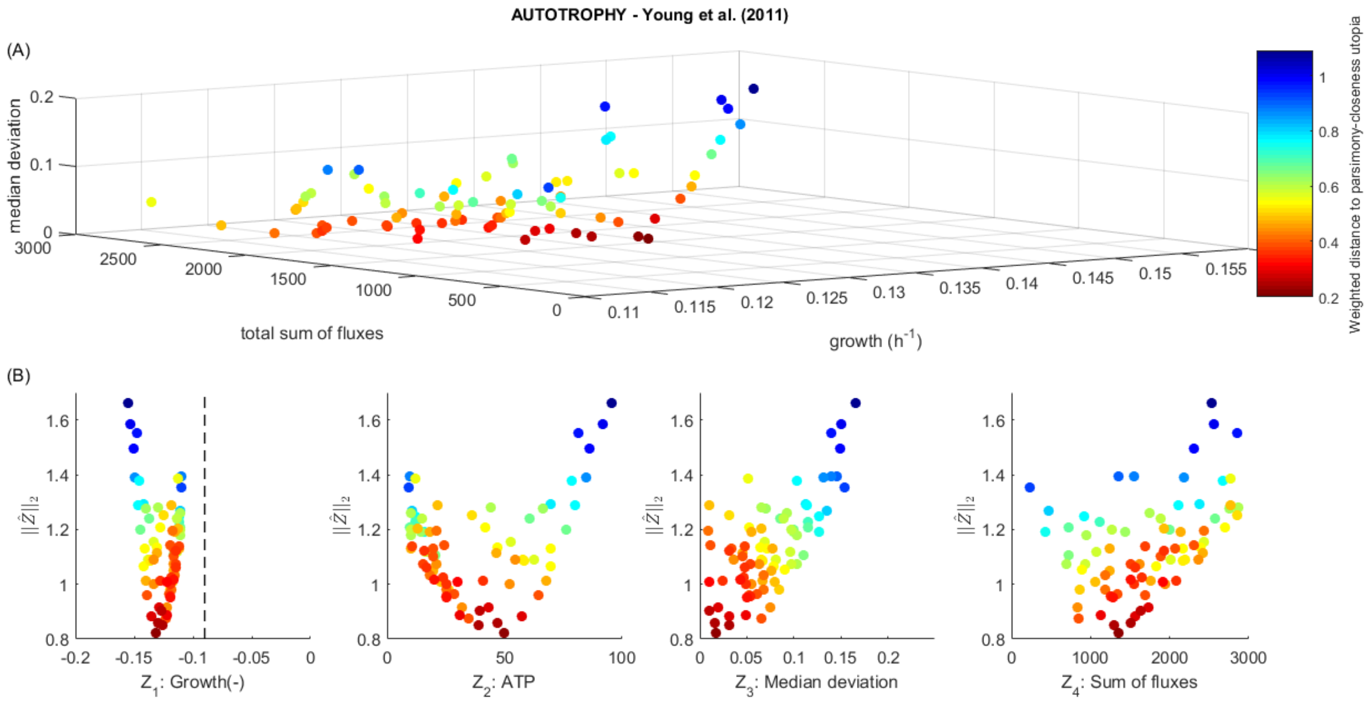


Figure 6.7: *Pareto front approximation found with Meta-MODE under autotrophic conditions from Young et al. (2011)*
(A) 3-D plot showing three of the four objectives: $Z_1\,(\boldsymbol{v})$*: growth;* $Z_3\,(\boldsymbol{v})$*: closeness;* $Z_4\,(\boldsymbol{v})$*: parsimony;*
(B) Level Diagrams of the four objectives. Vertical line in the diagram of growth objective shows experimental growth.

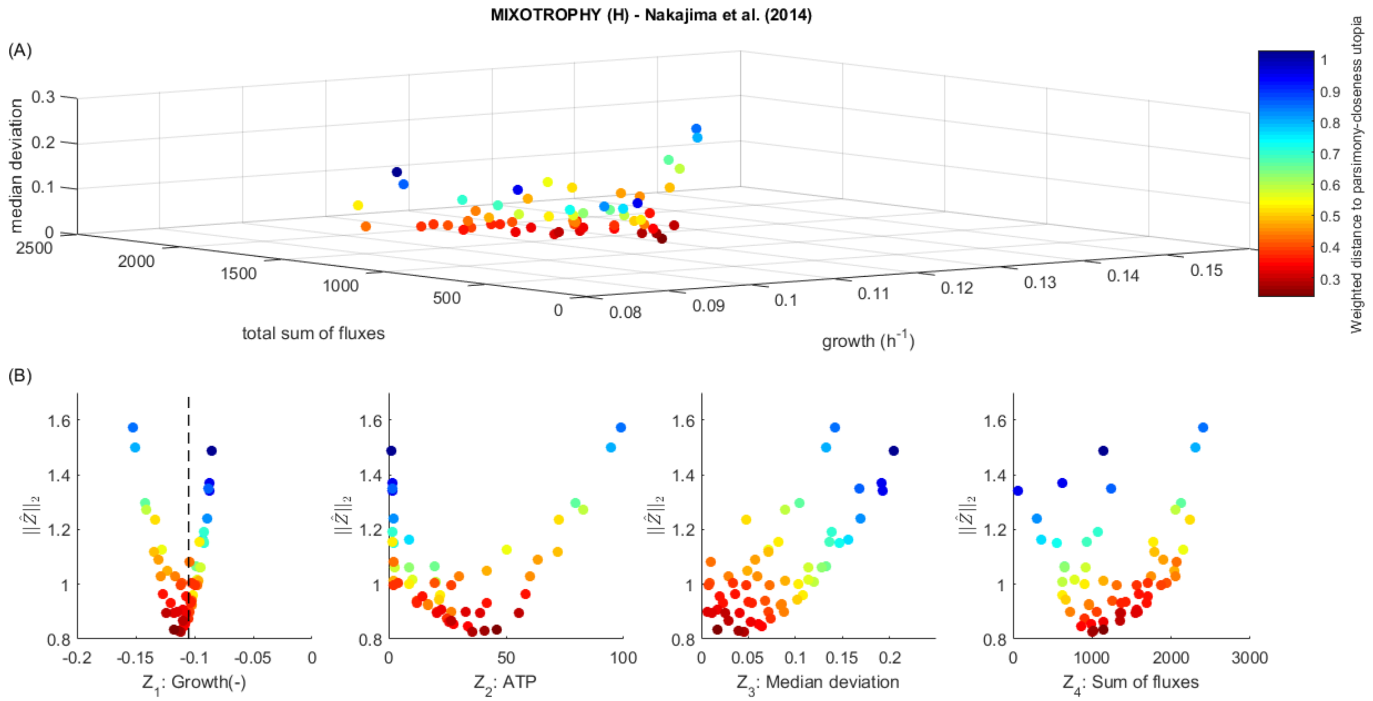


Figure 6.8: *Pareto front approximation found with Meta-MODE under mixotrophic (H) conditions from Nakajima et al. (2014)*
(A) 3-D plot showing three of the four objectives: $Z_1\,(\boldsymbol{v})$: growth; $Z_3\,(\boldsymbol{v})$: closeness; $Z_4\,(\boldsymbol{v})$: parsimony;
(B) Level Diagrams of the four objectives. Vertical line in the diagram of growth objective shows experimental growth.

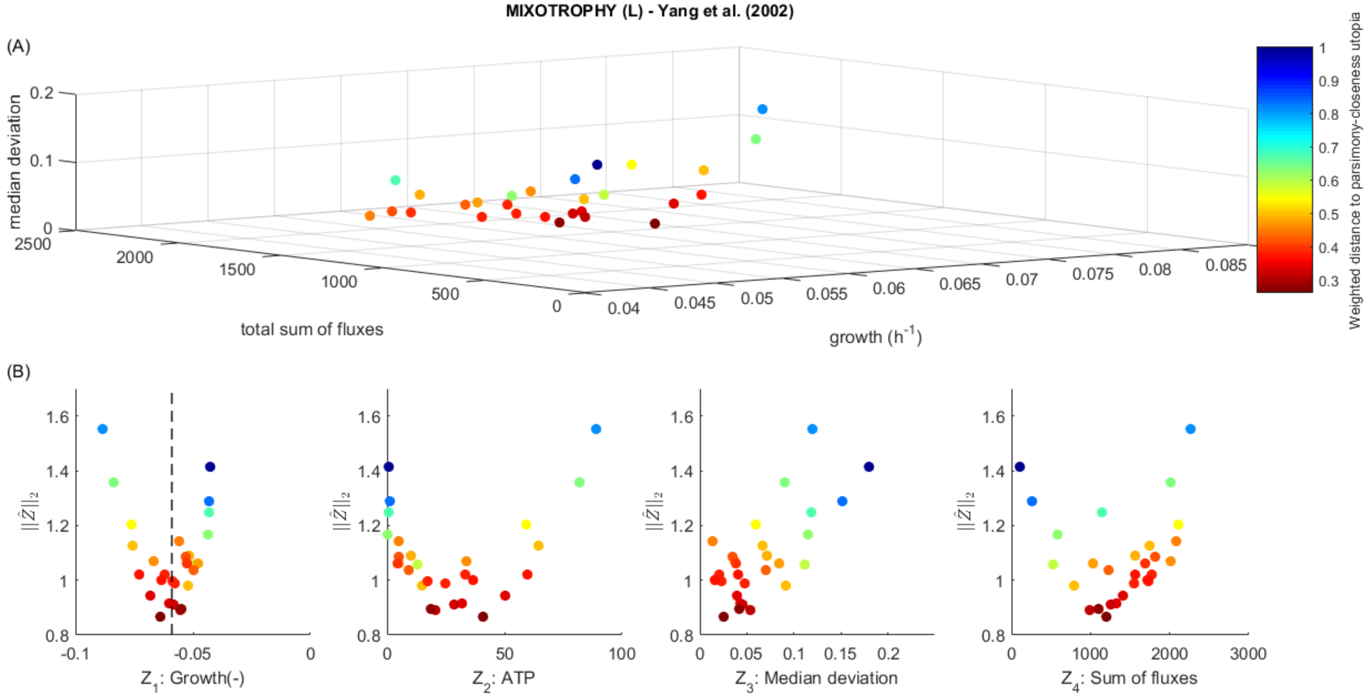


Figure 6.9: *Pareto front approximation found with Meta-MODE under mixotrophic (L) conditions from Yang et al. (2002)*
(A) 3-D plot showing three of the four objectives: $Z_1\,(\boldsymbol{v})$: growth; $Z_3\,(\boldsymbol{v})$: closeness; $Z_4\,(\boldsymbol{v})$: parsimony;
(B) Level Diagrams of the four objectives. Vertical line in the diagram of growth objective shows experimental growth.

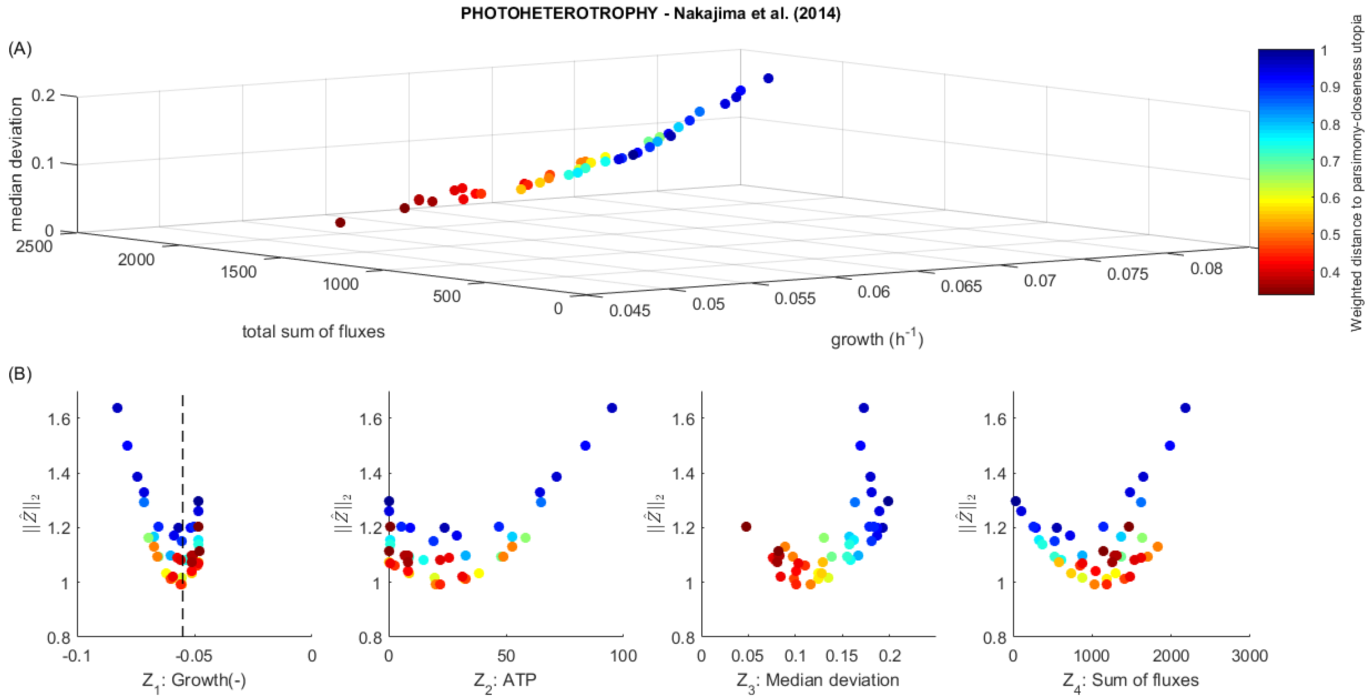


Figure 6.10: *Pareto front approximation found with Meta-MODE under photoheterotrophic conditions from Nakajima et al. (2014)*
(A) 3-D plot showing three of the four objectives: $Z_1(\boldsymbol{v})$: growth; $Z_3(\boldsymbol{v})$: closeness; $Z_4(\boldsymbol{v})$: parsimony;
(B) Level Diagrams of the four objectives. Vertical line in the diagram of growth objective shows experimental growth.

