University of Catania

Department of Mathematics and Computer Science

Jole Costanza

Biological Circuit Design via BioCAD Tools

A thesis submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy

PhD Course in Computer Science – XXVI Cycle

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Prof. Giuseppe Nicosia) Advisor

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Prof. Vincenzo Cutello) Director of Graduate Studies

Approved for the University Committee on Graduate Studies.

Acknowledgements

Here, I would to thank all persons that helped me to realize this PhD thesis, they always encouraged and supported me.

First of all, I say thanks to my family: my parents, my amazing sisters Elisa, Daniela and Claudia, to my Roberto, Linda and grandma Caterina.

I have no words to express my gratitude to Professor Giuseppe Nicosia, for his enthusiasm, professionalism and love towards science. Thanks for all the precious advices and for giving me the opportunity to participate at wonderful and stimulating international conferences.

A great acknowledgment is for Professor Pietro Lió and for his advise during my research periods at the Computer Laboratory of the University of Cambridge.

I would to thank also my friends Anna and Ilenia: even if we are physically distant, we always supported each other.

Abstract

In this thesis, will be presented BioCAD, a novel computational tool able to design optimal and robust biological circuits. In BioCAD, the main idea is to use Pareto optimality and the Electronic Design Automation methods for Systems and Synthetic Biology. However, BioCAD is a general purpose tool and can be seen as well as a black box able to receive in input a generic model and analyze its components and submodules, estimate its parameters, or optimize specific functions. BioCAD implements novel and state-of-the-art algorithms performing: (i) Optimization, by analyzing continuous, discrete or hybrid (continuous and discrete) variable spaces, for Single- and Multi-objective optimization problems and for local or global search; (ii) Sensitivity Analysis, for evaluating the importance of the parameters by ranking them according to their influence on the model; (iii) Robustness Analysis, for estimating the global and local fragility and robustness of optimal synthetic circuits; (iv) Identifiability Analysis, that finds functional relations among parameters, by analyzing the values of the decision variables after and before the optimization. Additionally, BioCAD implements the ϵ-dominance analysis, able to relax the Pareto condition and expand the solution space to neighborhood region of the Pareto surface. Optimization core contains novel tools for engineering enzymes, genes and fluxes in biological systems, while Sensitivity Analysis can reveal the main genes, enzymes, species or pathways. BioCAD can be adopted and used with various modeling techniques: flux balance analysis with or without the gene protein reaction mappings, ordinary differential equations, differential algebraic equations and partial differential equations. In this thesis will be reported several experiments applied on Synthetic Biology, such as the design of the novel 1,4-butanediol synthetic pathway.

Contents

List of Figures

List of Tables

Chapter 1

Prologue

In the last decade, computational methods have been revealed very useful for understanding the metabolic precesses inside biological systems. With the advent of high-throughput technologies, scientists have used computational tools in order to menage big mole of data, therefore disciplines like Mathematics and Computer Science have quite changed the study of biology. Additionally, by means of biological models, we can understand the behavior of biological systems and for instance to reconstruct the metabolic network of biochemical reactions or the gene regulatory networks for the study of pathologies. A *System* is an object that interacts with the external environment and/or with other systems. Biological Systems are systems which components are biological entities, as well as genes, metabolites, enzymes and so on.

In the book "*Systems Biology: Properties of Reconstructed Networks*" [1], the author Bernhard Palsson, the father of the *Flux Balance Analysis* mathematical approach [2], states that Systems Biology is not focused on the biological components themselves, but on the nature of the *links* that connect them and the functional states of the networks that result from the assembly of all such links. Furthermore, the advent of high-throughput experimental technologies is forcing biologists to view cells as systems, rather than focusing their attention on individual cellular components. Not only are high-throughput technologies forcing the systems point of view, but they also enable us to study cells as systems [1]. The delineation of the chemical interactions of these components gives rise to reconstructed biochemical reaction networks that underlie various cellular functions.

Since Systems Biology does not investigate individual genes or proteins one at a time and investigates the behavior and relationships of all of the elements, data have to be integrated, graphically displayed, and ultimately modeled computationally. Together with colleagues in computer science, mathematics, statistics and biologists, researchers

are developing the necessary tools to acquire, store, analyze, graphically display, model, and distribute this information. An enormous challenge for the future is how to integrate the different levels of information pertaining to genes, mRNAs, proteins, and pathways [3].

Systems Biology stems from advances in technology, particularly in genome sequencing, computing and in analytical platforms such as mass spectrometry. In order to study a large system in its entirety, one requires the ability to model and measure it in its entirety. Until the advent of whole genome sequencing, this was an insurmountable experimental challenge for biologists. With the advancements in computing power, genomics, transcriptomics, proteomics, metabolomics and fluxomics, it is becoming possible to profile and model a complete biological system [4]. The last year, the group of Markus W. Covert published the first in silico whole cell [5]. Their model is based on a synthesis of over 900 publications and includes more than 1,900 experimentally observed parameters. They implemented 28 different submodules to represent the entire cell of the human pathogen *Mycoplasma genitalium*. Additionally, the most important feature is that each module was modeled using the most appropriate mathematical representation. For example, metabolism was modeled using flux balance analysis [2], whereas RNA and protein degradation were modeled as Poisson processes. Furthermore, the whole cell model provided insights into many previously unobserved cellular behaviors, including in vivo rates of protein-DNA association and an inverse relationship between the durations of DNA replication initiation and replication. As a result, comprehensive whole-cell models can be used to facilitate biological discovery.

Then, understanding biological systems requires the integration of experimental and computational research. Computational biology, through pragmatic modeling and theoretical exploration, provides a powerful foundation from which to address critical scientific questions head-on. *Computational Systems Biology* addresses questions fundamental to our understanding of life, yet progress here will lead to practical innovations in medicine, drug discovery and engineering [6].

Computational Biology has two distinct branches: (i) knowledge discovery, or data-mining and (ii) simulation-based analysis, which tests hypotheses with *in silico* experiments, providing predictions to be tested by *in vitro* and *in vivo* studies. Knowledge discovery is used extensively within bioinformatics for such tasks as the inference of gene regulatory networks from gene expression profile. These methods typically use predictions based on sophisticated statistical discriminators. Instead, simulation attempts to predict the dynamics of systems so that the validity of the underlying assumptions can be tested. Models must be validated by means experimental observations. Inconsistency

at this stage means that the assumptions that we are adapting on the system under consideration are incomplete or incorrect. Models that overpass initial validation can then be used to make predictions to be tested by experiments, as well as to explore questions that are not amenable to experimental inquiry. Combined with rapid progress of genome and proteome projects, *simulation-based research* is convincing increasing numbers of researchers of the importance of a system-level approach. Additionally, advances in computational power have enabled the creation and analysis of reasonably realistic intricate biological models. There are still issues to be resolved that computational modeling and analysis are now able to provide useful biological insights and predictions for well understood targets [6]. The main task that Computational Systems Biology is facing is the development of efficient algorithms, tools for the visualization and communication, data structures and data bases with the target of computer modeling of biological systems. To date, a vast number of biological databases are used to collect a big mole of information; for example KEGG PATHWAY Database [7], used to cluster metabolic pathway information, Gene Expression Omnibus for microarray collection and EcoCyc database, dedicated to describe the genome and the biochemical machinery of the model organism *E. coli* K-12 and so on. Additionally, a new research topic concerns the integration of all the information stored in the biological databases, such as Meta-Base [8], the wiki-database containing more than 2000 commonly used biological databases.

In this thesis, will be presented *BioCAD*, a novel computational framework able to design *Biological Systems*. Studying a biological system is more complicated than a physical or technological system due to the biological complexity and the limitation of experimental data. But in the other side, modeling a biological system is very important since allows to predict the system behavior in function of other stimuli, without doing again experiments or when experiments are not allowable. In BioCAD, the main idea is to use Pareto optimality and the Electronic Design Automation [9] methods for Systems Biology. However, BioCAD is a general purpose tool and can be seen as well as a black box able to receive in input a generic model and analyze its components and submodules, estimate its parameters, or optimize specific functions. BioCAD implements novel and state-of-the-art algorithms performing: (i) *Optimization*, by analyzing continuous, discrete or hybrid (continuous and discrete) variable spaces, for Single and Multi-objective optimization problems and for local or global search; (ii) *Sensitivity Analysis*, for evaluating the importance of the parameters by ranking them according to their influence on the model; (iii) *Robustness Analysis*, for estimating the global and local fragility and robustness of optimal synthetic circuits; (iv) *Identifiability Analysis*, that finds functional relations among parameters, by analyzing the values of the decision variables after and before the optimization. Additionally, BioCAD implements the ϵ-*dominance Analysis*,

able to relax the Pareto condition and expand the solution space to neighborhood region of the Pareto surface. Optimization core contains tools for engineer enzymes, genes and fluxes in optimized organisms, while Sensitivity Analysis can reveal the main genes, enzymes, species or pathways.

BioCAD can be adopted and used with various modeling techniques: Flux Balance Analysis [2] with or without the gene protein reaction (GPR) mappings [10], Ordinary Differential Equations (ODEs), Differential Algebraic Equations (DAEs) and Partial Differential Equations (PDEs). More details about BioCAD and algorithms are reported in Chapter 2. Multi-objective optimization has been revealed very useful when coupled with flux balance analysis. I have optimized bacterial *Escherichia coli* strains in order to overproduce specific and important metabolites, as well as acetate and succinate natural metabolites (and other substances, such as ethanol, 1,2-propanediol, lactate, formate) or industrial synthetic chemicals such as 1,4-butanediol by inferring optimal genetic manipulations. The Genetic Design through Multi-objective Optimization (GDMO) method here used, has been compared with previous and recent methods. Results show that GDMO algorithm outperforms the GDLS [11], OptKnock [12], OptFlux [13] and OptGene [14] heuristics. Details can be found in Chapter 3 and 4. In Chapter 5 I report a detailed analysis about the study on three important metabolic networks for the Artificial Photosynthesis. In particular, I analyzed the metabolic capability of the *photosynthetic carbon metabolism* pathway in a general leaf, the *Rhodobacter spheroides* bacterium, and the *Chlamydomonas reinhardtii* alga. I used single and multi-objective optimization algorithm to maximize the CO_2 uptake rate from organisms in order to sequester atmospheric or industrially produced CO_2. Comparing states before and after optimization provides also attractive highlights on enzymatic variation. Additionally, I investigated the bioenergetic behavior in mitochondria in different conditions, since these organelles have a pivotal role in human diseases and pathologies. I evaluated the changing in ATP and NADH concentrations (that constitute the energy in a cell) in monogenic pathologies such as fumarate deficiency, when calcium concentration is disrupted and in cancer conditions. In this way, the BioCAD sampling has individuated and distinguished the pathological and healthy states (Chapter 6).

BioCAD and Pareto Optimality have been revealed appealing for the analysis of metabolism and for Synthetic biology applications for the design and construction of new biological parts and the re-design of natural biological systems for useful purposes. Pareto Optimality is useful to obtain not only a vast range of Pareto optimal solutions, but also the best *trade-off* design. Each feasible solution represents a particular optimal biological system, so the user can choose between several candidate optimal solutions. Additionally, the area underlying the Pareto curve and the first derivative, and in particular the presence of jumps (i.e., quick variations in the objective functions during

the optimization procedure), carry valuable biotechnological information. Remarkably, BioCAD can be used for each CAD problem, linked to biology but also to electronic devices design.

Publications

Journal Articles

1. C. Angione, G. Carapezza, Jole Costanza, P. Lió and G. Nicosia, **Pareto Optimality in Organelle Energy Metabolism Analysis**, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2013.

2. C. Angione, G. Carapezza, Jole Costanza, P. Lió, and G. Nicosia, **Design and strain selection criteria for bacterial communication networks**, *Nano Communication Networks*, 2013.

3. C. Angione, Jole Costanza, G. Carapezza, P. Lió and G. Nicosia, **A Design Automation Framework for Computational Bioenergetics in Biological Networks**, *Molecular BioSystems*, 2013.

4. G. Carapezza, R. Umeton, Jole Costanza, C. Angione, G. Stracquadanio, A. Papini, P. Lió and G. Nicosia, **Efficient Behavior of Photosynthetic Organelles via Pareto Optimality, Identifiability and Sensitivity Analysis**, *ACS Synthetic Biology Journal*, 2012.

5. Jole Costanza, G. Carapezza, C. Angione, P. Lió and G. Nicosia, **Robust Design of Microbial Strains**, *Bioinformatics - Oxford Journal* 2012.

6. R. Umeton, G. Stracquadanio, A. Papini, Jole Costanza, P. Lió, and G. Nicosia, **Identification of Sensitive Enzymes in the Photosynthetic Carbon Metabolism**, *Advances in experimental medicine and biology*, 2012, 736, 441-59.

Conference Proceedings

1. C. Angione, Jole Costanza, G. Carapezza, P. Lió, G. Nicosia, **Identifiability Analysis to Characterize Mitochondrial Diseases**, in Proceedings of the 5th *International Workshop on Bio-Design Automation* - IWBDA 2013 at Imperial College, London, UK, July 12-13, 2013.

2. Jole Costanza, C. Angione, G. Carapezza, P. Lió, and G. Nicosia, **Multi-objective Optimisation of Escherichia coli for Direct Production of 1,4-butanediol**, in Proceedings of the 5th *International Workshop on Bio-Design Automation* - IWBDA 2013 at Imperial College, London, UK, July 12-13, 2013.

3. C. Angione, G. Carapezza, Jole Costanza, P. Lió, G. Nicosia, **Multi-Objective Design of Bacterial Communication Networks**, in Proceedings of the 3rd

IEEE Int. *Workshop on Molecular and Nanoscale Communications* - IEEE MoNaCom 2013 held in conjunction with IEEE Int. Conference on Communications 2013 - IEEE ICC 2013, June 9-13, 2013, Budapest, Hungary, IEEE Press.

4. Jole Costanza, C. Angione, G. Carapezza, P. Lió, G. Nicosia, **Pareto epsilon-Dominance and Identifiable Solutions for BioCAD Modeling**, in Proceedings of the 50th *Design Automation Conference* - DAC 2013, 2-6 June 2013, Austin Texas, ACM Press, 43, 2013.

5. C. Angione, G. Carapezza, Jole Costanza, P. Lió, and G. Nicosia, **Rational Design of Organelle Compartments in Cells**, in Proceedings of the *Twelth International Workshop on Network Tools and Applications in Biology* - NETTAB 2012 - Focus on Integrated Bio-Search, November 14-16, 2012, Como, Italy. EMBnet Journal Volume 18 Supplement B, pp. 20-22, 2012

6. Jole Costanza, G. Carapezza, C. Angione, P. Lió and G. Nicosia, **Multi-Objective Optimization, Sensitivity and Robustness Analysis in FBA Modeling**, in Proceedings of the 10th *Conference on Computational Methods in Systems Biology* - CMSB 2012, October 3-5, 2012 The Royal Society, London, UK, Springer.

7. C. Angione, G. Carapezza, Jole Costanza, P. Lió and G. Nicosia, **The Role of the Genome in the Evolution of the Complexity of Metabolic Machines**, in Proceedings of the *European Conference in Complex Systems* - ECCS'12, Brussels, 3-7 September 2012, Springer Complexity.

8. C. Angione, G. Carapezza, Jole Costanza, P. Lió and G. Nicosia, **Computing with Metabolic Machines**, in Proceedings of the *Turing100, The Alan Turing Centenary Conference*, University of Manchester, June 22-25, 2012. *Best Paper Award presented by Sir Roger Penrose.*

9. Jole Costanza, C. Angione, P. Lió and G. Nicosia, **Are Bacteria Unconventional Computing Architectures?**, in Proceedings of the *Turing Centenary Conference* - CiE 2012 - How the World Computes, University of Cambridge, June 18-23, 2012.

10. Jole Costanza, G. Carapezza, C. Angione, R. Umeton, P. Lió and G. Nicosia, **metaDesign: Bacterial Strain Design Automation Software**, in Proceedings of the 4th *International Workshop on Bio-Design Automation* - IWBDA 2012 at the 49th ACM/EDAC/IEEE Design Automation Conference (DAC), June 3-7, 2012 at the Moscone Center, San Francisco, CA, USA.

11. D. Agostini, Jole Costanza, V. Cutello, L. Zammataro, N. Krasnogor, M. Pavone and G. Nicosia, **Effective Calibration of Artificial Gene Regulatory Networks**, in Proceedings of the 20th *European Conference on Artificial Life* - ECAL 2011, pp. 39-46, MIT Press, Paris, France, August 8-12, 2011.

12. Jole Costanza, V. Cutello and M. Pavone, **A Memetic Immunological Algorithm for Resource Allocation Problem**, in Proceedings of the 10th *International Conference on Artificial Immune Systems* - ICARIS 2011, LNCS 6825, pp. 308-320, Springer-Verlag, Cambridge, UK, July 18-21, 2011.

13. G. Stracquadanio, R. Umeton, Jole Costanza, **Large Scale Agent-Based Modeling of the Humoral and Cellular Immune Response**, V. Annibali, R. Mechelli, M. Pavone, L. Zammataro, and G. Nicosia, in Proceedings of the 10th *International Conference on Artificial Immune Systems* - ICARIS 2011, LNCS 6825, pp. 15-29, Springer-Verlag, Cambridge, UK, July 18-21, 2011.

Refereed Conference papers

1. C. Angione, G. Carapezza, Jole Costanza, P. Lió and G. Nicosia, **Pareto Front and Identifiability Analysis to Characterize Mitochondrial Diseases**, *Mitochondrial Disease: Translating biology into new treatments*, October 2-4, 2013, Wellcome Trust Conference Center, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

2. C. Angione, Jole Costanza, G. Carapezza, P. Lió and G. Nicosia, **Organelle as a New Frontier in Computational Medicine**, the 11th *International Conference on Artificial Immune Systems* - ICARIS 2012, Taormina, Italy, August 28-31, 2012.

3. Jole Costanza, G. Carapezza, C. Angione, P. Lió and G. Nicosia, **Design of Robust Metabolic Machines**, the *Biological System Design 2012* - BSD at ISMB, 13 July, 2012 - Long Beach, CA USA.

4. Jole Costanza, G. Carapezza, C. Angione, P. Lió and G. Nicosia, **Pareto Front Sensitivity in Large Scale Biological Networks**, NetSci, 18-22 June 2012, Evanston, IL

5. Jole Costanza, L. Zammataro, P. Lió and G. Nicosia, **The ATP, NADH and Calcium Trade-offs in the Mitochondrial Bioenergetics**, the 12th *International Conference on Systems Biology* - ICSB 2011, Heidelberg/Mannheim,Germany August 28 - September 1, 2011

6. Jole Costanza, L. Zammataro, P. Lió and G. Nicosia, **Pareto Fronts in Genetic Design Strategies using Multi-Objective Optimization**, the 12th *International Conference on Systems Biology* - ICSB 2011, Heidelberg/Mannheim,Germany August 28 - September 1, 2011

7. Jole Costanza, L. Zammataro, P. Lió and G. Nicosia, **High-dimensional Pareto Surfaces in the Genetic Design of Escherichia Coli**, the 8th *Annual Meeting of the Bioinformatics Italian Society* - BITS 2011, June 20-22, Pisa, Italy.

8. Jole Costanza, L. Zammataro, P. Lió and G. Nicosia, **Multi-objective Optimization for the Mitochondrial Bioenergetics**, the 8th *Annual Meeting of the Bioinformatics Italian Society* - BITS 2011, June 20-22, Pisa, Italy.

9. Jole Costanza, L. Zammataro, P. Lió and G. Nicosia, **Pareto Optimal Trade-offs in Genetic Design Strategies using Global Search**, the 5th *International Meeting on Synthetic Biology* - SB5.0 2011, Stanford University, Stanford, California USA, June 15-17, 2011.

10. Jole Costanza, L. Zammataro, P. Lió and G. Nicosia, **Pareto Optimal Fronts in Bacterial Knockout Strategies**, the 3rd *International Workshop on Bio-Design Automation* - IWBDA 2011 at DAC, San Diego Convention Center, San Diego, California USA, June 6 - 7, 2011.

11. Jole Costanza, L. Zammataro, P. Lió and G. Nicosia, **Optimal Design of the Mitochondrial Bioenergetics**, the 3rd *International Workshop on Bio-Design Automation* - IWBDA 2011 at DAC, San Diego Convention Center, San Diego, California USA, June 6 - 7, 2011.

Working papers

1. Jole Costanza, C. Angione, G. Carapezza, P. Lió and G. Nicosia, **Comparative Analysis and Design of Mitochondrial Metabolism**

2. Jole Costanza, C. Angione, G. Carapezza, P. Lió and G. Nicosia, **Multi-objective Optimization of Production of 1,4-butanediol**

3. C. Angione, Jole Costanza, G. Carapezza, P. Lió and G. Nicosia, **Designing and Programming Molecular Machines**

4. C. Angione, Jole Costanza, G. Carapezza, P. Lió and G. Nicosia, **Optimizing constraint-based models through gene expression and codon usage information**

Chapter 2

CAD for the Design of Biological Circuits

In this Chapter of my thesis, I will present *BioCAD*, a new computational framework able to analyze, optimize and re-design biological systems. BioCAD faces Computational Systems Biology problems, by designing new modified and robustness metabolic systems and by providing comprehensive information about the structure or composition of the biological system or of its subsystems. The BioCAD framework includes many novel and state-of-the-art tools and in particular the (i) *Single* and *Multi-Objective Optimization* coupled with the novel ϵ-*dominance Analysis*, (ii) *Sensitivity*, (iii) *Identifiability* and (iv) *Robustness analysis*. BioCAD takes as input a general biological network (that can be modeled by means different mathematical approaches) and gives in output an other biological network, derived from the input network, but optimized according to the chosen targets. By means of the optimization core is possible, for instance, to perform the genetic design in order to recreate organisms that produce biodiesel or other interesting substances. In the work titled *Robust Design of Microbial Strains* [15], BioCAD was tested for the first time on the genome-scale metabolic network of the *Escherichia coli* bacterium for searching the best genetic manipulations in order to lead the bacterium to overproduce chemicals or biochemicals. In this case, I have used the Flux Balance Analysis (FBA) [2] mathematical approach to simulate the bacterium metabolism.

In the following sections, the general framework of BioCAD and its potentiality are described in details.

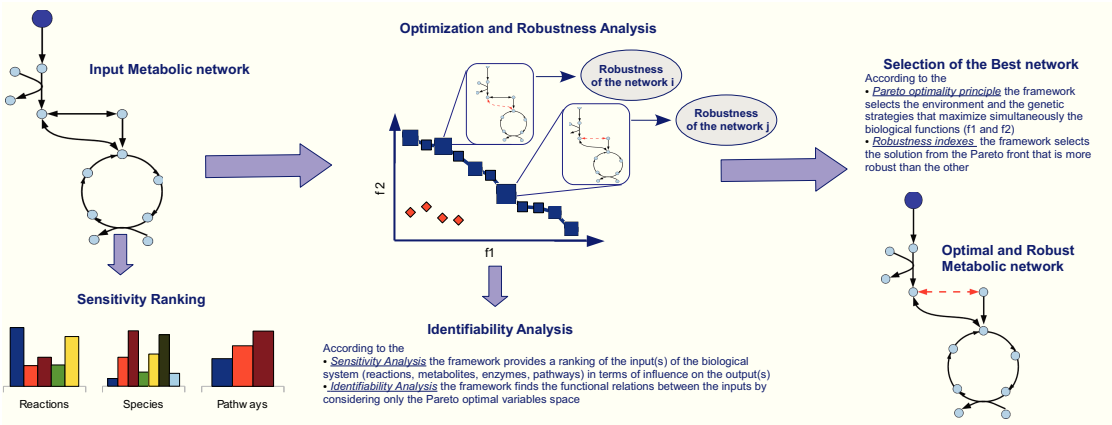


FIGURE 2.1: BioCAD flowchart. In the first step, a general Input Metabolic network is analyzed according to the Sensitivity Analysis that ranks the components of the circuit (reactions, species or pathways) according to their influence on the output(s) of the model. This first step does not change the conformation of the network, but only evaluates the importance of its components in the model. In the optimization procedure (single- or multi-objective), the algorithm optimizes the decision variables to maximize or minimize one or more objectives (in the flowchart, two objectives f_1, f_2 are being optimized). The method tunes the variables in an appropriate way (described in the text) for optimizing the objectives (chosen by the user). The result of the multi-objective optimization is the Pareto front (blue points). Each point of the front represents a particular conformation of the network. Additionally, by investigating the variables space of the Pareto front, BioCAD calculates the Identifiability relationships, i.e., the functional relations between the decision variables. The Pareto optimality is also coupled with the Robustness Analysis. For each Pareto point, the Global, Local and the Pathway-oriented Robustness is calculated (in the flowchart, the size of the Pareto point represents the robustness associated to the particular optimal network). The output of the method is therefore a modified network.

2.1 The Biological Computer-Aided Design Tool

BioCAD is composed of different parts, each of them performs a specific task. Specifically, there are three main parts: the pre-processing step, the optimization core and the post-processing step. The three parts can work together or in a separately procedure. BioCAD can be considered such as a black box, which receives a generic biologic circuit in input and returns the optimal and robust one(s). In the work "*Robust Design of Microbial Strain*" [15] BioCAD was tested on the genome-scale metabolic of *Escherichia coli* bacterium [16] and used to find the best genetic strategies in terms of knockout that lead the overproduction of succinate and acetate. The candidate optimal solutions was also analyzed by using Robustness Analysis and the new *Pathway-oriented Sensitivity analysis* (PoSA) was presented and applied for the first time. Details on this work can be found in section 2.5.2.

In Figure 2.1, the pipeline of BioCAD is shown. The computational method takes in input a metabolic network, which can be modeled by using ordinary differential equations (ODEs), differential algebraic equations (DAEs), partial differential equations (PDEs) or by using the Flux Balance Analysis framework (FBA) [2] containing or not the gene

protein reaction (GPR) mappings [10]. Models can be also uploaded by using systems biology markup language (SBML) [17]. In this way, the method is able to menage different mathematical models and to perform different analysis. Additionally, the framework here presented can be used in different areas, as in Electronic Design Automation (EDA), making it suitable for general purposes. The method can perform *Single* and *Multi-Objective Optimizations* to reach desired targets. The aim is to find the values of the decision variables in order to obtain one (single-objective) or more (multi-objectives) phenotypes. The user can analyze or optimize different components of the biological model. For instance, (i) perturbing enzymatic concentrations, (ii) turning off genes (genetic design), or (iii) searching for the optimal nutrients (fluxes design). Enzymes concentrations, genes and nutrients represent the decision variables of the optimization problem.

As above introduced, the framework is composed of three blocks. The first is constituted by the *Sensitivity Analysis* block, able to find the most sensitive parameters of the model. The reactions and the species can be analyzed in terms of sensitivity by using the *Reaction-oriented Sensitivity Analysis* (RoSA) and *Species-oriented Sensitivity Analysis* (SoSA) methods. Furthermore, the novel PoSA is able to identify the most sensitive metabolic pathways by ranking them according to gene knockouts or fluxes through each metabolic pathway. We can consider a single pathway as an input/parameter of the model. Each pathway (that is a set of reactions converting particular substrates in specific final products) is perturbed by deleting genes that control its biochemical reactions or by changing fluxes that occur in its reactions. PoSA ranks the pathways according to their influence on the outputs of the model. Moreover, the *Sensitivity* results can be coupled with *Optimization* process since the set of the decision variables can be resized and reduced by considering only the most sensitive parameters. As a matter of fact, in the photosynthetic carbon metabolism analysis, described in Chapter 5, I have considered only the most sensitive parameters/enzymes and optimized them for minimize the carbon dioxide uptake rate.

The second and the main part of BioCAD is the *Optimization* core. In Figure 2.1, it's reported the result of the *Multi-Objective Optimization* when the function f_1 and f_2 are maximized. The results of a Multi-Objective Optimization (MOO) problem is not a single solution (such as in a Single Optimization problem), but a set of non-dominated points, which form the *Pareto front* (the blue points of the central plot of Figure 2.1). A point is called "non-dominated" if there are no points that outperform it in all the objective functions. Instead, the dominated points (represented in red) are feasible points, but are less good with respect to the blue points of the graph, then not optimal. All the dominated points and the non-dominated points, that satisfy all of the constraints, and all of the variable bounds, constitute the observed feasible region.

Pareto optimality is useful to obtain not only a vast range of Pareto optimal solutions, but also the best *trade-off* design. Each feasible point represents a particular network, so the user can choose between several candidate optimal solutions. Pareto Optimality has been revealed appealing for the analysis of metabolism, as reported in the previous works [18, 19], where the authors used deterministic multi-objective approaches to evaluate the fluxes distributions in the *E. coli* wild-type network. In this thesis, I want to remark the usefulness of Pareto optimality and adopt effective and state-of-the-art algorithms to investigate the knockout space. After the optimization, the *ϵ-dominance Analysis* relaxes the Pareto constraints and can search accurately near the edge of the Pareto-optimal region. The ϵ-dominance Analysis and the Multi-Objective Optimization core are crossed together. In the section 2.3 definitions and many details about the Multi-Objective Optimization can be found . In addition, the shape of the front and the number of Pareto solutions give an idea of the behavior of the biological circuit with respect to a particular phenotype optimization.

One of the novelty is the genetic design, where each strain (a particular phenotype) is identified by a binary "knockout vector" (which represents the genotype), whose elements are 1 when the corresponding *gene set* is turned off. The importance of the knocked out genes can be evaluated by means of the ranking provided by Gene sets-PoSA. A gene set can be composed of a single gene, when it synthesizes for an enzyme, or can be associated with more genes, that synthesize for enzymes to form enzyme complexes and enzyme subunits. The relation between genes in a gene set is regulated by means of a Boolean relationship. When all the genes are necessary to catalyze the corresponding reactions (a single gene set can regulate more reactions), genes are linked by the "AND" operator; otherwise, if at least a gene is necessary to catalyze the reactions, genes are linked by the "OR" operator. For more details about the modeling and the managing of knockout strategies, the reader is reported to the sections 2.2.3 and 2.4. In addition, through the Multi-Objective Optimization, BioCAD is also able to find the favorable nutrients set (Flux Design) to optimize the wild-type or strains yield and evaluate the over/under investment of nutrients (uptake rate of fluxes) or the metabolic changing.

The *Robustness Analysis* is the third task of the BioCAD. For each phenotype (strain or wild type), in a post-processing step, BioCAD evaluates the fragility of the metabolic network when it is subjected to small perturbations, which can be exogen (for instance small perturbation of the external environment) or endogen (for instance small perturbation of the enzymatic concentrations). From the Pareto front, interesting solutions can be selected by using decision-making methods: for instance, by choosing points with suitable features or the solutions near to the ideal solution, the knee points, the

end points or by using the robustness analysis. For each solution, the *Global Robustness* (GR), the *Local Robustness* (LR) and the *Pathway-oriented Robustness* (PoRA) indexes values are calculated. These indexes indicate respectively the robustness of the whole network, the robustness of each single reaction and of each metabolic pathway. The introduction of robustness in the analysis should hence result in more reliable and realistic targets for biotechnology.

The computational analysis framework is extended also with the *Identifiability Analysis* that finds functional relations among enzymes, by analyzing the values of the decision variables after and before the optimization.

The output of the method is one or more biological circuits. In Figure 2.1, the output is chosen from the Pareto front according to trade-off and robustness values. Vertices of the network, figured as a graph, represent the metabolites and the edges are the relationships between metabolites, such as the biochemical and transport reactions. An edge represents a reversible reaction, while an oriented edge represents an irreversible reaction. The dashed line represents a modified reaction, for instance an intervention in the regulatory factors of the reaction or in the gene knockout array. The blue color of the vertex represents a change in the uptake rate, i.e., a different nutrient feed.

As introduced, BioCAD can be adopted and used with several modeling techniques and aims. In Table 2.1, have been showed advantages and limitations linked to the methods and models used in BioCAD. In particular, by using FBA and GPR associations, it is not possible to predict metabolite concentrations, since this method does not use kinetic parameters. However, it can determine fluxes at steady state. Additionally, FBA does not account for regulatory effects such as activation of enzymes or regulation of gene expression. Therefore, its predictions may not always be accurate. However, since FBA does not require kinetic parameters, it can be computed very quickly even for large networks. This makes it well suited to studies that characterize many different perturbations such as different substrates or genetic manipulations [2].