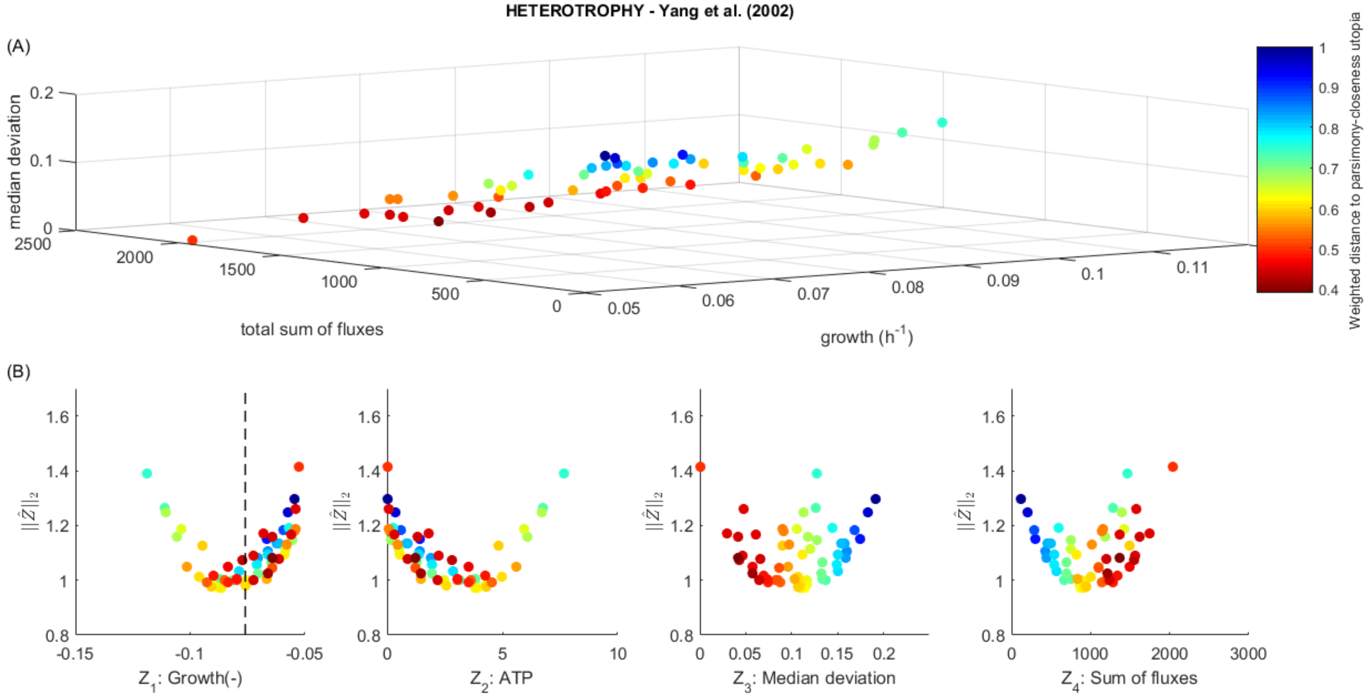


Figure 6.11: *Pareto front approximation found with Meta-MODE under heterotrophic conditions from Yang et al. (2002)*
(A) 3-D plot showing three of the four objectives: $Z_1(\boldsymbol{v})$: growth; $Z_3(\boldsymbol{v})$: closeness; $Z_4(\boldsymbol{v})$: parsimony;
(B) Level Diagrams of the four objectives. Vertical line in the diagram of growth objective shows experimental growth.

$Z_1(\boldsymbol{v})$ (growth) it can be seen that, in most of the cases, the best solutions appear grouped together within a narrow value range. Moreover, except in the case of autotrophic simulations (that will be discussed later in Section 6.4.4), these values are close to the experimental values. Even more, better solutions, both in terms of proximity to utopian solution (vertical axis in LD) and closeness-parsimony trade-off (dark red colour), better match the experimental growth values.

It is important to note that experimental information related with biomass formation or growth was not used during the optimisation process to tune or constrain the results; thus, this fitting between *in silico* and *in vivo* growth rates has to be underlined as a validation of the method and the set of constraints and objectives used.

Looking at objective $Z_3(\boldsymbol{v})$ (closeness), it can be seen that, as expected, solutions with higher deviation are painted in more blue colours (far from the defined weighted sum of metabolic pertinency). The same effect, although a bit less marked, is observed in the case of objective $Z_4(\boldsymbol{v})$ (parsimony), solutions with higher sum of fluxes appear more blue coloured. This was to be expected, since the index used to account for metabolic pertinency (and illustrated by the colouring) was defined based on these two objectives, with greater weight on the closeness objective.

6.4.2 Trade-off between objectives

Trade-off between closeness and parsimony as a measure of realism

From the Level Diagrams it can be seen that some kind of trade-off exists between closeness and parsimony objectives. In order to better observe this phenomenon, these two objectives have been represented in bi-dimensional objective spaces (Figure 6.12, B). Looking over each closeness-parsimony figure from left to right, it can be seen that in all cases, at the left part, it is not possible to reduce deviation without increasing the total sum of fluxes. This fact points out that, even when cells try to perform their metabolic functions aiming at greatest efficiency, the flux levels that have been reported from actual systems (closeness objective is based on internal fluxes measured from real cells) demand a minimum of metabolic activity to be achieved. In most cases (except in simulations under heterotrophic conditions) it can be observed that some solutions can be found that suffer degradation of both objectives. It can be seen that the defined weighted distance (encoded by the colour scale) has been designed in such a way that it favours the solutions that appear around the point in which minimal deviation is achieved without further degradation in parsimony.

Metabolic performance vs. metabolic pertinency

Figure 6.12, A, also shows the relation between objective $Z_1(\boldsymbol{v})$ (maximisation of growth) and objective $Z_2(\boldsymbol{v})$ (minimisation of ATP synthesis). As it is to be expected, greater growth rates imply greater energy demands. Even more, the relationship between these two objectives appears to be linear, with coefficient of determination (R^2) greater than 0.91 in all cases[26]. It can be seen that the solutions marked in dark red colours (*i.e.* solutions closer to the closeness-parsimony utopia) in general show medium levels of growth and ATP synthesis.

Altogether, the bi-objective spaces plotted in Figure 6.12 show that the quasi-optimal solutions proposed as better solutions based on weighted distance to the closeness-parsimony ideal (Equation (6.14)), exhibit good levels of fulfilment in both metabolic performance (growth-energy trade-off) and metabolic pertinency (closeness-parsimony trade-off). For the following analyses presented in this section, five solutions were selected from each Pareto front according to best metabolic pertinency criterion.

[26] Autotrophy: $R^2 = 0.9997$. Mixotrophy (H): $R^2 = 0.9105$. Mixotrophy (L): $R^2 = 0.9327$. Photoheterotrophy: $R^2 = 0.9668$. Heterotrophy: $R^2 = 0.9691$

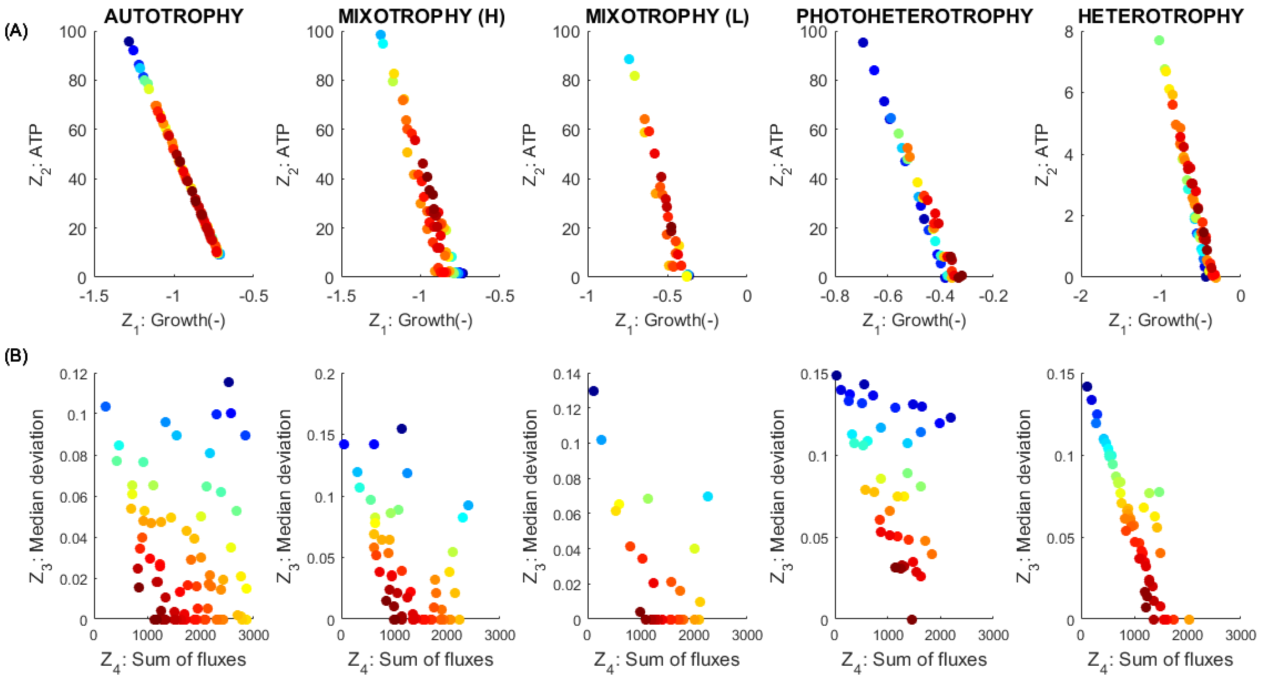


Figure 6.12: *Bi-dimensional objective spaces of the Pareto front approximation obtained with Meta-MODEunder all conditions.*

(A) Trade-off between metabolic performance objectives (growth - ATP synthesis).

(B) Trade-off between between metabolic pertinency objectives (parsimony - closeness).

Effect of flux trough central pathways on the objective functions

In order to analyse how the flux through the central metabolic pathways affects the values of the objective functions Level Diagrams have been plotted (Figures from 6.13 to 6.17) that show the values of the objectives as well as the values of flux at a set of selected reactions (Table 6.3).

In relation with metabolic performance, as it was to be expected, in general, greater fluxes through all pathways correspond to greater growth (as it is also observed when analysing the relationship between growth and parsimony). Particular differences between trophic conditions can be observed for example in the 6-phosphofructokinase (PFK) and RuBisCO (RBCO) reactions: the greater the contribution of glucose (sequenced from heterotrophy to autotrophy) the greater the dominance of PFK over RBCO and vice versa.

Regarding metabolic pertinency, it can be observed that, even when these reactions where not used as reference for the closeness objective, since they are closely related to other reactions at the same pathways that were among the experimental values used, the flux values that show good scores in terms of metabolic pertinency (dark red colours) are distributed around concrete values. This effect is specially visible in the case of the fumarate hydratase (FUMH), that shows a small interval of values in all trophic modes. This fact shows that using a small set of experimental values helps to adjust a greater set of fluxes, that is, the accuracy of the flux distribution, and as a consequence the metabolic pertinency, is improved by using experimental information from only a few reactions.

As for the distribution of fluxes observed through the different pathways under different growth conditions, it is remarkable the case of the enzyme RuBisCO (RBCO) which shows the activity of the Calvin-Benson-Bassham cycle. This enzyme shows greater flux in the case of autotrophy, and its flux decreases while the contribution of the autotrophic component decreases, reaching null values in the case of heterotrophy. Another reaction that displays a characteristic change among

Table 6.3: *Reactions selected to monitor the effect of the flux at certain metabolic pathways over the values of the objective functions. None of the selected reactions has been used for closeness objective or closeness constraints. See page xx for the list of acronyms.*

Reaction	**Name in *i*Syn842**	**Pathway**	**Stoichiometry**
PFK	2.7.1.11	Glycolysis/gluconeogenesis	$ATP + F6P \rightarrow ADP + FBP$
FUMH	4.2.1.2	Citric Acid Cycle	$MAL \rightleftarrows FUM + H_2O$
TAL	2.2.1.2	Pentose Phosphate Pathway	$GAP + S7P \rightleftarrows F6P + E4P$
RBCO	4.1.1.39	Calvin-Benson-Bassham Cycle	$RUBP + CO_2 + H2O \rightarrow 2\ 3PG + 2\ H^+$

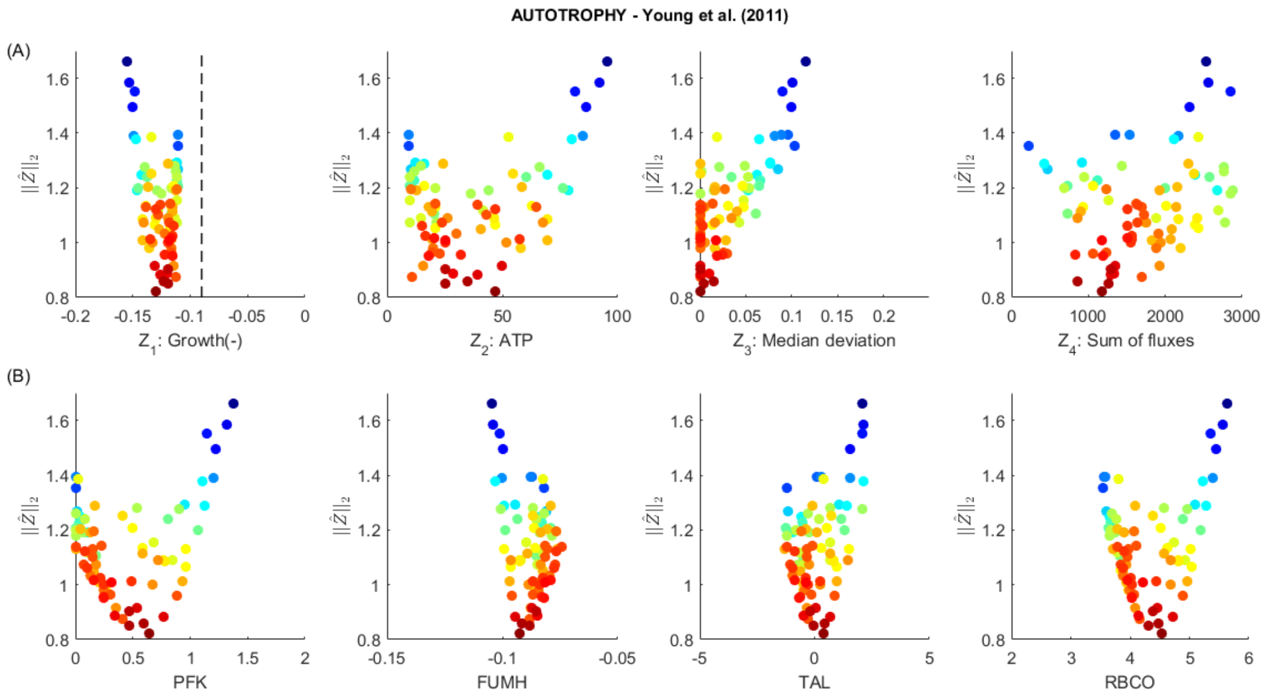


Figure 6.13: *Pareto front and Pareto set approximations found with Meta-MODE under autotrophic conditions from Young et al. (2011)*
(A) Level Diagrams of the four objectives: $Z_1(\boldsymbol{v})$*: growth;* $Z_2(\boldsymbol{v})$*: min. ATP;* $Z_3(\boldsymbol{v})$*: closeness;* $Z_4(\boldsymbol{v})$*: parsimony. Vertical line in the diagram of growth objective shows experimental growth.*
(B) Level Diagrams of the flux values at four selected reactions from different central carbon pathways (Table 6.3).

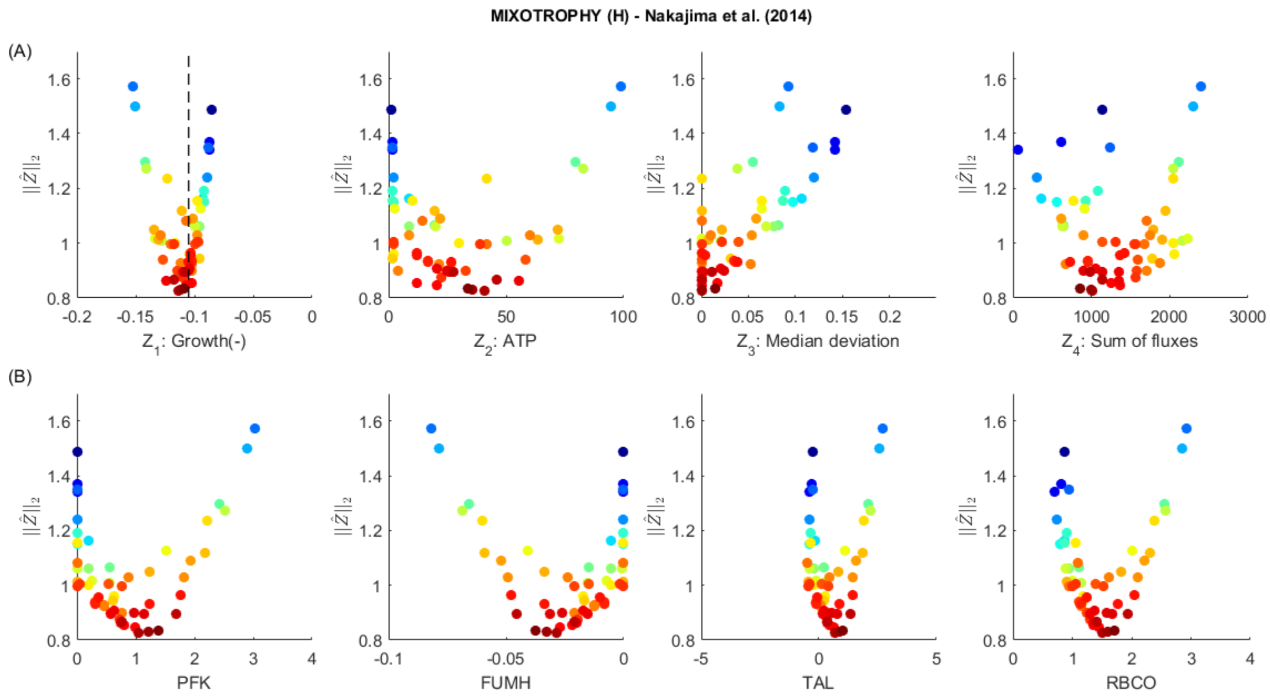


Figure 6.14: *Pareto front and Pareto set approximations found with Meta-MODE under mixotrophic (H) conditions from Nakajima et al. (2014)*

(A) Level Diagrams of the four objectives: $Z_1(\boldsymbol{v})$: growth; $Z_2(\boldsymbol{v})$: min. ATP; $Z_3(\boldsymbol{v})$: closeness; $Z_4(\boldsymbol{v})$: parsimony. Vertical line in the diagram of growth objective shows experimental growth.

(B) Level Diagrams of the flux values at four selected reactions from different central carbon pathways (Table 6.3).

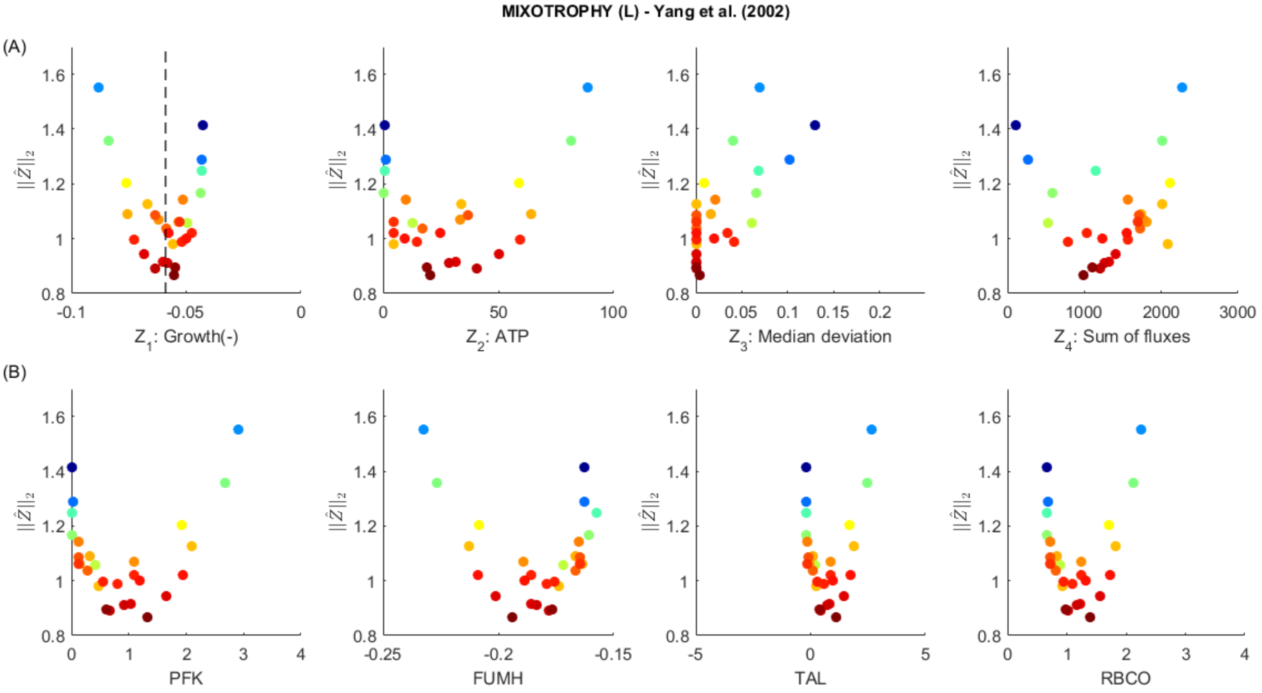


Figure 6.15: *Pareto front and Pareto set approximations found with Meta-MODE under mixotrophic (L) conditions from Yang et al. (2002)*
(A) Level Diagrams of the four objectives: $Z_1(\boldsymbol{v})$*: growth;* $Z_2(\boldsymbol{v})$*: min. ATP;* $Z_3(\boldsymbol{v})$*: closeness;* $Z_4(\boldsymbol{v})$*: parsimony. Vertical line in the diagram of growth objective shows experimental growth.*
(B) Level Diagrams of the flux values at four selected reactions from different central carbon pathways (Table 6.3).

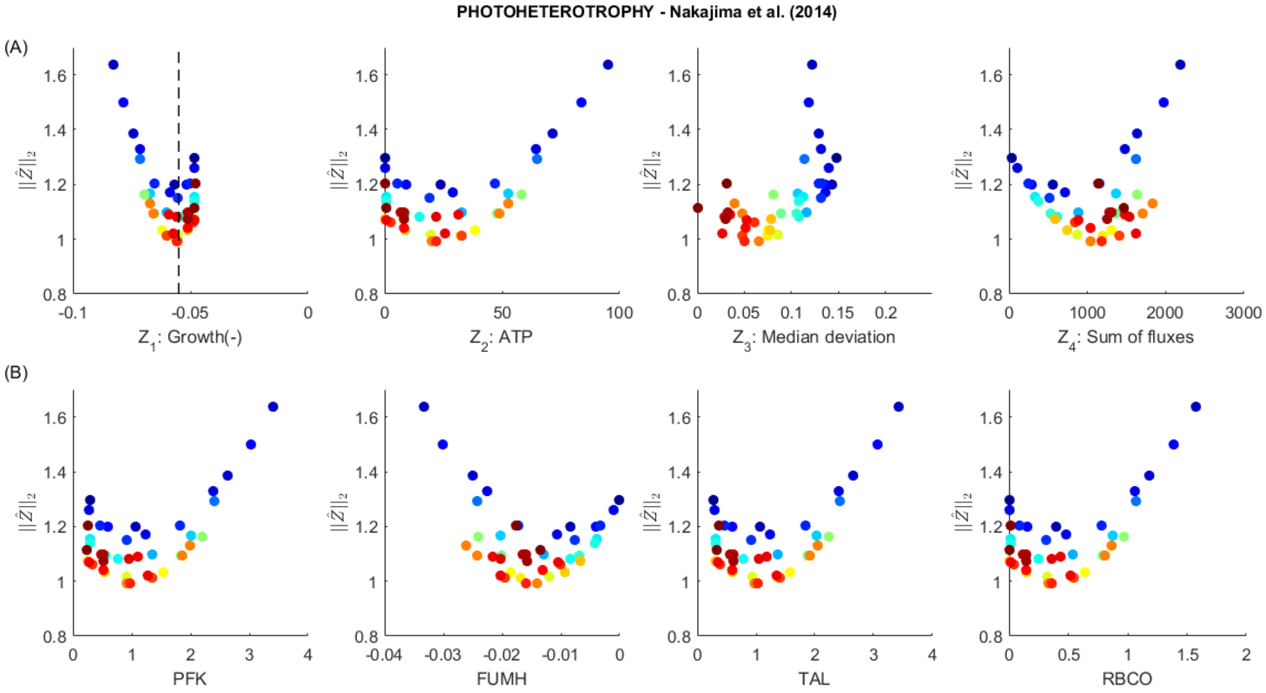


Figure 6.16: *Pareto front and Pareto set approximations found with Meta-MODE under photoheterotrophic conditions from Nakajima et al. (2014)*

(A) Level Diagrams of the four objectives: $Z_1\,(\boldsymbol{v})$: growth; $Z_2\,(\boldsymbol{v})$: min. ATP; $Z_3\,(\boldsymbol{v})$: closeness; $Z_4\,(\boldsymbol{v})$: parsimony. Vertical line in the diagram of growth objective shows experimental growth.

(B) Level Diagrams of the flux values at four selected reactions from different central carbon pathways (Table 6.3).

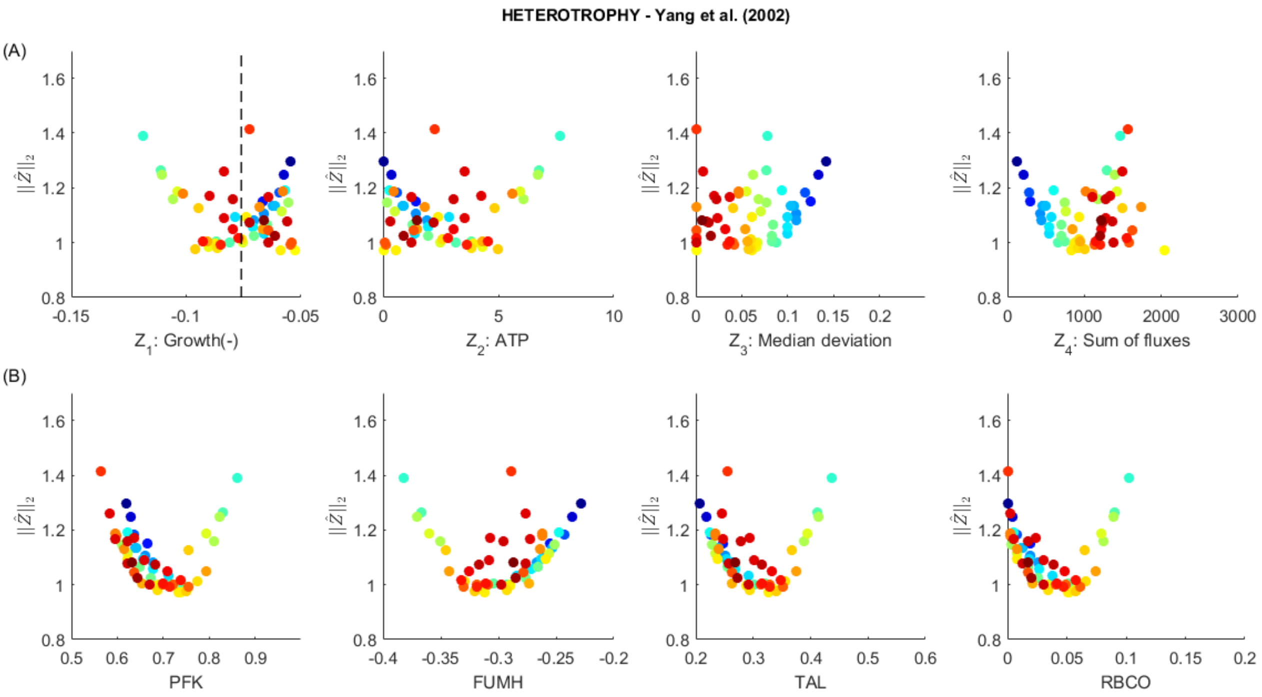


Figure 6.17: *Pareto front and Pareto set approximations found with Meta-MODE under heterotrophic conditions from Yang et al. (2002)*
(A) Level Diagrams of the four objectives: $Z_1(\boldsymbol{v})$: growth; $Z_2(\boldsymbol{v})$: min. ATP; $Z_3(\boldsymbol{v})$: closeness; $Z_4(\boldsymbol{v})$: parsimony. Vertical line in the diagram of growth objective shows experimental growth.
(B) Level Diagrams of the flux values at four selected reactions from different central carbon pathways (Table 6.3).

trophic conditions is the transketolase (TAL) that shows the inversion of the pentose phosphate pathway according to the availability of light and CO_2: the flux at this reaction moves more to the positive values as the heterotrophic component is more important.

6.4.3 Comparison between mono and multi-objective optimisation approaches

In order to benchmark the algorithm proposed in this work, the results of Meta-MODE have been compared to the results of the commonly used FBA, as well as to an FBA that incorporates as constraints the information on 10 internal reactions for which experimental values were available (Table 6.2), termed FBA_exp (see Section 6.3).

Metabolic efficiency and agreement with observed fluxes

Figure 6.18 shows a comparison among the values obtained from the five solutions selected from Meta-MODE's Pareto set for each of the four objectives and the corresponding values calculated from the solutions obtained from the two mono-objective methods. It is important to highlight that since FBA and FBA_exp are mono-objective optimisation algorithms not all the four values shown for these methods are "objective" values. In these mono-objective methods, the optimised function was maximisation of growth (first value from the left in Figure 6.18) defined as flux through the biomass equation. The other three values were calculated from the flux distribution of the mono-objective optimal solutions. The values shown in these plots have been normalised with respect to maximum and minimum values in the set of solutions of all three methods, so that minimum value translates to 0 and maximum value translates to 1. In the case of growth, also the experimental value was added.