On the other hand, by using ODEs-DAEs-PDEs, the time to solve the system increases, though the metabolic system is not large and the precision depends on the computational solver. Instead, FBA uses a linear programming approach to find the solution of the problem, therefore the solution is equal using also different libraries (glpk, Gurobi Optimizer, LINDO Systems and so on). The advantage of ODEs-DAEs-PDEs models lies on the use of kinetic parameters, allowing to investigate several features, such as the regulatory effect, the variation on time of the metabolite concentrations, and in some cases, thermodynamic constraints.

TABLE 2.1: Advantages and limitations of methods and models used in BioCAD.

Features	ODEs-DAEs-PDEs	FBA
Kinetic parameters	considered	not considered
Regulatory effects	modeled	not modeled
Metabolite concentrations	prediction allowable	steady state
Accuracy	not always accurate	not always accurate
Simulation time	long	short
Size	small network	large network
Precision	low	good

Remarkably, the general sensitivity- and robustness-based framework allows a detailed understanding and comparison of the roles played by each component in the models taken into account. The main goal of this work is proposing a pipeline for model-based in silico design based on the state-of-the-art Multi-Objective Optimization approaches. BioCAD is implemented in Matlab in a parallel computing version.

2.2 The Flux Balance Analysis

BioCAD is able to menage metabolic networks modeled by using Flux Balance Analysis (FBA). FBA is a widely used approach for studying biochemical networks, in particular the genome-scale metabolic network reconstructions that have been built in the past decade. These network reconstructions contain all of the known metabolic reactions in an organism and the genes that encode each enzyme. FBA calculates the flow of metabolites through the metabolic network e.g., their formation and degradation, transport and cellular utilization, thereby making it possible to predict the growth rate of an organism or the rate of production of a biotechnologically important metabolite.

The first step in FBA is to mathematically represent metabolic reactions: for every metabolite X_i, $i = 1, \ldots, m$ a material balance is derived as $\frac{dX_i}{dt} = \sum_{j=1}^{n} S_{ij} v_j$, where S_{ij} is the stoichiometric coefficient associated with each reaction flux v_j, $j = 1, \ldots, n$. If we consider this material balance under steady state conditions, we have $\sum_{j=1}^{n} S_{ij} v_j = 0$. By considering all the intermediates simultaneously at steady states the balance equation can be written in matrix form $S \times v = 0$, where S is the stoichiometric matrix of m rows and n columns and v is the vector of the metabolic and transport fluxes. In any realistic large-scale metabolic model, the matrix S is not square and $n > m$, so we have a plurality of solutions. That is, there can be a number of feasible flux distributions satisfying these stoichiometric constraints, each representing a particular metabolic state. Therefore, the null space, or the set of all feasible flux distributions, represents the capabilities of the metabolic genotype. The transport fluxes represent environmental conditions that, along with the genotype, define the metabolic state.

The FBA approach finds the metabolic state in order to optimize a particular objective function, such as the maximization of growth rate or ATP production. Consequently, the problem can be formulated as a linear programming problem:

$$\begin{aligned} \text{maximise (or minimise)} \quad & f'v, \\ & \text{such that } Sv = 0 \\ & v_j^L \leq \ v_j \leq \ v_j^U, \quad j = 1, \ldots, n, \end{aligned} \tag{2.1}$$

where f is a vector of weights (n dimensional) that are either costs of or benefits derived from the fluxes. v_j^L and v_j^U are the lower and upper bound values (thermodynamic constraints) of the flux v_j. The output of FBA is a particular distribution of fluxes, denoted by v, that optimizes the objective function. Remarkably, FBA does not describe how a certain flux distribution is realized (by kinetics or enzyme regulation), but which flux distribution is optimal for the cell. Grown rate, also called biomass can be defined in terms of the biosynthetic requirement for the cell, and is represented by a dummy reaction formulated according to experiments found in literature. Simulating the generation of cellular biomass products from available inputs using the biomass objective function allows for the prediction of allowable growth rates for given substrate uptake rates and maintenance requirements.

2.2.1 The biomass function

The formulation of a detailed biomass objective function for use in examining metabolic networks is dependent on knowing the composition of the cell and energetic requirements necessary to generate biomass content from metabolic precursors. In this work, the advanced biomass objective function [20] is used and it is formed by detailing the necessary vitamins, elements, and cofactors required for growth as well as determining core components necessary for cellular viability. Inclusion of vitamins, elements, and cofactors allow for the analysis of a broader coverage of network functionality and required network activity. Another advanced approach is to not only define the *wild-type* biomass content of the cell, but to generate a separate biomass objective function that contains the minimally functional content of the cell. This objective function, referred to as the *core* biomass objective function, can result in increased accuracy when predicting gene, reaction, and metabolite essentiality and is formulated using experimental data from genetic mutants and knockout strains. For this reason in this work, for the genetic design is used the *core* biomass instead of the *wild-type* biomass [20]. It should be noted that with full reconstructions of the entire protein synthesis machinery, that the level and detail in biomass objective functions can continue to grow.

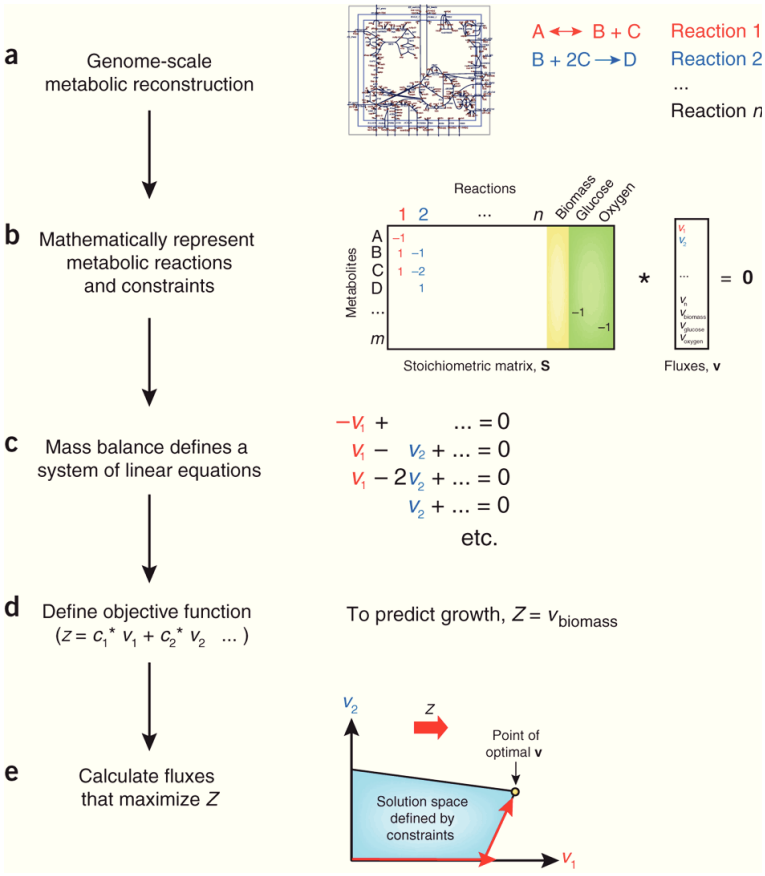


FIGURE 2.2: In figure have been showed the steps for the formulation of a FBA problem. Firstly, (a) the metabolic network reconstruction is performed by clustering the list of stoichiometrically balanced biochemical reactions. These information are extracted from data of literature or for data base on line, such as KEGG. (b) This reconstruction is converted into a mathematical model by forming the matrix S, in which each row represents a metabolite and each column represents a reaction. Growth is incorporated into the reconstruction with a biomass reaction, which simulates metabolites consumed during biomass production. Biomass production is mathematically represented by adding an artificial biomass reaction that consumes precursor metabolites at stoichiometries that simulate biomass production. Also exchange reactions are used to represent the flow of metabolites, such as glucose and oxygen, in and out of the cell. (c) At steady state, the flux through each reaction is given by $S \times v = 0$, which defines a system of linear equations. As large models contain more reactions than metabolites, there is more than one possible solution to these equations, therefore (d) a natural objective function (maximum growth rate) is chosen to predict the metabolite flow throughout the network. (e) Linear programming is used to identify a flux distribution that maximizes or minimizes the objective function within the space of allowable fluxes (blue region) defined by the constraints imposed by the mass balance equations and reaction bounds. The thin red arrows depict the process of linear programming, which identifies an optimal point at an edge or corner of the solution space. The output of FBA is a particular flux distribution, v, which maximizes or minimizes the objective function. Figure extracted from Palsson paper [2].

FBA can be used to perform simulations under different conditions by altering the constraints on a model. To change the environmental conditions (such as substrate availability), we can change the bounds on exchange reactions (that is, reactions representing metabolites flowing into and out of the system). Substrates that are not available are constrained to an uptake rate of 0 mmolh^{-1} gdW^{-1} (millimoles per gram dry cell weight per hour). Constraints can also be tailored to the organism being studied, with lower bounds of 0 mmolh^{-1} gdW^{-1} used to simulate reactions that are irreversible in some organisms. Nonzero lower bounds can also force a minimal flux through artificial reactions such as the *ATP maintenance reaction*, which is a balanced ATP hydrolysis reaction used to simulate energy demands not associated with growth [21]. Constraints can even be used to simulate gene knockouts by limiting reactions to zero flux. The main advantage of FBA method is the capability to menage large networks and in a faster way. However, FBA has limitations. Because it does not use kinetic parameters, it cannot predict metabolite concentrations. It is also only suitable for determining fluxes at steady state. Except in some modified forms, FBA does not account for regulatory effects such as activation of enzymes by protein kinases or regulation of gene expression. Therefore, its predictions may not always be accurate [2].

2.2.2 Gene Protein Reaction (GPR) associations

In order to allow BioCAD algorithms to work at the genetic level, I employ the Gene-Protein-Reaction (GPR) association, that indicates which gene has what function by means of a list of Boolean rules that dictate which genes are connected with each reaction in the model. When a reaction is catalyzed by isozymes (two different enzymes that catalyze the same reaction), the associated GPR contains an "OR" rule, where either of two or more genes may be knocked out but the reaction will not be constrained. The verification and refinement necessary in this step includes determining: (i) if the functional protein is a heteromeric (composed of two distinct gene products) enzyme complex; (ii) if the enzyme (complex) can carry out more than one reaction and (iii) if more than one protein can carry out the same function (i.e., isozymes exist). For the first case (i), the genome annotation often has refined information, which indicates that there is at least one more subunit needed for the function of the protein complex. Furthermore, KEGG (Kyoto Encyclopedia Genes and Genomes) [7] database lists subunits in some cases. Often, a more comprehensive database and/or literature search is required. Also, the protein-complex composition may differ between organisms. The second case (ii) can also be identified from biochemical databases and/or literature. Multitasking of enzymes may also differ between organisms [10]. Therefore, GPR associations distinguish between single and multi-functional enzymes, isoenzymes, enzyme complexes, enzyme subunits,

so that they capture the complexity and diversity of the biological relationships through the Boolean model above described [22].

2.2.3 Genetic Strategies modeling

Once constructed, GPR associations can be used to relate various data types, including transcriptomic, proteomic and flux data. In this thesis, the GPR mappings provide the links between each gene set and the reactions v_j that depend on it and define how certain genetic manipulations affect reactions in the metabolic network. For a set of L genetic manipulations, the GPR mappings are represented by a $L \times n$ matrix G, where the *(l,j)*-th element is 1 if the *l*-th genetic manipulation maps onto the reaction j, and is 0 otherwise.

Knockouts are modeled in terms of gene sets that can affect one or more reduced reactions using GPR mappings, which give a many-to-many mapping of genes to reactions. The *knockout cost* is defined according to the Boolean relationship between genes modeled by means of GPR map. If a gene set is composed of two genes linked by "AND", the cost to ensure the turning off of the corresponding reactions (knockout cost) is 1. Instead, the cost to ensure the catalysis of the corresponding reactions is 2, since both genes are necessary to turn on the reactions associated with that gene set. For example, the GPR for phosphofructokinase (PFK) is "b1723 (pfkB) or b3916 (pfkA)", where "b1723" and "b3916" are the gene IDs, "pfkB" and "pfkA" are the gene names. So according to this Boolean rule, both pfkB and pfkA must be knocked out to restrict this reaction. When a reaction is catalyzed by a protein with multiple essential subunits, the GPR contains an "AND" rule, and if any gene of the set is knocked out the corresponding reactions will be constrained to 0 flux. Succinyl-CoA synthetase (SUCOAS), for example, has the GPR "b0728 (sucC) and b0729 (sucD)" so knocking out either of these genes will restrict this reaction. Some reactions are catalyzed by a single gene product, while others may be associated with ten or more genes in complex associations [2].

I used the approach implemented in OptKnock method [12] to find the fluxes distribution that reproduces the desired productions (synthetic objectives) and achieves the maximal growth. The bi-level problem is mathematically formulated as:

$$
\begin{aligned}
&\max && g'v \\
&\text{such that} && \sum_{l=1}^{L} y_l \leq C \\
& && y_l \in \{0,1\} \\
& \max && f'v \\
& \text{such that} && Sv = 0 \\
& && (1-y)'G_j v_j^L \leq \ v_j \leq \ (1-y)'G_j v_j^U, \\
& && j = 1, \ldots, n,
\end{aligned}
\tag{2.2}
$$

where g is a vector of weights (n dimensional) associated with the synthetic objectives, and g' is its transpose. For example, when the synthetic objectives v_j and v_h have to be maximized, the weights g_j and g_h are equal to 1. The genetic strategies searching consists in finding the best knockout strain. I used the y knockout vector of L integers. If there are no impaired reactions in the metabolic network, y contains only zeros and the metabolic network is "wild-type". Conversely, when $y_l = 1$, the gene set embroiled in the l-th manipulation is turned off, and the corresponding reactions are in the absent status (the lower and upper bounds are set to zero, resulting in a modified metabolic network). C is an integer representing the maximum number of knockout allowed.

The bi-level problem can be converted to a mixed integer linear programming (MILP) problem (for a detailed description, see the original work [12]). I implemented and solved the problem using Matlab and the glpk (Gnu linear programming kit) solver [1].

2.3 BioCAD: the Multi-Optimization core

Multi-objective optimization is with no doubt a very important research topic both for scientists and engineers because of the multi-objective nature of most real-world problems.

Multi-objective optimization can be defined as the problem of finding a vector of decision variables which satisfies constraints and optimizes a vector function whose elements represent the objective functions [23]. These functions form a mathematical description of performance criteria which are usually in conflict with each other. Hence, the term "optimize" means finding such a solution which would give the values of all the objective functions acceptable to the designer.

Formally, we can state it as follows:

[1] Gnu linear programming kit, version 4.47. http://www.test.org/doe/

Find the vector $\overline{x}^* = [x_1^*, x_2^*, \ldots, x_n^*]^T$ which will satisfy the m inequality constraints:

$$g_i(\overline{x}) \geq 0, \ i = 1, 2, \ldots, m \tag{2.3}$$

the p equality constraints

$$h_i(\overline{x}) = 0, \ i = 1, 2, \ldots, p \tag{2.4}$$

and optimizes the vector function

$$\overline{f}(\overline{x}) = [f_1(\overline{x}), f_2(\overline{x}), \ldots, f_k(\overline{x})]^T \tag{2.5}$$

where $\overline{x} = [x_1, x_2, \ldots, x_n]^T$ is the vector of decision variables.

In other words, we wish to determine from among the set $\mathcal{F}$ of all numbers which satisfy 2.3 and 2.4 the particular set $x_1^*, x_2^*, \ldots, x_n^*$ which yields the optimum values of all the objectives functions.

The constraints given by 2.3 and 2.4 define the *feasible region* $\mathcal{F}$ and any point $\overline{x}$ in $\mathcal{F}$ defines a *feasible solution*. The vector function $\overline{f}(\overline{x})$ is a function which maps the set in $\mathcal{F}$ in the set $\mathcal{X}$ which represents all possible values of the objective functions. The k components of the vector $\overline{f}(\overline{x})$ represent the *non-commensurable* criteria which must be considered. The constraints $g_i(\overline{x})$ and $h_i(\overline{x})$ represent the restriction imposed on the decision variables. The vector $\overline{x}^*$ will be reserved to denote the optimal solutions. The problem is that the meaning of optimum is not well defined in this context, since we rarely have an $\overline{x}^*$ such that for all $i = 1, 2, \ldots, k$

$$\forall x \in \mathcal{F}, \ f_i(\overline{x}^*) \geq f_i(\overline{x}). \tag{2.6}$$

If this was the case, then $\overline{x}^*$ would be a desirable solution, but in real-world problems we never have a solution like this, in which all the $f_i(\overline{x})$ have a maximum in $\mathcal{F}$ at a common point $\overline{x}$. Therefore, we have to establish a certain criteria to determine what would be considered as an optimal solution.

As above introduced, when we have to perform a multi-objective optimization problem, the solution is not a single point (such as in the case of a single-objective optimization problem), but is a set of optimal points. This set of solutions is called *Pareto Optimal Set*. The concept of *Pareto optimum* was formulated by the Italian economist Vilfredo Pareto and constitutes by itself the origin of research in multi-objective optimization.

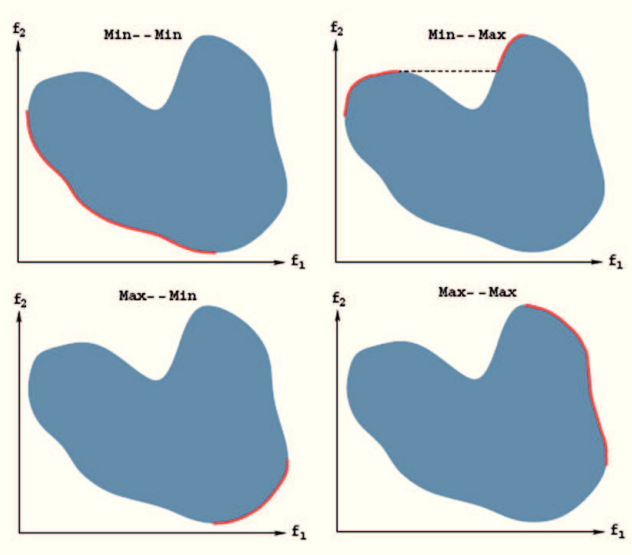


FIGURE 2.3: Pareto-optimal solutions are marked with continuous curves for four combinations of two type of objectives. Figure extracted from Kalyanmoy Deb book [24].

We say a point $\overline{x}^* \in \mathcal{F}$ is *Pareto optimal* if for every $\overline{x} \in \mathcal{F}$ either $\forall i \in I,\ f_i(\overline{x}^*) = f_i(\overline{x})$ or, there is at least one $i \in I$ such that $f_i(\overline{x}^*) > f_i(\overline{x})$.

In world, this definition says that $\overline{x}^*$ is *Pareto optimal* if there exists no feasible vector $\overline{x}$ which would increase some criterion without causing a simultaneous decrease in at least one other criterion [23]. Unfortunately, the *Pareto optimum* almost always gives not a single solution, but rather a set of solutions called *non-dominated* solutions.

In a maximization problem, the maxima in the Pareto sense are going to be in the boundary of the design region, or in the locus of the tangent points of the objective functions. In Figure 2.3, a bold line is used to mark this boundary for a bi-objective problem in four cases: to minimize both the objective functions, to minimize and maximize respectively the two objective functions, to maximize and minimize respectively the two objective functions and to maximize both the objective functions. The region of points defined by this bold line is called the *Pareto Front*.

Most real-world search and optimization problems involve multiple objectives. The extremist principle of finding the optimum solution cannot be applied to one objective alone when the rest of the objectives are also important. The presence of multiple conflicting objectives gives rise to a set of *trade-off* optimal solutions, i.e., the Pareto-optimal solutions. Different Pareto-optimal solutions produce a trade-off (conflicting scenarios) among different objectives. A solution that is better with respect to one objective requires a compromise in at least one other objective. Since many such solutions are targets here, clearly, there are two goals of multi-objective optimization: find a set

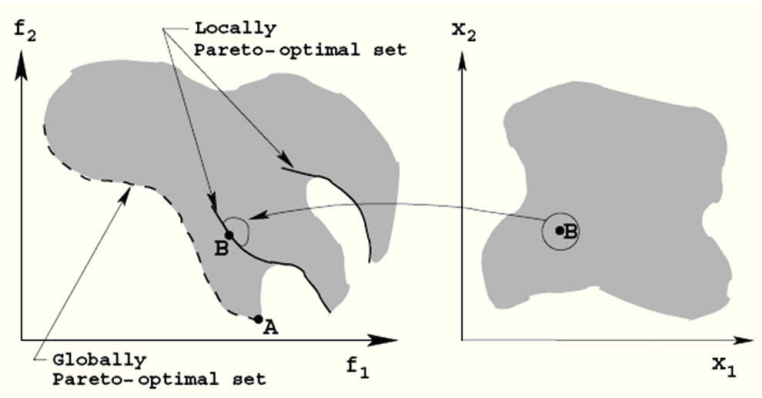


FIGURE 2.4: Solution space and variable space mapping. In the solution space we can observe the Local versus Global Pareto-optimal Fronts when we have the maximization of two objectives. Figure extracted from Kalyanmoy Deb book [24].

of solutions close to the Pareto-optimal solutions and find a set of solutions which are diverse enough to represent the entire spread of the Pareto-optimal front.

2.4 Genetic Design through Multi-objective Optimization Algorithm

Genetic Design through Multi-objective Optimization (GDMO) algorithm [15] is a combinatorial global search method that finds the genetic manipulation strategies to simultaneously optimize multiple cellular functions. As described in the section 2.3, the simultaneous optimization of multiple objectives differs from the single-objective optimization because the solution is not unique when the objectives are in conflict with each other. For instance, the knockout strategy able to improve the production of a metabolite alters the biomass formation and the ability of the organism to reproduce itself. Therefore, metabolite production and biomass formation are strongly in conflict.

GDMO implements a genetic algorithm inspired by the Non Dominated sorting Genetic Algorithm II (NSGA-II) [25]. A genetic algorithm is a search technique used in computing to find exact or approximate solutions to optimization and search problems. Genetic algorithm are categorized as global search heuristics and are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as inheritance, mutation, selection and crossover. In contrast to local search methods, genetic algorithms are based on a set of independent computations controlled by a probabilistic strategy. This is a simulation of natural selection of best individuals inside successive generations. Following the classical terminology, a solution for a problem under consideration is called an individual. The set of considered individuals is called a population. Each individual has one chromosome string encoding its data characteristics [26].

GDMO is composed of four key steps. It starts with the initialization of the population *Pop*. The population can be initialized in different ways: randomly, or assigning present status to all genes, or selecting a set of knocked out genes. In my work, I do not remove genes from the possible knockout list based on experimental predictions of lethality, as the lethality of a single knockout may not hold when combined with other knockouts. The possible knockout list is, however, easily set by the user. The population *Pop* is represented by a $I \times (L + K + 2)$ matrix, where I is the number of individuals, L is the number of the decision variables and K is the number of the objective functions. The last two columns are used to store two parameters of the algorithm linked to each individual: the fitness or rank, and the crowding distance [25]. The values of the objective functions are calculated solving the combinatorial problem 2.2 of the section 2.2.3.

Once the population is initialized the population is sorted based on non-domination into each front. The first front being completely non-dominant set in the current population and the second front being dominated by the individuals in the first front only and the front goes so on. Each individual in the each front are assigned rank or fitness values based on front in which they belong to. Individuals in first front are given a fitness value of 1 and individuals in second are assigned fitness value as 2 and so on. In addition to fitness value the crowding distance is calculated for each individual. The crowding distance is a measure of how close an individual is to its neighbors. The crowding distance computation requires sorting the population according to each objective function value in ascending order of magnitude. Thereafter, for each objective function, the boundary solutions (solutions with smaller and largest function values) are assigned an infinite distance value. All other intermediate solutions are assigned a distance value equal to the absolute normalized difference in the function values of two adjacent solutions. This calculation is continued with other objective functions. The overall crowding-distance is calculated as the sum of individual distance values corresponding to each objective. Each objective function is normalized before calculating the crowding-distance. Large average crowding distance will result in better diversity in the population.

Each individual represents a feasible solution, composed of the proposed $\tilde{y}$ knockout strategy array. Successively, three steps are iteratively carried out. In a *binary tournament selection* process, two individuals are selected at random, and their fitness is compared. The individual with the best fitness is selected as a parent. GDMO algorithm selects by default a number of parents (i.e., the best individuals) equal to $\frac{I}{2}$.

Parents are mutated using a *combinatorial mutation operator* to create an offspring population. Mutation represents a switch, from 0 to 1, or from 1 to 0. The process is randomly executed; for each parent individual, ten offspring are created and only the best one is chosen. Mutations can achieve the maximum knockout number equal to

the parameter C (equal to 50 by default). A new population of I individuals is formed selecting the best individuals from the set of parents of the previous generation and the novel offspring. The new population undergoes a new round of evaluation. Finally, a *selection* operator is performed in order to reach the last front. For each generation of the algorithm, the Pareto optimal solutions are provided.

This cycle is repeated until the solutions set does not improve, or until an individual with a desired phenotype is achieved or when the number of generations reaches its upper bound. The number of generations D and individuals I are parameters chosen by the user. After calculating the Pareto-optimal solutions, a post-processing filtering is performing in order to eliminate redundant knockout that are not, in fact, necessary for the achievement of the selected production and biomass level. The time-complexity of the genetic algorithm is $O(KDI^2)$, where K is the number of the objectives, D is the number of generations and I the population size.

2.5 Sensitivity Analysis

The Sensitivity Analysis (SA) is an indispensable tool for the analysis of complex systems used to discover which parameters are the most important and the most likely to affect predictions of the model, that is the parameters/inputs that have a substantial influence on the output(s) of the model. Following a sensitivity analysis, values of critical parameters can be refined while parameters that have little effect can be simplified or even ignored. In order to decide which parameters can be discarded in models with redundant parameters the importance of the parameters given by a ranking based on the overall sensitivity can be used. Therefore, in the contest of numerical modeling, sensitivity indices play an important role in uncertainty analysis, parameter estimation, optimization, experimental data analysis and model discrimination.

Local sensitivity indices are computed at the nominal values used for the parameters and the behavior of the response function is described only locally in the input space. More in details, local sensitivity coefficients are the partial derivatives of the model state variables to the model parameters evaluated at the nominal operating point where all the parameters have their nominal values. There are several numerical methods for the calculation of local sensitivities but, these methods are linear thus they are not sufficient for dealing with complex models, especially those in which there are nonlinear interactions between parameters. In contrast, global sensitivity analysis methods evaluate the effect of a parameter while all other parameters are varied simultaneously, accounting for interactions between parameters without depending on the stipulation of a nominal point (they explore the entire range of each parameter) [27]. Global sensitivity indices

Algorithm 1 GDMO Algorithm pseudo-code

Require: $[f, y, I, D, C,]$
/* f output of the model */
/* y knockout vector */
/* I is the number of the individuals */
/* C is the maximum knockout number allowed*/
Generate Initial Population P(y) with random binary vector $\tilde{y}$ or null wild type vector y
Evaluate the rank (fitness) of all individuals
for $i \leftarrow 1$ to D **do**
/* Define $p_i(y)$ selecting the best I/2 individuals of population $P_i(y)$ according Tournament selection (lower rank and higher crowding distance)*/
    $p_i(y) \leftarrow$ Tournament Selection($P_i(y)$)
    $p_i(y) \leftarrow$ best I/2 individuals from $p_i(y)$
    for $h \leftarrow 1$ to $I/2$ **do**
        /* perform Mutation Operator for the single parent $\mathrm{p}_{i,h}(y)$ */
        $\tilde{y}_{i,h,1} \leftarrow$ Mutation($y_{i,h}, C$)
        for $j \leftarrow 2$ to 10 **do**
        /* perform Mutation Operator for each child $\tilde{y}_{i,h,j}$*/
            $\tilde{y}_{i,h,j} \leftarrow$ Mutation($\tilde{y}_{i,h,j-1}, C$)
        end for
        Evaluate the rank (fitness) of the K children $p_{i,h,j}(\tilde{y})$)
        Select the best child $p_{i,h}(\tilde{y})$) from $p_{i,h,j}(\tilde{y})$)
        /* $\tilde{p}_{i,h}(y)$ is the mutation of $\mathrm{p}_{i,h}(y)$*/
    end for
    Merge $P_i(y)$ with $\tilde{p}_i(y)$ in $Pp_i(y)$
/* Perform Selection on $Pp_i(y)$ and obtain the new $P_i(y)$ */
    $P_i(y) \leftarrow$ Tournament Selection($Pp_i(y)$)
    $P_i(y) \leftarrow$ best I individuals from $P_i(y)$

    if finds desired solutions **then**
        return $P_i(y)$
    end if
end for
return $P_i(y)$

should be regarded as a tool for studying the mathematical model rather then its specified solution. The most widely used methods in global sensitivity analysis are the Morris method [28] and the Sobol' method [29].

In 1991, Morris published a method that perturbs in a random way the input(s) of a model in order to obtain for each input a distribution of *elementary effects* of the perturbation. An elementary effect is calculated by comparing (for instance by means of Euclidian distance) the output(s) of the model when the input is perturbed with the output(s) of the model without perturbation. The plot of the distribution gives an idea of how the perturbation of the input affects the output(s) of the model. If the curve has a large spread, the input has an high influence dependent on the values

of the other inputs; instead, if the mean of the distribution is large, the input has an important overall influence on the output(s) [28]. SA is frequently used for the in-silico design of electronic devices, and in the last decade it has been also used in Systems Biology. While for electronic design automation, input(s) of the model can be gain and tension, or resistance and conductance, in a biological model the input(s) of the model can be: (i) nutritive substances of a cell (for instance the uptake rate of glucose or oxygen); (ii) gene knockouts in the genome of a bacterium (for instance the knockout of pyruvate dehydrogenase complex); (iii) enzyme concentrations in a metabolic pathway (for instance the concentration of RuBisCO in a plant cell). According to the modeling technique and the parameters included in the model, SA provides a ranking of selected input(s), based on their importance.

SA indices have been adopted for interrogating the reactions space (RoSA - Reactions oriented Sensitivity Analysis) [30] and species space (SoSA - Species oriented Sensitivity Analysis) to find their influence on the response of the system [31]. Moreover, in order to associate biochemical pathways with sensitivity values, BioCAD implements the novel Pathway-oriented Sensitivity Analysis (PoSA) methods [15], able to evaluate the importance of a pathway in terms of genetic knockout or fluxes through the pathway. BioCAD framework includes Morris [28], Sobol' [29] and the PoSA methods.

2.5.1 Morris method

The Morris method [28] is traditionally used as a screening method for problems with high number of variables and for which function evaluations are CPU-time consuming. It is composed of individually randomized "one-factor-at-a-time" experiments. Each input factor may assume a discrete number of values, called levels, which are chosen within the factor range of variation. Consider a computational model for which the output is a deterministic function of k inputs denoted by $x = [x_1, x_2, \ldots, x_k]$. The range of each input variable x_i (called region of interest Ω) is divided into p levels, then the region of experimentation ω is a k-dimensional p-level grid. Each x_i may take on values from $\{0, 1/(p-1), 2/(p-1), \ldots, 1\}$, assuming that each x_i is scaled to take on values in the interval [0,1]. The sensitivity measures are based on what is called an elementary effect. The elementary effects method is a simple but effective screening strategy. Starting from a number of initial points, the method creates random trajectories to then estimate factor effects. In turn, those estimates are used for factor screening. For a given value of x^*, the elementary effect of the i-th input is defined as:

$$EE_i(x^*) = \frac{\Big[f(x_1^*, ..., x_{i-1}^*, x_i^* + \Delta, x_{j+1}^*, ..., x_k^*) - f(x^*)\Big]}{\Delta}, \tag{2.7}$$

where Δ is a predetermined multiple of $1/(p-1)$ and point $x^* \in \omega$ is such that $x^* + \Delta$ is still in ω. The distribution of elementary effects F_i is obtained by randomly sampling n points from ω. The mean μ_i and the standard deviation σ_i are the two sensitive indexes for the input x_i. When the output of a system is non-monotonic function, that has regions of positive and negative values of partial derivatives $EE_i(x^*)$, μ_i can be very small or even zero. For this reason, in Campolongo et al. [32] has been considered another sensitivity measure μ_i^* , which is an estimate of the mean of the distribution of the absolute values of the elementary effects and has been showed that μ_i^* gives a better estimate of the order of importance than μ_i.