It can be observed in Figure 6.18 that the behaviour of the two mono-objective methods has in general an opposite trend to the Meta-MODE:

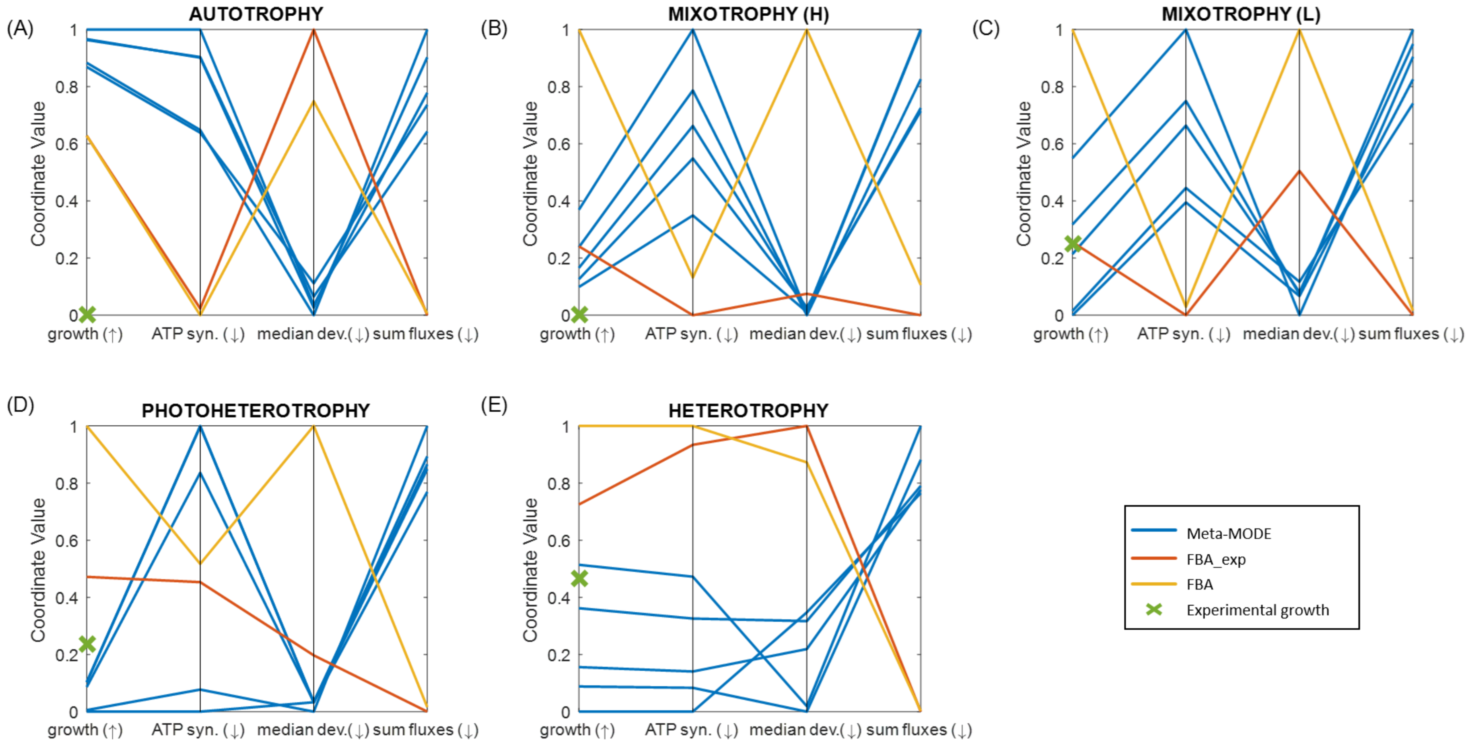


Figure 6.18: ***(A-E)*** *Comparison of growth rate, ATP synthesis, deviation from experimental fluxes and total sum of fluxes, shown by solutions from the different simulation methods. Coordinate values are normalised with respect to maximum and minimum values in the set of solutions of all three methods. (A) Autotrophy; (B) Mixotrophy (H); (C) Mixotrophy (L); (D) Photoheterotrophy; (E) Heterotrophy. Arrows indicate the direction of the desired values. Experimental values are added in the case of growth rate.*

- Regarding growth, in general, FBA tends to yield solutions with growth rate values quite higher than the obtained with Meta-MODE. FBA_exp results in intermediate growth values.
- In terms of ATP synthesis, even for greater growth values, both FBA and FBA_exp simulations predict lower energy requirements than Meta-MODE.
- Deviation from experimental fluxes in the case of FBA is always much greater than the Meta-MODE results, while FBA_exp sometimes achieves intermediate values.
- And finally, relative to parsimony, FBA and FBA_exp show also lower values for the total sum of fluxes, even with greater growth rates than Meta-MODE.

Overall, taking into account the values for the four functions, FBA solutions show very high metabolic efficiency, they achieve high growth rates with low energy (ATP) and activation (sum of fluxes) requirements. However, regarding closeness to observed behaviour, they fail in approximating experimental flux values. In the case of FBA_exp, the general trend of the solutions is quite similar to those of FBA, which was expectable since they use the same principles. However, the addition of the experimental measurements to the set of constraints moderately moves the results from extreme metabolic performance towards a better metabolic pertinency.

Estimation of growth rate and yield

If the focus is put on growth (left value in plots from Figure 6.18), taking into account the experimental values, it can be seen that, in general, FBA, and to a lesser extent FBA_exp, tend to overestimate growth rate. And, as it was already remarked, they predict low energy requirements to achieve those growth rates. To evaluate the realism of the metabolic efficiency displayed by each kind of simulations, biomass-glucose yields were calculated (Table 6.4 and Figure 6.19). In the case of Meta-

MODE values, the mean and standard deviation of the five selected values are shown.

Table 6.4: *Biomass/glucose yield ($Y_{X/S}$) shown by solutions from the different simulation methods under different trophic conditions. Standard deviations are shown between parentheses for Meta-MODE.*

	mixo (H)	mixo (L)	photohetero	hetero
Experimental	1.10	0.87	0.53	0.50
Meta-MODE	1.21 (0.06)	0.85 (0.05)	0.48 (0.01)	0.45 (0.04)
FBA_exp	1.23	0.86	0.59	0.55
FBA	1.63	1.04	0.74	0.61

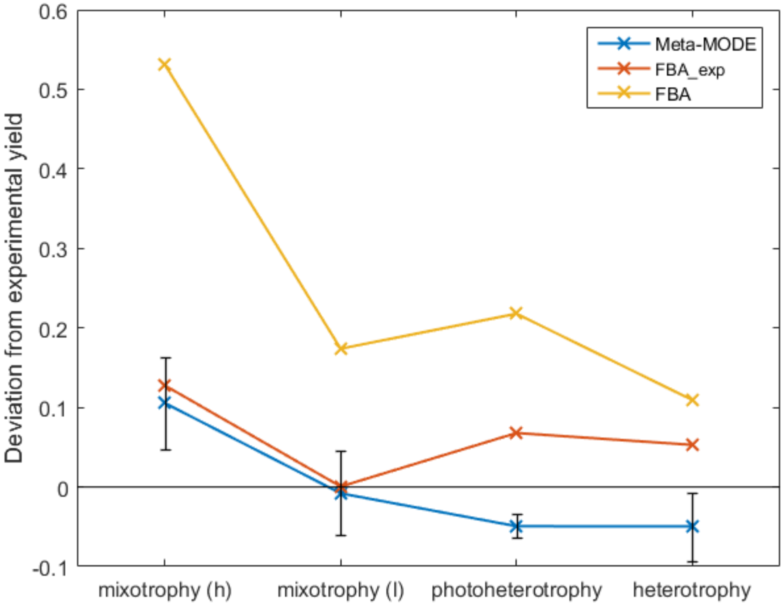


Figure 6.19: *Deviation from experimental value of biomass/glucose yield ($Y_{X/S}$) shown by solutions from the different simulation methods under different trophic conditions. Standard deviations are shown by error bars for Meta-MODE.*

Both Figures 6.18 and 6.19 show that the results of the multi-objective simulation are as closer to the experimental values of yield and growth than the ones from FBA and FBA_exp. Meta-MODE is closer to the experimental values of biomass-glucose yield and growth in the heterotrophic, photoheterotrophic and mixotrophic (H) conditions. In the particular case of autotrophic simulations (Figure 6.18, A), none of the tested methods is able to approximate the experimental growth, which will be discussed later (Section 6.4.4).

Interestingly, Meta-MODE always produces lower values of yield and growth than FBA or FBA_exp. These mono-objective algorithms use techniques to bring their growth and yield values closer to the experimental ones by adding maintenance costs that are meant to reflect energy costs (for maintenance, cell division, *etc.*) not contained in the metabolic model. Meta-MODE algorithm is able to better reflect realistic metabolic efficiency without those mathematical constructs.

Closeness to measured internal fluxes

Using the fuzzy metric described in Section 6.3 (Equation (6.15)), the similarity between flux vectors obtained from different methods and sets of experimental fluxes was evaluated. These similarity values were then used to plot the clustering dendrograms shown in Figures 6.20 and 6.21. Figure 6.20 shows the dendrograms obtained considering only the subset of 10 fluxes, common to all conditions (Table 6.2), that have been used to define closeness constraints and objective. Figure 6.21 shows the clusters obtained with the extended sets of measured fluxes (see Additional file 6.4 for the complete sets of measured fluxes).

As it can be seen, mono-objective simulations based on classic FBA methodology appear in all cases far from experimental values, while multi-objective simulations performed with Meta-MODE algorithm are close to the experimental set and always cluster in the same group as experimental values. FBA_exp simulations display an intermediate agreement with experimental fluxes. It must be taken into account that information about part of these experimental fluxes (the subset of 10) was included during the optimisation process in Meta-MODE and FBA_exp and not in normal FBA.

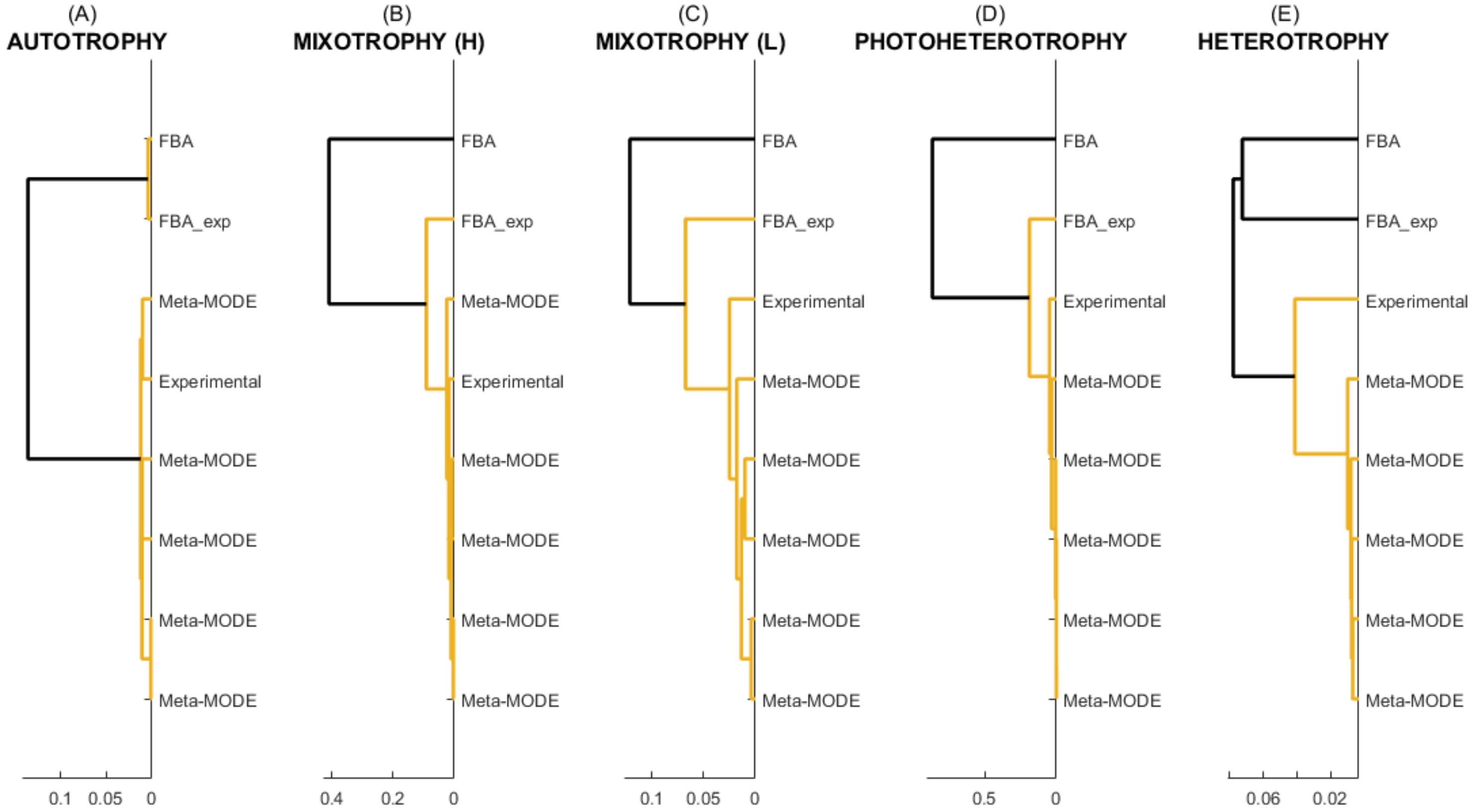


Figure 6.20: ***(A-E)*** *Simulation methods clustered according to similarity in flux values of the subset of 10 internal fluxes considered for closeness objective and constraints. (A) Autotrophy; (B) Mixotrophy (H); (C) Mixotrophy (L); (D) Photoheterotrophy; (E) Heterotrophy.*

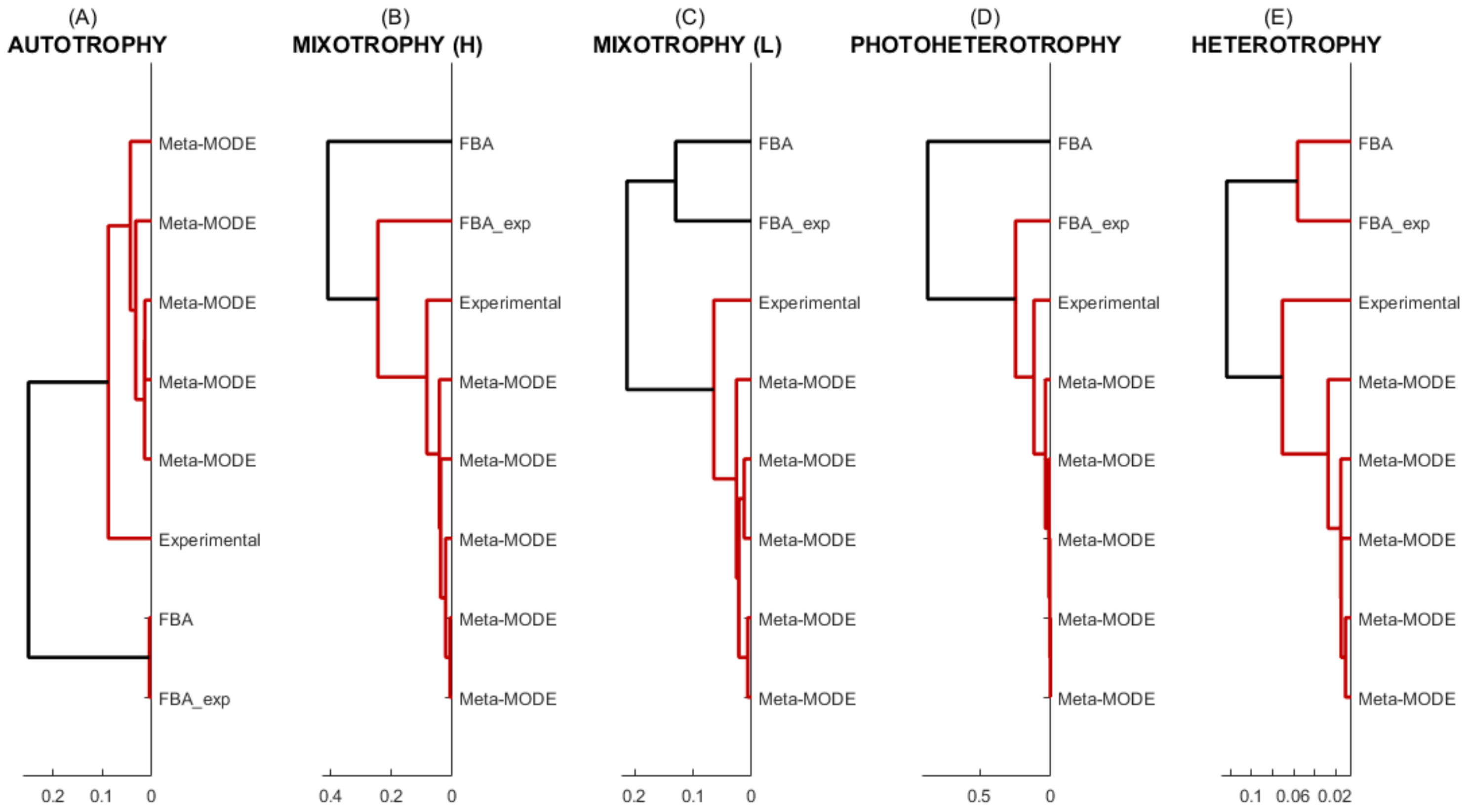


Figure 6.21: ***(A-E)*** *Simulation methods clustered according to similarity in flux values of extended sets of internal fluxes (see Additional file 6.4 for the list of internal fluxes measured in each condition). (A) Autotrophy; (B) Mixotrophy (H); (C) Mixotrophy (L); (D) Photoheterotrophy; (E) Heterotrophy.*

Condition-specific biomass composition of simulations based on multi-objective optimisation

As it was explained in Chapter 1, mono-objective linear optimisation algorithms for simulating metabolic fluxes rely on definition of biomass equations describing fixed proportions of biomass components. On the contrary, in Meta-MODE biomass evolution is considered as the independent production of biomass elements, without the description of a specific biomass equation. In fact, this algorithm allows the study of the proportion of those biomass components as a result of the simulation.

In this study, five different environmental conditions were considered for which internal flux measurements were available from literature. Using this information, and considering closeness to measured fluxes as an objective, the proportion of biomass components observed in the selected solutions appear to be adjusted to match the flux distributions observed *in vivo*.

Figure 6.22 shows the proportions of biomass components obtained from the simulations carried out with the different methods. FBA and FBA_exp always exhibit the same biomass composition, since it was previously fixed through the biomass equation. However, the solutions obtained using Meta-MODE display a range of different biomass compositions for the different simulated environmental conditions. Furthermore, the multiple solutions obtained from the Pareto front approximation for each simulation show slightly different compositions, while they present similar performance in terms of the four objectives considered. This fact highlights the interest of considering multiple non-dominated solutions, instead of one single mono-objective optimal flux distribution, which better captures the diversity observed in actual cells.

Some observations can be extracted from the biomass compositions shown in Figure 6.22. In all conditions the greatest part of the biomass is made of amino acids, and the distribution among them is not even, some amino acids appear to have greater contribution than others. In the cases of autotrophic and the two mixotrophic simulations, the con-

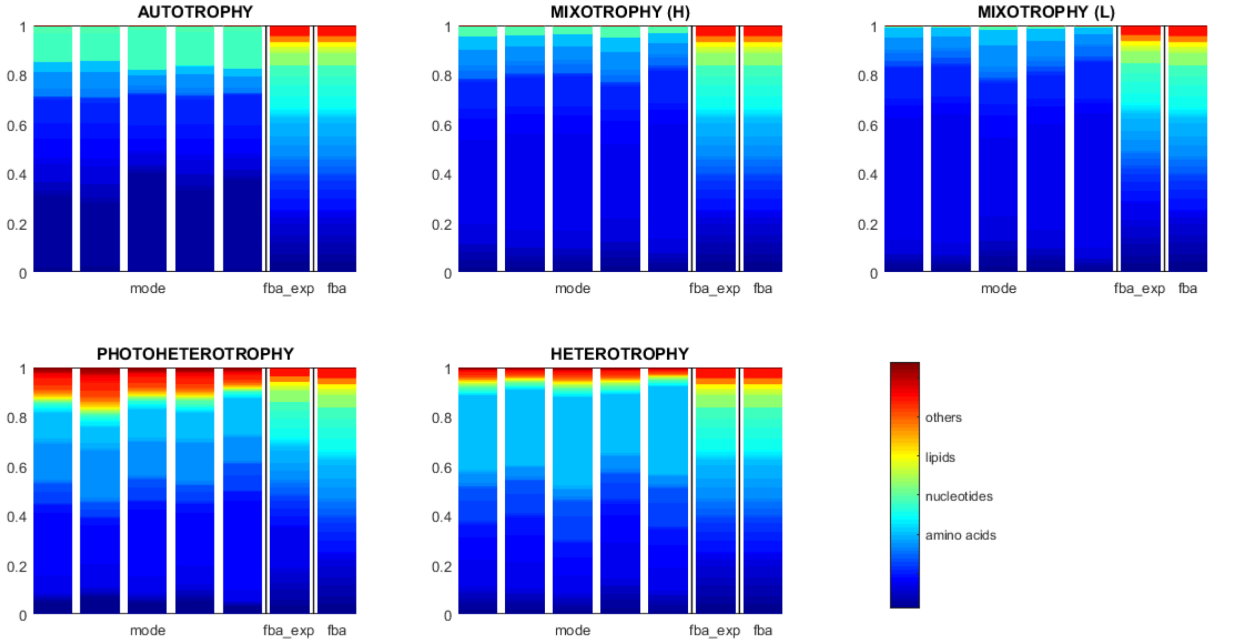


Figure 6.22: *Biomass composition of the solutions obtained with different simulation methods under different trophic conditions.*

tribution of other metabolites is very low (they are present, even when the small proportion is not perceptible in the bar plot). In the cases of heterotrophic and photoheterotrophic conditions the contribution of lipids, nucleotides and trace elements is more noticeable, especially in the case of photoheterotrophy. Comparing the biomass compositions obtained from multi-objective simulations with the biomass equation fixed in the case of mono-objective algorithms, it can be concluded that the proposed biomass equation in mono-objective simulations (taken from Montagud et al. (2011)) is more adequate for simulations under heterotrophy and, with some limitations, photoheterotrophy.

6.4.4 Different trophic conditions exhibit different behaviour in simulation results

In this chapter the multi-objective optimisation algorithm proposed in this work for metabolic simulations has been tested under five trophic conditions. The results from those simulations shown at this section present some common characteristics that have been exposed. However, they present some particularities depending on the studied conditions.

The first thing that must be taken into account is the degree of flexibility that each condition implies. Those conditions in which the cell can make use of glucose, CO_2 and sunlight at the same time, *i.e.* mixotrophic conditions, give the system much flexibility, and allow for a better reorganisation of internal fluxes. This makes it easier for the system to achieve states closer to the ideal state, and thus both metabolic performance and metabolic pertinency requirements can be fulfilled in an easier way. This fact can be observed in Figures 6.8 and 6.9, where the agreement between distance to the utopian solution (vertical axis in Level Diagrams) and weighted distance to the ideal closeness-parsimony trade-off (colour code) is almost perfect.

On the other hand, under more restricted situations, like the case of heterotrophic and photoheterotrophic conditions, the system is more constrained, and moving toward more pertinent solutions has a bigger cost on metabolic performance and thus, it is more difficult to approach the ideal solution (see Figures 6.11 and 6.10).

The case of autotrophic conditions needs its own separated analysis. It can be seen in Figures 6.7 (Pareto front in LD) and 6.18 (comparison of methods), that none of the tested methods was able to appropriately estimate experimental growth rate since all methods overestimated this objective. This fact contrasts with all the other conditions studied in which Meta-MODE, and FBA_exp in some cases, approximated the values quite well. Some reasons for these discrepancies can be related to the model, the experimental fluxes and the intrinsic complexity of autotrophic metabolism. With respect to the model, it could be possible that the pathways implied in autotrophy need further revision and verification in order to ensure that they represent adequately autotrophic metabolism. However, if the case would be that the metabolic model is incomplete or inaccurate in describing light-driven reactions, it would also affect, even when in a lesser extent, mixotrophic simulations.

Furthermore, apart from agreement between measured and estimated growth, the performance of the solutions of autotrophic simulations in terms of distance to the ideal solutions and metabolic pertinency are as good as the ones obtained from mixotrophic simulations (see LD in Figure 6.7). This fact suggests that the simulation was able to fulfil the optimisation requirements, which were high metabolic efficiency (maximisation of growth and minimisation of ATP synthesis), moderated metabolic activity (minimisation of total sum of fluxes) and closeness to experimental fluxes (minimisation of deviation); but the emerging value of growth did not concur with the experimental one. At this point it could be interesting to mention that the technique used to retrieve internal metabolic fluxes at the laboratory under autotrophic conditions differ from the standard technique applied when glucose (or other organic molecules) is used as carbon source. The fact that the only carbon input under autotrophy is CO_2, which is a single carbon molecule, makes it necessary to turn to a special procedure called *Isotopically Non-Stationary Metabolic Flux Analysis* (INST-MFA) (see Young et al. (2011)

for details), which measures carbon fluxes under metabolic steady state but not isotopic steady state. It could be possible that the cause of the discrepancy between measured and simulated growth could reside, at least in part, in the method used to measure the fluxes, which could have an effect on the accuracy of the data obtained. However, further simulations and analyses are needed to assess the sources of this discrepancy.

6.5 Conclusions

The aim of this chapter was to prove the use and application of the simulation tool based on multi-objective evolutionary optimisation described in the previous chapter and to assess its performance in comparison to well-established methods. For this, it has been applied to simulate *Synechocistis*' metabolic network under five trophic conditions, and it has been compared with the classic FBA method based on mono-objective linear optimisation.

The results provided in this chapter show that using the multi-objective algorithm proposed in this work, metabolic simulations can be performed in which the inclusion of a few internal fluxes obtained from measurements allows to readjust the whole flux distribution, thus avoiding the necessity of hard, sometimes difficult to validate, restrictions that try to reflect costs and limitations that the cells must face.

The solutions chosen from simulations with Meta-MODE show a quasi-optimal state in which a balance is found between metabolic performance and metabolic pertinency. Thanks to the observation of the Pareto front, instead of obtaining a single mono-objective optimal solution, it was possible to analyse the trade-off between objectives and to realise that solutions less optimal in terms of metabolic efficiency better describe the observed behaviours. This also raises the question if this optimality, even assuming it is a biological objective, is something ever reached in nature.

This methodology also avoids the necessity of a biomass equation to account for biomass formation. Instead, under this scheme, the biomass composition arises as a consequence of the studied conditions, instead of being imposed beforehand. As a result of the simulation, slightly different phenotypes, including biomass composition, will appear that match the known internal fluxes and show quasi-optimal metabolic performance.

With the use of measured fluxes the results of these simulations have been validated. It has been shown that, with the inclusion of a reduced subset of experimental fluxes, the obtained flux distributions resemble those described in literature, and they approximate the experimental data closer than other classic methods that have already proved their applicability.

Finally, it has been shown that the algorithm proposed here is flexible to tally experimental observations under different environmental conditions adapting the flux behaviour to the specific situation. This way, it has been seen that trophic modes that offer the cell more varied sources of carbon and energy confer it more plasticity, which allows reaching more efficient states closer to the ideal performance. On the other hand, more restricted conditions limit the operability of the cell and lead to less optimal phenotypes.

Overall, this chapter shows that the use of multi-objective approaches to simulate metabolic phenotypes offers interesting opportunities to further analyse implications of different factors over the metabolic performance of the cell. The simulations performed in this way relax some impositions needed in classic approaches, which gives them more plasticity to adapt to different simulation conditions. The final aim of metabolic modelling is to describe, as accurately as possible, the behaviour of metabolic systems, and their response to perturbed conditions. Thus, it is advantageous to have a simulation tool that appropriately describes real metabolic traits, such as growth, and shows plasticity to perform well under changing conditions, when some experimental information is available.

However, the tool proposed in the present work is, in its current state, limited to analysis purposes and could not be used for prediction, since experimental fluxes are needed from *in vivo* experiments. Further work is needed to include additional simulation techniques in order to make possible the prediction of the metabolic response to genetic and/or environmental perturbations taking advantage of the *wild type* measurements as a reference. Some ideas have been planned in this sense that make use of techniques similar to MOMA and ROOM (see Section 1.2.2) that can be easily combined with the current objectives and constraints due to the multi-objective and evolutionary characteristics of the algorithm. These ideas will be the following areas of research of present PhD candidate.

Conclusions and closing remarks

Conclusions and closing remarks

The main objective of this thesis was to contribute, with models and algorithms, to metabolic modelling techniques in cyanobacteria. These organisms are photoautotrophic bacteria that can perform oxygenic photosynthesis to obtain energy from sunlight and carbon from inorganic CO_2. Their low energetic requirements, make them very interesting organisms for the design of cell factories. For this purpose, the natural organisms have to be modified to enhance their productive capacities and improve their robustness. But rational design needs planning: it is crucial to assess the effect that genetic and environmental modifications will have on the whole metabolism of the organisms and how it will affect their performance. Constraint-based metabolic modelling offers the opportunity to perform genome-scale simulations that contribute to the understanding of the metabolic behaviour of the cells and their response under perturbed conditions.

In order to obtain reliable simulation results, both accurate metabolic models and functional simulation algorithms are needed. The aim of this thesis was to provide tools to enhance the plasticity and the predictive capacities of the constraint-based simulations of cyanobacteria. Following this goal, in **Chapter 3** reconstruction of metabolic models of two cyanobacterial species was addressed. Consideration of as much accurate information as possible during the reconstruction process is fundamental to obtain a functional model that enables the analysis of genome-scale flux distributions. Besides, continuous update and improvement of the models with knowledge retrieved from last discoveries is important to enhance the simulations and ensure their applicability. The metabolic model of *Synechocystis* sp. PCC 6803 obtained in this chapter includes important improvements with respect to the previous version, that lead to better description of the energy metabolism, which turns into more realistic energy costs, and fewer amount of constraints, which contributes to greater plasticity. The metabolic model of *Synechococcus elongatus* PCC 7942 presented in this chapter was the first genome-scale metabolic reconstruction of this organism. It has been qualitative and quantitatively compared with *Synechocystis*' network and it was observed that these two organisms have many common traits. Further improvement of these models would involve, besides continuous inclusion of up-to-date metabolic knowledge, a more accurate definition of biomass components, including classification of the proteins by function. Such level of detail exceeds the published experimental information available by now in these organisms, but when possible, it would notably improve the mathematical definition and simulation of cell growth.

In **Chapter 4** Flux Balance Analysis methodology was applied to perform simulations with the models reconstructed in the previous chapter. Using this algorithm several studies were conducted to assess metabolic robustness, natural and perturbed physiological states and productive capacities of the two cyanobacteria. It would be extremely interesting to have experimental data available in order to verify the results of the simulations performed in this chapter. Nevertheless the experiments needed to obtain these data are beyond the scope of present PhD thesis. This would however make an excellent future collaborative project with experimental colleagues.

Despite the variety of analyses that can be performed with FBA, important limitations of this approach were detected that motivated the research made in the following part of this thesis. Among these limitations, it is important to note the limited range of mathematical functions that can be used under this linear approach to describe informative biological constraints and objectives, the reliance on hard constraints to tailor simulations, and the strong dependency of the simulation results on a single objective function that must describe the complexity of metabolic behaviour. Altogether these limitations motivate the search of new approaches that circumvent them without losing the convenient simplicity of constraint-based simulation techniques. In this PhD thesis fundamentals of multi-objective optimisation procedures are applied to the field of constraint-based metabolic modelling with the aim of improving plasticity and reliability of the obtained solutions. For this purpose, in **Chapter 5** a multi-objective framework is presented to define and optimise multi-objective problems that ensure the metabolic pertinency of the solutions. The optimisation algorithm presented in this chapter includes a set of mechanisms to deal with the resulting optimisation statements. This multi-objective evolutionary tool avoids some of the main drawbacks of classic mono-objective linear optimisation methods: it allows definition of multiple biologically significant objectives, it enables the use of non-linear functions to describe metabolic constrains and objectives, and it allows the integration of experimental data that serve to define soft constraints which relaxes the impositions made in classical approaches to obtain practical solutions.