By using Morris method [28], we can calculate the distribution of elementary effects for each metabolite or enzyme of a biological system. In SoSA and RoSA the analyzed inputs/parameters are real-valued variables. For instance, in FBA models the parameters I take into account are the uptake rate and the metabolic fluxes of the network.

A large (absolute) central tendency μ^* indicates an input with an important overall influence on the output. A large spread σ^* indicates an input whose influence is highly dependent on the values of the inputs [28].

2.5.2 Pathway-oriented Sensitivity Analysis

BioCAD includes two different version of PoSA: the *Gene sets*-Posa and *Fluxes*-PoSA.

Unlike other SA methods applied in biological modeling, whose inputs (reactions or species) are valued in a *continuous* region of interest, Gene sets-PoSA is applied when inputs are valued in a *discrete* region of interest and finds the genetic manipulations that have the largest influence on the response of the system.

Gene sets-PoSA investigates the knockout solution space and determines the influence of a pathway on the output(s) of a FBA model. Since GDMO algorithm provides a set of feasible solutions with different genetic manipulations, it is worth seeking a relationship between the sensitivity indices and the proposed manipulations. This way, we can select only the best manipulations after GDMO searching (section 2.4). In particular, thanks to the information provided by PoSA, we can prefer that knockout strategies (between the Pareto-optimal candidate solutions) that affect genes belonging to *insensitive pathways*.

In Gene sets-PoSA, the knockout vector y used to represent the genetic manipulations is partitioned in p subsets of bits $\{b_1, b_2, \ldots, b_s, \ldots, b_p\}$. Each subset b_s represents a pathway and includes the genetic manipulations linked to the reactions involved in the s-th metabolic pathway of the network. Each subset b_s has a cardinality W_s, where $W_s < L$, $\forall\ s = 1, \ldots, p$. Each pathway performs a particular task in the metabolism, e.g.,

the *Citric Acid Cycle*, the *Oxidative Phosphorylation*, the *Pentose Phosphate Pathway*, and so on. Gene sets-PoSA takes also into account the eventuality that a reaction could belong to different pathways: when the gene responsible for that reaction is knocked out, the reaction is impaired in all its pathways. The gene-pathway mappings (GP) is implemented and is defined by the $L \times p$ matrix P, where the *(l,s)*-th element of P is 1 if the l-th genetic manipulation is linked to the reactions involved in the s-th functional pathway, and 0 otherwise. I also adopted the reaction-pathway mappings (RP), mathematically described by the $n \times p$ matrix R, where the *(j,s)*-th element of R is 1 if the j-th reaction belongs to the s-th functional pathway, and 0 otherwise. For the combinatorial problem described in equation 2.2, the elementary effect (EE_s) [28] for the input b_s is defined as:

$$EE_s = \left[F(b_1, b_2, \ldots, b_{s-1}, \tilde{b}_s, b_{s+1}, \ldots, b_p) - F(\tilde{y})\right] / \Delta_s, \tag{2.8}$$

where $\tilde{b}_s$ is the mutation on the input b_s, and consists of the *switching* of bits chosen randomly in b_s: if a bit is equal to 0 (or 1), the permutation turns it into 1 (or 0). Δ_s is a scale factor defined as:

$$\Delta_s = \frac{1}{W_s} \sum_{i=1}^{W_s} \tilde{b}_s(i), \ s = 1, \ldots, p. \tag{2.9}$$

The output $F(y)$ considered in Gene sets-PoSA is the vector v of fluxes. $\tilde{y}$ is the global mutation carried out on the knockout vector y defined in the Boolean region of interest $\Omega = \{0,1\}^L = \{(y_1, \ldots, y_l, \ldots, y_L) | y_l \in \{0,1\}\}$.

The distribution of effects EE_s is obtained by permuting y through a random sampling of KQ points from Ω and permuting b_s by randomly sampling KQN points from Ω. If the procedure was performed for each input, the result would be a random sample at a total cost of KQ for calculating $F(\tilde{y})$ and KQN for $F(b_1, b_2, \ldots, \tilde{b}_s, \ldots, b_p)$, with a total cost of $pKQ(N+1)$ evaluates of function. As regards the details, following the pseudo-code of the algorithm is reported.

The estimation of the mean μ^* and the standard deviation σ^* of the distribution of the elementary effects will be used to detect those inputs that should be considered influent in the model.

In Fluxes-PoSA method, the perturbation is applied on the value of fluxes through the metabolic pathway. For each pathway of the system, a global random perturbation is performed. The global perturbation consists on applying a random noise on the exchange fluxes. After that, FBA is processed and the fluxes distribution $F(\tilde{\overline{v}}_{ex})$ is obtained, where $\tilde{\overline{v}}_{ex}$ represents the random perturbation on the exchange fluxes $\overline{v}_{ex}$. The response $F(\tilde{\overline{v}}_{ex})$

Algorithm 2 Gene-sets PoSA Algorithm

Require: $[f, y, Q, N, K, \alpha, \beta]$
/* f output of the model */
/* y knockout vector */
/* $[Q, N, K, \alpha, \beta]$ parameters of PoSA*/

Given $Y = \{b_1, b_2, \ldots, b_s, \ldots, b_p\}$
for $s \leftarrow 1$ to p **do**
  Select the pathway b_s
  for $q \leftarrow 1$ to Q **do**
    for $\alpha \leftarrow 1$ to K **do**
    /* perform Mutation Operator on y */
      $\tilde{y}(q, \alpha) \leftarrow \text{Mutation}(y, \beta \cdot W_s)$
      for $h \leftarrow 1$ to N **do**
      /* perform Mutation Operator on b_s */
        $\tilde{b}_s(\alpha, q, h) \leftarrow \text{Mutation}\Big(b_s, \frac{\alpha}{K} \cdot W_s\Big)$
      /* evaluate scale factor Δ */
        $\Delta_s(h, \alpha, q) = \frac{\sum \tilde{b}_s(\alpha, q, h)}{W_s}$
      /* evaluate an elementary effect on pathway b_s */
        $EE_s(h, \alpha, q) = \frac{f(\tilde{y}(q, \alpha)) - f(b_1, b_2, \ldots, \tilde{b}_s, \ldots, b_p)}{\Delta_s(h, \alpha, i)}$
      end for
    end for
  end for
  $\mu_s{}^* \leftarrow \text{mean}(EE_s)$ /* evaluate the $\mu_s{}^*$ sensitivity index */
  $\sigma_s{}^* \leftarrow \text{var}(EE_s)$ /* evaluate the $\sigma_s{}^*$ sensitivity index */
end for
return $[\mu*\ , \sigma*\ ]$

Algorithm 3 Mutation Operator

Require: $[x, a \cdot W_x]$
/* $a \in \{0, 1\}$ is a real constant and defines the percentage of mutations in the Boolean vector x of W_x elements*/

/* Select randomly an integer value A in $[1, a \cdot W_x]$ */
$A \leftarrow \text{random } (1, aW_x)$

/* Select randomly A bits on x vector */
$ind \leftarrow \text{random } (1, W_x, A)$

$\tilde{x} \leftarrow \text{not}(x\,[ind]\ )$
return $\tilde{x}$

gives us the new distribution of all the fluxes $v_j,\ j = 1, \ldots, n$ of the systems. From these, a random perturbation is applied on the values of fluxes $\overline{v}_s,\ s = 1, \ldots, p$ where $\overline{v}_s$ is the vector of fluxes belonging to the pathway b_s. For perturbing fluxes, a Δ uniformly noise is added to the nominal values of the set $\overline{v}_s$, obtaining the modified vector $\tilde{\overline{v}}_s$. For Fluxes-PoSA, the elementary effect for the generic pathway b_s is calculated as:

$$EE_s = \left[F(\overline{v}_1, \overline{v}_2, \ldots, \overline{v}_{s-1}, \tilde{\overline{v}}_s, \overline{v}_{s+1}, \ldots, \overline{v}_p) - F(\tilde{\overline{v}}_{ex})\right] / \Delta. \tag{2.10}$$

By sampling EE_s several times, the distribution of elementary effect for the pathway b_s is estimated. The Δ noise is randomly chosen for each evaluation of the elementary effect.

2.5.3 Sobol' method

The method of global sensitivity indices developed by Sobol' is based on ANOVA decomposition. Assume that the model under study is described by the function $f(x)$, where the input $x = (x_1, \ldots, x_n)$ and $x \in I^n$, where I is the unit interval [0,1], and I^n the n-dimensional unit hypercube. Consider an integrable function $f(x)$ defined in I^n and its representation in the form

$$f(x) = f_0 + \sum_{s=1}^{n} \sum_{i_1 < \ldots < i_s}^{n} f_{i_1, \ldots, i_s}(x_{i_1}, \ldots, x_{i_s}), \tag{2.11}$$

where $1 \leq\ i_1 < \ldots < i_s \leq\ n$. Equation 2.11 can be also written as

$$f(x) = f_0 + \sum_i f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) + \ldots + f_{12\ldots n}(x_1, x_2, \ldots, x_n), \tag{2.12}$$

and the number of summands is 2^n. Equation 2.11 is called ANOVA (Analysis Of Variances) representation of f(x) if

$$\int_0^1 f_{i_1, \ldots, i_s}(x_{i_1}, \ldots, x_{i_s})\, dx_k = 0,\ for\ k = i_1, \ldots, i_s. \tag{2.13}$$

It follows from Equation 2.12 that the member in 2.11 are orthogonal and can be expressed as integrals of $f(x)$ and:

$$\int_0^1 f(x)\, dx = f_0. \tag{2.14}$$

Assume now that $f(x)$ is square integrable. Then all the $f_{i_1,\ldots,i_s}$ in Equation 2.11 are square integrable also, therefore we get:

$$\int_0^1 f^2(x)\,dx - f_0 = \sum_{s=1}^{n} \sum_{i_1<\ldots<i_s}^{n} \int_0^1 f^2_{i_1,\ldots,i_s}, dx_{i_1}, \ldots\ dx_{i_s}. \tag{2.15}$$

The constants $D = \int_0^1 f^2\,dx - f_0$, and $D_{i_1,\ldots,i_s} = \int_0^1 f^2_{i_1,\ldots,i_s}, dx_{i_1}, \ldots, dx_{i_s}$ are called variances and $D = \sum_{s=1}^{n} \sum_{i_1<\ldots<i_s}^{n} D_{i_1,\ldots,i_s}$. The ratios

$$S_{i_1,\ldots,i_s} = \frac{D_{i_1,\ldots,i_s}}{D} \tag{2.16}$$

are called global sensitivity indices and all the $S_{i_1,\ldots,i_s}$ are nonnegative and their sum is $\sum_{s=1}^{n} \sum_{i_1<\ldots<i_s}^{n} S_{i_1,\ldots,i_s} = 1$.

Consider two complementary subsets of variables $y = (x_{k_1}, \ldots, x_{k_m})$ of m variables and z of $n-m$ variables, thus $x = (y, z)$. Let $K = (k_1, \ldots, k_m)$. The variance corresponding to the subset y can be defined as

$$D_y = \sum_{s=1}^{m} \sum_{(i_1<\ldots<i_s)\in K} D_{i_1,\ldots,i_s}. \tag{2.17}$$

The sum in Equation 2.17 is extended over all groups $(i_1, \ldots, i_s)$ where all the $i_1, \ldots, i_s$ belong to K. Similarly, the variance D_z can be introduced. Then the total variance corresponding to the subset y is $D_y^{tot} = D - D_z$. The two global sensitivity indices for the subset y are

$$S_y = \frac{D_y}{D} \quad and \quad S_y^{tot} = \frac{D_y^{tot}}{D}. \tag{2.18}$$

Clearly, $S_y^{tot} = 1 - S_z$ and always $0 \leq S_y \leq S_y^{tot} \leq 1$. The most informative are the extreme situations: $S_y = S_y^{tot} = 0$ means that $f(x)$ does not depend on y and $S_y = S_y^{tot} = 1$ means that $f(x)$ depend on y only.

Sobol' [29] found an elegant way of computing these indices directly from the model $f(x)$:

$$S_y = \frac{\int_0^1 f(x)f(y,z')\,dx\ dz' - f_0^2}{\int_0^1 f^2(x)\,dx - f_0^2} \tag{2.19}$$

$$S_y^{tot} = \frac{1}{2}\frac{\int_0^1 [f(y,z) - f(y',z)]^2 \, dx \; dy'}{\int_0^1 f^2(x)\, dx - f_0^2}. \tag{2.20}$$

S_y and S_y^{tot} can be used as sensitivity indices and to provide a parameter ranking.

2.6 ϵ-dominance Analysis

The *ϵ-dominance Analysis*, inspired by Laumanns et al. [33], is a technique that improves the diversity of the solutions and the convergence of the optimization algorithm. This method, together with the Identifiability analysis and the Robustness analysis, is part of the post-processing core of BioCAD framework. The aim of the ϵ-dominance analysis is to extend the solutions space and search for other interesting solutions.

In section 2.3, I highlighted that a point $\overline{x}^*$ in the solution space is said to be Pareto optimal if there does not exist a point $\overline{x}$ such that $f_i(\overline{x}) > f_i(\overline{x}^*), \forall\, i = 1, ..., k$, where f is the vector of k objective functions to optimize in the objective space. The ϵ-dominance technique applies a "relaxed" condition of dominance. That is, a point $\overline{x}^*$ in the solution space is said to be ϵ-non-dominated if there does not exist a point $\overline{x}$ such that $f(\overline{x})$ dominates $f(\overline{x}^*)$ of a value higher than ϵ. Formally, $\overline{x}^*$ is said to be ϵ-non-dominated if $\nexists\ \overline{x} : f_i(\overline{x}) > f_i(\overline{x}^*) + \epsilon_i,\ \epsilon_i > 0,\ \forall i = 1, ..., k$. This "relaxed" condition captures both the "ϵ-non-dominate" solutions and the non-dominated ones (Pareto-optimal). This technique allows inferior solutions to remain in the population, increasing diversity and helping obtain multi-modal solutions also in single objective optimization problems.

In my works, I use this method in multi-objective optimization problems to seek solutions that may have been discarded because they are dominated by a small amount ϵ that, for biological purposes, can be considered negligible. After the optimization, I perform an ϵ-dominance analysis to search accurately near the edge of the Pareto-optimal region. Formally, let f be the array of the k objective functions, and suppose that all the objective functions are positive and must be maximized. Let $\epsilon_i > 0$ be the tolerance of the relaxed condition. For each objective function $f_i(\overline{x})$, there are different ϵ_i tolerances. I seek all points (solutions) $\overline{x}^*$ belonging to the set $\{\overline{x}^* : f_i(\overline{x}^*) + \epsilon \geq f_i(\overline{x}),\ \forall\ i = 1, ..., k\}$,where f is the vector of the k objective functions and $\overline{x}$ represents all the others sampled points. The "ϵ-non-dominated" solutions can be considered suboptimal solutions because they are close to the Pareto-optimal region.

2.7 Identifiability Analysis

A biological model is made up of many components (e.g., parameters, variables) estimated through fitting to experiments. The Identifiability Analysis (IA) seeks the functional relations underlying the components of a given system, and can be used after the multi-objective optimization. Coupled with the sensitivity analysis, it gives insight into the model under investigation.

A component is said to be *non-identifiable* if there is no unique solution for its estimation. The non-identifiability can be (i) *structural*, when there are relations among components and therefore they cannot be determined unambiguously; (ii) *practical*, when the low amount or quality of data available does not allow to have a good estimate for the component. Using repeated fitting to data and estimations of components, the IA is aimed at finding the structural non-identifiable components of a model.

Specifically, let m be the number of decision variables $\{x_1, ..., x_m\}$ of the model, which are related by unknown linear or non-linear functional relations. Let n be the number of estimates available for each variable. These estimates are usually organized into a table $K = [v_1, ..., v_m] \in \mathbb{R}^{n \times m}$, where each column $v_i \in \mathbb{R}^n$ consists of the n estimates for the i-th variable

Let us denote by α and β_j the true transformations that linearize the relations among variables:

$$\alpha(x_i) = \sum_{j \neq i}^{m} \beta_j(x_j) + \xi, \tag{2.21}$$

where ξ is a Gaussian noise. The alternating conditional expectation (ACE) algorithm [34] estimates the optimal transformations $\hat{\alpha}(x_i)$ and $\hat{\beta}_j(x_j),\ j \neq i$, such that

$$\hat{\alpha}(x_i) = \sum_{j \neq i}^{m} \hat{\beta}_j(x_j), \tag{2.22}$$

where x_i is the response, while all the other variables are the predictors.

The process of repeating estimates in the matrix K is replaced by taking into account all the non-dominated points of the Pareto front. In other words, a single fitting sequence K is obtained by considering the entire front. Thus, the problem of identifiability analysis is mapped onto the problem of detecting groups of the functionally related decision variables that produce that Pareto front.

Specifically, the connection between the identifiability analysis and a constraint structure stems from the fact that a non-identifiable constraint involving decision variables causes them to be functionally related.

2.8 Robustness Analysis

The ability of a system to adapt to perturbations due to internal or external agents, aging, temperature, environmental changes and, in our case, also due to molecular noise and mutation is one of evolutionism guidelines and should also be a fundamental design principle. After the optimization, the validity of the biological system, designed in-silico, must be tested, and this is performed by the *robustness analysis* post-processing step. In this way, we can assess the ability of a system to adapt to small perturbations that can occur at any stage of the biochemical processes, either within the metabolic network or caused by the external environment. As we shall see, by the term "adaptive capacity" we can indicate the ability to maintain "acceptable" the performances previously optimized.

There are numerous methods that can be used to fulfil this task. Among these, in Callaway et al. [35] the authors consider a big network (in this case the Internet network) and use the *theory of percolation on random graphs* to test the robustness of the network in case of random or targeted node deletion, or in case of random link deletion. In another work [36], the relationship between the general characteristics of a chemical reactions network and the sensitivity of its equilibrium is investigated according to changes in the overall supply of reagents. The authors define *the sensitivity of a species* as the variation of it with respect to the element concentration one, and they find a lower bound to such sensitivity that depends on the network structure alone. In particular, they argue that a strong robustness of the equilibrium against element variations is likely only if the various species are constructed from building block highly gregarious (i.e., each one binds with many others) or present in some species with high multiplicity. Finally, in Wagner et al. [37] the authors use a combined approach of global and local robustness that they call *Glocal Robustness*. The global analysis investigates the parameter space with the aim of finding where a circuit cell shows experimental observed features (global), while the local one determines the robustness of parameter sets sampled during the previous phase. Similar works making use of the robustness analysis for parameter estimation are also present in Gilbert works [38, 39]. In my works, however, I have used very simple robustness analysis that shows a high degree of transversality because easily applicable in other fields, as was done in Nicosia et al. works [40, 41].

The basic principle of this analysis is as follows. Firstly, the perturbation is defined as a function $\tau = \gamma(\Psi, \sigma)$ where γ applies a stochastic noise σ to the system Ψ and generates

a trial sample τ. The γ-function is called γ-perturbation. Without loss of generality, assume the noise is defined by a random distribution. In order to make statistically meaningful the calculation of robustness, a set T of trial samples τ is generated. Each element τ of the set T is considered robust to the perturbation, due to stochastic noise σ, for a given property (or metric) ϕ if the following condition is verified:

$$\rho\left(\Psi, \tau, \phi, \epsilon\right) = \begin{cases} 1, & \text{if } \left|\phi\left(\Psi\right) - \phi\left(\tau\right)\right| \leq \epsilon \\ 0, & \text{otherwise} \end{cases} \tag{2.23}$$

where Ψ is the *reference system*, ϕ is a *metric* (or property), τ is a *trial sample* of the set T and ϵ is a *robustness threshold*. The definition of this condition makes no assumptions about the function ϕ. It can be anything (not necessarily related to properties or characteristics of the system); however, it is implicitly assumed that it is quantifiable. The robustness of a system Ψ is the number of robust trials of T, with respect to the property ϕ, over the total number of trials ($|T|$). In formal terms:

$$\Gamma\left(\Psi, T, \phi, \epsilon\right) = \frac{\sum_{\tau \in T} \rho\left(\Psi, \tau, \phi, \epsilon\right)}{|T|} \tag{2.24}$$

where Γ is a dimensionless quantity that states, in general, the robustness of a biological system.

Robustness index is a function of ϵ, so the choice of this parameter is crucial and not a trivial task. Since I am interested in the behavior of the system when it is subjected to small perturbations and because the behavior is acceptable when the deviation from the original value is as small as possible, I choose the values of epsilon equal to 1% of the metric and sigma equal 1% of the perturbed variable. In Figure A.5 of Appendix A I report the study conducted on bacterial strains and the analysis for choosing ϵ and σ. More tests have been also conducted in [42].

Based on this principle, in BioCAD three robustness methods are implemented, each of them evaluate respectively the *Global Robustness* (GR), the *Local Robustness* (LR) and the *Pathway-oriented Robustness* (PoRA) indexes. Additionally, in BioCAD is implemented the *Normalized Feasible parameter Volume* also called *Glocal Analysis* [37], a global robustness technique, for giving a comparison between Glocal results and the GR/LR values.

The first three values only differ in the perturbation kind, in particular, chosen σ, it will differ the set of variables that will be perturbed. For a FBA model, the values of the fluxes v_j, $j = 1, \ldots, n$ of the metabolic fluxes are perturbed. In particular, in GR the perturbation is carried out simultaneously for all the fluxes of the network to evaluate the fragility of the complete organism. In LR, the perturbation is carried out for each

flux (hence I have a robustness index for each flux), whereas in PoRA the perturbation is carried out simultaneously for all the fluxes clustered in a metabolic pathway obtaining a robustness index for each pathway.

Regarding the Global Robustness of a strain [15], I chosen δ equal to 1% of $\phi(\Psi)$ and I performed the perturbation with Gaussian noise with zero mean and standard deviation equal to 1% of the perturbed variable. In particular, I perturbed the metabolic fluxes. Hence, a trial τ is created by perturbing all the v_j, $j = 1, \ldots, n$ of the metabolic fluxes. In order to satisfy the constraints of the FBA model I choose to apply the perturbation to subsets of parameters chosen randomly for each iteration. At the end of the analysis all the parameters/fluxes underwent to perturbation. A set T of trials is created and for each of them, the fluxes are perturbed in order to evaluate the property $\phi(\tau)$ (by means of flux balance analysis [2]), and calculate the function ρ. Once a value of ρ is obtained for each of the trials, the value of robustness (Equation 2.24) is calculated. This value indicates the *Global Robustness* index.

For the Local Robustness, I perturb again the v_j fluxes, $j = 1, \ldots, n$, but in this case I create a sample trial perturbing a single flux, evaluate the property $\phi(\tau)$ and calculate the function ρ. For each flux I have a set T of trials and the robustness (Equation 2.24) is calculated. Therefore, in this case, we can obtain the LR value for each metabolic flux. In Pathway-oriented Robustness Analysis, since the reactions are clustered in many pathways, all the flows belonging to a pathway are perturbed and ρ is calculated. Such as in the previously GR and LR, a set T of trials is obtained and the Pathway-oriented Robustness Analysis is performed.

2.8.1 Normalized Feasible parameter Volume: the Glocal Analysis

BioCAD also implements the analysis described in the work of Andreas Wagner et al. [37]. In the *Glocal Analysis*, the authors implement a procedure that calculates the volume occupied by those parameters such that the system maintains the desired characteristics. The volume is computed in the n-dimensional parameter space. In our case, the volume is such that Equation 2.23 holds. Since this research requires a huge computational effort, given the high number of dimensions (for FBA models, R^n, where n is the number of parameters), it is guided by an iterative procedure that involves the Principal Component Analysis (PCA). In the second part, they calculate local coefficients, and from these they derive which parameters are influential on the robustness (by Spearman's partial correlation coefficient).

In particular, the first part requires two steps. The first is a Monte Carlo sampling obtained with n-dimensional Gaussian random variations centered around a parameter

vector (known in advance). In this case this vector is represented by the n parameters: v_j. Then a set $T_\tau^{(1)}$, $n \times K$ is created, that contains K parameter vectors. Among these, only a fraction will satisfy the Equation 2.23, the set comprising this fraction is the set of the feasible parameter vectors $V^{(1)}$. The second step begins with a principal component analysis of $V^{(1)}$; this analysis allows to identify statistical linear structures within high-dimension data sets. Here, instead, it is used to guide the sampling of the parameter vectors in subsequent iterations. In particular, $T_\tau^{(2)}$ and the subsequent sets $T_\tau^{(h)}$ are generated from $V^{(1)}$ and, in general, from $V^{(h-1)}$, where $h = 1, \ldots, H$, and H is the number of iterations. In particular the generic element $\tau_{j,k}$ of $T_\tau^{(h)}$ is generated as:

$$\tau_{j,k} = \frac{\sum_{t^*=1}^{T^*} V_{j,t^*}^{(h-1)}}{|T^*|} + \lambda^{(h-1)} \cdot \xi_{j,k}, \tag{2.25}$$

where $j = 1, \ldots, n$, since the columns of $T_\tau^{(h)}$ contain the perturbed values of the fluxes v_j, $j = 1, \ldots, n$; $k = 1, \ldots, K$ is the cardinality of $T_\tau^{(h)}$; the first term, on the right side, is the average of the elements for each perturbed parameter (that is the average for each row) of the set $V^{(h-1)}$ obtained in the previous iteration; $\xi_{j,k}$ is a Gaussian noise with zero mean and standard deviation equals to the $(j,k)^{th}-$element of the covariance matrix $\Sigma^{(h-1)}$, i.e., the pair wise covariance calculated for all vectors τ_a and τ_b of $V^{(h-1)}$ (the eigenvectors of this matrix are the principal axes of the $V^{(h-1)}$ set by PCA); finally, the real value $\lambda^{(h-1)}$ guides the h^{th} Gaussian process by scaling the standard deviations of the distribution along the PCA directions. The purpose of Equation 2.25 is to avoid unnecessary sampling in a parameter space region where there are no probably feasible vectors. At the end of this procedure, a hyper-box B is constructed in the parameter space, whose axes are parallel to the PCA axes of the last iteration. The bounds of this box, for each direction, are given by the more extreme elements in the set V^H of the last iteration. Then B is uniformly sampled constructing the final set T_τ; a subset V of T_τ will verify the Equation 2.23. Finally, the feasible parameter volume will be calculated as $R^n = (|V| \setminus |T_\tau|) * Vol(B)$, where $|.|$ determines the cardinality. The logic of this measure is that as the value of R^n increases, the likelihood to generate another feasible robust parameters vector increases. Finally, for comparing systems with different number of parameters the *normalized feasible parameter volume* R is defined as $R = \sqrt[n]{R^n}$. R can be considered as the *permissible average variation per-parameter* that leaves intact the system performance.

The second part of this analysis is connected to the global part. The authors take into account the final set of the feasible parameter vectors V and for each parameter vector produces Q sample trials perturbing the n parameters by Gaussian noise with zero mean and standard deviation equal to 0.2; then, they calculate the fraction of robust trials;

after that, they repeat the calculations for all vectors. Finally, for the n-parameters, they calculate the Spearman partial correlation coefficient with respect to the robust trial fraction values and the different values assumed by the parameters $\delta_j(V(j), X)$, where $j = 1, \ldots, n$; $V(j)$ is the j^{th} row of V (containing the observations of the j^{th}-parameter) and X is a vector (containing the values of the robust trial fractions).

Chapter 3

Robust Design of Microbial Strains

In this Chapter, I will report the work titled "Robust Design of Microbial Strains" [15], where BioCAD was applied for the *Genetic Design* in *Escherichia coli* bacteria. Genetic Design stands for searching for the best genetic manipulations in terms of knockout. The aim is to find the optimal knockout strategy able to outperform the production of specific chemical and biochemicals from the bacterium, as well as identifying novel and non-native synthesis pathways. By engineering bacteria, we can obtain substances suitable for industrial or biotechnology purposes, such as biofuel or vitamins and drugs.

Metabolic engineering is becoming central in basic and applied biological fields and requires mathematical models for accurate design purposes. Many organisms are used to analyze the metabolite production potential and identify the metabolic interventions needed to produce the metabolite of interest. Thus, strains have been systematically designed through in silico analysis to overproduce target metabolites, such as lycopene [43], ethanol [44], isobutanol [45]. The efforts are particularly focused on predicting flux distributions and network capabilities, most notably Flux Balance Analysis (FBA) [2]. Recent FBA models incorporate also information on enzymes and genome, integrating the relationships among genes, enzymes and reactions. This makes it well suited to studies that characterize many different perturbations such as different substrates or genetic manipulations (knockouts). By using computational metabolic engineering methods, it is possible to explore the reaction network and search for the solutions that satisfy the objectives.

In the past years, a variety of methods has been implemented to search for the genetic manipulations that optimize a cellular function of interest. These methods, such as OptKnock [12], OptFlux [14], OptGene [13] and GDLS [11], have been tested in FBA

organism models. However, all these methods require high computational efforts: the execution times grow exponentially [12–14] or linearly [11] as the number of manipulations allowed in the final designs increases. Because of the large number of reactions occurring in the cellular metabolism, the dimension of the solution space is very large and finding genetic manipulations is quite expensive.

In this work, a multi-objective optimization algorithm has been implemented to seek the genetic manipulations that optimize multiple cellular functions. The algorithm implements a global search with a heuristic and combinatorial method called *Genetic Design through Multi-objective Optimization* (GDMO). The idea is to use and improve the Pareto optimal solutions. Pareto optimality is important to obtain not only a wide range of Pareto optimal solutions, but also the *best trade-off design*, as reported in [46] for the protein structure prediction problem. Moreover, the multi-objective optimization turns out to aid in the automatic design in several biological problems [30].

The area underlying the Pareto curve and the first derivative, and in particular the presence of jumps (i.e., quick variations in the objective functions during the optimization procedure), carry valuable biotechnological information. For the first time, the ϵ-*dominance analysis* has been used in systems biology so as to consider all the solutions obtained by GDMO that are dominated with a tolerance ϵ by the Pareto-optimal solutions. Multi-objective optimization provides more insights than single-objective optimization on the capability of these organisms to adapt to the simultaneous presence of different conditions and constraints. By combining multiple-target optimization with knockout parameter space we are able to investigate the most complete available metabolic data and search for the optimal nutrients in strains that allow the maximization or minimization of metabolic targets.