This multi-objective metabolic modelling methodology is used in **Chapter 6** in order to validate its applicability and to benchmark the results with the obtained from widely applied methods. The results obtained in this chapter prove that the proposed methodology is applicable to genome-scale metabolic simulations and it presents important advantages. Importantly, it allows the analysis of a collection of Pareto optimal solutions, and permits the exploration of trade-off and relation between objectives, which further enriches the analysis. Thanks to this property it was possible to observe that the simulation results present an equilibrium between metabolic performance and metabolic pertinency, which points out that metabolic systems operate in a quasi-optimal metabolic state that cannot be fully described from a mono-objective perspective. Besides, the non-linear constraints and objective functions defined in this work ease the description of meaningful rules that govern metabolic behaviour. The obtained results have proved to better approach important metabolic characteristics like growth rate and biomass yield.

In order to further validate the applicability and interest of the multi-objective evolutionary algorithm proposed in this thesis to perform metabolic simulations, it would be interesting, in future projects, to take advantage of the enormous amount of experimental information available for better studied model organisms, like *Escherichia coli* and *Saccharomyces cerevisiae*. In these organisms several works have been conducted towards strain characterisation; using data from such studies would allow to analyse in more detail the characteristics of the solutions obtained with this simulation tool and compare them with the well-characterised metabolic behaviour of these organisms.

Regarding the expansion and improvement of the current simulation tool, the future perspectives for this work are promising. The multi-objective approach described in this thesis can be extended to broaden the range of possibilities that it can offer to the metabolic simulation toolbox. Integration of other types of experimental data, such as metabolomic, transcriptomic, and proteomic data, have to be in the scope of future works. New strategies must be defined to appropriately include these different kinds of measurements, keeping in mind the inherent complexity of the relationship between these different levels of information.

Furthermore, the line of *analysis* must be crossed to contribute to the set of tools that address the tuning and redesign of engineered organisms. Additional simulation techniques must be added in order to make possible the prediction of the metabolic response to genetic and/or environmental perturbations taking advantage of the wild type measurements as a reference. Some ideas have been planned in this sense that make use of techniques similar to those used by algorithms like *Minimisation Of Metabolic Adjustment* (MOMA) and *Regulatory On/Off Minimisation* (ROOM) that can be easily combined with the current objectives and constraints due to the multi-objective and evolutionary characteristics of the algorithm. Multi-objective optimisation and multi-criteria decision making approaches can be applied to evaluate, compare and actively search metabolic engineering strategies that guide experimental efforts in order to design efficient and environmentally-friendly cell factories for the production of high-value products, such as biofuels, drugs or refined chemicals.

From the algorithmic point of view there is space to notable improvements in terms of performance. In this work the attention was focused on verification of the quality of the final solutions, in terms of metabolic pertinency, to validate the multi-objective problem defined in this thesis. The next step is now to conduct further investigations to assess the performance of different algorithms to solve this multi-objective problem. A collaborative work in this sense is now starting.

In this PhD thesis the groundwork was laid for the application of a new optimisation and analysis framework to the field of constraint-based metabolic modelling. From this point on, new strategies and techniques can be developed that should contribute to the set of tools available for simulation and design of living metabolic systems.

Bibliography

Abed, R. M. M., Dobretsov, S., Sudesh, K., 2009. Applications of cyanobacteria in biotechnology. Journal of applied microbiology 106 (1), 1–12.

Alon, U., 2007. An introduction to systems biology : design principles of biological circuits. Chapman & Hall/CRC.

Anderson, S. L., Mcintosh, L., 1991. Light-activated heterotrophic growth of the Light-Activated Heterotrophic Growth of the Cyanobacterium Synechocystis sp . Strain PCC 6803 : a Blue-Light-Requiring Process. Journal of Bacteriology 173 (9), 2761–2767.

Angermayr, S. A., Hellingwerf, K. J., Lindblad, P., de Mattos, M. J. T., 2009. Energy biotechnology with cyanobacteria. Current opinion in biotechnology 20 (3), 257–63.

Baebprasert, W., Jantaro, S., Khetkorn, W., Lindblad, P., Incharoensakdi, A., 2011. Increased H(2) production in the cyanobacterium Synechocystis sp. strain PCC 6803 by redirecting the electron supply via genetic engineering of the nitrate assimilation pathway. Metabolic engineering 13 (5), 610–616.

Barabasi, A. L., Oltvai, Z. N., 2004. Network Biology: Understanding the Cell's Functional Organization. Nature Review 5 (February), 101–113.

Baroukh, C., Muñoz-Tamayo, R., Steyer, J.-P. P., Bernard, O., 2015. A state of the art of metabolic networks of unicellular microalgae and cyanobacteria for biofuel production. Metabolic Engineering 30, 49–60.

Beard, D. A., Liang, S.-d., Qian, H., 2002. Energy Balance for Analysis of Complex Metabolic Networks. Biophysical Journal 83 (1), 79–86.

Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. Ø., , Herrgard, M. J., 2007. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nature protocols 2 (3), 727–38.

Becker, S. A., Palsson, B. O., 2008. Context-specific metabolic networks are consistent with experiments. PLoS computational biology 4 (5), e1000082.

Blasco, X., Herrero, J. M., Sanchis, J., Martínez, M., 2008. A new graphical visualization of n-dimensional Pareto front for decision-making in multiobjective optimization. Information Sciences 178 (20), 3908–3924.

Bonissone, P. P., Subbu, R., Lizzi, J., 2009. Multicriteria decision making (MCDM): a framework for research and applications. Computational Intelligence Magazine, IEEE 4 (3), 48–61.

Bordbar, A., Monk, J. M., King, Z. A., Palsson, B. Ø., 2014. Constraint-based models predict metabolic and associated cellular functions. Nature Reviews Genetics 15 (2), 107–120.

Brandes, A., Lun, D. S., Ip, K., Zucker, J., Colijn, C., Weiner, B., Galagan, J. E., 2012. Inferring carbon sources from gene expression profiles using metabolic flux models. PloS one 7 (5), e36947.

Branke, J., Deb, K., Miettinen, K., Slowinskyi, R., 2008. Multiobjective Optimization - Interactive and Evolutionary Approaches. Vol. 5252 LNCS. Springer-Verlag.

Braunstein, A., Mulet, R., Pagnani, A., 2008. Estimating the size of the solution space of metabolic networks. BMC Bioinformatics 9 (1), 240.

Broddrick, J. T., Rubin, B. E., Welkie, D. G., Du, N., Mih, N., Diamond, S., Lee, J. J., Golden, S. S., Palsson, B. O., 2016. Unique attributes of cyanobacterial metabolism revealed by improved genome-scale metabolic modeling and essential gene analysis. Proceedings of the National Academy of Sciences 113 (51), E8344–E8353.

Burgard, A. P., Nikolaev, E. V., Schilling, C. H., Maranas, C. D., 2004. Flux coupling analysis of genome-scale metabolic network reconstructions. Genome Research 14 (2), 301–312.

Burgard, A. P., Pharkya, P., Maranas, C. D., 2003. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnology and bioengineering 84 (6), 647–57.

Burrows, E. H., Wong, W.-K., Fern, X., Chaplen, F. W. R., Ely, R. L., 2009. Optimization of pH and nitrogen for enhanced hydrogen production by Synechocystis sp. PCC 6803 via statistical and machine learning methods. Biotechnology progress 25 (4), 1009–17.

Caspeta, L., Shoaie, S., Agren, R., Nookaew, I., Nielsen, J., 2012. Genome-scale metabolic reconstructions of Pichia stipitis and Pichia pastoris and in silico evaluation of their potentials. BMC systems biology 6 (1), 24.

Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., Karp, P. D., 2016. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Research 44 (D1), D471–D480.

Chandrasekaran, S., Price, N. D., 2010. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. Proceedings of the National Academy of Sciences of the United States of America 107 (41), 17845–50.

Chavali, A. K., D'Auria, K. M., Hewlett, E. L., Pearson, R. D., Papin, J. A., 2012. A metabolic network approach for the identification and prioritization of antimicrobial drug targets. Trends in Microbiology 20 (3), 113–123.

Chen, Y., Holtman, C. K., Magnuson, R. D., Youderian, P. A., Golden, S. S., 2008. The complete sequence and functional analysis of pANL, the large plasmid of the unicellular freshwater cyanobacterium Synechococcus elongatus PCC 7942. Plasmid 59 (3), 176–92.

Chuang, H.-Y., Hofree, M., Ideker, T., 2010. A decade of systems biology. Annual review of cell and developmental biology 26, 721–44.

Coello Coello, C. A., 2011. An Introduction to Multi-Objective Particle Swarm Optimizers. Soft Computing in Industrial Applications 96 (103570), 3–12.

Colijn, C., Brandes, A., Zucker, J., Lun, D. S., Weiner, B., Farhat, M. R., Cheng, T. Y., Moody, D. B., Murray, M., Galagan, J. E., 2009. Interpreting expression data with metabolic flux models: Predicting Mycobacterium tuberculosis mycolic acid production. PLoS Computational Biology 5 (8).

Covert, M. W., Schilling, C. H., Palsson, B. Ø., 2001. Regulation of gene expression in flux balance models of metabolism. Journal of theoretical biology 213 (1), 73–88.

Covert, M. W., Xiao, N., Chen, T. J., Karr, J. R., 2008. Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli. Bioinformatics (Oxford, England) 24 (18), 2044–50.

Cvijovic, M., Olivares-Hernandez, R., Agren, R., Dahr, N., Vongsangnak, W., Nookaew, I., Patil, K. R., Nielsen, J., Olivares-Hernández, R., Agren, R., Dahr, N., Vongsangnak, W., Nookaew, I., Patil, K. R., Nielsen, J., 2010. BioMet Toolbox: Genome-wide analysis of metabolism. Nucleic Acids Research 38 (SUPPL. 2), 144–149.

Das, I., Dennis, J. E., 1998. Normal-Boundary Intersection: A New Method for Generating the Pareto Surface in Nonlinear Multicriteria

Optimization Problems. SIAM Journal on Optimization 8 (3), 631–657.

Das, S., Mullick, S. S., Suganthan, P. N., 2016. Recent advances in differential evolution-An updated survey. Swarm and Evolutionary Computation 27, 1–30.

Das, S., Suganthan, P. N., 2010. Differential Evolution: A Survey of the State-of-the-Art. IEEE Transactions on Evolutionary Computation PP (99), 1–28.

Dersch, L. M., Beckers, V., Wittmann, C., 2016. Green pathways: Metabolic network analysis of plant systems. Metabolic Engineering 34, 1–24.

Dismukes, G. C., Carrieri, D., Bennette, N., Ananyev, G. M., Posewitz, M. C., 2008. Aquatic phototrophs: efficient alternatives to land-based crops for biofuels. Current opinion in biotechnology 19 (3), 235–40.

Dong, H., Wei, J.-M., 2004. Low Concentrations of Tetracycline Enhance the Photophosphorylation and P/O Ratio of Chloroplasts. Photosynthesis research 79 (2), 201–208.

Douglas, S. E., 1998. Plastid evolution: origins, diversity, trends. Current Opinion in Genetics & Development 8 (6), 655–61.

Ducat, D. C., Way, J. C., Silver, P. a., 2011. Engineering cyanobacteria to generate high-value products. Trends in biotechnology 29 (2), 95–103.

Ebrahim, A., Lerman, J. A., Palsson, B. O., Hyduke, D. R., 2013. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC systems biology 7, 74.

Edwards, J. S., Covert, M., Palsson, B. Ø., 2002a. Metabolic modelling of microbes: The flux-balance approach. Environmental Microbiology 4 (3), 133–140.

Edwards, J. S., Ramakrishna, R., Palsson, B. Ø., 2002b. Characterizing the metabolic phenotype: a phenotype phase plane analysis. Biotechnology and Bioengineering 77 (1), 27–36.

Erdrich, P., Knoop, H., Steuer, R., Klamt, S., 2014. Cyanobacterial biofuels: new insights and strain design strategies revealed by computational modeling. Microbial Cell Factories 13 (1), 128.

Erdrich, P., Steuer, R., Klamt, S., 2015. An algorithm for the reduction of genome-scale metabolic network models to meaningful core models. BMC Systems Biology 9 (1), 48.

Eriksen, N. T., 2008. Production of phycocyanin–a pigment with applications in biology, biotechnology, foods and medicine. Applied microbiology and biotechnology 80 (1), 1–14.

Famili, I., Forster, J., Nielsen, J., Palsson, B. Ø., 2003. Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. Proceedings of the National Academy of Sciences 100 (23), 13134–13139.

Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., Palsson, B. Ø., 2009. Reconstruction of biochemical networks in microorganisms. Nature reviews. Microbiology 7 (2), 129–43.

Feist, A. M., Palsson, B. Ø., 2010. The biomass objective function. Current opinion in microbiology 13 (3), 344–9.

Fischer, E., Sauer, U., 2005. Large-scale in vivo flux analysis shows rigidity and suboptimal performance of Bacillus subtilis metabolism. Nature Genetics 37 (6), 636–640.

Fu, P., 2009. Genome-scale modeling of *Synechocystis* sp. PCC 6803 and prediction of pathway insertion. Journal of Chemical Technology & Biotechnology 84 (4), 473–483.

Furusawa, C., Horinouchi, T., Hirasawa, T., Shimizu, H., jun 2012. Systems Metabolic Engineering: The Creation of Microbial Cell Factories by Rational Metabolic Design and Evolution. Advances in biochemical engineering/biotechnology.

Gamermann, D., Montagud, A., Conejero, J. A., Urchueguía, J. F., de Córdoba, P. F., Fernández de Córdoba, P., 2014a. New Approach for Phylogenetic Tree Recovery Based on Genome-Scale Metabolic Networks. Journal of computational biology : a journal of computational molecular cell biology 21 (00), 1–12.

Gamermann, D., Montagud, A., Infante, R. A. J., Triana, J., Urchueguía, J. F., Fernández de Córdoba, P., 2014b. PyNetMet: Python tools for efficient work with networks and metabolic models. Computational and Mathematical Biology 3 (3), 1–19.

Garcia-Albornoz, M., Thankaswamy-Kosalai, S., Nilsson, A., Väremo, L., Nookaew, I., Nielsen, J., 2014. BioMet Toolbox 2.0: genome-wide analysis of metabolism and omics data. Nucleic acids research, 1–7.

García Martín, H., Kumar, V. S., Weaver, D., Ghosh, A., Chubukov, V., Mukhopadhyay, A., Arkin, A., Keasling, J. D., 2015. A Method to Constrain Genome-Scale Models with 13C Labeling Data. PLOS Computational Biology 11 (9), e1004363.

George, A., Veeramani, P., 1994. On some results in fuzzy metric spaces. Fuzzy Sets and Systems 64 (3), 395–399.

Goelzer, A., Fromion, V., Scorletti, G., 2011. Cell design in bacteria as a convex optimization problem. Automatica 47 (6), 1210–1218.

Greenwald, B. C., Stiglitz, J. E., 1986. Externalities in Economies with Imperfect Information and Incomplete Markets. The Quarterly Journal of Economics 101 (2), 229–264.

Güell, O., Sagués, F., Serrano, M. Á., 2014. Essential Plasticity and Redundancy of Metabolism Unveiled by Synthetic Lethality Analysis. PLoS Computational Biology 10 (5), e1003637.

Gutthann, F., Egert, M., Marques, A., Appel, J., 2007. Inhibition of respiration and nitrate assimilation enhances photohydrogen evolution under low oxygen concentrations in Synechocystis sp. PCC 6803. Biochimica et biophysica acta 1767 (2), 161–9.

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

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., Steinbeck, C., 2013. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Research 41 (D1), D456–D463.

He, L., Wu, S. G., Wan, N., Reding, A. C., Tang, Y. J., 2015. Simulating cyanobacterial phenotypes by integrating flux balance analysis, kinetics, and a light distribution function. Microbial Cell Factories 14 (1), 206.