Additionally, I relate pathways to Sensitivity Analysis. In modeling, Sensitivity Analysis (SA) is a method used to discover the main inputs, that is the inputs that have a substantial influence on the outputs of the model. In the last years, SA indices have been adopted in systems biology interrogating the reactions space (RoSA - Reactions oriented Sensitivity Analysis) [30] and species space (SoSA - Species oriented Sensitivity Analysis) to find their influence on the outputs of the system [31]. In this work, I perform SA to find the most sensitive pathways in the FBA model of *E. coli*. In particular, I present the novel Gene sets Pathway-oriented Sensitivity Analysis (Gene sets-PoSA), to find the genetic manipulations that have the largest influence on the output of the model. Unlike other SA methods applied in biological modeling, whose inputs (reactions or species) are valued in a *real* region of interest, Gene sets-PoSA is applied when inputs are valued in a *binary* region of interest. Each input of the model is represented through a set of binary variables whose values can assume only two values: 0 or 1. Gene sets-PoSA

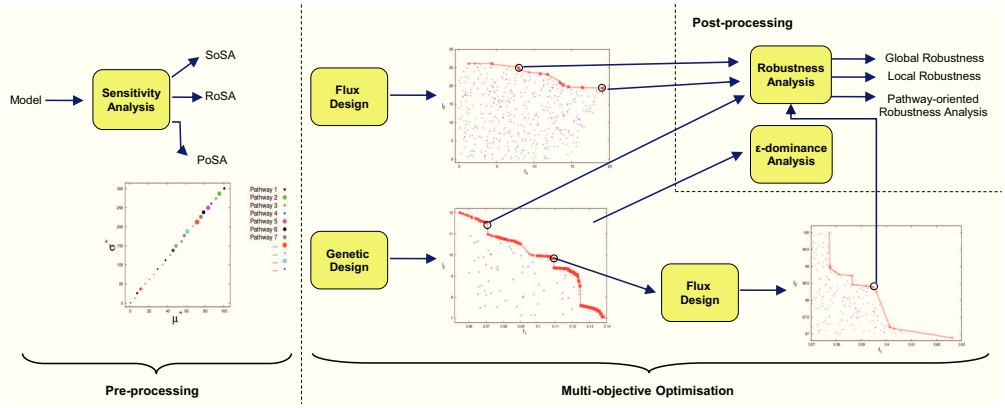


FIGURE 3.1: A schematic representation of the automatic framework for optimal bacterial metabolism. In the pre-processing step the Species (SoSA), Reaction (RoSA) and Pathway-oriented Sensitivity Analysis (PoSA) are applied to the metabolic model. Then, the Multi-objective Optimisation allows genetic and flux design. In the post-processing step, suitable solutions (selected from the Pareto front) are subjected to Global, Local and Pathway-oriented Robustness Analysis. The ϵ-dominance Analysis is performed to investigate the neighbourhood of the suitable genetic designs.

investigates the knockout solution space and determines the influence of the pathways on the outputs of an FBA model. Since this search-and-optimize algorithm provides a set of feasible solutions with different genetic manipulations, it is worth seeking a relationship between the sensitivity indices and the proposed manipulations. This way, we are able to select only the best manipulations. In particular, thanks to the information provided by Gene sets-PoSA, we can choose the GDMO knockout strategies that affect genes belonging to *insensitive pathways*.

Each point of the Pareto front represents a strain, i.e., an *E. coli* with specific genetic manipulations, and it is also associated with three Robustness Analysis (RA) indices that BioCAD computes. The *Robustness* estimates how much is robust a strain obtained by GDMO when it undergoes small perturbations, which can be *external* (changes in the nutrients) or *internal* (changes in the metabolism). Among the strains proposed by GDMO, we are able to choose the most robust one. In particular, BioCAD implements three robustness methods to evaluate different components of the model. For more details on BioCAD structure and Multi-objective optimization, definitions, Sensitivity and other methods please see the previously Chapter 2.

3.1 BioCAD for the Metabolic Engineering of Robust strains

In Figure 3.1 it's shown the layout of BioCAD computational framework applied to organisms modeled through FBA. The framework is composed of three blocks. The first is constituted by the Sensitivity Analysis, able to find the most sensitive parameters of

the model. We can investigate the reactions and the species in the metabolic model in terms of sensitivity, using the RoSA and PoSA methods. Furthermore, the novel Gene sets-PoSA is able to identify the most sensitive metabolic pathways by ranking them according to knockouts. We can consider a single pathway as an input of the system. Each pathway (that is a set of reactions converting particular substrates in specific final products) is perturbed by mutating genes that control its biochemical reactions. PoSA ranks the pathways according to their influence on the outputs of the model. Pathways with important influence have large sensitivity index (μ^* and σ^*), as reported in the pre-processing part of Figure 3.1. Each pathway in the graph is represented by a circle and its size indicates the number of genes belonging to it.

The multi-objective optimization algorithm searches both for the genetic manipulations (through gene deletions) and for nutrients with respect to defined target functions. In this Chapter, I will show both the genetic design and the flux design in microbial strains. The result of the multi-objective optimization is a set of non-dominated points, called Pareto front (or Pareto surface). The non-dominated points are shown in red in Figure 3.1, while all the dominated points are shown in blue. All the dominated points and the non-dominated points, that satisfy all of the inequality and equality constraints, and all of the variable bounds, constitute the observed feasible region.

In the genetic design, each strain (a particular phenotype) is identified by a binary y "knockout vector" (which represents the genotype), whose elements are 1 when the corresponding gene set is turned off. The importance of the knocked out genes can be evaluated by means of the ranking provided by Gene sets-PoSA. A gene set can be composed of a single gene, when it synthesizes for an enzyme, or can be associated with more genes, that synthesize for enzymes to form enzymatic complexes and enzymatic subunits. The relation between genes in a gene set is regulated by means of a Boolean relationship. When all the genes are necessary to catalyze the corresponding reactions (a single gene set can regulate more reactions), genes are linked by the "AND" operator; otherwise, if at least a gene is necessary, genes are linked by the "OR" operator. Details about the genetic modeling in FBA can be found in section 2.2.3 in Chapter 2. In addition, through the multi-objective optimization we are also able to find the favorable nutrients set (Flux Design) to optimize the wild-type/strains yield and evaluate the over/under investment of nutrients (uptake rate of fluxes). After the optimization, we can perform the ϵ-dominance analysis to search accurately near the edge of the Pareto-optimal region.

The Robustness Analysis is the third task of BioCAD computational framework. For each phenotype (strain or wild type), in a post-processing step BioCAD processes the fragility of the metabolic network when it is subjected to small perturbations, which can

be exogen or endogen. From the Pareto fronts we can select interesting solutions using decision-making methods: for instance, solutions near to the ideal solution, the knee points, the end points, or points with suitable features. For each solution, I calculate the Global Robustness (GR), the Local Robustness (LR) and the Pathway-oriented Robustness (PoRA) values, indicating respectively the robustness of the whole network, of each single reaction and of each metabolic pathway.

Here, BioCAD is tested on the genome-metabolic network of *E. coli* *i*AF1260 [47], composed of 2382 reactions, in order to maximize one or more metabolites of interest and simultaneously ensure the biomass formation, with the minimum knockout cost. The knockout cost is defined according to the Boolean relationship between genes. For example, if a gene set is composed of two genes linked by "AND", the cost to ensure the catalysis of the corresponding reactions is 2, since both genes are necessary to turn on the reactions associated with that gene set. Instead, the cost to ensure the turning off of the corresponding reactions (knockout cost) is 1.

3.1.1 Implementation

PoSA, GDMO and the Robustness Analysis are implemented in MATLAB and GLPK (GNU Linear Programming Kit). Here, have been illustrated the capabilities of GDMO by applying it to several overproduction problems in *i*AF1260 *E. coli* [47]. In a pre-processing analysis, a reduction of the FBA network has been performed to remove duplicate and dead-end reactions as described in [48, 49]. After the reduction, the resulting metabolic network is mathematically identical to the original network. In addition, the reduction in terms of reactions, metabolites and genes improves the numerical stability. Initially, in *i*AF1260, there are n=2382 reactions, m=1668 metabolites and L=913 gene sets; after the reduction, the network changes in n=959, m=483 and L=632 in anaerobic conditions, and n=1019, m=506 and L=663 in aerobic conditions. In particular, for acetate and succinate production, I carried out experiments in both anaerobic and aerobic conditions, with 10 and 5 $\text{mmolh}^{-1}\text{gDW}^{-1}$ of available glucose (GLC).

3.2 Results

Taking into account that each gene in the *i*AF1260 *E. coli* is assigned to at least one of the 36 different pathways, (e.g., all the genes involved in Krebs cycle, in Pentose Phosphate Pathway, and so on), PoSA (from here referred to Gene-sets PoSA) evaluates the importance of a pathway on the basis of the knockouts that are involved in its

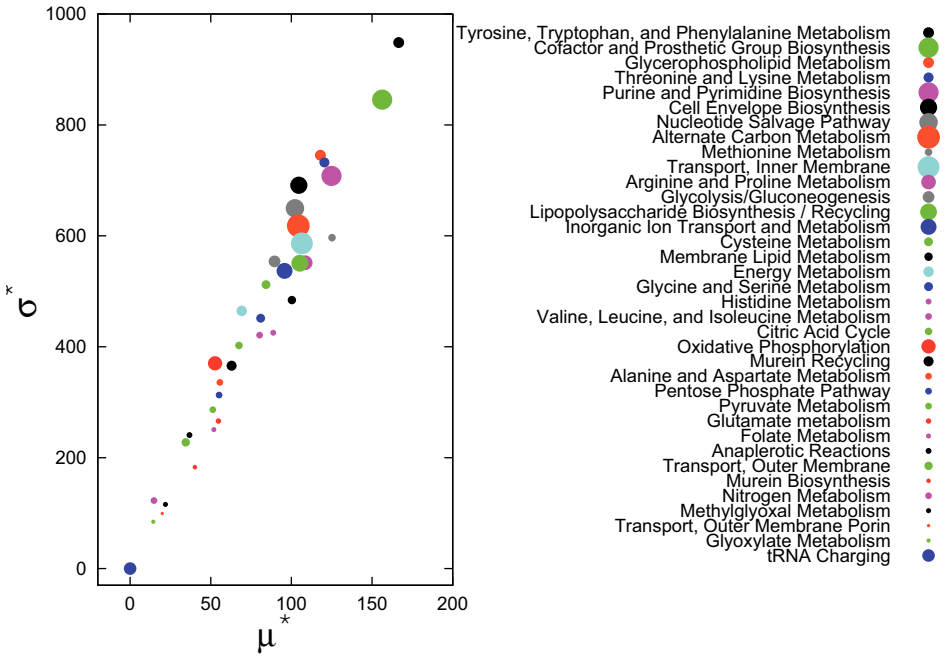


FIGURE 3.2: Pathway-oriented Sensitivity Analysis (PoSA) for the model of *E. coli* *i*AF1260. In y and x axes I report the sensitive indexes found through PoSA. The higher are indexes, higher is the sensitivity associated to the parameter. The model is composed of 36 pathways whose reactions and genes are clustered according to the functionality of the pathway they belong. The size of a sign is proportional to the number of genes involved in the pathway.

metabolism and indicates a ranking of the metabolic pathways in the (μ^*,σ^*) space reported in Figure 3.2.

Moreover, the study of the variance-to-mean ratio (VMR) is a good measure of the degree of randomness of a given phenomenon. In the *E. coli* analysis, PoSA μ^* and σ^* indices are linked with a linear relationship and the VMR is larger than 1; thus, the elementary effects set is said to be over-dispersed, highlighting the presence of great variability. We can deduce that the elementary effects of the 36 pathways are sampled from a negative binomial distribution. The VMR is linked to the Pareto front and can be harnessed to explore the solution space, since it describes the probability distribution of the phenomenon.

In general, highly networked cell components (such as those for nucleic acids, amino acid, cofactors and energetic metabolism) are in top right corner, while specific, often single reaction very abundant components (such as those for bacterial walls, nitrogen, glutamic and carbohydrates) are in a bottom left position. tRNA changing pathway results to have a sensitivity equal to zero. This is because it is not associated to any gene in the model. By exploiting clusters positioning of pathways we can deduce some metabolic relevancy. Although Glycolisis/Gluconeogenesis should be considered the central core of bacterial biomass and energetic storage, five of these pathways are localized in the

same cluster: 1) Alanine and Aspartate Metabolism, 2) Pentose Phosphate Pathway, 3) Glutamate Metabolism, 4) Folate Metabolism, and 5) Pyruvate Metabolism (Figure 3.2). All these pathways are conceptually connected to Citric Acid Cycle and to Purine and Pyrimidine Biosynthesis, which are two of the pathways occupying the top of the (μ^*,σ^*) space. PoSA automatically generates biological information locating these pathways in same area, and indicating them as components of an unique metabolic mechanism. These results are also confirmed by literature, since there are biochemical evidences that Pyruvate and Folate metabolism must be considered strictly connected with the Purine metabolism. Moreover, the Pyrimidine metabolism becomes the principal collector of Alanine and Aspartate Metabolism. Results in Figure 3.2 were obtained by evaluating more 3 million calls to the function F of Equation 2.8.

After SA, GDMO algorithm was initialized by setting the *E. coli* network with an empty set of knockout, i.e., in wild type configuration, and setting the population size $I = 1000$ and the number of generations $D = 1500$. Table 3.3 reports the best solutions in terms of acetate and succinate obtained by the previous methods, along with GDMO proposed solutions extracted from the Pareto front. Details of the genetic strategies, the knockout cost, the genes, reactions names are reported in Table 3.1 and Table 3.2.

The multi-objective optimization method is used to maximize the production of acetate (or succinate) and biomass, with the smaller knockout cost. Table 3.3 also reports the robustness indices for the wild type organism and strains. The effect of the knockouts on the robustness of the network can be noticed by comparing the GR, LR and PORA values of strains with those of the wild type. The values of GR and LR are of the same order of magnitude probably because the robustness of the network is strongly linked to the glucose uptake rate.

As described in previous sections, the result of a multi-objective optimization problem is a set of optimal points that form the Pareto front. For succinate and acetate maximization, I have conducted several experiments by setting different initial conditions. Experiments have been conducted in anaerobic and aerobic conditions and with different glucose feed. Pareto fronts obtained by GDMO are shown in Figure 3.3. We can also see that for each Pareto front, the wild type point corresponds with the right end point. Indeed, the deletion of a gene in a metabolic circuit leads to a decrement of biomass formation. The maximization of a synthetic objective is strongly in conflict with the grown rate, therefore a multi-objective optimization technique is suitable to solve this kind of problem. Pareto optimality is also suitable to evaluate the behavior of the organism when it is in different environment conditions. In Figure 3.3-A-B we can see how the area under the front grows when oxygen and glucose increase. Additionally, the number of points gives an idea of the capabilities of an organism to produce a specific

TABLE 3.1: Knockout strategies obtained through GDMO for maximizing acetate production [mmolh^{-1} gDW^{-1}] in *i*AF1260 *E. coli*. For each strategy (Str), I have the biomass (Biom) formation [h^{-1}], the knockout cost (kcost) and the genes and reactions switched off. The variation of acetate (Ac) and biomass in comparison with the wild type is enclosed in brackets. In Table is provided a comparison between the robustness analysis methods. The Glocal R value [37] and GR value are the global robustness indexes. The strain is more robust when R and GR are high values. For PoRA and LR, it's reported the minimum values found, which are respectively associated with the less robust flux (glucose uptake rate) and the less robust pathway (energy metabolism). *In the table, due to space limitation, I report the the enzymatic complex 1) "(FdhF AND Hyd4) OR (FdhF AND HycB)" associated with the gene set (b4079 AND (b2481 AND b2482 AND b2483 AND b2484 AND b2485 AND b2486 AND b2487 AND b2488 AND b2489 AND b2490) OR (b4079 AND (b2719 AND b2720 AND b2721 AND b2722 AND b2723 AND b2724))) and the protein 2) "Nuo" associated with the gene set (b2276 AND b2277 AND b2278 AND b2279 AND b2280 AND b2281 AND b2282 AND b2283 AND b2284 AND b2285 AND b2286 AND b2287 AND b2288).

Str	Ac	Biom	kcost	GR	LR	PoRA	R	Genes	Reactions
A_1	13.791	0.130	3	45.32%	39.33%	81.33%	0.78	(b0351) OR (b1241)	acetaldehyde dehydrogenase (acetylating)
	(66.13%)	(-43.72%)						(b1539)	L-allo-threonine dehydrogenase
									D-serine dehydrogenase
									L-serine dehydrogenase
A_2	19.150	0.053	10	27.60%	24.00%	43.33%	0.44	(b0351) OR (b1241)	acetaldehyde dehydrogenase (acetylating)
	(130.7%)	(-77.10%)						(b3945)	aldose reductase (acetol)
									Glycerol dehydrogenase
									D-Lactaldehyde:NAD+ 1-oxidoreductase
								(b4381)	deoxyribose-phosphate aldolase
								(FdhF AND Hyd4) OR (FdhF AND HycB)*	Formate-hydrogen lyase
								(b3617)	glycine C-acetyltransferase
								(b1380) OR (b2133)	D-lactate dehydrogenase
								(b3236)	malate dehydrogenase
A_3	18.532	0.096	9	40.72%	35.33%	72%	1.27	(b0351) OR (b1241)	acetaldehyde dehydrogenase (acetylating)
	(123.2%)	(-58.6%)						(b0910)	cytidylate kinase (CMP)
									cytidylate kinase (dCMP)
								(b2975) OR (b3603)	D-lactate transport via proton symport
									glycolate transport via proton symport
									L-lactate transport via proton symport
								(b4381)	deoxyribose-phosphate aldolase
								(b3617)	glycine C-acetyltransferase
								(b0963)	methylglyoxal synthase
								Nuo*	NADH dehydrogenase
A_4	14.046	0.104	5	41.52%	36.0%	75.33%	1.74	(b0351) OR (b1241)	acetaldehyde dehydrogenase (acetylating)
	(69.20%)	(-55.14%)						(b3617)	glycine C-acetyltransferase
								(b4025)	glucose-6-phosphate isomerase
								(b3708)	Tryptophanase (L-tryptophan)

metabolite. For example, *E. coli* seems more able to produce acetate than succinate. For succinate maximization (Figure 3.3-B), Pareto points are less than acetate points. It's also important to evaluate the presence of essential genes, i.e., genes that lead the biomass formation to zero. By analyzing the variables space and the knocked out gene sets, BioCAD also provides a list of (i) *essential genes*, also called destructive genes, of (ii) *neutral genes* and of (iii) *trade off genes*. Neutral genes include all that genes that do not improve any objective functions, instead trade off genes include all genes that improves at least one objective function (for more details see the section 3.5).

In order to compare GDMO performance with previous methods, I have considered GDLS [11], OptFlux [13], OptGene [14], OptKnock [12] algorithms, where the goal is to

TABLE 3.2: Knockout strategies obtained through GDMO for maximizing succinate production [mmolh^{-1} gDW^{-1}] in *i*AF1260 *E. coli*. For each strategy (Str), I have the biomass (biom) formation [h^{-1}], the knockout cost (kcost) and the genes and reactions switched off. The variation of succinate (Succ) and biomass compared with the wild type is enclosed in brackets. I also provide a comparison between the robustness analysis methods. The Glocal R value [37] and GR value are the global robustness indexes. The strain is more robust when R and GR are high values. For PoRA and LR I report the minimum values found, which are respectively associated with the less robust flux (glucose uptake rate) and the less robust pathway (energy metabolism).

Str	Succ	Biom	kcost	GR	LR	PoRA	R	Genes	Reactions
B_1	12.012	0.055	15	44.60%	44.67%	84.67%	0.15	(b0351) OR (b1241)	acetaldehyde dehydrogenase (acetylating)
	(15476%)	(-76.33%)						(b2587)	2-oxoglutarate reversible transport via symport
								(b0870) OR (b2551)	D-alanine transaminase
									alanine transaminase
									L-allo-Threonine Aldolase
									Threonine aldolase
								(b1852)	glucose 6-phosphate dehydrogenase
								(b1849)	GAR transformylase-T
								(b1380) OR (b2133)	D-lactate dehydrogenase
								(b2463)	malic enzyme (NADP)
								(b0963)	methylglyoxal synthase
								(b4388)	phosphoserine phosphatase (L-serine)
								(b2661)	succinate-semialdehyde dehydrogenase (NADP)
								(b1602 AND b1603)	NAD(P) transhydrogenase (periplasm)
								(b3708)	Tryptophanase (L-tryptophan)
B_2	11.530	0.070	10	43.48%	42.0%	80.67%	0.92	(b0351) OR (b1241)	acetaldehyde dehydrogenase (acetylating)
	(14875%)	(-69.3%)						(b2587)	2-oxoglutarate transport via symport
								(b3945)	aldose reductase (acetol)
									Glycerol dehydrogenase
									D-Lactaldehyde:NAD+ 1-oxidoreductase
								(b1852)	glucose 6-phosphate dehydrogenase
								(b1380) OR (b2133)	D-lactate dehydrogenase
								(b2463)	malic enzyme (NADP)
								(b2661)	succinate-semialdehyde dehydrogenase (NADP)
								(b1602 AND b1603)	NAD(P) transhydrogenase
B_3	10.610	0.087	8	40.40%	46.0%	83.33%	1.32	(b0351) OR (b1241)	acetaldehyde dehydrogenase (acetylating)
	(13659%)	(-62%)						(b3945)	aldose reductase (acetol)
									Glycerol dehydrogenase
									D-Lactaldehyde:NAD+ 1-oxidoreductase
								(b1380) OR (b2133)	D-lactate dehydrogenase
								(b2463)	malic enzyme (NADP)
								(b0767)	6-phosphogluconolactonase
								(b1602 AND b1603)	NAD(P) transhydrogenase
B_4	9.037	0.123	3	44.64%	44.0%	84.0%	1.25	(b0356) OR (b1241)	alcohol dehydrogenase (ethanol)
	(11619%)	(-46.7%)						OR (b1478)	

optimize succinate and acetate productions in the *i*AF1260 *E. coli* metabolic network. Results are reported in Table 3.3.

In Figure 3.5 we can show the comparison between the results obtained by the method proposed by Lun et al. [11], and the Pareto solutions for optimizing acetate and succinate production. The solutions provided by GDLS do not outperform Pareto fronts, since they occupy positions in the area under the Pareto curves. In the best cases, they lie on the Pareto fronts. Other experiments are reported in Appendix A.

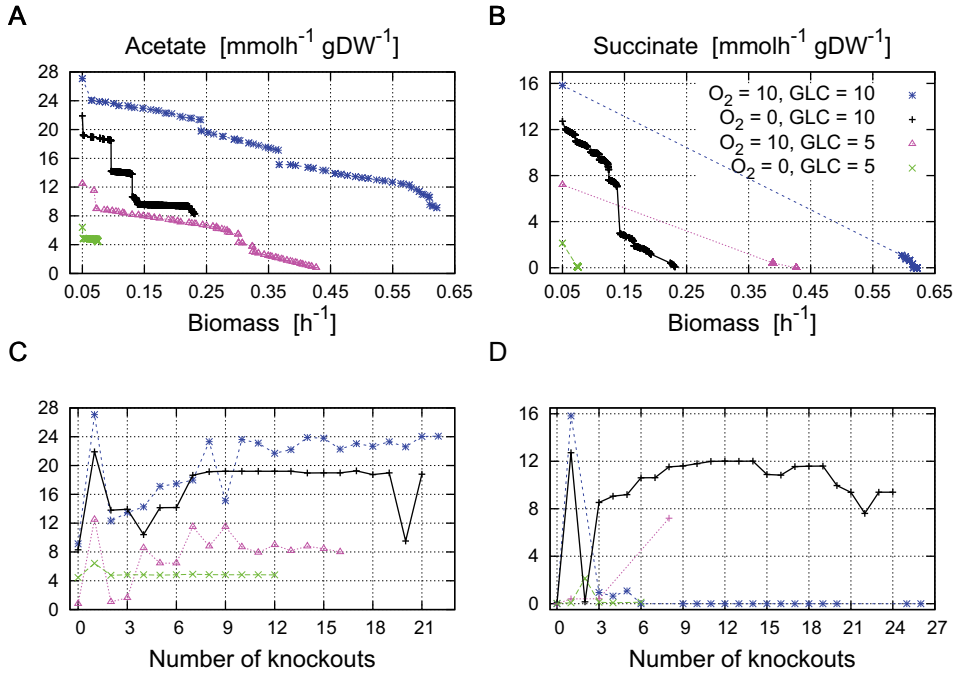


FIGURE 3.3: In A and B, respectively the Pareto fronts for the maximization of the biomass [h^{-1}] and the acetate production rate and succinate production rate [mmolh^{-1} gdW^{-1}] in the recent genome-scale model of *E. coli*, *i*AF1260, in four different environmental conditions: aerobic condition (O_2 = 10 mmolh^{-1} gDW^{-1}) with glucose uptake rate 10 mmolh^{-1} gDW^{-1} (blue signs), anaerobic condition with glucose uptake rate 10 mmolh^{-1} gDW^{-1} (black signs), aerobic condition (O_2 = 10 mmolh^{-1} gDW^{-1}) with glucose uptake rate 5 mmolh^{-1} gDW^{-1} (purple signs) and anaerobic condition with glucose uptake rate 5 mmolh^{-1} gDW^{-1} (green signs). Figures C and D refer to A and B respectively, and show the number of knockouts associated with acetate and succinate production rates.

In addition, the ϵ-dominance analysis reveals other interesting points. For instance, I have found 14.05 mmolh^{-1} gDW^{-1} of acetate with a knockout cost equal to 5, and 9.26 mmolh^{-1} gDW^{-1} of succinate (biomass = 0.096 h^{-1}) with a knockout cost equal to 4. Such as we can see in Figure 3.4-D, the Pareto point in purple of succinate (9.94 mmolh^{-1} gDW^{-1} with a biomass equal to 0.1 h^{-1}) has a greater knockout cost corresponding to 12. In this case, ϵ-dominance analysis was useful to explore the solutions and variables space highlighting better solutions in terms of knockout. ϵ-succinate level is less than Pareto point succinate level, but the variables space is clearly more suitable (ϵ-dominance point in red).

3.2.1 Flux Design in *i*AF1260 *E. coli*

In order to study the favorable environmental conditions (flux design), i.e., nutrients for *E. coli*, I performed the simultaneous optimization of acetate, succinate and biomass on the complete circuit, i.e., without knockouts. I considered the anaerobic and aerobic condition (O_2 uptake rate = 10 $mmolh^{-1}gDW^{-1}$) and maintained fixed the glucose

TABLE 3.3: Comparison between GDMO and previous genetic design methods: OptFlux [13], OptGene [14], GDLS [11], OptKnock [12] to maximize acetate (Ac) and succinate (Succ) production [mmolh^{-1} gdW^{-1}]. The third and fourth rows show the biomass (biom) [h^{-1}] and the knockout cost (kc). The last four rows show a comparison between the robustness analysis methods: Global (GR), Local (LR), Glocal (R) and Pathway-oriented (PoRA) Robustness analysis. The two values reported for wild type are referred respectively to the productions of Ac and Succ. For PoRA and LR, I report the minimum value found, which is associated with the less robust flux (glucose uptake rate) and the less robust pathway (energy metabolism). "n.a." stands for *not applicable*.

	Wild type	**OptFlux**	**OptGene**	**GDLS**	**GDLS**	**OptKnock**	**OptKnock**	**GDMO**	**GDMO**	**GDMO**
Ac	8.30	15.129	15.138	15.914	n.a.	n.a.	12.565	13.791	19.150	n.a.
		(+82.3%)	(+82.4%)	(+91.7%)	n.a.	n.a.	(+51.4%)	(+66.13%)	(+130.7%)	n.a.
Succ	0.077	10.007	9.874	n.a.	9.727	9.069	n.a.	n.a.	n.a.	10.610
		(+12877%)	(+12704%)	n.a.	(+12514%)	(+12362%)	n.a.	n.a.	n.a.	(+13659%)
Biom	0.23	n.a.	n.a.	0.0500	0.0500	0.1181	0.1165	0.130	0.053	0.087
		n.a.	n.a.	(-78.4%)	(-78.4%)	(-77.9%)	(-49.6%)	(-43.72%)	(-77.10%)	(-62%)
kc	n.a.	n.a.	n.a.	14	26	54	53	3	10	8
GR (%)	54.76/53.68	n.a.	n.a.	13.76	16.6	43.24	43.08	45.32	27.6	40.40
LR (%)	54.0/54.67	n.a.	n.a.	16.0	21.33	40.0	40.60	39.33	24.0	46.0
R	1.30/1.34	n.a.	n.a.	1.45	1.45	1.18	1.02	0.78	0.44	1.32
PoRA (%)	100.0/99.33	n.a.	n.a.	19.33	28.67	87.33	76.67	81.33	43.33	83.33

uptake rate at 10 $mmolh^{-1}gDW^{-1}$. I used NSGA-II [25] to perform the optimization by exploring the continuous space of exchange fluxes. For this design, I perturbed the thermodynamics constrains $\overline{v}_{ex}^{L}$, where $\overline{v}_{ex}$ is vector of n_{ex} exchange fluxes. The decision variables are real values from 0 to -100 (0 when the uptake is not allowed, -100 when the potential uptake rate is 100 mmolh^{-1} gDW^{-1}). Only glucose and oxygen were kept constant. Setting the population size at 100, I ran NSGA-II for 500 generations. In Figure 3.6 it is shown the results of the optimization in aerobic and anaerobic conditions (Pareto fronts and feasible points). In anaerobic condition, I have found 100 mmolh^{-1} gDW^{-1} of acetate, 42.918 mmolh^{-1} gDW^{-1} of succinate and 3.6204 h^{-1} of biomass (the trade-off point). In this condition, I have noticed a significant increment in the L-Aspartate, Citrate, Lactose, Fumarate and Malate uptake rates. In aerobic condition, I have found 100 mmolh^{-1} gDW^{-1} of acetate, 21.889 mmolh^{-1} gDW^{-1} of succinate, 4.16 h^{-1} of biomass, and a significant increment in the L-Asparagine, 1, 4-alpha-D-glucan, Fe(III)dicitrate, 2-Oxoglutarate uptake rates. Here, I perturbed simultaneously almost all the exchange fluxes, but it is possible to select a smaller set of nutrients according to experimental requirements.

3.3 FBA using experimental conditions

The gene expression data provide several information on the activation of genes when the organism undergoes specific external stimuli. In a first approximation we may transform microarray data matrix in a Boolean matrix, where 0 represents the knockout condition for a gene, and 1 represents the activation. BioCAD framework is able to read gene expression data, transfer them to a metabolic model, and evaluate in silico the metabolic

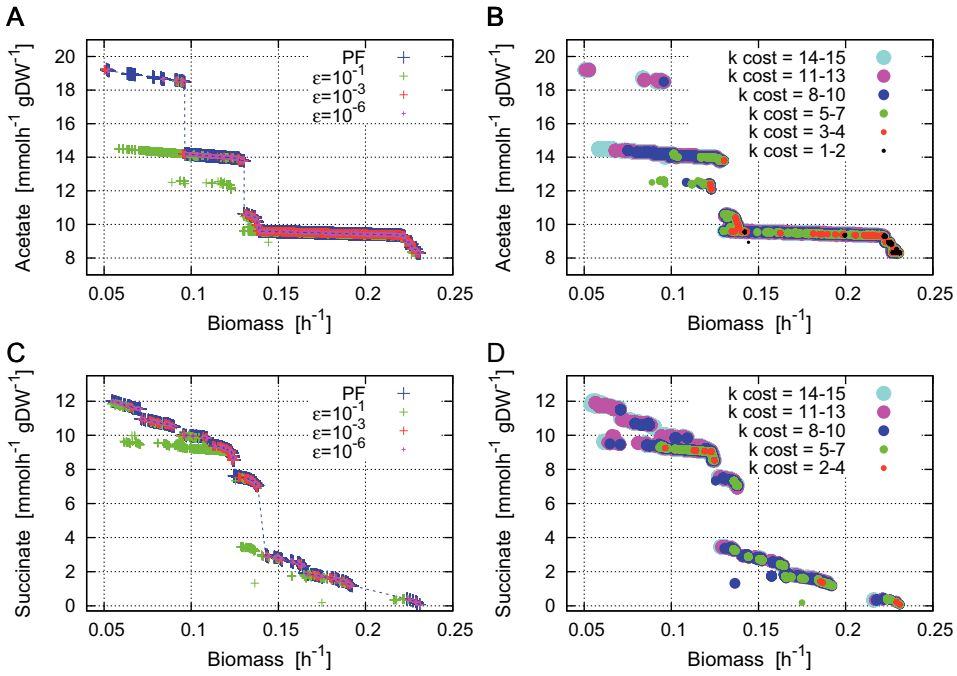


FIGURE 3.4: ϵ-dominance analysis results in *i*AF1260 *E. coli* network for acetate (A) and succinate (C) multi-objective optimization for different ϵ values. In blues Pareto points (PF). Figures (B) and (D) report the knockout cost associated with the solutions points. The size of the circle is proportional to the knockout cost of the solution.