Heinemann, M., Sauer, U., 2010. Systems biology of microbial metabolism. Current Opinion in Microbiology 13 (3), 337–343.

Herrero, A., Flores, E., 2008. The Cyanobacteria: Molecular Biology, Genomics, and Evolution. Horizon Scientific Press.

Hess, W. R., 2011. Cyanobacterial genomics for ecology and biotechnology. Current opinion in microbiology 14 (5), 608–14.

Hong, S. J., Lee, C. G., 2007. Evaluation of central metabolism based on a genomic database of Synechocystis PCC6803. Biotechnology and Bioprocess Engineering 12 (2), 165–173.

Ibarra, R. U., Edwards, J. S., Palsson, B. Ø., 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420 (6912), 186–189.

Ideker, T., Galitski, T., Hood, L., 2001. A new approach to decoding life: systems biology. Annual Review of Genomics and Human Genetics 2 (1), 343–372.

Ishibuchi, H., Tsukamoto, N., Nojima, Y., 2008. Evolutionary many-objective optimization: A short review. In: 2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence). IEEE, pp. 2419–2426.

Jensen, P. A., Papin, J. A., 2011. Functional integration of a metabolic network model and expression data without arbitrary thresholding. Bioinformatics (Oxford, England) 27 (4), 541–7.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., Barabasi, A.-L., 2000. The large-scale organization of metabolic networks. Nature 407 (6804), 651–654.

Jerby, L., Shlomi, T., Ruppin, E., 2010. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. Molecular systems biology 6, 401.

Johannsen, W., 1911. The Genotypic Conception of Heredity. The American Naturalist 45 (531), 129–159.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., Morishima, K., 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research 45 (D1), D353–D361.

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

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research 44 (D1), D457–D462.

Kaneko, T., Nakamura, Y., Sasamoto, S., Watanabe, A., Kohara, M., Matsumoto, M., Shimpo, S., Yamada, M., Tabata, S., 2003. Structural analysis of four large plasmids harboring in a unicellular cyanobacterium, Synechocystis sp. PCC 6803. DNA research : an international journal for rapid publication of reports on genes and genomes 10 (5), 221–8.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., Tabata, S., 1996. Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions (supplement). DNA research : an international journal for rapid publication of reports on genes and genomes 3 (3), 185–209.

Kaneko, T., Tanaka, A., Sato, S., Kotani, H., Sazuka, T., Miyajima, N., Sugiura, M., Tabata, S., 1995. Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. I. Sequence features in the 1 Mb region from map positions 64% to 92% of the genome. DNA research : an international journal for rapid publication of reports on genes and genomes 2 (4), 153–66, 191–8.

Karp, P. D., Latendresse, M., Paley, S. M., Krummenacker, M., Ong, Q. D., Billington, R., Kothari, A., Weaver, D., Lee, T., Subhraveti, P., Spaulding, A., Fulcher, C., Keseler, I. M., Caspi, R., 2016. Pathway tools version 19.0 update: Software for pathway/genome informatics and systems biology. Briefings in Bioinformatics 17 (5), 877–890.

Karp, P. D., Paley, S., Romero, P., 2002. The Pathway Tools software. Bioinformatics (Oxford, England) 18 Suppl 1, S225–32.

Khanna, N., Lindblad, P., 2015. Cyanobacterial hydrogenases and hydrogen metabolism revisited: recent progress and future prospects. International journal of molecular sciences 16 (5), 10537–61.

Kim, W. J., Lee, S.-M., Um, Y., Sim, S. J., Woo, H. M., 2017. Development of SyneBrick Vectors As a Synthetic Biology Platform for Gene Expression in Synechococcus elongatus PCC 7942. Frontiers in Plant Science 8, 293.

King, Z. A., Lloyd, C. J., Feist, A. M., Palsson, B. Ø., 2015. Next-generation genome-scale models for metabolic engineering. Current Opinion in Biotechnology 35, 23–29.

Kitano, H., 2001. Foundations of systems biology. MIT Press.

Kitano, H., 2002. Systems biology: a brief overview. Science 295 (5560), 1662–1664.

Klipp, E., Liebermeister, W., Wierling, C., Kowald, A., 2016. Systems biology : a textbook. Wiley-Blackwell.

Knoop, H., Gründel, M., Zilliges, Y., Lehmann, R., Hoffmann, S., Lockau, W., Steuer, R., 2013. Flux Balance Analysis of Cyanobacterial Metabolism: The Metabolic Network of Synechocystis sp. PCC 6803. PLoS computational biology 9 (6), e1003081.

Knoop, H., Steuer, R., apr 2015. A Computational Analysis of Stoichiometric Constraints and Trade-Offs in Cyanobacterial Biofuel Production. Frontiers in Bioengineering and Biotechnology 3, 47.

Knoop, H., Zilliges, Y., Lockau, W., Steuer, R., 2010. The metabolic network of Synechocystis sp. PCC 6803: systemic properties of autotrophic growth. Plant physiology 154 (1), 410–22.

Knorr, A. L., Jain, R., Srivastava, R., 2007. Bayesian-based selection of metabolic objective functions. Bioinformatics (Oxford, England) 23 (3), 351–7.

Koksharova, O. A., Wolk, C. P., 2002. Genetic tools for cyanobacteria. Applied microbiology and biotechnology 58 (2), 123–37.

Konak, A., Coit, D. W., Smith, A. E., 2006. Multi-objective optimization using genetic algorithms: A tutorial. Reliability Engineering and System Safety 91 (9), 992–1007.

Kun, Á., Papp, B., Szathmáry, E., 2008. Computational identification of obligatorily autocatalytic replicators embedded in metabolic networks. Genome Biology 9 (3), R51.

Kung, H. T., Luccio, F., Preparata, F. P., 1975. On finding the maxima of a set of vectors. Journal of the Association for Computing Machinery 22 (4), 469–476.

Kwon, J.-H. H., Bernát, G., Wagner, H., Rögner, M., Rexroth, S., 2013. Reduced light-harvesting antenna: Consequences on cyanobacterial metabolism and photosynthetic productivity. Algal Research 2 (3), 188–195.

Lau, N. S., Matsui, M., Abdullah, A. A. A., 2015. Cyanobacteria: Photoautotrophic Microbial Factories for the Sustainable Synthesis of Industrial Products. BioMed Research International 2015, 754934.

Lee, J. M., Min Lee, J., Gianchandani, E. P., Eddy, J. A., Papin, J. A., 2008. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. PLoS computational biology 4 (5), e1000086.

Lerman, J. A., Hyduke, D. R., Latif, H., Portnoy, V. A., Lewis, N. E., Orth, J. D., Schrimpe-Rutledge, A. C., Smith, R. D., Adkins, J. N., Zengler, K., Palsson, B. Ø., 2012. In silico method for modelling metabolism and gene product expression at genome scale. Nature communications 3, 929.

Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. A., Charusanti, P., Polpitiya, A. D., Adkins, J. N., Schramm, G., Purvine, S. O., Lopez-Ferrer, D., Weitz, K. K., Eils, R., König, R., Smith, R. D., Palsson, B. Ø., 2010. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Molecular systems biology 6, 390.

Lewis, N. E., Nagarajan, H., Palsson, B. Ø., 2012. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nature reviews. Microbiology 10 (4), 291–305.

Li, P., Harding, S. E., Liu, Z., 2001. Cyanobacterial exopolysaccharides: their nature and potential biotechnological applications. Biotechnology & genetic engineering reviews 18, 375–404.

Llaneras, F., Picó, J., 2010. Which metabolic pathways generate and characterize the flux space? A comparison among elementary modes, extreme pathways and minimal generators. Journal of Biomedicine and Biotechnology 2010.

Lotov, A. V., Miettinen, K., 2008. Visualizing the pareto frontier. In: Multiobjective Optimization - Interactive and Evolutionary Approaches. Vol. 5252 LNCS. Springer, Heidelberg, pp. 213–243.

Lun, D. S., Rockwell, G., Guido, N. J., Baym, M., Kelner, J. A., Berger, B., Galagan, J. E., Church, G. M., 2009. Large-scale identification of genetic design strategies using local search. Molecular Systems Biology 5 (1).

Maarleveld, T. R., Boele, J., Bruggeman, F. J., Teusink, B., mar 2014. A data integration and visualization resource for the metabolic network of Synechocystis sp. PCC 6803. Plant physiology 164 (3), 1111–21.

Mahadevan, R., Edwards, J. S., Doyle, F. J., 2002. Dynamic flux balance analysis of diauxic growth in Escherichia coli. Biophysical journal 83 (3), 1331–40.

Mahadevan, R., Schilling, C. H., 2003. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metabolic Engineering 5 (4), 264–276.

Mardinoglu, A., Gatto, F., Nielsen, J., 2013. Genome-scale modeling of human metabolism - a systems biology approach. Biotechnology Journal 8 (9), 985–996.

Mardinoglu, A., Nielsen, J., 2012. Systems medicine and metabolic modelling. Journal of Internal Medicine 271 (2), 142–154.

Marler, R. T., Arora, J. S., 2004. Survey of multi-objective optimization methods for engineering. Structural and multidisciplinary optimization 26, 369–395.

Martínez-Iranzo, M., Herrero, J. M., Sanchis, J., Blasco, X., García-Nieto, S., 2009. Applied Pareto multi-objective optimization by stochastic solvers. Engineering Applications of Artificial Intelligence 22 (3), 455–465.

Mattson, C. A., Messac, A., 2005. Pareto Frontier Based Concept Selection Under Uncertainty, with Visualization. Optimization and Engineering 6 (1), 85–115.

McEwen, J. T., Machado, I. M. P., Connor, M. R., Atsumi, S., 2013. Engineering Synechococcus elongatus PCC 7942 for continuous growth under diurnal conditions. Applied and Environmental Microbiology 79 (5), 1668–1675.

Messac, A., Ismail-Yahaya, A., Mattson, C. A., 2003. The normalized normal constraint method for generating the Pareto frontier. Structural and Multidisciplinary Optimization 25 (2), 86–98.

Messac, A., Mattson, C. A., 2002. Generating well-distributed sets of Pareto points for engineering design using physical programming. Optimization and Engineering 3 (4), 431–450.

Metallo, C. M., Vander Heiden, M. G., 2013. Understanding Metabolic Regulation and Its Influence on Cell Physiology.

Mezura-Montes, E., Reyes-Sierra, M., Coello Coello, C. A., 2008. Multi-objective Optimization Using Differential Evolution: A Survey of the State-of-the-Art. In: Chakraborty, U. K. (Ed.), Advances in Differential Evolution. Vol. 143 of Studies in Computational Intelligence. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 173–196.

Miettinen, K., 1999. Nonlinear Multiobjective Optimization. Vol. 12. Kluwer Academic Publishers, Boston.

Mohammadi, R., Fallah-Mehrabadi, J., Bidkhori, G., Zahiri, J., Javad Niroomand, M., Masoudi-Nejad, A., 2016. A systems biology approach to reconcile metabolic network models with application to Synechocystis sp. PCC 6803 for biofuel production. Molecular BioSystems 12 (8), 2552–2561.

Montagud, A., Gamermann, D., de Córdoba, P., Urchuegu\'\ia, J. F., Fernández de Córdoba, P., Urchueguía, J. F., 2015. Synechocystis sp. PCC6803 metabolic models for the enhanced production of hydrogen. Critical reviews in biotechnology 35 (2), 184–198.

Montagud, A., Navarro, E., Fernández de Córdoba, P., Urchueguía, J. F., Patil, K. R., 2010. Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. BMC systems biology 4 (1), 156.

Montagud, A., Zelezniak, A., Navarro, E., Fernández de Córdoba, P., Urchueguía, J. F., Patil, K. R., 2011. Flux coupling and transcriptional regulation within the metabolic network of the photosynthetic bacterium Synechocystis sp. PCC6803. Biotechnology journal 6 (3), 330–42.

Mori, M., Hwa, T., Martin, O. C., De Martino, A., Marinari, E., 2016. Constrained Allocation Flux Balance Analysis. PLoS Computational Biology 12 (6), e1004913.

Nagrath, D., Avila-Elchiver, M., Berthiaume, F., Tilles, A. W., Messac, A., Yarmush, M. L., 2010. Soft constraints-based multiobjective framework for flux balance analysis. Metabolic engineering 12 (5), 429–45.

Nagrath, D., Avila-Elchiver, M., Tilles, F. B., W., A., Messac, A., Yarmush, M. L., Berthiaume, F., Tilles, A. W., Messac, A., Yarmush, M. L., 2007. Integrated energy and flux balance based multiobjective framework for large-scale metabolic networks. Ann Biomed Eng 35 (6), 863–885.

Nakajima, T., Kajihata, S., Yoshikawa, K., Matsuda, F., Furusawa, C., Hirasawa, T., Shimizu, H., 2014. Integrated metabolic flux and omics analysis of Synechocystis sp. PCC 6803 under mixotrophic and photoheterotrophic conditions. Plant & cell physiology 55 (9), 1605–12.

Navarro, E., Montagud, A., Fernández de Córdoba, P., Urchueguía, J. F., 2009. Metabolic flux analysis of the hydrogen production potential in Synechocystis sp. PCC6803. International Journal of Hydrogen Energy 34 (21), 8828–8838.

NCBI Resource Coordinators, N. R., 2016. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 44 (D1), D7–D19.

Nielsen, J., 1997. Metabolic engineering: techniques for analysis of targets for genetic manipulations. Biotechnology and bioengineering 58 (2-3), 125–32.

Nogales, J., Gudmundsson, S., Knight, E. M., Palsson, B. Ø., Thiele, I., 2012. Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis. Proceedings of the National Academy of Sciences of the United States of America 109 (7), 2678–83.

Oliveira, A. P., Patil, K. R., Nielsen, J., 2008. Architecture of transcriptional regulatory circuits is knitted over the topology of biomolecular interaction networks. BMC systems biology 2, 17.

Orth, J. D., Thiele, I., Palsson, B. Ø., 2010. What is flux balance analysis? Nature Publishing Group 28 (3), 245–248.

Palsson, B. Ø., 2009. Metabolic systems biology. FEBS Letters 583 (24), 3900–3904.

Papin, J. A., Stelling, J., Price, N. D., Klamt, S., Schuster, S., Palsson, B. O., 2004. Comparison of network-based pathway analysis methods. Trends in Biotechnology 22 (8), 400–405.

Parmar, A., Singh, N. K., Pandey, A., Gnansounou, E., Madamwar, D., 2011. Cyanobacteria and microalgae: A positive prospect for biofuels. Bioresource Technology 102 (22), 10163–72.

Patil, K. R., Akesson, M., Nielsen, J., Åkesson, M., Nielsen, J., Akesson, M., Nielsen, J., 2004. Use of genome-scale microbial models for metabolic engineering. Current opinion in biotechnology 15 (1), 64–9.

Patil, K. R., Nielsen, J., 2005. Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proceedings of the National Academy of Sciences of the United States of America 102 (8), 2685–9.

Patil, K. R., Rocha, I., Förster, J., Nielsen, J., Forster, J., Nielsen, J., 2005. Evolutionary programming as a platform for in silico metabolic engineering. BMC bioinformatics 6 (1), 308.

Pharkya, P., Burgard, A. P., Maranas, C. D., 2004. OptStrain: a computational framework for redesign of microbial production systems. Genome research 14 (11), 2367–76.

Pharkya, P., Maranas, C. D., 2006. An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. Metabolic Engineering 8 (1), 1–13.