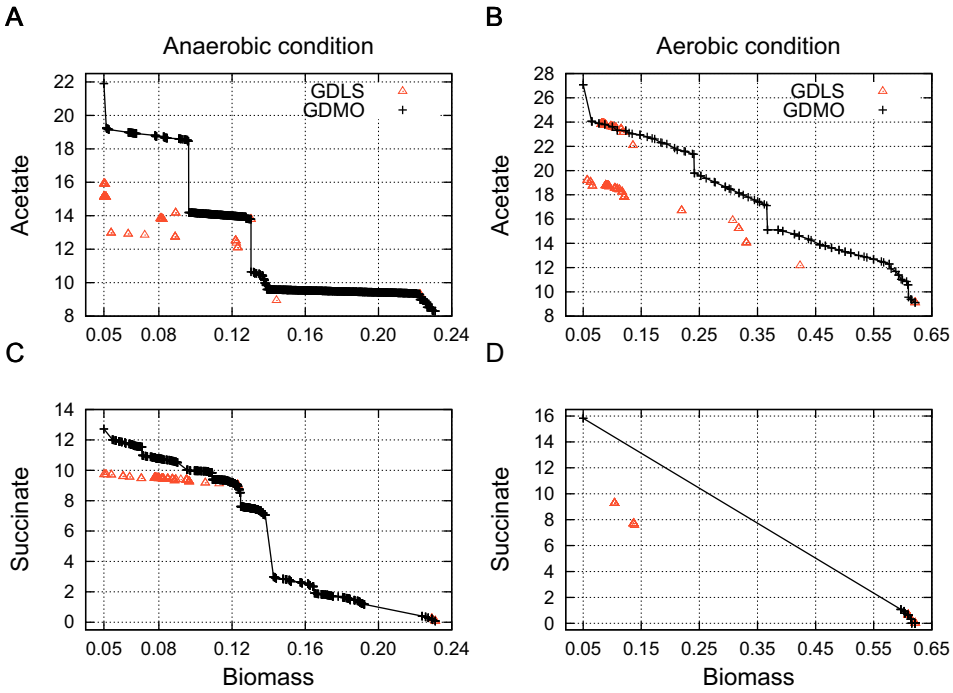


FIGURE 3.5: Maximization of biomass and acetate production in anaerobic (A) and aerobic conditions(B) and maximization of biomass and succinate production in anaerobic (C) and aerobic (D) conditions in *i*AF1260 *E. coli*. In black, Pareto fronts obtained by GDMO, in red GDLS [11] results.

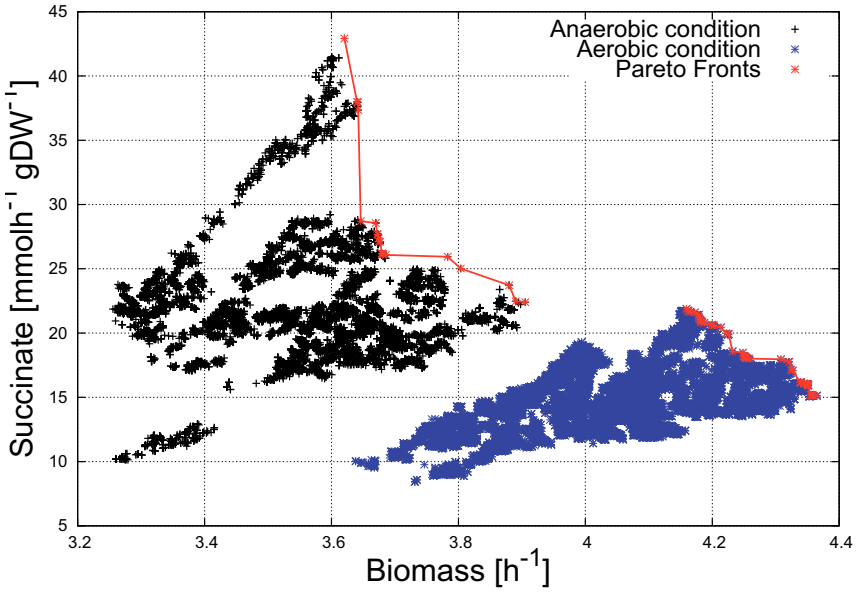


FIGURE 3.6: Three-objective optimization for maximizing acetate, succinate (y axis) and biomass (x axis). I considered the wild-type bacteria (i.e., knockout zero) and performed the maximization in aerobic (blue signs, O2 = 10 mmolh^{-1} gDW^{-1}) and anaerobic conditions (black signs) on a basis of 10 mmolh^{-1} gDW^{-1} glucose fed to identify favorable nutrients (input fluxes). The algorithm reaches the maximum production of acetate (100 mmolh^{-1} gDW^{-1}). In red, the Pareto fronts.

fluxes distribution using FBA. In this way, it is possible to investigate the behavior of an organism, as well as to compare different experimental conditions. It could be interesting, in future developments, extending the exploration analysis from a binary domain to a continuous domain, evaluating the gene expression in the metabolic network. In addition, through the optimization method, we can deduce how the growth of the organism improves in a given experimental condition, when additional genes are turned off (or on).

3.4 Quantitative and Qualitative Knockout Analysis

Pareto optimality gives information about the trend of organisms in their ability to produce particular metabolites, as reported in the previous sections. In addition, we can color the Pareto points in order to obtain a map of the knockout cost dispersion (Figure 3.4-B-D). In this way, each point characterizes both the phenotype of an organism (for instance, the amount of acetate and biomass) and the genotype (in terms of how many genes are knocked out). Nevertheless, it is also important to give a qualitative score for each knockout strategy. This can be calculated by means of the two sensitivity measures: μ^* (mean), σ^* (standard deviation) obtained from *Gene Sets*-PoSA (Figure 3.2). Large μ^* indicates high overall influence, high linear effect, while large σ^* indicates

that either the specific input is involved with other inputs, or its effect is nonlinear or non-additive. According to the μ^* sensitive index obtained by PoSA, I assign a quality score (QS) for each strain. Strains that have genetic manipulations involved in pathways with low μ^* values are preferred, and thus get a high score. The score ranges from 0 to 1; 0 when the genetic strategy involves gene sets linked to the pathway with the largest μ^* index, that is the most sensitive, and 1 when the genetic strategy involves gene sets linked to the pathway with the lowest μ^* index. The score is normalized by the square root of the number of samples, since manipulations involve different knocked out genes. Consequently, if I find two Pareto solutions through GDMO that have the same phenotype and different knockout manipulations, we are able to choose the best solution in terms of knockout, according to the QS calculated using PoSA. The greater is QS, better is the associated strain. For strains reported in Table 3.3, I obtained from left to right respectively a QS equal to 0.285, 0.063 and 0.223.

3.5 Inferring neutral, trade off and destructive strategies

By using the Pareto solutions obtained from the multi-objective optimization, a statistical analysis has been performed in order to cluster genetic strategies in three groups: (i) *neutral*, (ii) *trade-off* and (iii) *destructive strategies*. A genetic strategy can be considered *neutral* when the objective functions do not improve (in terms of increment or decrement, in the maximization or minimization problem respectively) with respect to the nominal value. Here, the nominal value is the wild type configuration of the metabolic network, i.e., when all genes are working.

The *trade-off genetic strategies* are knockout combinations that improve an objective function and disadvantage the others one. The *destructive genetic strategies* are those involving the essential genes, i.e., all the genes that are key to the biomass formation. In the knockout optimization, constructive genetic strategies do not exist, since a knockout cannot improve the wild type biomass. Conversely, when the decision variables are the uptake rates (nutrients optimization) I also have constructive solutions, since all the objective functions can be improved.

3.5.1 Case study 1: *i*AF1260 *E. coli* model.

The network of the *E. coli* model of Feist et al. [47] contains 1260 genes and 913 gene sets. In wild type, in anaerobic condition and with a glucose feed of 10 mmol h^{-1} gDW^{-1}, the bacterium grows at 0.231 h^{-1}. When I maximize the production of acetate

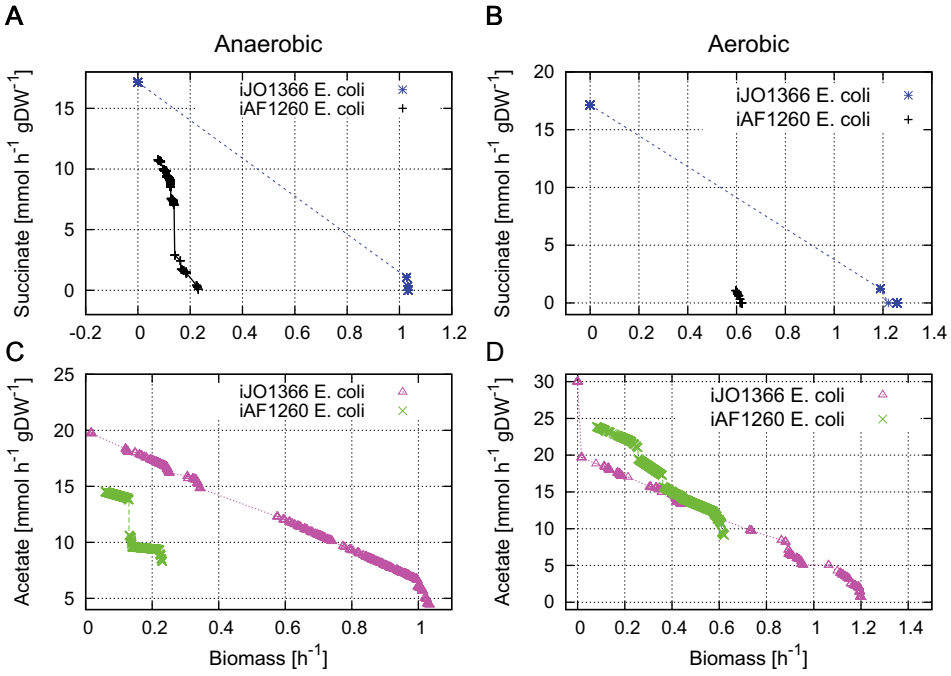


FIGURE 3.7: GDMO results for acetate and succinate maximization in different conditions by using the *i*AF1260 [47] and *i*JO1366 [50] *E. coli* genome-scale metabolic networks.

(or succinate) searching for best knockout strategy, the biomass rate always decreases, as shown in Figure 3.7-C, Pareto front in green.

The maximum level of acetate in anaerobic conditions, is 14.519 mmol h^{-1} gDW^{-1}, corresponding to the minimum value of biomass (0.057 h^{-1}). The corresponding *knockout cost*, namely the minimum number of genes to be turned off, is 19. In wild type, acetate production is 8.301 mmol h^{-1} gDW^{-1}.

By considering the point (0.1303; 13.7911) of the Pareto front in green of Figure 3.7-C that has a knockout cost 3 (YdfG; MhpF OR AdhE), I have found that 88 gene sets can be considered neutral genetic strategies, i.e., if I turn off these gene sets, the level of acetate and biomass do not change. For instance, Aas; rffT; AtoB; rfe; Acs; Lpd and SucA and SucB; Amn; and AdiA are just some of the neutral genetic strategies. Instead, (YdfG; MhpF OR AdhE) is the trade-off strategy. I have also discovered that the genes MraY and murG are essential (destructive strategies), i.e., turning them off causes a null biomass, and the organism dies. Moreover, I have found the gene tnaA is involved in 796 Pareto manipulations (out of 1000) when the acetate is maximized in anaerobic conditions. Moreover, this gene is also involved in most genetic manipulations in aerobic conditions (721 out of 1000 strategies).

As regards the succinate versus biomass optimization, the maximum synthetic production is equal to 10.757 mmol h^{-1} gDW^{-1}, with a biomass 0.076 h^{-1} and a knockout

cost equal to 15. A suitable solution can be obtained knocking out the gene set (isoenzymes) "AdhP OR AdhE OR FrmA", reaching 9.0373 mmol h^{-1} gDW^{-1} of succinate and 0.1231 h^{-1} of biomass.

An interesting result is that in anaerobic conditions the gene set "(SapD AND TrkA AND TrkH) OR (Kch) OR (SapD AND TrkA AND TrkG) OR (Kup)" is the most frequent both for maximizing acetate and for maximizing succinate. In Table 3.4, I report the most significant results.

TABLE 3.4: Frequency of gene sets in the knockout genetic strategies for acetate maximization and succinate maximization in the *i*AF1260 *E. coli* metabolic network [47]. The values reported in the second and third columns, between 0 and 1, represent the percentage of occurrences of the gene set (first column) in all the Pareto manipulations obtained with GDMO method.

Acetate	Anaerobic	Aerobic
tnaA	0.796	0.721
GuaB	0.728	0.310
SurE	0.568	0.077
Succinate	Anaerobic	Aerobic
Apt	0.950	0.510
DeoD	0.610	0.230
DeoD or DeoA	0.260	0.320

3.5.2 Case study 2: *i*JO1366 *E. coli* model.

The last FBA *E. coli* model was published in [50] by Palsson group in October 2011. The new metabolic network contains 1366 genes, 1041 gene sets, 2251 reactions and 1136 metabolites. In the wild type configuration, the organism grows at 1.033 h^{-1}.

By using GDMO algorithm, the maximum level of acetate in anaerobic conditions is 19.789 mmol h^{-1} gDW^{-1}, corresponding to the minimum value of biomass (0.016 h^{-1}). The knockout cost is 19. Instead, if we consider a trade-off between biomass and acetate, I suggest the solution with a knockout cost equal to 1 (Mdh, malate dehydrogenase) that reaches 16.209 mmol h^{-1} gDW^{-1} of acetate and 0.252 h^{-1} of biomass.

Although the organism modeled in the two flux balancing networks is the same and the environmental conditions are identical, the Pareto fronts are different. There are significant differences in *aerobic conditions*, especially for succinate. In wild type conditions, in *i*JO1366 *E. coli* network, the larger and more recent network, the succinate in the cytoplasm compartment is involved in 26 metabolic reactions, and in anaerobic conditions only five reactions are activated. In particular, succinate is produced by the reactions succinyl-diaminopimelate desuccinylase, O-succinylhomoserine lyase and Fumarate depended Dihydroorotate, and consumed by succinate dehydrogenase and succinyl-CoA synthetase. The succinate can be transferred in the periplasm and in the extracellular

space through ten transport fluxes. In anaerobic conditions, succinate is transported out via proton antiport. In this case, succinate production is equal to zero. Additionally, the search of knocked out genes for maximizing succinate production gives only four Pareto solutions (showed in Figure 3.7-A in blue): the maximum level of succinate is 17.142 mmol h^{-1} gDW^{-1}, but the biomass formation is null (the bacterium dies). So this solution cannot be considered biologically feasible. Another solution is equal to the wild type condition: succinate is zero, and biomass 1.033 h^{-1}. The third solution reaches 1.071 mmol h^{-1} gDW^{-1} of succinate and 1.027 h^{-1} of biomass, knocking out the gene set "FumA OR FumB OR FumC", linked to the reaction "Fumarase" of the Citric Acid Cycle. The forty solution reaches succinate equal to 0.34 mmol h^{-1} gDW^{-1} and biomass equal to 1.03 h^{-1} with 2 knockout.

Instead, for the *E. coli* network of Feist et al. [47], succinate in cytoplasm is involved in 22 reactions, and in anaerobic conditions five reactions are activated: fumarate reductase, succinyl-diaminopimelate desuccinylase, O-succinylhomoserine lyase, succinate transport out via proton antiport and succinyl-CoA synthetase. In this model, the reaction Fumarate depended Dihydroorotate is missing and the succinate production in the wild type condition is equal to 0.077 mmol h^{-1} gDW^{-1}. Unlike the *i*JO1366 *E. coli* network, the maximization of succinate and biomass produces a Pareto front with a high number of non-dominated solutions (see Figure 3.7-A in black). The two models represent the same organism, the environmental conditions are identical and ATP maintenance requirement flux is set for both networks to 8.39, but Pareto fronts depict apparently two different behaviors. The most recent network contains additional reactions with respect to the older model, and a glucose feed equal to 10 mmol h^{-1} gDW^{-1} is not sufficient for producing succinate; indeed, by incrementing glucose feed, the Pareto front contains more points and it is more dense. Unlike the acetate production, which depends mostly on the oxygen provision, the succinate production is sensitive to the glucose feed.

In this experimental protocol, I have obtained only one acceptable solution that maximizes succinate in anaerobic conditions. Analyzing all genetic strategies obtained by the front in purple in Figure 3.7-C, the gene set SerA results the most involved in the knockout manipulations (554 manipulations out of 1000). In the *i*AF1260 *E. coli* model, this gene has a frequency equal to 510 out of 1000. Additional details about the solutions are reported in Table A.1 in Appendix A.

3.6 Discussion

Pareto fronts provide significant information in metabolic design automation. The number of non-dominated solutions, the first derivative and the area under the curve are important markers for the best design within the same organism or between different organisms. Jumps correspond to sudden decreases in the availability of entire pathways and sub-networks when a crucial hub is eliminated, for instance the elimination of Krebs cycle or other key biosynthetic hubs. They result in decreasing the area under the Pareto front. The area under the Pareto front provides an estimate of number of intermediates which may be exploited for biotechnology purposes (optimization of an additional objective) or to build synthetic pathways (synthetic biology). Given two bacteria or two conditions for the same bacterium, the highest Pareto front would probably represent the best conditions for adding or optimizing pathways leading to new biotechnology products. Pareto optimality is useful to compare the ability of different organisms for optimizing specific metabolites (Figures 3.8 and Figure A.6 in Appendix A). I ran GDMO to compare the behavior of *E. coli*, *Y. pestis* [51], *G. sulfurreducens* [52] and *M. barkeri* models [53] for the maximization of acetate/succinate and biomass. For *Y. pestis* I analyzed its behavior in two different temperature conditions: in environment and in a human host.

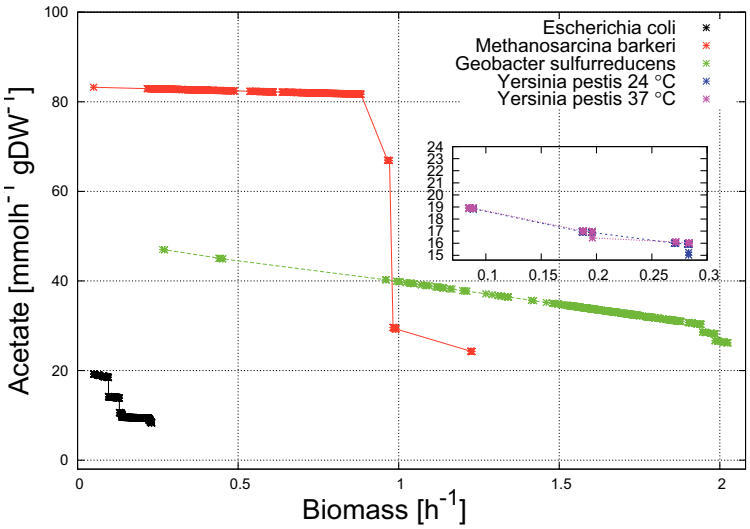


FIGURE 3.8: Pareto fronts obtained by optimizing the acetate production [mmolh^{-1} gDW^{-1}] and the biomass formation [h^{-1}] using GDMO in four organisms in the same environmental conditions : *E. coli* [47], *M. barkeri* [53], *G. sulfurreducens* [52] and *Y. Pestis* [51]. For *Y. Pestis* I consider two biomass compositions: at 24-28℃ and 37℃. The significance of these two temperatures stems from the two types of hosts that *Y. Pestis* infects in the natural environment, namely insect vectors at ambient temperature and mammalian hosts with regulated body temperatures of about 37℃.

Through the framework here discussed we can program bacteria in order to obtain desired outputs, thus framing them as living computers [54]. The goal is to provide a

simple tool to search and propose to the biotechnologist the best and suitable solutions in silico, so as to reproduce them in vivo. BioCAD framework investigates nutrients, reactions, metabolic pathways and knockouts for bacteria in an efficient automatic design. We are able to present several proposals, and indicate the best in terms of environmental conditions, knockout cost, robustness and sensitivity. Knockout strategies are useful in synthetic biology, while simulating the flux balance analysis in a particular experimental condition using the gene expression values is important for providing an optimization of bacteria in a given environment and biotechnological/medical condition.

The slope of the Pareto front reflects the progressive lack of pathways able to sustain the production of one component when we are optimizing the metabolism to maximize the other.

Therefore, the area, the slope and the position of jumps are meaningful parameters to characterize a Pareto front. It is noteworthy that the Pareto optimality could act as a parameter describing the improvement of a model for a bacterium with respect to a previous model for the same species. Indeed, the incompleteness of the model (number of reactions) may be identified by the reduced size of the Pareto front. The first derivative in the Pareto front, and in particular its discontinuity, indicates the preferable conditions for metabolite production. The multi-objective optimization method here presented provides a set of optimal candidate solutions, thus PoSA and RA allow us to choose the best genetic design.

GDMO scales effectively as the size of the metabolic system and the number of genetic manipulations increase. GDMO considerably outperforms the GDLS heuristic, OptFlux, OptGene and other heuristics, search methods, global and local optimization algorithms.

For sake of simplicity I have considered double target optimization. The methodology and the software implementation could be easily extended to optimize simultaneously several biotechnological targets. Optimizing two or more properties is of interest because all organisms experience oscillating conditions ranging from starvation to food richness. In particular, environmental changes present the availability of different sources of food. Therefore, the simultaneous optimization allows to evaluate the capability of the organism to cope with these changes.

Chapter 4

Multi-objective optimization for the production of 1,4-butanediol

In the last 20 years, using of microorganisms, microbes and algae for the synthesis of natural or synthetic substances is increased and revealed crucial for industrial, biotechnological and natural processes. The synthesis of pharmaceutical molecules is usually very costly, therefore their production by means of genetic and metabolic engineering of microorganisms could be cost effective [55]. Additionally, the decrease of oil reserves and the request of new sources of energy have moved the interest of many researchers to the study of biofuels production techniques and other petroleum-derived chemicals from biological systems. Many microorganisms have been used to produce alcohols such as ethanol and butanol. In the last years, the study on the optimization of production of specific substances has widely expanded. Moreover, through metabolic engineering, many strains have been systematically designed for producing natural and synthetic products. Synthetic products are obtained by inserting in an organism enzymes that naturally the organism does not own. The addition of new genes and enzymes allows the activation of new reactions and then new synthetic pathways. Many researcher groups have realized in vitro strains of *Escherichia coli* able to overproduce lycopene [56] and lactic acid [57]. Others have engineered *E. coli* strains through synthetic pathways to produce isobutanol [45]. In the recent work of Yim et al. [58], the authors have engineered the *E. coli* in order to produce 1,4-butanediol (BDO), a chemical compound industrially used and not produced naturally by any organisms. BDO is currently manufactured entirely from petroleum-based feedstocks such as acetylene, butane, propylene and butadiene.

In an in-silico step, strains are designed by computational techniques and successively

realized in vitro. Genome-scale metabolic networks are useful to investigate the capability of organisms, since they include information about metabolite interactions, reactions, genes and enzymes control.

In this Chapter, will be reported the genetic strategies optimization in order to maximize BDO production from the genome-scale metabolic network of *E. coli*. Genetic strategy stands for knockout manipulation, i.e., turning off one or more genes. Searching for the optimal knockout set is a hard problem, since an organism such as a bacterium, contains more than 4000 genes. Importantly, a knockout strategy must also guarantee the survival of the organism. By using Pareto optimality and multi-objective optimization, I simultaneously ensure the biomass formation and find several solutions that represent trade-offs [15]. Knockout manipulations are represented in the model by means of a Boolean string y, where the lth element is 1 if the corresponding gene set is turned off. The aim is to find the optimal Boolean string y that maximizes or minimizes simultaneously two or more biological functions (in this case, BDO production and biomass formation). The *E. coli* model is optimized by using GDMO algorithm and the metabolic network is analyzed with flux balance methos, as described in details in the sections 2.2 and 2.4.

4.1 Synthesis of BDO from glucose in *E. coli*

The authors of Yim et al. [58] used the *i*JR904 *E. coli* model and added the BDO synthetic pathway obtaining a network composed of 904 genes, 941 biochemical reactions and 625 metabolites. BDO is a non-natural compound not synthesized by any known organism. Therefore, by using *SimPheny Biopathway Predictor* software, Yim et al. authors studied and analyzed all potential pathways from *E. coli* central metabolites to BDO. *SimPheny Biopathway Predictor* is a computational tool implemented in Genomatica's SimPheny platform for enumerating and evaluating metabolic networks, with the goal of identifying novel pathways for producing a chemical of interest. This software found over 10,000 pathways for the synthesis of BDO from common central metabolites such as acetyl-CoA, α-ketoglutarate, succinyl-CoA and glutamate. Among all pathways, authors selected the BDO production pathways proceeding through the 4-hydroxybutyrate (4HB) intermediate. Researchers at Genomatica report on their metabolic engineering of *E. coli* for the direct production of 1,4-butanediol in a paper in the journal Nature Chemical Biology.

The resulted BDO designed pathway starts from the tricarboxylic acid cycle intermediate succinate, which is activated as succinyl-CoA by the *E. coli* enzyme succinyl-CoA synthetase (SucCD). After two sequential reduction steps catalyzed by CoA-dependent

succinate semialdehyde dehydrogenase (SucD) and 4HB dehydrogenase (4HBd), respectively, the CoA derivative converts to 4HB via succinate semialdehyde. To synthesize 4HB, Yim et al., consider also the forming from α-ketoglutarate.

The conversion of 4HB to BDO requires two reduction steps, catalyzed by dehydrogenases. Alcohol and aldehyde dehydrogenases (ADH and ALD, respectively) are NADH- and/or NADPH-dependent enzymes that together can reduce a carboxylic acid group (derivatized with Coenzyme A) to an alcohol group.

Authors show the results conducted in in-vivo experiments after engineering the bacterium with knockout strategies found in silico by OptKnock algorithm [12]. I want to remark that this is the first industrial application of FBA in-silico analysis, and the positive results confirm the success of this method to investigate big metabolic networks. The resulted BDO synthetic pathway is figured in 4.1.

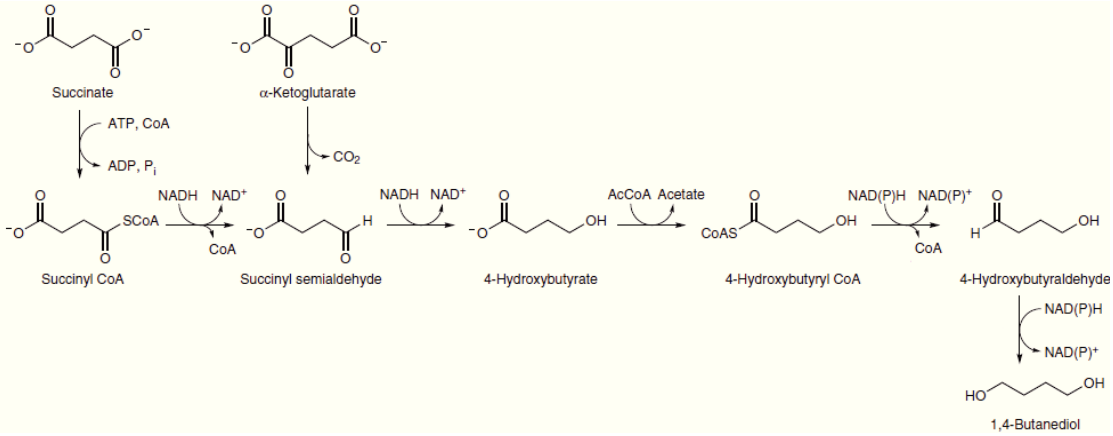


FIGURE 4.1: The BDO synthetic pathway included in *E. coli* organism. Figure extracted from [58].

Evaluation of the OptKnock results proposes a knockout strategy removing alcohol dehydrogenase (adhE), pyruvate formate lyase (pfl), lactate dehydrogenase (ldh), and malate dehydrogenase (mdh) genes.

The Genetic Design through Multi-objective Optimization method is used here to investigate other suitable knockout strategies. The original *i*JR904 *E. coli* model contains also the *GTP amine hydrolysis* (GTHP) reaction, that authors do not consider in their in-silico model. In my experiments, I performed simulations both including and excluding the GTHP reaction. Additionally, for each experiment I used two value of the parameter C (section 2.2.3 in Chapter 2). C indicates the maximum knockout number allowable and was set to 50 and 10.

4.2 Overproduction of BDO in *E. coli*

In a first phase of my study on BDO overproduction, I have considered the complete metabolic network published in [59] that includes the engineered synthetic pathway. I have used BioCAD in order to perform the maximization of BDO via facilitated diffusion and search for the optimal genetic manipulations. Results are shown in Figure 4.2-A. The two Pareto fronts have been obtained by GDMO with a population composed of 1000 individuals. After 3000 generations, by analyzing the whole solutions space, therefore all the 3,000,000 feasible points, in anaerobic conditions I have selected 6,757 Pareto strains, i.e., Pareto points, when C is 10. Instead, if the search space is expanded by setting C=50, the Pareto strains increment to 19,152. Moreover, by fixing C=50, the Pareto front on the left section owns better solutions in terms of BDO with respect to Pareto front obtained with C=10. BioCAD does not limit the Pareto analysis to the last population, but extract all the Pareto solutions through non-dominated sorting. In Figure 4.2-A, the points in red represent solutions proposed in [58], conversely in green and black results obtained from GDMO. In Figure 4.2-B-C the knockout cost associated to each synthetic production in the Pareto fronts are reported. In the first case, with C=10, we can observe that a good maximum approximation level of BDO is reached with a knockout cost equal to 10, when BDO results 15.176 $mmolh^{-1}gDW^{-1}$ and biomass = 0.332 h^{-1}. Indeed, with a knockout cost from 11 to 16, we can see that the BDO level is almost stationary. After, the BDO production decreases. A good trade off could be the point with a knockout cost equal to 8, that reaches BDO = 14.625 $mmolh^{-1}gDW^{-1}$ and biomass=0.504 h^{-1}.

In a second stage of my study on BDO overproduction, in order to compare GDMO results with Yim et al. ones, I have considered the same *E. coli* network, that does not consider the *GTP amine hydrolysis* (GTHP) reaction. I performed GDMO and also in this case I used the same previously experimental protocol. Results are reported in Figure 4.3-A. Here, I have found 12,836 Pareto strains with C=10 and 49,876 Pareto strains with C=50. Each of this point is an optimal strain able to outperform BDO production. The decision-maker can select one or more of these solutions according to its interests. For example, one can choose solutions that have a particular knockout cost, due to biotechnological difficulties; or one can choose that particular strain that reaches the highest BDO level with the maximum robustness. Figures 4.3-B-C help the decision-maker to select best and suitable points. In Figure 4.3-B we can examine that with a cost of 6 knockout, GDMO finds 15.168 mmolh^{-1} gDW^{-1} of BDO removing Acetaldehyde dehydrogenase (ADHer), lactaldehyde dehydrogenase (LCADi), malate dehydrogenase (Mdh), D-lactate, glycolate and formate transport. Instead, in Figure

TABLE 4.1: Maximum BDO production in the modified *i*JR904 *E. coli* for each knockout cost resulted from GDMO by considering 50 maximum allowable knockout (Figure 4.3-C in red). In bold the best trade-off designs. In particular, GDMO finds the same solution proposed in [58] with a minor knockout cost, that is 6. In bold the best trade-off designs. Details can be found in Table 4.4.

BDO (mmolh^{-1} gDW^{-1})	*Biomass* (h^{-1})	*Knockout cost*
16.229	**0.105**	**12**
16.222	0.105	15
16.222	0.105	13
16.221	0.105	11
16.219	0.105	16
16.121	0.111	10
16.079	0.115	19
16.079	0.115	17
16.077	0.115	18
16.074	**0.115**	**9**
15.974	0.117	26
15.974	0.117	23
15.974	0.117	25
15.974	0.118	22
15.974	0.118	20
15.974	0.118	21
15.586	0.120	29
15.586	0.120	27
15.586	0.120	28
15.586	0.120	24
15.523	0.133	8
15.521	**0.133**	**7**
15.168	**0.140**	**6**
10.938	0.231	32
10.938	0.231	30
10.936	0.232	31
10.933	0.234	35
10.930	0.235	33
10.925	0.237	34
10.765	0.292	5
10.294	0.299	4
7.882	0.336	3
6.945	0.350	2
5.716	0.378	1
0.000	0.418	0 (Wild type)

4.3-C, an interesting point is obtained by deleting 9 genes, reaching 16.073 mmolh^{-1} gDW^{-1} of BDO production.

After optimization and post-processing analysis, BioCAD software provides a detailed list of genetic strategies, with information about pathways and reactions associated to.

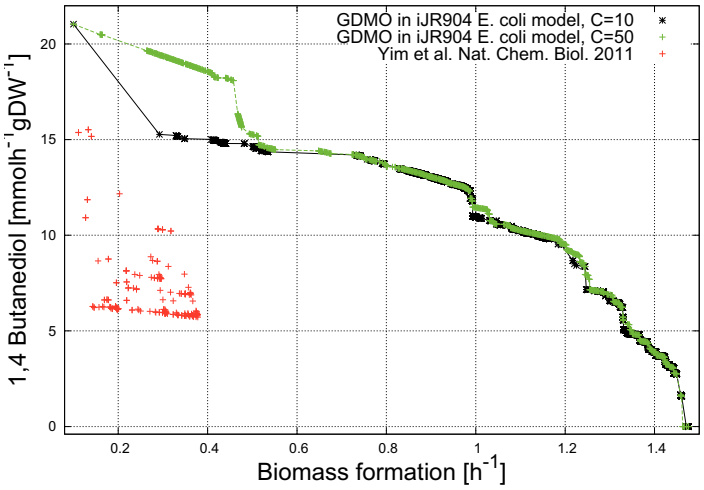
In Figure 4.3-A, we can notice that some points in green overcome Pareto fronts in red and blue. This is because OptKnock algorithm [12] used by Yim et al. authors, does not take into account the GPR mappings and it apply knockout directly on the fluxes of the metabolic network. Instead, in GDMO we can consider isoenzymes, enzymatic complexes and remove reactions starting from the genome level (for details, reader is remanded to section 2.2.2). Indeed, a reaction can be catalyzed by an enzyme linked to other reactions. Therefore, the single removing is not correct and the resulting flux distribution can be unreal. Yim et al. solutions that overcome Pareto fronts are obtained turning off reactions that in the FBA model are not linked to any gene, therefore not included on the variable space. On the contrary, GDMO considers gene sets and implements GPR relationship. In Table 4.4 reader can find supplementary information and detailed information on knockout strategies.

BioCAD software implements also a method able to infer how many times a gene is involved in Pareto genetic manipulations. An interesting result is that the gene set b1241 (adhE), associated to acetaldehyde dehydrogenase is present in almost all knockout solutions. In Table 4.2, I report the most involved genes (on 101) turned off in GDMO points for C=50. I want to remark that Pareto optimality is a good tool in biological design automation and allows to find a big mole of solutions to propose to biologists or biotechnologists. As a matter of fact, In Yim et al. work, authors provide 203 solutions against 49,876 Pareto strains if we consider C=50.

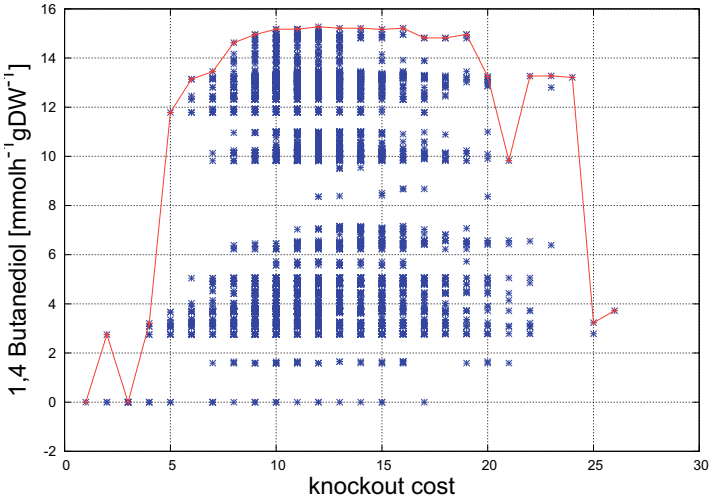
The method in Yim et al. does not consider the GPR map and turns off the flux of the reactions. In order to compare GDMO method with Yim et al. results, I have also performed knockout research in the reaction space. In Figure 4.4 we can see the results obtained by GDMO and in this case green points belong to the underlying Pareto front area. GDMO finds several interesting solutions (Table 4.3). With a four-reaction knockout GDMO finds the same strategy proposed in [58]: BDO equal to 15.166 mmol h^{-1} gdW^{-1} and biomass 0.140 h^{-1}. The real knockout cost associated to this genetic manipulation is 7. GDMO reveals points reaching the same level of BDO and biomass with better knockout cost (Tables 4.3 and 4.4.) ϵ-dominance analysis and experiments in aerobic conditions are shown in Appendix A.

TABLE 4.2: Knocked out gene sets occurrences for maximizing BDO production in the modified *i*JR904 *E. coli* with respect to 49,876 Pareto strains (front in green, C=50, Figure 4.3).

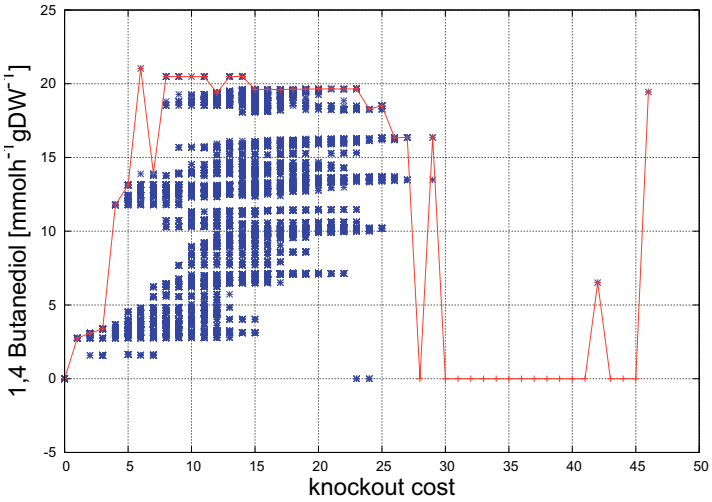
Gene sets ID	*Occurrences*	*Frequency*
b1241	49824	0.999%
b2975, b3603	46167	0.926%
b3708	31242	0.626%
b1602+b1603	30289	0.607%
b0507	28322	0.568%
b2276+b2277+b2278+b2279+b2280+ +b2281+b2282+b2283+b2284+ b2285+b2286+b2287+b2288	25019	0.502%
b3926	24000	0.481%
b2492, b0904	23174	0.465%
b2965, b0693	20413	0.409%
b4384, b2407	17807	0.357%
b0004	17085	0.343%
b0003	15004	0.301%
b0429+b0430+b0431+b0432	14551	0.291%
b4301, b3386	12595	0.253%
b3236	12161	0.244%
b0171	11541	0.231%
b1091	11492	0.230%
b0469	11339	0.227%
b1849	11304	0.227%
b1232	10980	0.220%
b0910	10927	0.219%
b3942, b1732	10432	0.209%



A)

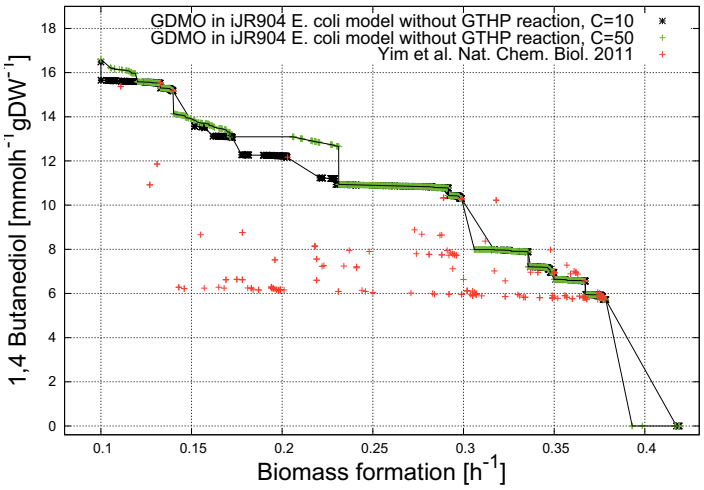


B)

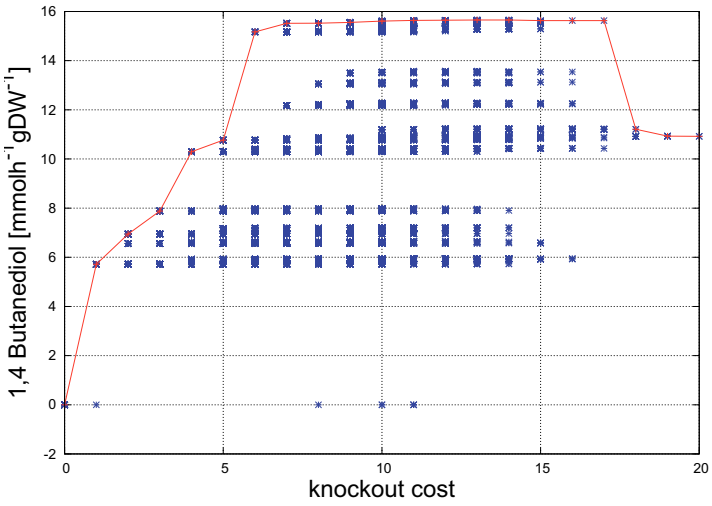


C)

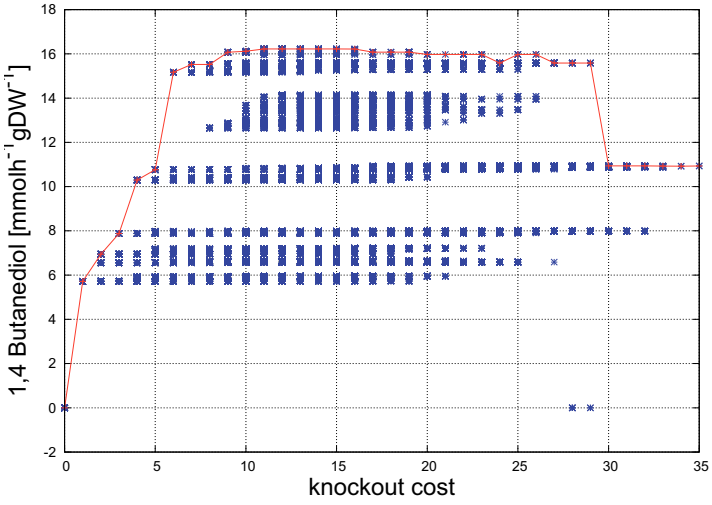
FIGURE 4.2: (A) Genetic strategies optimization to maximize BDO and biomass in anaerobic conditions and glucose feed equal to 20 mmolh^{-1} gDW^{-1} in original *i*JR904 *E. coli* by using C=10 and C=50, a population I=1000 and gen=3000. Knockout cost versus BDO production for Pareto strategies with C=10 (B) and C=50 (C).



A)

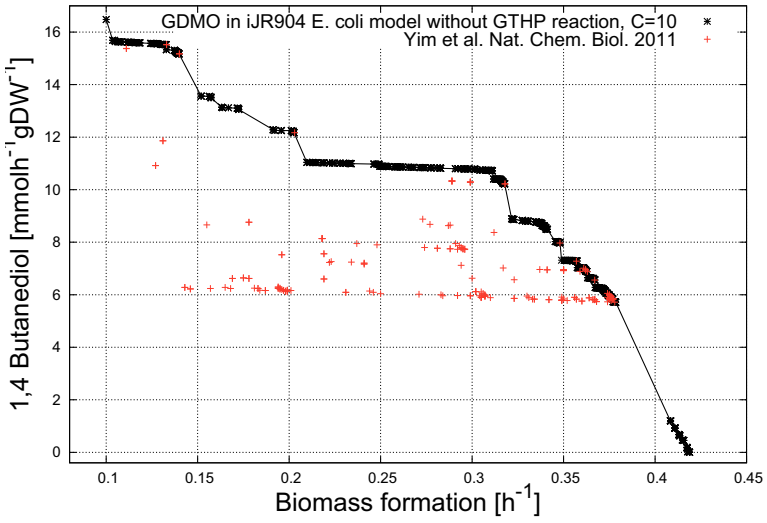


B)

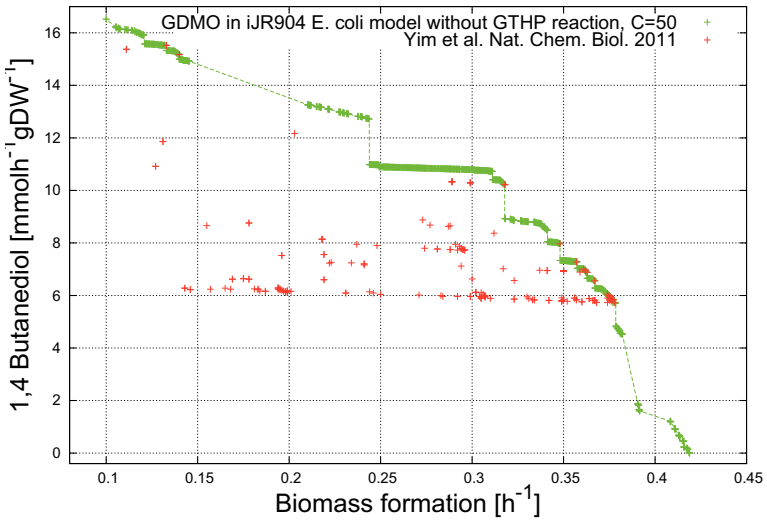


C)

FIGURE 4.3: (A) Genetic strategies optimization to maximize BDO and biomass in anaerobic conditions and glucose feed equal to 20 mmolh^{-1} gDW^{-1} in the modified *i*JR904 *E. coli* (without the reaction GTHP such as in Yim et al. [58] simulations) by using C=10 and C=50, a population I=1000 and gen=6000. Knockout cost versus BDO production for Pareto strategies with C=10 (B) and C=50 (C).



A)



B)

FIGURE 4.4: Genetic strategies optimization by deleting reactions to maximize BDO and biomass in anaerobic conditions and glucose feed equal to 20 mmolh^{-1} gDW^{-1} in the modified *i*JR904 *E. coli* (without the reaction GTHP such as in Yim et al. [58] simulations) by using a population I=1000 and gen=6000, C=10 in (A) and C=50 in (B).

TABLE 4.3: Maximum BDO production in the modified *i*JR904 *E. coli* for each knocked out reaction resulted from GDMO by considering 50 maximum allowable knockout (Figure 4.4-B in green). In bold the best trade-off designs.

BDO ($mmolh^{-1}$ gDW^{-1})	*Biomass* (h^{-1})	*Reaction cost*
16.223	**0.105**	**11**
16.222	0.105	10
16.220	0.105	9
16.219	0.105	13
16.215	0.106	8
16.072	0.116	7
15.969	0.118	18
15.969	0.118	17
15.968	0.118	15
15.968	0.118	16
15.968	0.118	14
15.924	0.119	21
15.924	0.119	20
15.924	0.119	19
15.584	0.121	22
15.522	0.133	6
15.518	0.133	5
15.166	**0.140**	**4**
13.193	0.215	27
10.886	0.256	23
10.865	0.266	25
10.864	0.266	24
10.853	0.272	26
10.220	0.318	3
6.882	0.363	2
5.716	0.378	1
0.000	0.419	0 (Wild type)

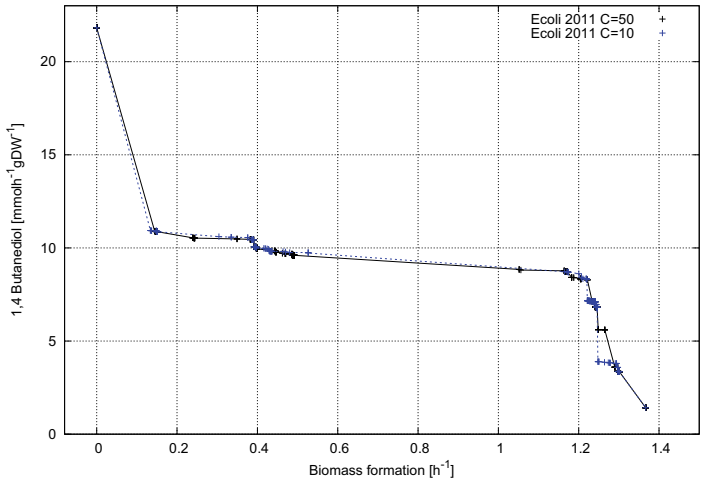
To consider more genes, I also take into account of the most recent genome-scale metabolic network of *E. coli* [50]. I have added the synthetic pathway composed of 10 new reactions that forms BDO (Figure 4.1), starting form Succinyl-CoA and α-ketoglutarate, produced naturally in *E. coli* metabolism. The *i*JO1366 *E. coli* model is composed of 1366 genes, 1041 gene sets, 1136 metabolites and 2261 reactions. By maintaining the same environment conditions used by Yim et al. [58] (a glucose uptake rate equal to 20 mmol h^{-1} gdW^{-1}), I have maximized BDO production and biomass formation in anaerobic and aerobic conditions. With a knockout cost equal to 6, I have obtained an overproduction of +662.669% of BDO (from 1.425 mmol h^{-1} gdW^{-1} to 10.869 mmol h^{-1} gdW^{-1}). This genetic strategy represents the best trade off design. The genes knocked out are: (i) the set (PflBec) or (TdcEec) or (PflDec) or (PflBec and YfiD) belonging to the Pyruvate Metabolism (pyruvate formate lyase reaction), (ii) the gene YgiN belonging to the Oxidative Phosphorylation (quinol monooxygenase reactions) and (ii) the gene Tpi belonging to the Glycolysis/Gluconeogenesis (triose-phosphate isomerase reaction). The maximum BDO production is reached with a knockout cost equal to 10, providing 10.933 mmol h^{-1} gdW^{-1} BDO production and 0.134 h^{-1} biomass formation.

Figure A.11 in Appendix A shows the evolution of BDO and biomass maximization during the optimization of the environment conditions. In blue the dominated solutions, and in black the final Pareto front. Environment conditions are represented by 330 uptake rate fluxes. In this experiment, I have maintained glucose and oxygen uptake rates respectively fixed at 20 mmol h^{-1} gdW^{-1} and 0 mmol h^{-1} gdW^{-1}. Changing uptake rates is a crucial element and can significantly improve the production.

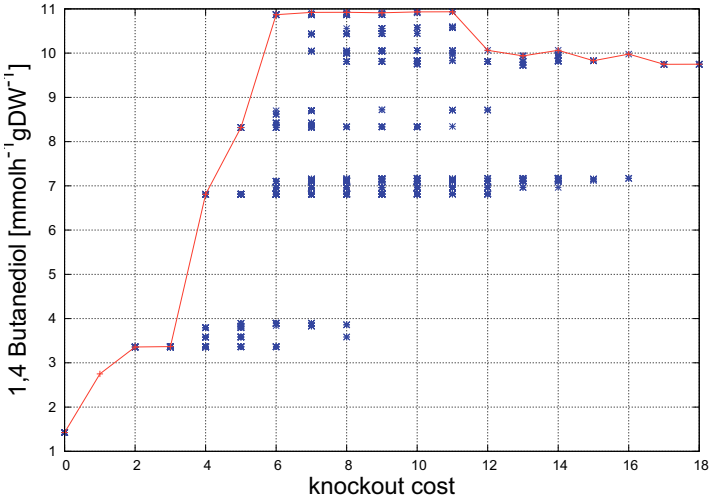
In Figures 4.6 and 4.7, have been reported the results obtained from Pathway oriented sensitivity analysis in terms of knockout and exchange fluxes respectively for *i*JR904 and *i*JO1366 networks. The higher are the sensitive indexes, higher is the influence of the pathway/parameter on the outputs of the model.

TABLE 4.4: Knockout strategies obtained through GDMO for maximizing BDO production by using C=50 in the modified *i*JR904 *E. coli* [mmolh^{-1} gDW^{-1}].

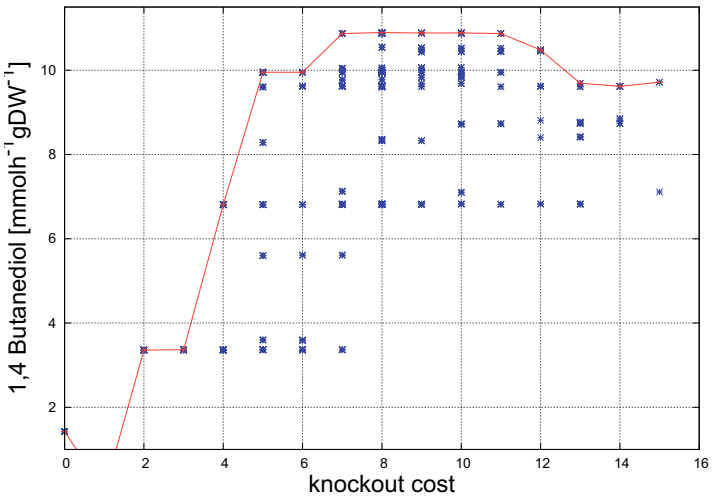
Str	BDO	Biom	kcost	Genes	Pathways	Reactions
A_1	16.2228	0.10519				
	(-74.8873%)		12	b1241	Alternate Carbon Metab., Pyruvate Metab.,	LCADi
						ADHEr
				b2661	Arginine and Proline Metab.,	SSALy
				b3236	Citrate Cycle (TCA),	MDH
				b2551	Cofactor and Prosthetic Group Biosynthesis,	ALATA-D2
					Glycine and Serine Metab.	
						ALATA-L2
						GHMT2
				b3708	Cysteine Metab., Tyrosine, Tryptophan,	TRPAS1
					and Phenylalanine Metab.	
						TRPAS2
				b1602+b1603	Oxidative Phosphorylation,	THD2
				b0767	Pentose Phosphate Pathway,	PGL
				b1849	Purine and Pyrimidine Biosynthesis,	GART
				b2975, b3603	Transport, Extracellular,	D-LACt2
						GLYCLTt2r
						L-LACt2
				b2492, b0904	Transport, Extracellular,	FORt
A_2	16.0736	0.11551				
	(-72.4245%)		9	b1241	Alternate Carbon Metab., Pyruvate Metab.,	LCADi
						ADHEr
				b2661	Arginine and Proline Metab.,	SSALy
				b3236	Citrate Cycle (TCA),	MDH
				b1602+b1603	Oxidative Phosphorylation,	THD2
				b0767	Pentose Phosphate Pathway,	PGL
				b2975, b3603	Transport, Extracellular,	D-LACt2
						GLYCLTt2r
						L-LACt2
				b2492, b0904	Transport, Extracellular,	FORt
A_3	16.0736	0.11551				
	(-72.4245%)		9	b1241	Alternate Carbon Metab., Pyruvate Metab.,	LCADi
						ADHEr
				b2661	Arginine and Proline Metab.,	SSALy
				b3236	Citrate Cycle (TCA),	MDH
				b1602+b1603	Oxidative Phosphorylation,	THD2
				b1852	Pentose Phosphate Pathway,	G6PDHy
				b2975, b3603	Transport, Extracellular,	D-LACt2
						GLYCLTt2r
						L-LACt2
				b2492, b0904	Transport, Extracellular,	FORt
A_4	15.5209	0.13277				
	(-68.3043%)		7	b1241	Alternate Carbon Metab., Pyruvate Metab.,	LCADi
						ADHEr
				b3236	Citrate Cycle (TCA),	MDH
				b1602+b1603	Oxidative Phosphorylation,	THD2
				b2975, b3603	Transport, Extracellular,	D-LACt2
						GLYCLTt2r
						L-LACt2
				b2492, b0904	Transport, Extracellular,	FORt
				b2492, b0904	Transport, Extracellular,	FORt
A_5	15.1683	0.13977				
	(-66.6334%)		6	b1241	Alternate Carbon Metab., Pyruvate Metab.,	LCADi
						ADHEr
				b3236	Citrate Cycle (TCA),	MDH
				b2975, b3603	Transport, Extracellular,	D-LACt2
						GLYCLTt2r
						L-LACt2
				b2492, b0904	Transport, Extracellular,	FORt



A)



B)

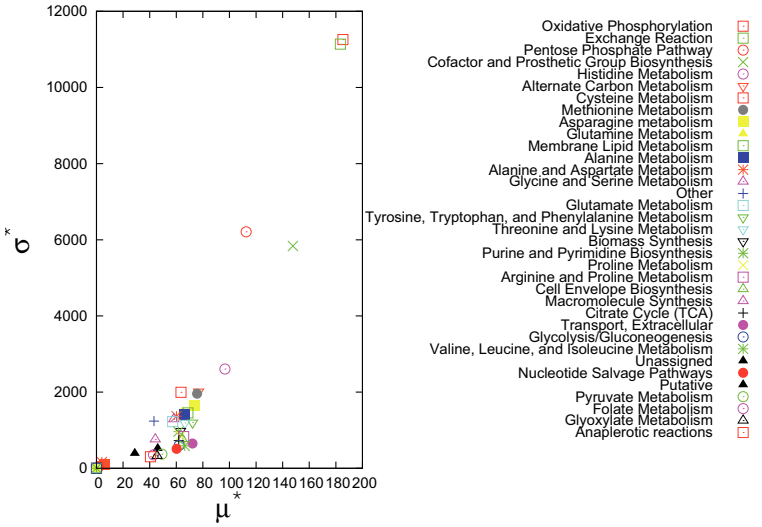


C)

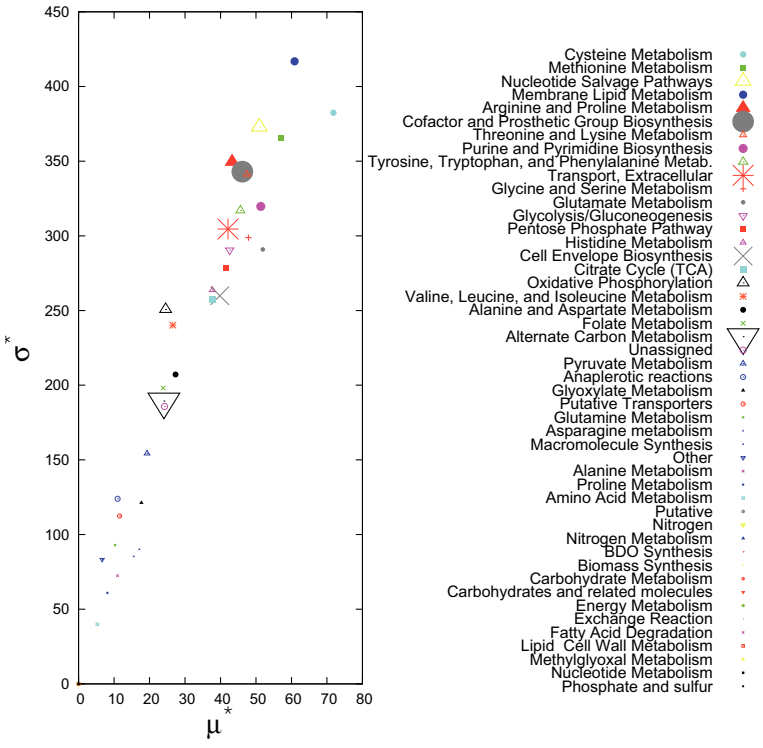
FIGURE 4.5: (A) Genetic strategies optimization to maximize BDO and biomass in anaerobic conditions and glucose feed equal to 20 mmolh^{-1} gDW^{-1} in *i*JO1366 *E. coli* by using C=10 and C=50, a population I=1000 and gen=6000. Knockout cost versus BDO production for Pareto strategies with C=10 (B) and C=50 (C).

TABLE 4.5: Maximum BDO production in *i*JO1366 *E. coli* for each knockout cost resulted from GDMO by considering 10 maximum allowable knockout (Figure 4.5-B in red). In bold the best trade-off designs. Details can be found in Table A.3 of Appendix A.

BDO ($mmolh^{-1}$ gDW^{-1})	*Biomass* (h^{-1})	*Knockout cost*
10.933	0.134	10
10.923	0.136	8
10.922	0.137	7
10.912	0.148	9
10.869	**0.151**	**6**
10.063	0.392	14
10.063	0.392	12
9.978	0.416	16
9.937	0.428	13
9.829	0.431	15
9.750	0.526	18
9.746	0.527	17
8.319	**1.220**	**5**
6.809	1.246	4
3.367	1.298	3
3.358	1.301	2
1.425	1.367	0 (Wild type)

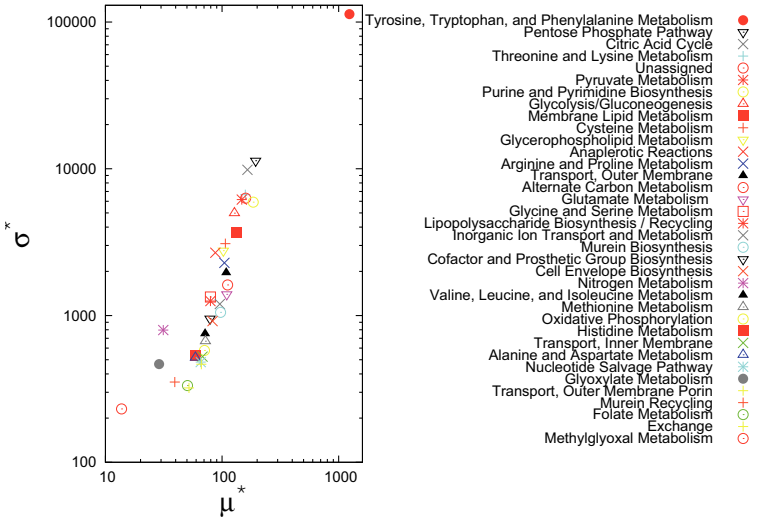


A)

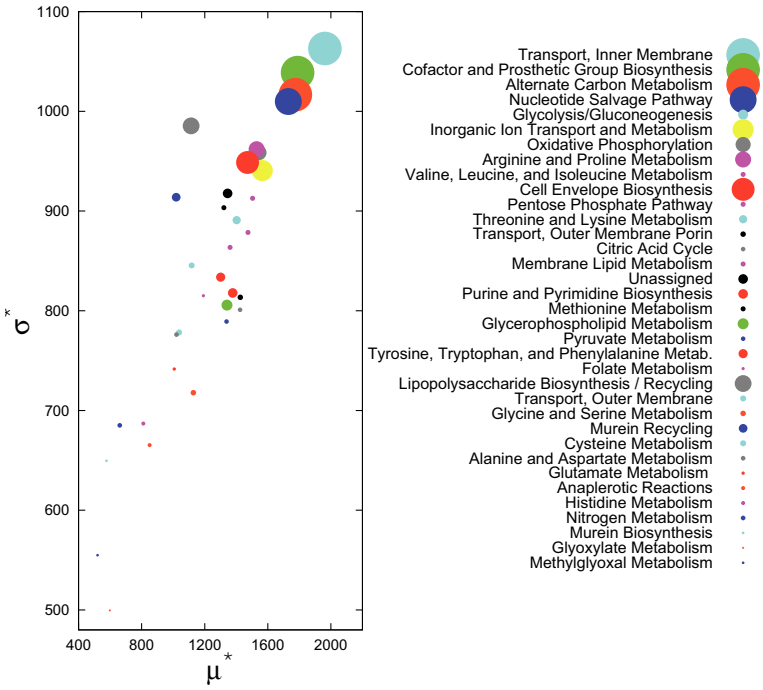


B)

FIGURE 4.6: (A) Fluxes-PoSA and (B) Gene set-PoSA simulations in *i*JR904 *E. coli* obtained with respectively 10^4 and 1.5×10^6 function evaluations.



A)



B)

FIGURE 4.7: (A) Fluxes-PoSA and (B) Gene set-PoSA simulations in *i*JO1366 *E. coli* obtained with respectively 10^4 and 1.5×10^6 function evaluations

Chapter 5

Artificial Photosynthetic Organisms

Developing models to simulate and predict the dynamic responses of metabolic networks has always been a challenging aim of systems biology. This goal is reached through the analysis of the main pathways involved in the metabolism of an organism. In particular, photosynthetic organisms perform many important functions for the planet, e.g. absorbing atmospheric CO_2, harvesting solar energy and generating O_2 instead of processing oxygen.