Pinto, F., Pacheco, C. C., Oliveira, P., Montagud, A., Landels, A., Couto, N., Wright, P. C., Urchueguía, J. F., Tamagnini, P., 2015. Improving a Synechocystis-based photoautotrophic chassis through systematic genome mapping and validation of neutral sites. DNA research : an international journal for rapid publication of reports on genes and genomes 22 (6), 425–37.

Pinto, F., van Elburg, K. a., Pacheco, C. C., Lopo, M., Noirel, J., Montagud, A., Urchueguía, J. F., Wright, P. C., Tamagnini, P., 2012. Construction of a chassis for hydrogen production: physiological and molecular characterization of a Synechocystis sp. PCC 6803 mutant lacking a functional bidirectional hydrogenase. Microbiology (Reading, England) 158 (Pt 2), 448–64.

Placzek, S., Schomburg, I., Chang, A., Jeske, L., Ulbrich, M., Tillack, J., Schomburg, D., 2017. BRENDA in 2017: New perspectives and new tools in BRENDA. Nucleic Acids Research 45 (D1), D380–D388.

Price, N. D., Reed, J. L., Palsson, B. Ø., 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. Nature reviews. Microbiology 2 (11), 886–97.

Raman, K., Chandra, N., 2009. Flux balance analysis of biological systems: applications and challenges. Briefings in bioinformatics 10 (4), 435–49.

Ranganathan, S., Suthers, P. F., Maranas, C. D., 2010. OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. PLoS computational biology 6 (4), e1000744.

Rastogi, R. P., Sinha, R. P., 2009. Biotechnological and industrial significance of cyanobacterial secondary metabolites. Biotechnology advances 27 (4), 521–39.

Raven, J. A., Allen, J. F., 2003. Genomics and chloroplast evolution: what did cyanobacteria do for plants? Genome biology 4 (3), 209.

Reed, J. L., Patel, T. R., Chen, K. H., Joyce, A. R., Applebee, M. K., Herring, C. D., Bui, O. T., Knight, E. M., Fong, S. S., Palsson, B. Ø., 2006. Systems approach to refining genome annotation. Proceedings of the National Academy of Sciences of the United States of America 103 (46), 17480–4.

Reyes, R., Gamermann, D., Montagud, A., Fuente, D., Triana, J., Urchueguía, J. F., Fernández de Córdoba, P., 2012. Automation on the generation of genome-scale metabolic models. Journal of computational biology : a journal of computational molecular cell biology 19 (12), 1295–306.

Reynoso-Meza, G., Blasco, X., Sanchis, J., Herrero, J. M., 2013a. Comparison of design concepts in multi-criteria decision-making using level diagrams. Information Sciences 221, 124–141.

Reynoso-Meza, G., Blasco, X., Sanchis, J., Martínez, M., 2014. Controller tuning using evolutionary multi-objective optimisation: Current trends and applications. Control Engineering Practice 28 (1), 58–73.

Reynoso-Meza, G., Blasco Ferragud, X., Sanchis Saez, J., Herrero Durá, J. M., 2017. Controller tuning with evolutionary multiobjective optimization : a holistic multiobjective optimization design procedure. Springer.

Reynoso-Meza, G., García-Nieto, S., Sanchis, J., Blasco, X., 2013b. Controller tuning using multiobjective optimization algorithms: a global tuning framework. IEEE Transactions on Control Systems Technology 21 (2), 445–458.

Reynoso-Meza, G., Sanchis, J., Blasco, X., Herrero, J. M., 2011. Hybrid DE algorithm with adaptive crossover operator for solving real-world numerical optimization problems. In: Evolutionary Computation (CEC), 2011 IEEE Congress on. pp. 1551–1556.

Reynoso-Meza, G., Sanchis, J., Blasco, X., Herrero, J. M., 2012. Multiobjective evolutionary algortihms for multivariable PI controller tuning. Expert Systems with Applications 39 (9), 7895–7907.

Reynoso-Meza, G., Sanchis, J., Blasco, X., Martínez, M., 2010. Design of continuous controllers using a multiobjective differential evolution algorithm with spherical pruning. In: Di Chio, C., Cagnoni, S., Cotta, C., Ebner, M., Ekárt, A., Esparcia-Alcazar, A. I., Goh, C.-K., Merelo, J. J., Neri, F., Preuß, M., Togelius, J., Yannakakis, G. N. (Eds.), Applications of Evolutionary Computation. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 532–541.

Rippka, R., Deruelles, J., Waterbury, J. B., Herdman, M., Stanier, R. Y., 1979. Generic Assignments, Strain Histories and Properties of Pure Cultures of Cyanobacteria. Journal of General Microbiology 111 (1), 1–61.

Rocha, I., Maia, P., Evangelista, P., Vilaça, P., Soares, S., Pinto, J. P., Nielsen, J., Patil, K. R., Ferreira, E. C., Rocha, M., 2010. OptFlux: an open-source software platform for in silico metabolic engineering. BMC systems biology 4 (1), 45.

Rodionova, M., Poudyal, R., Tiwari, I., Voloshin, R., Zharmukhamedov, S., Nam, H., Zayadan, B., Bruce, B., Hou, H., Allakhverdiev, S., 2016. Biofuel production: Challenges and opportunities. International Journal of Hydrogen Energy 42 (12), 8450–8461.

Saha, R., Verseput, A. T., Berla, B. M., Mueller, T. J., Pakrasi, H. B., Maranas, C. D., 2012. Reconstruction and comparison of the metabolic potential of cyanobacteria Cyanothece sp. ATCC 51142 and Synechocystis sp. PCC 6803. PloS one 7 (10), e48285.

Savinell, J. M., Palsson, B. Ø., 1992a. Network analysis of intermediary metabolism using linear optimization .I. Development of mathematical formalism. Journal of theoretical biology 154 (4), 421–454.

Savinell, J. M., Palsson, B. Ø., 1992b. Network analysis of intermediary metabolism using linear optimization. II. Interpretation of hybridoma cell metabolism. Journal of theoretical biology 154 (4), 455–73.

Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., Kang, J., Hyduke, D. R., Palsson, B. Ø., 2011. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nature protocols 6 (9), 1290–1307.

Schmidt, B. J., Ebrahim, A., Metz, T. O., Adkins, J. N., Palsson, B. Ø., Hyduke, D. R., 2013. GIM3E: Condition-specific models of cellular metabolism developed from metabolomics and expression data. Bioinformatics 29 (22), 2900–2908.

Schomburg, I., Hofmann, O., Baensch, C., 2000. Enzyme data and metabolic information : BRENDA , a resource for research in biology , biochemistry , and medicine. Gene Funct. Dis 1 (3-4), 109–118.

Schopf, J. W., 2000. The fossil record: tracing the roots of the cyanobacterial lineage. In: Whitton, B. A., Potts, M. (Eds.), The ecology of cyanobacteria. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 13–35.

Schuetz, R., Kuepfer, L., Sauer, U., 2007. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Molecular systems biology 3 (119), 119.

Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M., Sauer, U., 2012. Multidimensional Optimality of Microbial Metabolism. Science 336 (6081), 601–604.

Schwender, J., Hay, J. O., 2012. Predictive Modeling of Biomass Component Tradeoffs in Brassica napus Developing Oilseeds Based on in Silico Manipulation of Storage Metabolism. PLANT PHYSIOLOGY 160 (3), 1218–1236.

Segrè, D., Vitkup, D., Church, G. M., 2002. Analysis of optimality in natural and perturbed metabolic networks. Proceedings of the National Academy of Sciences of the United States of America 99 (23), 15112–7.

Shastri, A. a., Morgan, J. a., 2005. Flux balance analysis of photoautotrophic metabolism. Biotechnology progress 21 (6), 1617–26.

Shestakov, S. V., Khyen, N. T., 1970. Evidence for genetic transformation in blue-green alga Anacystis nidulans. Molecular & general genetics : MGG 107 (4), 372–5.

Shlomi, T., Berkman, O., Ruppin, E., 2005. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. Proceedings of the National Academy of Sciences of the United States of America 102 (21), 7695–7700.

Shlomi, T., Cabili, M. N., Herrgård, M. J., Palsson, B. Ø., Ruppin, E., 2008. Network-based prediction of human tissue-specific metabolism. Nature biotechnology 26 (9), 1003–10.

Shlomi, T., Eisenberg, Y., Sharan, R., Ruppin, E., 2007. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. Molecular systems biology 3 (1), 101.

Stephanopoulos, G., 1999. Metabolic fluxes and metabolic engineering. Metabolic engineering 1 (1), 1–11.

Stephanopoulos, G., Aristidou, A., Nielsen, J., 1999. Metabolic Engineering: Principles and Methodologies.

Storn, R., 2008. SCI: Differential Evolution Research: Trends and Open Questions. In: (Ed.), U. K. C. (Ed.), Advances in Differential Evolution. Vol. LNCS 143. Springer, Heidelberg, pp. 1–31.

Storn, R., Price, K., 1997. Differential Evolution: A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. Journal of Global Optimization 11, 341–359.

Takahashi, H., Kopriva, S., Giordano, M., Saito, K., Hell, R., 2011. Sulfur assimilation in photosynthetic organisms: molecular functions and regulations of transporters and assimilatory enzymes. Annual review of plant biology 62, 157–84.

Tamagnini, P., Leitão, E., Oliveira, P., Ferreira, D., Pinto, F., Harris, D. J., Heidorn, T., Lindblad, P., 2007. Cyanobacterial hydrogenases: diversity, regulation and applications. FEMS microbiology reviews 31 (6), 692–720.

Teusink, B., Wiersma, A., Jacobs, L., Notebaart, R. A., Smid, E. J., 2009. Understanding the adaptive growth strategy of Lactobacillus plantarumby in silico optimisation. PLoS Computational Biology 5 (6).

Thiele, I., Palsson, B. Ø., 2010. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature protocols 5 (1), 93–121.

Tiwari, A., Pandey, A., 2012. Cyanobacterial hydrogen production - A step towards clean environment. International Journal of Hydrogen Energy 37 (1), 139–150.

Triana, J., Montagud, A., Siurana, M., Fuente, D., Urchueguía, A., Gamermann, D., Torres, J., Tena, J., de Córdoba, P. F., Urchueguía, J. F., 2014. Generation and Evaluation of a Genome-Scale Metabolic Network Model of Synechococcus elongatus PCC7942. Metabolites 4 (3), 680–698.

Van Berlo, R. J. P., De Ridder, D., Daran, J. M., Daran-Lapujade, P. A. S., Teusink, B., Reinders, M. J. T., 2011. Predicting metabolic fluxes using gene expression differences as constraints. IEEE/ACM Transactions on Computational Biology and Bioinformatics 8 (1), 206–216.

van den Hondel, C. A., Verbeek, S., van der Ende, A., Weisbeek, P. J., Borrias, W. E., van Arkel, G. A., 1980. Introduction of transposon Tn901 into a plasmid of Anacystis nidulans: preparation for cloning in cyanobacteria. Proceedings of the National Academy of Sciences of the United States of America 77 (3), 1570–4.

Van der Plas, J., Oosterhoff-Teertstra, R., Borrias, M., Weisbeek, P., 1992. Identification of replication and stability functions in the complete nucleotide sequence of plasmid pUH24 from the cyanobacterium Synechococcus sp. PCC 7942. Molecular microbiology 6 (5), 653–64.

Varma, A., Boesch, B. W., Palsson, B. Ø., 1993. Stoichiometric interpretation of Escherichia coli glucose catabolism under various oxygenation rates. Applied and Environmental Microbiology 59 (8), 2465–2473.

Varma, A., Palsson, B. Ø., 1994a. Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. Bio/Technology 12 (10), 994–998.

Varma, A., Palsson, B. Ø., 1994b. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. Applied and environmental microbiology 60 (10), 3724–31.

Vermaas, W., 1996. Molecular genetics of the cyanobacteriumSynechocystis sp. PCC 6803: Principles and possible biotechnology applications. Journal of Applied Phycology 8, 263–273.

Vermaas, W. F. J., 2001. Photosynthesis and Respiration in Cyanobacteria.

Wang, Y., Eddy, J. A., Price, N. D., 2012. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. BMC Systems Biology 6 (1), 153.

Watson, M. R., 1984. Metabolic maps for the Apple II. Biochemical Society Transactions 12 (6), 1093–1094.

Yang, C., Hua, Q., Shimizu, K., 2002. Quantitative analysis of intracellular metabolic fluxes using GC-MS and two-dimensional NMR spectroscopy. Journal of bioscience and bioengineering 93 (1), 78–87.

Yang, X.-S., 2010. Nature-Inspired Metaheuristic Algorithms. Luniver Press.

Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., Khandurina, J., Trawick, J. D., Osterhout, R. E., Stephen, R., Estadilla, J., Teisan, S., Schreyer, H. B., Andrae, S., Yang, T. H., Lee, S. Y., Burk, M. J., Van Dien, S., 2011. Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. Nature Chemical Biology 7 (7), 445–452.

Yoshikawa, K., Kojima, Y., Nakajima, T., Furusawa, C., Hirasawa, T., Shimizu, H., 2011. Reconstruction and verification of a genome-scale metabolic model for Synechocystis sp. PCC6803. Applied microbiology and biotechnology 92 (2), 347–58.

Young, J. D., Shastri, A. a., Stephanopoulos, G., Morgan, J. a., 2011. Mapping photoautotrophic metabolism with isotopically nonstationary (13)C flux analysis. Metabolic engineering 13 (6), 656–65.

Zhang, Y. H. P., Sun, J., Ma, Y., 2016. Biomanufacturing: history and perspective. Journal of Industrial Microbiology and Biotechnology 44 (4), 1–12.

Zhou, A., Qu, B.-Y., Li, H., Zhao, S.-Z., Suganthan, P. N., Zhang, Q., 2011. Multiobjective evolutionary algorithms: A survey of the state of the art. Swarm and Evolutionary Computation 1 (1), 32–49.

Zomorrodi, A. R., Suthers, P. F., Ranganathan, S., Maranas, C. D., 2012. Mathematical optimization applications in metabolic networks. Metabolic Engineering 14 (6), 672–686.

Zur, H., Ruppin, E., Shlomi, T., 2010. iMAT: an integrative metabolic analysis tool. Bioinformatics (Oxford, England) 26 (24), 3140–2.

Appendices

Additional file 3.1 - iSyn842

Text file with the stoichiometric model of *Synechocystis* sp. PCC 6803, *i*Syn842, in OptGene (Patil et al., 2005) format.

Additional file 3.2 - iSyf715

Text file with the stoichiometric model of *Synechococcus elongatus* PCC 7942, *i*Syf715, in OptGene (Patil et al., 2005) format.

Additional file 4.1 - Flux landscapes syn-syf

Excel file with the flux distributions resulting from the simulations of *i*Syn842 and *i*Syf715 models under autotrophic conditions (see Table 4.1 at page 96 for constraints), and the list of common reactions between them.

Additional file 4.2 - Substrate study

Excel file with the flux distributions resulting from all the simulations performed for the substrate study to enhance H_2 production (Section 4.4.3).

Additional file 6.1 - Constraints FBA

Excel file with all the constraints used during the simulations performed with FBA algorithm under five growth conditions.

Additional file 6.2 - Constraints FBA_exp

Excel file with all the constraints used during the simulations performed with FBA including information about internal flux measurements under five growth conditions.

Additional file 6.3 - Constraints Meta-MODE

Excel file with all the constraints used during the simulations performed with Meta-MODE algorithm under five growth conditions.

Additional file 6.4 - Fluxes, objective values and growth yield

Excel file with the results of the simulations performed with the three methods compared in Chapter 6: growth and biomass/glucose yield values, objective values and flux distributions.