In this work, I focused my studies on the investigation of three metabolic networks: (i) the photosynthetic carbon metabolism pathway [60] of a generic leaf, (ii) the *Rhodobacter spheroides* [61] and (iii) the light-driven algal metabolism of *Chlamydomonas reinhardtii* [62].

The carbon metabolism is a process that takes place in chloroplasts, which are organelles present in the cells of plants and eukaryotic algae and represents the site of the photosynthesis. The energy from light is captured by chlorophyll pigments and is converted into chemical energy (ATP and NADH). Chloroplasts produce glucose from sunlight energy. The glucose then transfers to the mitochondrion for aerobic respiration. The function of chloroplasts is basically to make food through the photosynthesis, i.e., by trapping light energy to convert water and carbon dioxide to form oxygen and glucose. During the photosynthesis, carbon is used for growth and some excess carbon can be fixed and stored in compact polymers as starch. The latter is stored in the constitute of granules made up of both linear and branched polymers of glucose [63].

The *R. sphaeroides* system by Imam et al. [61] models all the most interesting features of photosynthesis, as well as the metabolic capabilities of this kind of organisms. Interestingly, when the *R. sphaeroides* lacks oxygen intake, it can process light energy through a photosynthetic electron transport chain, whose features are similar to those found in plants [64]. Moreover, during its photosynthetic growth, *R. sphaeroides* uses either CO_2 as the sole carbon source or other organic carbon sources in order to grow autotrophically or heterotrophically. The autotrophic metabolism of *R. sphaeroides* makes it a potential organism for use in the synthesis of chemicals or polymers that can serve as raw materials in the production of biofuels, or as a means of sequestering atmospheric or industrially produced CO_2. Therefore, a comprehensive analysis of the *R. sphaeroides* model may prove very useful for the understanding of both the lifestyle and the mechanisms underlying transitions between these different metabolic states.

The model of the *Chlamydomonas reinhardtii* metabolism [62] allows the investigation of photosynthesis in algae, and in particular of light regulation. The advantages of this model and its optimization are evident from the perspective of the biofuel production. The organisms and pathways above mentioned cover an important task by using photosynthesis process. Increasing the ability of these organisms to consume CO_2 can be very interesting. For this reason, in this work I investigated the photosynthesis process and optimized it.

5.1 Photosynthetic carbon metabolism

The concept of robustness is extremely pervasive in nature, and seems to be one of the driving force of evolution [42]; moreover, the ability of a system to preserve its behavior, despite internal or external perturbations, is a crucial design principle for any biological and synthetic system [42, 65–67]. Applying the concept of robustness to the Calvin Cycle and to the pathways involved in photosynthesis process allows BioCAD method to calculate the limits of enzymes perturbation at which the system property of interest (a given level of CO_2 uptake) is maintained. The estimation of the robustness of in silico designed pathways has been performed using the methodology proposed by Stracquadanio et al. [42] and described in details in section 2.8 of Chapter 2. A Monte-Carlo algorithm applies a Gaussian noise to the enzyme concentrations and then estimates the variation of CO_2 uptake. A robust system is characterized by small fluctuations of its quantitative behavior under investigation, which means that a robust pathway will ensure the same uptake rate even if the enzyme concentrations differ from the nominal values.

Although it is possible to perform the in-silico design and verification of a biological system, it is still impracticable to edit long regions of a genome in an arbitrary way; the intrinsic structure of the genetic information introduces a number of constraints that must hold in order not to decrease the fitness of a living organism. From this point of view, it is extremely important to focus the design on a set of restricted significant parameters, in order to decrease the complexity of a biological implementation. However, identifying a set of crucial genes encoding for important enzymes is an open problem. The sensitivity analysis tries to correlate the uncertainty in the output of a model with the uncertainty in the input; it is important to note that while the robustness analysis performs a local estimation of the output variation in a limited input range, the sensitivity analysis aims to study the output variation at a global level by investigating all the parameter space [30]. CO_2 uptake was simulated as the solution of the complete set of linked differential equations representing concentration change of the substrates of each reaction of the Calvin Cycle and related cycles related to the carbon metabolism.

The model and the chosen algorithms allow to find the optimized concentration of the enzymes in order to obtain the highest increase in CO_2 uptake, keeping constant the total amount of protein nitrogen. The Parallel optimization algorithm (PAO) [30], a single-objective optimization algorithm described in depth in section 5.5, allows to identify solutions consisting in an optimized set of enzymatic concentration capable of reaching a theoretical CO_2 uptake rate of 36.382 μmol m^{-2} s^{-1} at a level of carbonate ions (c_i) of 270 μmoles $moles^{-1}$ (fourth column in Table 5.1). The CO_2 uptake at the initial enzyme concentrations was 15.486 μmol m^{-2} s^{-1} (second column in Table 5.1). This solution showed that six enzymes were particularly enhanced: cytosolic FBP aldolase, cytosolic FBPase, UDPGPP , SBPase, RuBisCO and ADPGPP (Figure 5.1). The complete name of enzymes is reported in Table A.6 in Appendix A. The method obtains a theoretical CO_2 uptake increase corresponding to +134% with respect to the initial enzymes concentration. An increase in theoretical CO_2 fixation rates obtained by varying the enzyme concentrations of the Calvin Cycle starting from the current experimentally determined values was shown already by various authors as Zhu et al. [60] and Stracquadanio et al. [30]. The PAO algorithm allowed to obtain a theoretical CO_2 uptake increase corresponding to 134% with respect to the initial enzymes concentration. This result is even higher with respect to Zhu et al. [60] solutions based on an evolutionary algorithm, that leads to an increase of 76% (from about 16 to 28 μmol m^{-2} s^{-1}). By analyzing the enzymatic variation in Figure 5.1, when all enzymes are perturbed, we can see that three enzymes are particularly involved in the maximization process and change much more with respect to their initial nominal values: cytosolic FBP aldolase, cytosolic FBPase and UDP-Glc pyrophosphorylase. Therefore in another simulation,

TABLE 5.1: Photosynthetic carbon metabolism results. Concentrations of the enzymes, individual robustness, CO_2 uptake rate (at $c_i = 270\ \mu mol\ mol^{-1}$, reflecting current CO_2 atmospheric concentration), global and local robustness values. The second column reports the touchstone concentrations used in simulations: the initial/natural leaf (modeled by Zhu et al.[60]). The third column reports the results of the optimization in which only the eleven sensitive enzymes are altered, while all of the others are kept at their nominal values. The fourth column reports the best-known leaf design, in terms of CO_2 uptake and robustness. The fifth column reports the results of a simulation where the enzymes cytosolic FBP aldolase, cytosolic FBPase and UDP-Glc pyrophosphorylase have been maintained to their initial values. The last column reports the most efficient known point in terms of CO_2, but corresponds to a highly instable solution.

Enzyme Name	Initial Conc. $mg\ N\ m^{-1}$ (the natural leaf)	Optimal Conc. of 11 Sensitive Enz. $mg\ N\ m^{-1}$	Optimal and Robust Conc. $mg\ N\ m^{-1}$	Optimal and Robust Conc. $mg\ N\ m^{-1}$ (3 fixed enz.)	Optimal but Not Robust Conc. $mg\ N\ m^{-1}$
RuBisCO	517.00 (100)	784.27 **(84.5)**	860.226 (100)	856.44 (100)	861.93 **(39)**
PGA kinase	12.20 (100)	4.66 (100)	3.989 (100)	3.63 (100)	3.98 **(0)**
GAP DH	68.80 (100)	69.03 **(81.5)**	64.483 (100)	65.08 (100)	63.55 **(17)**
FBP aldolase	6.42 (100)	10.40 (100)	9.050 (100)	10.86 (100)	9.29 **(30.5)**
FBPase	25.50 (100)	29.44 (100)	26.889 (100)	32.24 (100)	27.03 **(0)**
Transketolase	34.90 (100)	*34.90* (100)	8.247 (100)	16.93 (100)	16.98 *(100)*
SBP aldolase	6.21 (100)	5.55 (100)	6.661 (100)	5.75 (100)	5.94 **(0)**
SBPase	1.29 (100)	4.70 (100)	4.397 (100)	4.43 (100)	4.31 **(1)**
PRK	7.64 (100)	7.04 (100)	7.007 (100)	6.38 (100)	7.99 **(22.5)**
ADPGPP	0.49 (100)	2.12 (100)	0.721 (100)	5.09 (100)	1.22 **(0)**
PGCA Pase	85.20 (100)	0.95 (100)	0.325 (100)	0.20 (100)	0.00 **(0)**
Glycerate kinase	6.36 (100)	*6.36* (100)	0.005 (100)	0.00 (100)	0.00 *(100)*
Glycolate oxidase	4.77 (100)	*4.77* (100)	0.019 (100)	0.16 (100)	0.00 *(100)*
GSAT	17.30 (100)	*17.30* (100)	0.027 (100)	0.00 (100)	0.00 *(100)*
Glycer. dehyd.	2.64 (100)	*2.64* (100)	0.003 (100)	0.00 (100)	0.00 *(100)*
GGAT	21.80 (100)	*21.80* (100)	0.00005 (100)	0.00 (100)	0.00 *(100)*
GDC	179.00 (100)	0.02 (100)	0.00003 (100)	0.00 (100)	0.00 *(100)*
Cyt. FBP ald.	0.57 (100)	*0.57* (100)	2.127 (100)	*0.57* (100)	2.03 **(0.5)**
Cyt. FBPase	2.24 (100)	*2.24* (100)	5.554 (100)	*2.24* (100)	5.27 **(30.5)**
UDPGPP	0.07 (100)	*0.07* (100)	0.531 (100)	*0.07* (100)	0.50 **(0)**
SPS	0.20 (100)	*0.20* (100)	0.034 (100)	0.01 (100)	0.03 **(30.5)**
SPP	0.13 (100)	*0.13* (100)	0.031 (100)	0.01 (100)	0.03 **(0)**
F26BPase	0.02 (100)	*0.02* (100)	0.00 (100)	0.00 (100)	0.00 *(100)*
CO_2 Uptake $\frac{\mu mol}{m^2 s}$	15.486	33.317	36.382	36.197	36.495
(Local R. %, Global R. %)	(100, 81.80)	(81.5, 78.3)	(100, 97.2)	(100, 92.6)	(0, 39.18)

CO_2 uptake is maximized by perturbing all the enzymatic concentration and maintaining fixed the three involved enzymes. Results are reported on the fifth column of Table 5.1. The perturbation of parameters (concentrations) allows to understand the level of sensitivity of each of the considered enzymes involved in CO_2 fixation. By means of Morris method, eleven enzymes are found to be sensitive and two of them fragile (Table A.4 in Appendix A). For this reason, PAO algorithm was processed in order to find the maximum CO_2 level by perturbing only the sensitive eleven enzymatic concentration. This designed optimization reached a level of 33.317 μmol m^{-2} s^{-1} of CO_2 (third column in Table 5.1).

Since biotechnological techniques are currently incapable of treating many enzymes at the same time, I have simulated the effect of perturbing six enzymes only (RuBisCO,

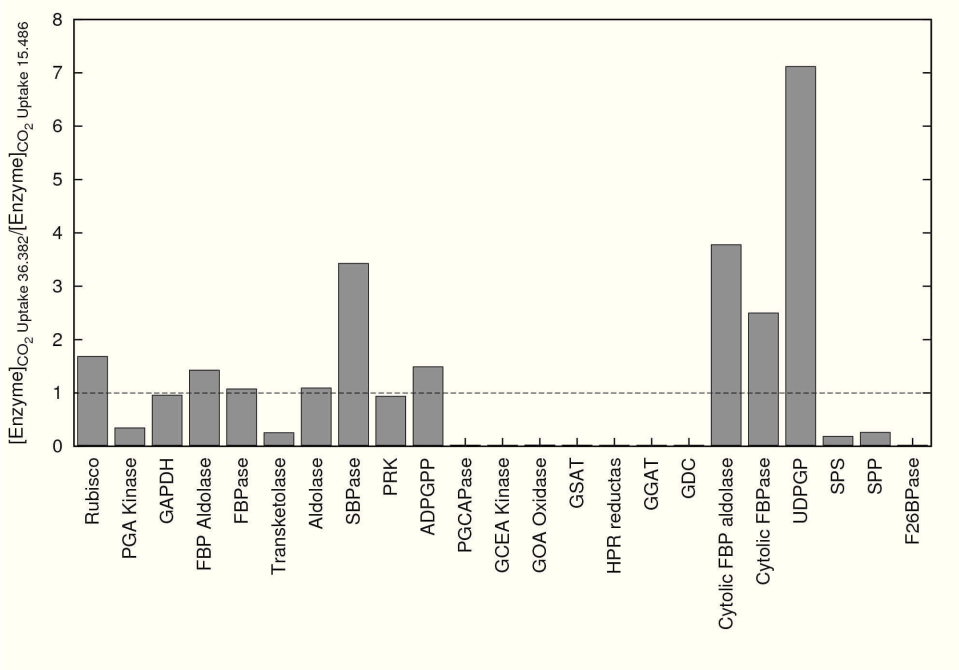


FIGURE 5.1: Optimization in photosynthetic carbon metabolism. The ratio of the enzyme concentrations optimized by the PAO algorithm (36.382 μmol m^{-2} s^{-1}) at $c_i = 270$ μmol mol$^{-}1$ compared to the initial concentrations (15.486 μmol m^{-2} s^{-1})

FBP aldolase, SBPase, ADPGPP, Phosphoglycolate phosphatase, and Gly decarboxylase (GDC)) while the remaining nineteen enzymes are maintained at their initial concentrations. I have chosen the sensitive enzymes which values change out of the range 0.2x-1.5x. For this set of six enzymes, I have defined the following constraint: the concentration must be ≥ 0.02 mg N m^{-1}. RuBisCO, FBP aldolase, SBPase, ADPGPP resulted overexpressed, while Phosphoglycolate phosphatase and GDC resulted almost switched off. Nitrogen is kept constant. This configuration obtained CO_2 uptake rate of 32.828 μmol m^{-2} s^{-1} (that is only 3.492 μmol m^{-2} s^{-1} less than the best solution), perturbing only six enzymes.

A still more refined analysis used always the same set of enzymes but allowing RuBisCO to increase up to a maximum of 15% with respect to the initial values. This constraint has been inserted in order to have more feasible biotechnological results, since a higher RuBisCO concentration increase is quite unlikely to be obtained in vivo. Much higher increase in expression level are known (40 fold for codon biased sequences in *E. coli*[68]). However RuBisCO is already the most expressed enzyme in plants. FBP aldolase, SBPase, and particularly ADPGPP resulted overexpressed, while phosphoglycolate phosphatase was switched off and GDC was kept close to its initial value. This configuration obtained a CO_2 uptake rate of 31.819 μmol m^{-2} s^{-1} (that is only 4.501 μmol m^{-2} s^{-1} less than the best solution).

A further simulation attempted an optimization of the CO_2 uptake rate perturbing four enzymes only (FBP aldolase, SBPase, PGCAPase, and GDC) while the remaining 21 enzymes were maintained at their initial concentrations. This configuration obtained a CO_2 uptake rate of 22.4202 μmol m^{-2} s^{-1} with respect to the initial concentration of about 16 μmol m^{-2} s^{-1}. In a combinatorial approach, I have performed another optimization with a different set of four enzymes, which are FBP aldolase, ADPGPP, PGCAPase, and GDC. This configuration obtained a CO_2 uptake rate of 20.626 μmol m^{-2} s^{-1} (Table A.5 in Appendix A).

Another important biotechnological target is to check the possible increase of CO_2 uptake leaving RuBisCO constant. This limitation is appropriate: given that RuBisCO is the most abundant protein in nature, it has been considered also a nitrogen reservoir for the plant metabolism [69, 70]. For instance, in an experiment on the haptophyte alga *Isochrysis galbana* on the effects of nitrogen limitation, as cells became more nitrogen limited, the fraction of total cell nitrogen contained in RuBisCO decreased from 21.3% to 6.7%, whereas that of the light harvesting complex remained relatively constant. That means that RuBisCO quantity is not only linked to the CO_2 uptake, but it has a secondary function as nitrogen storage. Moreover, after some studies, the enzyme might already be naturally optimized under an evolutionary point of view [71]. Hence, further optimization of RuBisCO may prove difficult and lead to only marginal improvements [72]. Therefore, it is quite unlikely that models allowing free further increase of RuBisCO concentration, would be really feasible. The optimization of CO_2 uptake rate perturbing 24 enzymes leaving RuBisCO at its initial concentration leads to a theoretical optimized uptake rate of 22.2698 μmol m^{-2} s^{-1} with respect to the initial about 16 μmol m^{-2} s^{-1} of the natural leaf. The most influential enzyme in this analysis was ADPGPP showing a very high increase in concentration.

The results summarized in Table A.4, showed that eleven enzymes are found to be sensitive and two of them fragile. A first conclusion is that the six most sensitive enzymes are key enzymes that can strongly influence the CO_2 uptake with slight concentration variation. The fact that these enzymes are mostly light controlled confirms the strict control of light availability on the Calvin Cycle. Highly and moderately sensitive enzymes found by Sun et al.[73], on the basis of microarrays expression patterns, largely correspond with those indicated in this analysis, with the exception of tranketolase (moderately sensitive in [73] and at low sensitivity in this analysis). The proposed solution has a high level of robustness (third column of Table 5.1). GAPDH and PRK did not vary much their concentration during the optimization analysis. This result would fit well with the fact that the expression of these two enzymes is controlled by light, while specific chloroplast proteins as CP12 are capable of controlling their activity forming with them a complex PRK/GAPDH/CP12 with high molecular weight [74]. Such a refined control appears

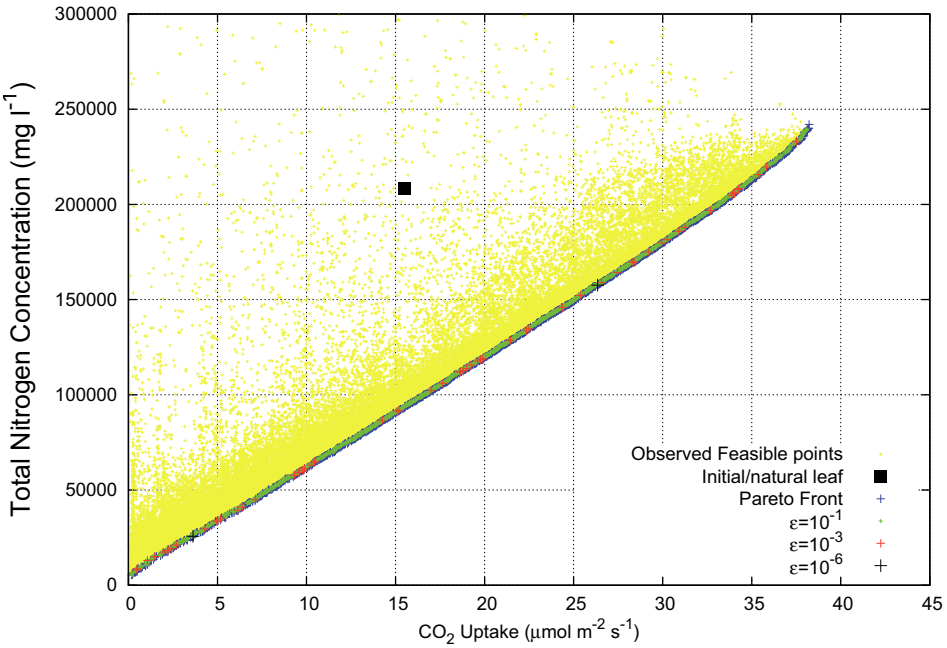


FIGURE 5.2: ϵ-dominance analysis for the photosynthetic carbon metabolism. In blue dots the Pareto fronts. In black the ϵ-dominance results choosing $\epsilon = 10^{-6}$, in red the ϵ-dominance results choosing $\epsilon = 10^{-3}$. Similarly, green points for $\epsilon = 10^{-1}$. Conversely, the yellow points are all the feasible points observed by the optimization algorithm, while the black square represents the initial/natural leaf (as modeled by Zhu et al.[60])

to be appropriate for sensitive enzymes and similar controls may be widespread for sensitive enzymes in vivo. RuBisCo also resulted the most sensitive enzyme confirmed also by Sobol' analysis (Figure A.13). Figure A.14 in Appendix A shows the interdependence between the enzymes.

The simple finding of an optimal solution with ideal concentrations could be not a sufficient task, since the transcription process of the genes and other control systems linked to changing environmental conditions and/or feedbacks coming from other biochemical pathways could vary the enzymes concentration or their activity with time. Moreover, biotechnological insertion of new promoters sequence is not able to produce an exact and foreseen amount of transcripts. Therefore it is clear that it is important to estimate how well the achieved CO_2 uptake is preserved under perturbation at the enzymes concentration level. Robustness can be defined as the persistence of a system property with respect to perturbations [42]. Such a property can be assessed in this analysis and can be fundamental to foresee the effect of a biotechnological genetic modification. The results of this analysis are shown in Tables 5.1 and Table A.5.

The analysis based on the evaluation of the nitrogen limitation effect showed that the minimal amount of nitrogen [75] allowed still a CO_2 uptake rate of 5.7. Such an amount could be taken into consideration as an assessment of the biomass growth limit of plants

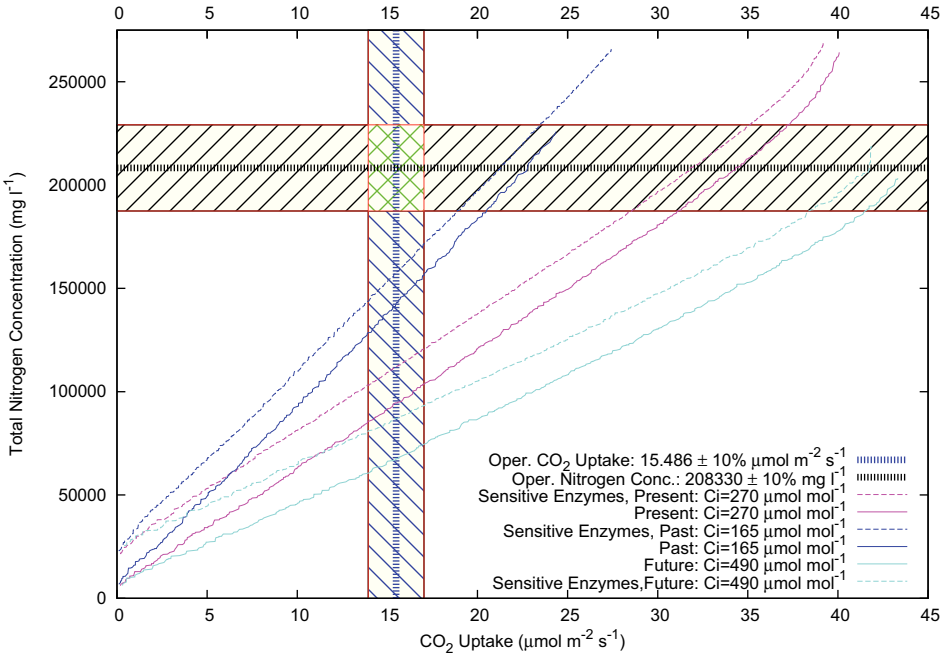


FIGURE 5.3: CO_2 uptake and protein-nitrogen concentration trade-off. Maximizing the CO_2 uptake while minimizing the total amount of protein-nitrogen concentration; the operative area of natural leaves is located in the green checked area. The label "Sensitive Enzymes" indicates the multi-objective optimization using the eleven most sensitive enzymes of the model. The three resulting Pareto Fronts have been dominated by the multi-objective optimization over all the enzymes of the model. This trade-off search has been carried out for the three c_i concentrations referring to the environmental conditions of 25 million years ago, nowadays and in 2100.

living in nitrogen limitations. The second result was shown as a result of the Pareto-Optimality analysis that should lead to the closest-to-ideal solution: in this case the CO_2 uptake rate is 21.213 (Figure A.15 of Appendix A). This value represents a theoretical limit for biotechnological targets leading to maximizing productivity with the minimum amount of nitrogen supply, which is the value close to the economical optimum. It is interesting to observe that, even in this case, the total CO_2 uptake resulted over 30% higher with respect to the natural CO_2 uptake rate at the natural enzymes concentration. Maximal CO_2 Uptake is 39.968. Figure 5.2 reports the ϵ-dominance analysis for the multi-objective optimization of CO_2 and nitrogen in photosynthetic carbon metabolism when the most sensitive eleven enzymes are considered in the optimization problem.

The model was used also to calculate the maximum CO_2 uptake rate at different atmospheric CO_2 concentrations: c_i of 270 μmol mol^{-1} corresponding to nowadays concentration of CO_2 in the atmosphere, c_i of 490 μmol mol^{-1} corresponding to a future concentration of CO_2 in the atmosphere and 165 μmol mol^{-1}, that is the CO_2 concentration estimated for 25M years ago ca. Results are report in Figure 5.3, where multi-objective optimization is performed to maximize CO_2 uptake rate and minimize

the nitrogen consumption, by perturbing all enzymes in carbon metabolism and by perturbing only the most eleven sensitive enzymes. Results showed that the main difference between the current CO_2 atmospheric concentration and that of the past regarded the optimization of ADPGPP, PGA kinase, HPR reductase, all much higher in the optimization at lower CO_2 concentration, and GCEA kinase, SPS and F26BPase, all much higher in the situation of high CO_2. These three last enzymes were not among the sensitive enzymes in the optimization at the current atmospheric CO_2 concentration. The results indicated that changing atmospheric conditions, particularly with respect to CO_2 amount, would produce very different evolutionary pressure on the enzymes. Concentration enhancement or reduction would affect one or the other enzyme, depending on the environmental conditions (at least relatively to CO_2).

5.1.1 Identifiability analysis for the carbon metabolism pathway

Here I considered the chloroplast model by Zhu et al. [60] and the 25 decision variables of the C3 cycle, namely the concentrations of its enzymes. I adopted the method proposed by Hengl et al. [76] to detect automatically structural identifiability consisting of functional relations between decision variables. These relations are detected by applying the alternating conditional expectation algorithm (ACE) [34].

I adopted the Mean Optimal Transformation Approach (MOTA) [76], by fixing at 5 the maximal number of parameters allowed to enclose a functional relation. The results are shown in Table 5.2. The "Groups" column indicates the functional relations between variables. For instance, RuBisCO and GAPDH are functionally related. In other words, the response variable x_1 is strongly related to the predictors x_3 and x_5. Conversely, the enzymes Transketolase type 1 and SBPase do not have any functional relation with any other enzyme (Table 5.2). The r^2 column indicates how much variance of the response can be explained by the predictors. A high amount of variance of the response that can be explained by the predictors indicates a large effect of the fixation of the predictors on the standard deviations of the response. The $cv(x) = std(x)/mean(x)$ helps to distinguish practical identifiable from non-identifiable parameters [76]. In case of practical non-identifiability, the choice of the parameter to fix depends on the experiments and on reference values found in the literature.

In Figure 5.4 has been shown the functional relations among RuBisCO, GAPDH and FBPase detected by the identifiability analysis applied to GAPDH.

It is noteworthy that, according to Table 5.2, RuBisCO belongs to the same functional group except for the presence of x_5 (FBPase). Indeed, Figure 5.5 shows that the optimal transformation β found for x_5 is different from the transformations found for x_1 and x_3,

TABLE 5.2: Identifiability Analysis applied on the 1903 non dominated points of the CO_2-nitrogen Pareto front. The enzymes are grouped according to functional relations. The r^2 column indicates how much variance of the response can be explained by the predictors. A high ratio $cv(x) = std(x)/mean(x)$ suggests that the data are scattered, and therefore there may be practical non-identifiability. The asterisk denotes the cases such that $r^2 > 0.9$ and $cv > 0.1$.

Variable	**Enzyme**	**Groups**	**Enzyme groups**	r^2	**cv**
x_1	RuBisCO	x_1, x_3*	RuBisCO, GAPDH *	1.000	0.517
x_2	PGA kinase	$x_2, x_{10}*$	PGA kinase, PRK *	0.930	0.340
x_3	GAPDH	x_1, x_3, x_5*	RuBisCO, GAPDH, FBPase *	0.999	0.502
x_4	FBP aldolase	x_1, x_4*	RuBisCO, FBP aldolase *	0.981	0.365
x_5	FBPase	x_1, x_5*	RuBisCO, FBPase *	0.995	0.553
x_6	Transketolase type 1	x_6*	Transketolase type 1 *	0.982	0.536
x_7	SBP aldolase	x_1, x_7*	RuBisCO, SBP aldolase *	0.961	0.354
x_8	SBPase	x_8	SBPase	0.981	0.053
x_9	Transketolase type 2	x_1, x_9*	RuBisCO, Transketolase type 2 *	0.991	0.597
x_{10}	PRK	$x_1, x_{10}*$	RuBisCO, PRK *	0.970	0.464
x_{11}	ATP	$x_{11}*$	ATP *	0.979	0.678
x_{12}	ADPGPP	$x_{12}*$	ADPGPP *	0.983	0.315
x_{13}	PGCA Pase	$x_{13}*$	PGCA Pase *	0.982	0.535
x_{14}	Glycerate kinase	$x_{14}*$	Glycerate kinase *	0.975	0.804
x_{15}	Glycolate oxidase	x_{15}	Glycolate oxidase	0.976	0.020
x_{16}	GSAT	$x_{16}*$	GSAT *	0.974	0.351
x_{17}	Glycerol dehydrogenase	$x_{17}*$	Glycerol dehydrogenase *	0.975	0.999
x_{18}	GGAT	$x_{18}*$	GGAT *	0.968	0.469
x_{19}	GDC	x_{19}	GDC	0.987	0.039
x_{20}	Cytosolic FBP aldolase	$x_{20}*$	Cytosolic FBP aldolase *	0.972	0.369
x_{21}	Cytosolic FBPase	$x_9, x_{21}*$	Transketolase type 2, Cytosolic FBPase *	0.909	0.445
x_{22}	UDPGPP	x_{22}	UDPGPP	0.983	0.050
x_{23}	SPS	$x_{23}*$	SPS *	0.969	0.208
x_{24}	SPP	$x_{24}*$	SPP *	0.969	0.584
x_{25}	F26BPase	$x_{25}*$	F26BPase *	0.974	0.690

although the IA applied to GAPDH has assigned x_5 to the same functional group of x_1 and x_3. This can happen when the variables taken into account are also practically non-identifiable (which is the case of these three enzymes, since their coefficient of variation (cv) is high).

The interdependent decision variables, which are non-identifiable, may be fixed at an arbitrary value in order to improve identifiability. Since the variables functionally related to the fixed variable change accordingly, the model's dynamical properties are not changed or restricted by the fixation.

5.2 Photosynthesis in *Chlamydomonas reinhardtii*

In order to analyze the photosynthetic capability of *C. reinhardtii*, a multi-objective optimization has been performed. Instead of the concentration values considered in the carbon metabolism work discussed in the previous section, in this case I took into account the genes as decision variables to optimize, and in particular their presence or not in the metabolic network.

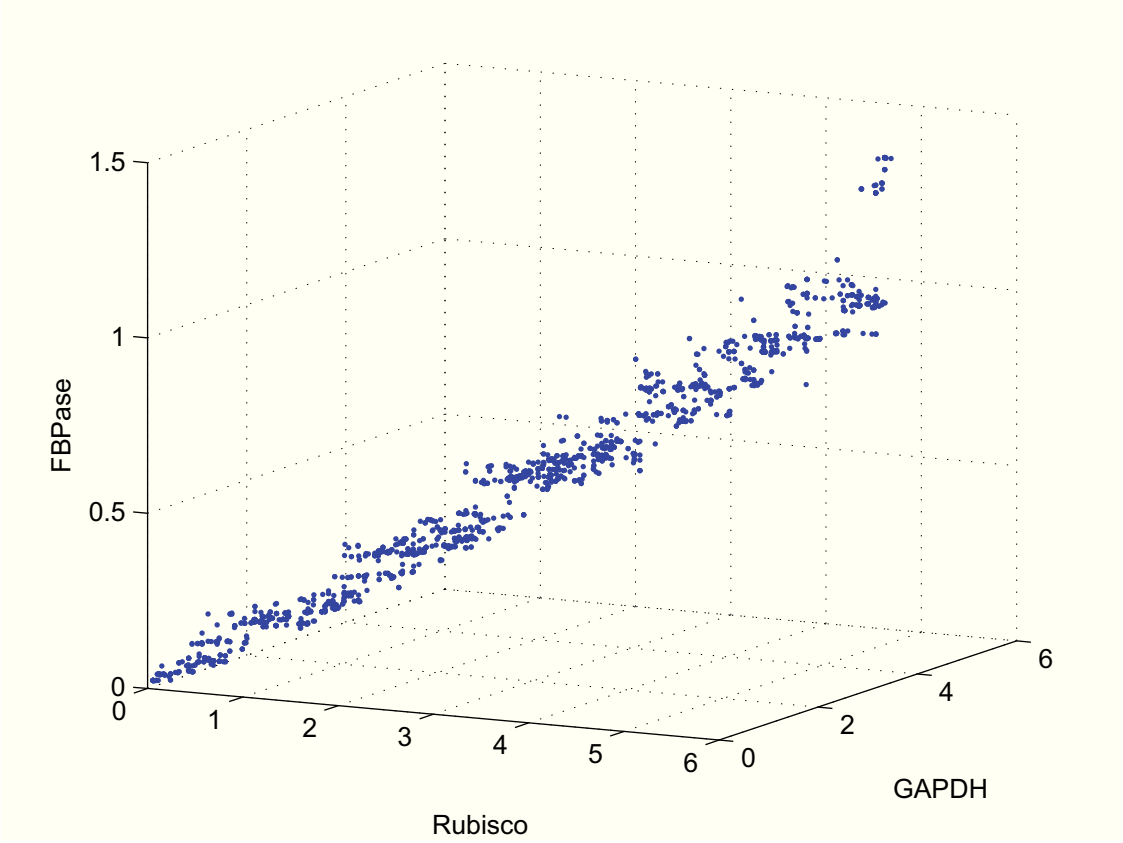


FIGURE 5.4: The plot shows the functional relation among the three decision variables RuBisCO, GAPDH and FBPase, thus highlighting the structural non-identifiability of these variables. This group has been detected for the GAPDH enzyme.

The model of *C. reinhardtii* is represented by using FBA framework. I set the maximum number of knockout allowed equal to 10 and considered both light and dark conditions. In Chang et al.[62], eleven windows of light spectrum can be chosen. Here, I used *Solar lithosphere spectrum*, that is the result of a composite analysis from several measurements taken from different locations under cloudless conditions in the 48 contiguous U.S. states and multiple data normalization procedures. In light conditions, I chose to minimize the CO_2 production (dual of maximization of CO_2 uptake rate) and the autotrophic biomass; then I performed the ϵ-dominance analysis. Figure 5.6 shows the results. In these conditions, the minimal CO_2 production is equal to $-6.7331 mmol h^{-1} gDW^{-1}$ (that corresponds to a maximal CO_2 uptake of $6.7331 mmol h^{-1} gDW^{-1}$) with a biomass formation equals to $0.1381 h^{-1}$ (Figure 5.6-C, red points). The organism is not able to absorb CO_2 from the atmosphere in dark conditions, indeed the CO_2 values in Figure 5.6-C in black points are positive, meaning only a production. The first two Figures (A,B) show the Pareto fronts and the related ϵ-dominance analysis in light and dark conditions, respectively. The ϵ-dominance analysis is a relaxed condition of dominance to select the Pareto optimal points observed by the optimization algorithm. In fact, if I consider the non-relaxed condition of dominance, some interesting solutions may be

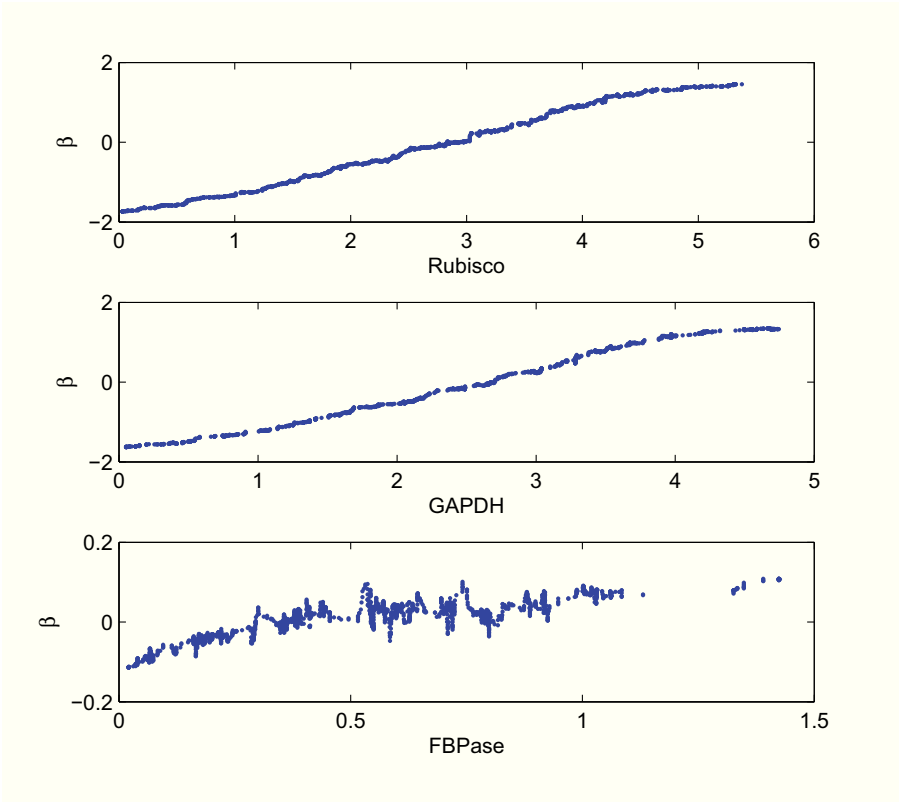


FIGURE 5.5: The plot shows the optimal transformations β (y axis) found for the three decision variables RuBisCO, GAPDH and FBPase (x axis). Although FBPase has been once assigned to the same functional group of RuBisCO and GAPDH, it shows a slightly different and noisier behavior.

TABLE 5.3: Global (GR) and Local (LR) Robustness Analysis in *C. reinhardtii* for two- and three-objective optimizations in light and dark conditions. For the LR values I report only the minimum and the related reactions.

Strain	CO_2	Biom	H_2O	GR (%)	LR (%)
Wild type (light)	285	6.15	n.a.	44.2	98(pyruvate transport by free diffusion, chloroplast) 72 (ammonia exchange) 54 (nitrate exchange)
Strain (light)	-6.733	0.138	n.a.	37.4	42 (na1 exchange)
Wild type (dark)	282	6.15	n.a.	43.52	69.23 (ammonia exchange) 61.53 (nitrate exchange)
Strain (dark)	0.286	2.492	n.a.	37.71	61.53 (nitrate exchange)
Wild type (light)	285	6.15	-10	43.8	98 (pyruvate transport by free diffusion, chloroplast) 72 (ammonia exchange) 70 (nitrate exchange)
Strain (light)	5.402	0.179	0.545	37.2	36 (nitrate exchange)

discarded although dominated by a small amount. The blue points belong to the Pareto optimal points obtained from a non-relaxed condition of dominance. With a relaxed condition, other acceptable solutions are added (in purple, red and green points).

Furthermore, Table 5.3 presents the robustness analysis results. I perturbed the metabolic fluxes. In particular, in the global robustness (GR) the perturbation is carried out simultaneously for all fluxes (rates of the reactions) of the network to evaluate the fragility

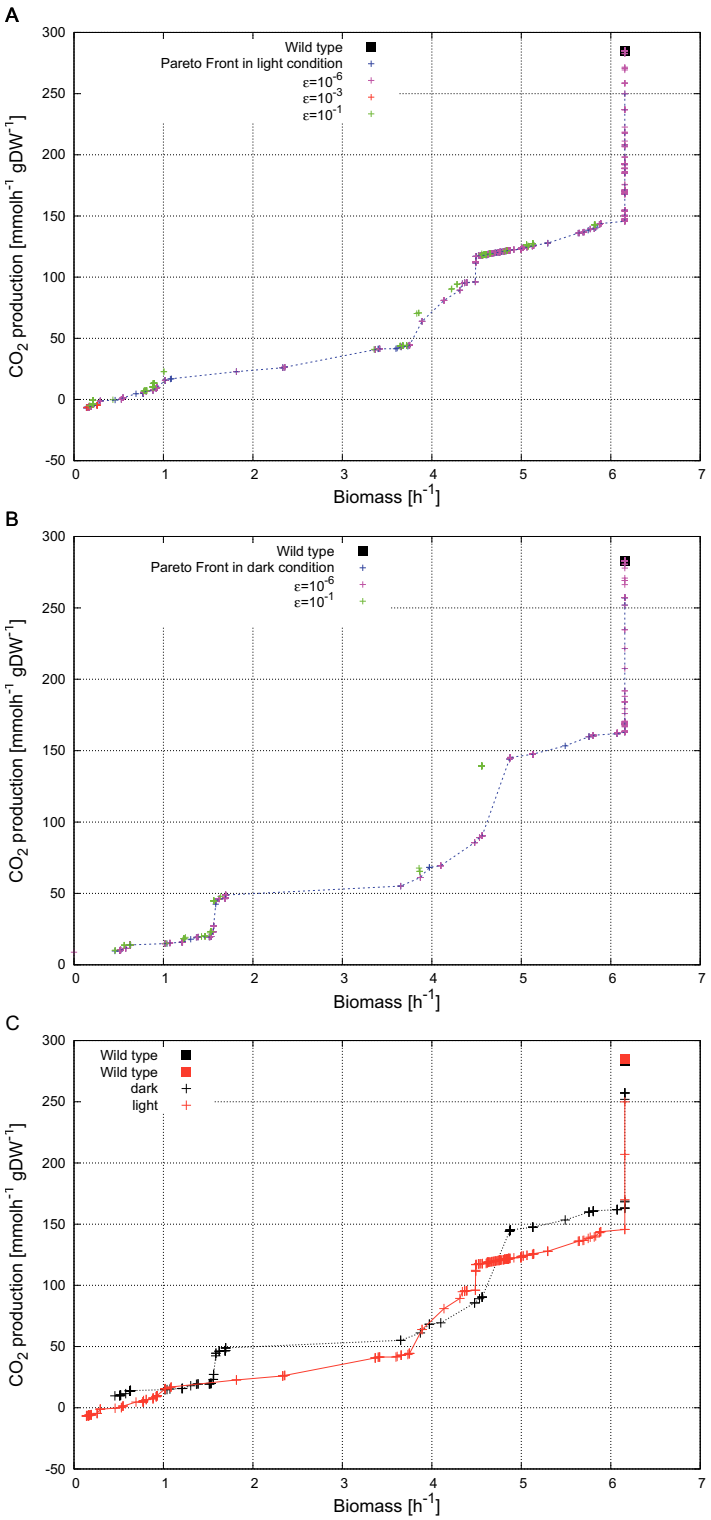


FIGURE 5.6: Minimization of CO_2 production (corresponding to a maximization of CO_2 uptake) and biomass formation in *C. reinhardtii* in light and dark condition and ϵ-dominance analysis with different ϵ values. The last Figure (C) shows a comparison between the Pareto fronts in A and B.

of the complete organism with respect to the metrics that are, in this case, the two objective functions. In the local robustness (LR), the perturbation is carried out for each flux (so, to have a robustness index for each flux). I selected from the Pareto front only one strain (one non-dominated solution), and compared it to the wild type (without knockout). I chose the strain that minimize the CO_2 production. If we consider the robustness index for each flux, the strain has one minimum, while the wild types in light and dark conditions have three and two minima, respectively, and the related fluxes are: the pyruvate transport by free diffusion (chloroplast), the nitrate exchange and the ammonia exchange in light condition and the nitrate exchange and the ammonia exchange in dark condition. In the same environmental conditions, but in light condition, in Figure 5.7 I chose to minimize the CO_2 production and simultaneously maximizing the H_2O production and autotrophic biomass. In this case, the algorithm found a minimum CO_2 production equal to 5.6268 $mmolh^{-1}$ gDW^{-1} with a biomass formation equal to 0.195 h^{-1} and the H_2O production equals to 9.8360 $mmolh^{-1}$ gDW^{-1}. A more interesting result is the good trade-off between the minimization of CO_2 and the maximization of H_2O production. In this case, a CO_2 production equal to 5.4024 $mmolh^{-1}$ gDW^{-1} has been obtained with a biomass formation equal to 0.179 h^{-1}, while the H_2O production is equals to 0.5455 $mmolh^{-1}$ gDW^{-1}. Furthermore, I performed the robustness analysis considering the three metrics (the three objective functions). The results are reported in Table 5.3. I chose the strain that obtains a good trade-off between the minimization of CO_2 production and the maximization of H_2O production and compare it to the wild type. The results are similar to those of two-objective optimization, so adding H_2O production does not causes variation in the global and local robustness.

5.3 Photosynthesis in *Rhodobacter spheroides*

In order to maximize the CO_2 uptake rate and biomass formation in *R. spheroides* [61], I used BioCAD method to find the best knockout strategies with the minimum knockout cost. I considered the photoautotrophic condition, i.e., a poor environment, where the only carbon source is CO_2. The exchange allowable fluxes are: sulfate, phosphate, ammonia, CO_2, magnesium, hydrogen, nicotinate and photon (light). Figure 5.8 shows the results of the multi-objective optimization. In wild type, *R. spheroides* grows with a biomass rate equal to 0.986 h^{-1}, and absorbing CO_2 to 44.705 $mmolh^{-1}$ gDW^{-1}. *R. spheroides* is able to absorb 57.452 $mmolh^{-1}$ gDW^{-1}, but reducing its growing to 0.418 h^{-1}, with a knockout cost equal to 14. The strain that absorbs until 44.7048 $mmolh^{-1}$ gDW^{-1} of CO_2 represents the tradeoff design, with a knockout cost equal to 8, turning off the following gene sets: RSP2138, RSP0361 or RSP2252, RSP0359, RSP0829,

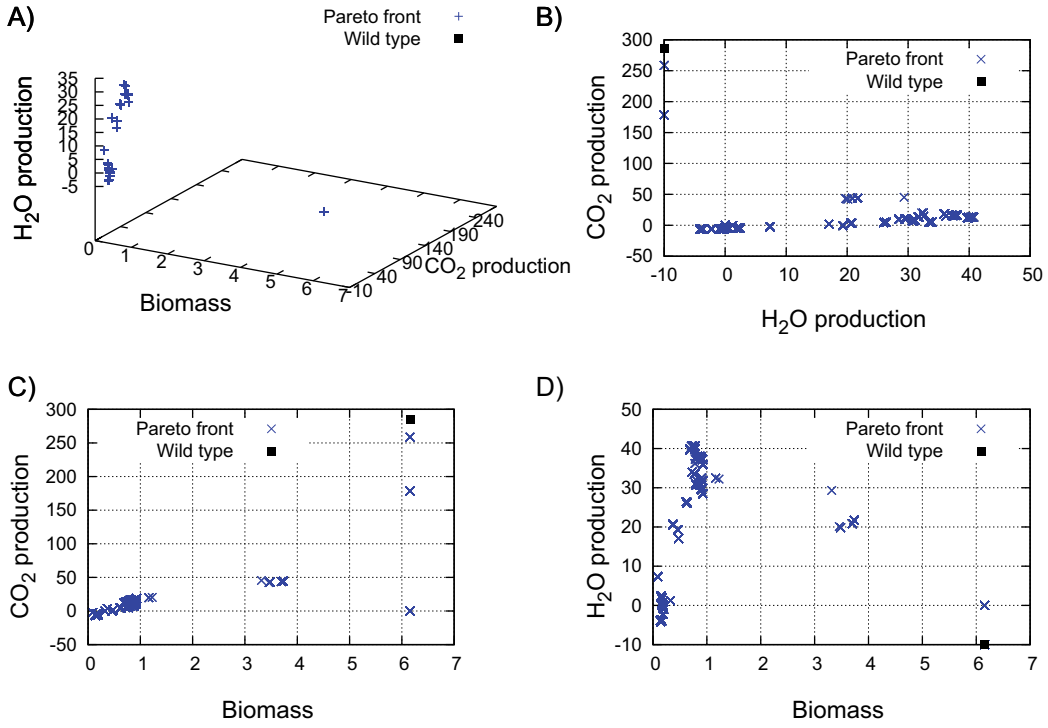


FIGURE 5.7: Simultaneous maximization of CO_2 uptake (corresponding to minimize CO_2 production), biomass formation and H_2O production in *C. reinhardtii*. I considered the photoautotrophic condition using the combinatorial optimization for searching gene knockout strategies. In (A) Pareto Front (blue points) obtained by the three-multi objective optimization. The results for each pair of two objective functions are shown in (B)-(C)-(D). Points in black indicate the amount of CO_2, H_2O production and biomass in wild type, i.e., without gene knockouts.

RSP3330 or RSP0656, RSP3142. In this configuration, six reactions are deleted: fumarate hydratase, L-serine ammonia-lyase, ribose-5-phosphate isomerase A, lactate dehydrogenase, sodium/sulfate symporter and acetate via Na+ symport. I performed other simulations and optimizations for *R. spheroides* in various photoautotrophic conditions. I optimized (i) biomass vs. H2O production, (ii) biomass vs. O_2 production and (iii) biomass vs. ethanol production. For all these experiments, the multi-objective optimization has identified only a Pareto solution very close to the wild type solutions. This could mean that in photoautotrophic conditions, the organism uses a metabolic pathway that is essential for its growth, and knockout genes are not feasible. I found H_2O production of 184.589 mmolh^{-1} gDW^{-1} with a biomass formation of 0.986 h^{-1}, and O_2 production of 1.2265×10^{-13} mmolh^{-1} gDW^{-1} and biomass 0.0099 h^{-1}. Conversely, in photoautotrophic conditions, *R. spheroides* does not produce ethanol, and even if I turn off genes, the result is always equal to zero for ethanol production. This means that ethanol is completely consumed in the metabolic network of the organism during its growth.

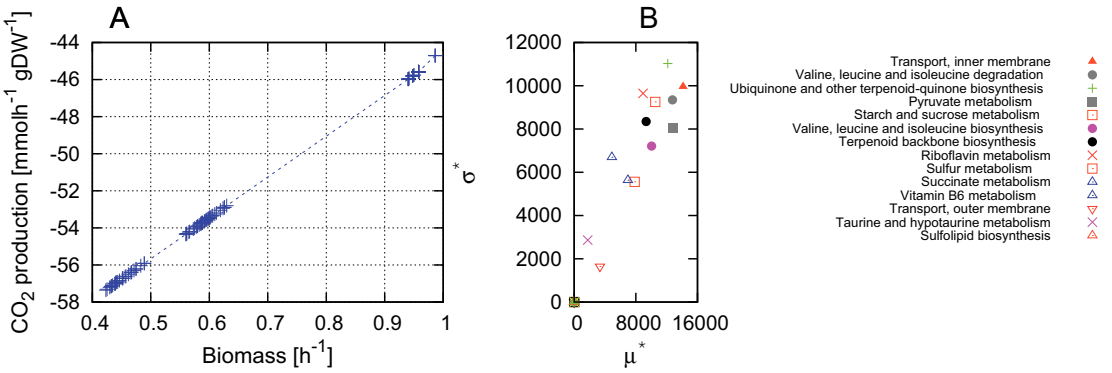


FIGURE 5.8: Results obtained by Sensitivity and optimization for *R. spheroides*. (A) Pareto front obtained maximizing biomass formation and minimizing CO_2 production (equivalent to maximize the CO_2 uptake rate) in *R. spheroides* using the multi-objective optimization to search for genetic knockout strategies in photoautotrophic conditions. Negative value of production correspond to positive values for uptake rate. (B) Pathway-oriented Sensitivity analysis for *R. spheroides*. The model includes 63 pathways, and only 14 pathways result to have sensitivity indexes greater than zero.

Figure 5.8-B reports the results of the Gene sets Pathway-oriented sensitivity analysis (PoSA), and the pathways which have sensitivity indexes greater than zero. Only 14 pathways (out of 63 pathways) are found to be sensitive, probably because in photoautotrophic conditions only the genes in these pathways have influence on the growth and metabolism of *R. spheroides*.

Furthermore, in Table 5.4 are reported the robustness analysis results. Similarly as in *C. reinhardtii*, the method acts perturbing the metabolic fluxes and calculating both the global and local robustness. I selected two strains from the Pareto front, and I compared them with the wild type. I chose the strains with good trade-off between the maximization of biomass formation and CO_2 uptake rate.

TABLE 5.4: Global (GR) and Local (LR) Robustness Analysis for *R. spheroides*.

	CO_2	Biom	GR (%)	LR (%)
Wild type	-44.705	0.986	38.65	53.84(Ammonia exchange) 53,84(Hydrogen exchange)
Rb1	-57.452	0.418	36.77	46.15(Hydrogen exchange)
Rb2	-44.708	0.9861	37.11	38.46(Hydrogen exchange)

5.4 Discussion

In this work, I used BioCAD methodologies for analyzing and cross comparing metabolic models. I analyzed in particular three metabolic networks because of their biotechnological and basic science importance. I adopted single and multi-objective optimization

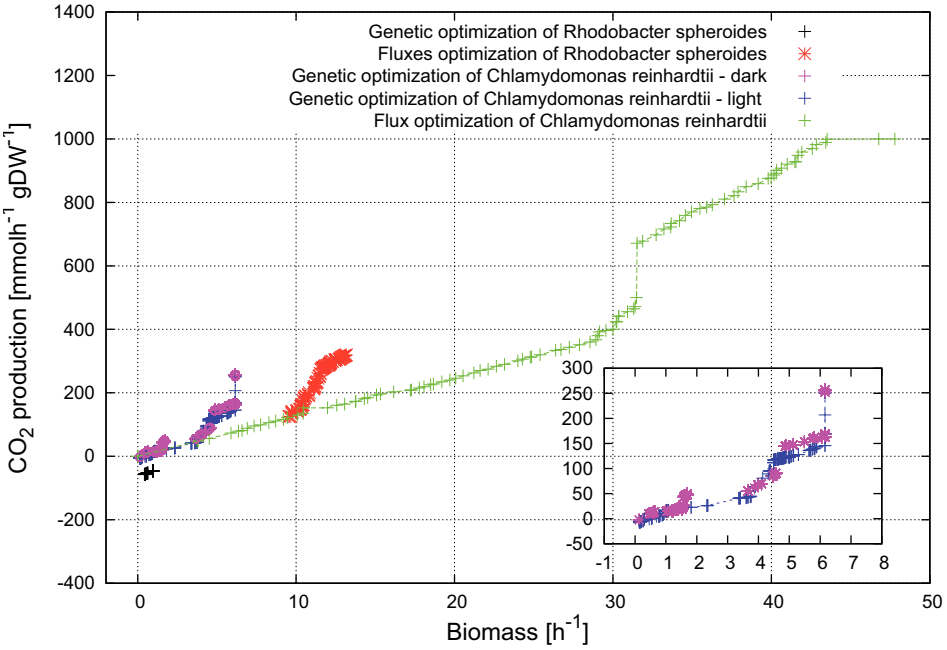


FIGURE 5.9: Pareto optimality results for *R. spheroides* and *C. reinhardtii*. Pareto fronts allow us to understand the behavior of a biological system by linking variables and responses. For instance, Genetic design provides different fronts according to the phenotypic response when genotype is mutated. In Figure we can observe that the genetic design in *R. Spheroides* does not provide a large Pareto front, probably due to the small network or unfavorable external conditions.

algorithms and I focused both on finding optimal knockout strategies, external environments or concentration enzymes for biotechnological or basic science purposes. The Pareto-optimality analysis is a useful tool for simulating biochemical pathways when contrasting objectives have to be considered simultaneously. By using this analysis, I found that *R. spheroides* is able to absorb an amount of CO_2 until 57.452 $mmolh^{-1}$ gDW^{-1} with a knockout cost equal to fourteen, deleting six reactions, while *C. reinhardtii* obtains only a CO_2 uptake rate value equal to 6.7331 in light condition.

The application of this analysis to the Calvin Cycle provided the best solution for the maximization of the CO_2 uptake rate and the minimization of the total nitrogen. The analysis was also used in order to understand which enzymes are the most important in CO_2 uptake rate and those whose modification is more robust, that is less prone to concentration fluctuation. This target is a fundamental biotechnological target, since it is not possible to engineer all the enzymes levels simultaneously and it is not currently possible even to work on transcription promoters so finely to obtain a completely definite final enzyme concentration in vivo. The finding of a limited number of targets (enzymes) sufficiently robust to obtain a working solution even in case of concentration fluctuations could lead to modified organisms whose activity could be better predicted. Furthermore, the optimization allowed to analyze the scenario foreseen for the end of the century,

when the atmospheric CO_2 will be much higher than nowadays, with an esteemed c_i of 490 μmol mol^{-1}. This simulation was carried out considering a case with minimal nitrogen availability and that with highest CO_2 uptake. Such simulation could foresee the response of the photosynthetic organisms to the increase in CO_2 concentration and the increase on agriculture productivity even with lower amount of available nitrogen. This approach is also useful to compare the genetic and metabolic ability of the organisms to absorb carbon dioxide by analyzing the Pareto features (Figure 5.9).

By using the sensitivity and robustness analysis, I have identified the most sensitive and fragile components of the biological systems I took into account. In *R. spheroides* I show that only fourteen pathways are sensitive, probably because in photoautotrophic conditions only the genes in these pathways have influence on the growth and metabolism. In *Chlamydomonas reinhardtii* alga, the method found that a flux perturbation of the reactions pyruvate transport by free diffusion (chloroplast), nitrate exchange or to ammonia exchange highlights the fragility of the organism with respect to the metrics chosen (CO_2 and biomass formation). The same behavior is shown by the *Rhodobacter spheroides* with respect to the ammonia or hydrogen exchange reactions.

In order to group enzymes according to functional relations, I applied the identifiability analysis (IA) to the chloroplast model. This approach allows to detect structural non-identifiability, i.e., some components of the model that cannot be determined unambiguously. The IA showed that RuBisCO, GAPDH and FBPase belong to the same functional group, i.e., they are interdependent decision variables. Interestingly, this bears out the results of the sensitivity analysis, which positioned these three enzymes in the most sensitive group of enzymes for the maximization of CO_2 consumption and the minimization of nitrogen consumption.

5.5 Material and Methods

The algorithms presented in this work include single and multi-objective optimization in a continuous or discrete research space. The algorithms are based on the evolutionary concept, where the solutions are calculated, compared and selected in each iteration/-generation of the algorithm. In particular, for the single-objective optimization the Differential Evolution (DE) Algorithm [77] was implemented, while for the multi-objective optimization I used two algorithms inspired by the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [25]: the Parallel optimization Algorithm (PAO), which performs optimization for continuous variables, and the Genetic Design through Multi-objective optimization (GDMO), which performs a combinatorial optimization search (described in section 2.4 in Chapter 2). NSGA-II [25] is a multi-objective evolutionary algorithm

(MOEA) designed to assure an efficient and effective approximation of the Pareto optimal set; NSGA-II belongs to the class of the evolutionary algorithms, which has been exploited by introducing the concept of non-dominated sorting and diversity preservation. The NSGA-II has been extended using an island-based model for parallel optimization; the new algorithm, called Parallel Algorithm for optimization (PAO) [78] performs parallel optimizations and swaps non-dominated solutions every given number of iterations. Decision variables have a continuous domain and, in this work, I considered the enzyme concentrations or the uptake rate of same metabolites that enter in biological systems.

In a multi-objective optimization problem, when the objectives functions are in conflict with each other, the output is a set of non-dominated solutions, called Pareto Front (section 2.3, Chapter 2). Pareto optimality proves very useful for bio-design automation, because it allows our method to obtain a wide range of optimal solutions and also the *best trade-off design*.

NSGA-II is characterized of four main steps. In a first step, a starting population is *initialized*. A population is formed by a set of individuals, each of which is represented by a decision variables set (whose values are chosen randomly or by the user) and the objective functions values obtained by using the corresponding decision variables. Decision variables are parameters of the system that we want to optimize. The value of the objective functions is strictly linked to the decision variables values. An individual represents a feasible solution. Once the first population is initialized, the algorithm enters in an evolutionary loop. A new population is created and updated for each iteration of the algorithm. Each iteration, called also *generation*, has the aim to improve the solution set and optimize the decision variables values, incorporating the evolutionary concept of Darwin. According to Darwin, the individuals of the population, from generation to generation evolve and only the best individuals survive. The same concept is incorporated in the evolutionary/genetic algorithm. By using the *crossover* and *mutation* operators new individuals are formed, and only the best individual are *selected* and inherited. An individual is better than another if the latter is dominated with respect to the first one. The loop terminates when a maximum generation number is reached, or when a particular solution is found.

Parallel Algorithms for optimization (PAO) are algorithms that exploit coarse-grained parallelism to let a pool of solutions exchange promising candidate solutions in an archipelago fashion. Using evolutionary operators such as recombination, mutation and selection, the framework completes with migration its approach based on islands. Each island is a virtual place where a pool of solutions is let evolve with a specific optimization algorithm; communications among islands in terms of solutions evolved by potentially different algorithms are arranged through a chosen archipelago topology. The island

TABLE 5.5: Methods and characteristics of the metabolic networks used for the analysis of the artificial photosynthetic organisms.

	Photosynthetic CM [60]	*R. sphaeroides* [61]	*C. reinhardtii* [62]
Modeling	ODEs	FBA-GPR	FBA-GPR
Reactions	39	1158	2190
Metabolites	38	796	1068
Enzymes	38	595	718
Genes	n.a.	1095	1080
Pathways	3	63	93
Optimization	PAO – NSGA II	GDMO	GDMO
Sensitivity	Morris [28] and Sobol' [28]	PoSA	Morris [28]
Robustness	GR/LR	GR/LR	GR/LR

model outlines an optimization environment in which different niches containing different populations are evolved by different algorithms and periodically some candidate solutions migrate in another niche to spread their building block. In this archipelago approach different topologies choices can raise to completely different overall solution introducing then another parameter that has to be chosen for each algorithm on each island. The PAO framework actually encloses two optimization algorithms and many archipelago topologies. The adopted configuration has two islands with two optimization algorithms, the Advanced CMA-ES algorithm (A-CMA-ES) and Differential Evolution algorithm (DE) [77], that exchange candidate solutions every 200 generations with an all-to-all (broadcast) migration scheme at a 0.5 probability rate. Even in its simplest configuration, this approach has shown enhanced optimization capabilities and an optimal convergence. After this phase, the NSGA-II multi-objective optimization algorithm has been used to tackle the problem relaxing the natural constraint about the fixed amount of protein-nitrogen. The goal is now to optimize two conflicting objectives, that are, to maximize the CO2 uptake and at the same time to minimize the total amount of protein-nitrogen. A-CMA-ES introduces a set of cut-off criteria to CMA-ES [79] and ensures with a constraint, a lower bound, for each enzyme concentration to be compatible with the smallest concentration observed in the natural leaf. In the case of biological networks modeled with ODEs, that is the case of photosynthetic carbon metabolism [60], the enzyme concentration values are optimized in each iteration/generation of PAO until a fixed number of generations is reached. The model is implemented in Matlab and the ODEs set is solved through the Matlab function ode15s. In the case of biological networks solved with FBA (*R. spheroides* and *C. reinhardtii*), the optimal genetic manipulations are searched through GDMO (section 2.4, Chapter 2) and models are implemented in Matlab and the FBA problem is solved by means of glpk [1].

[1] Gnu linear programming kit, version 4.47. http://www.test.org/doe/

In Table 5.5 the mathematical modeling approaches adopted for each biological systems here discussed are reported together with the number of reactions, metabolites, enzymes and genes. The photosynthetic carbon metabolism is modeled through a set of ODEs. I considered the model proposed by Zhu et al. [60]. The model takes into account rate equations for each discrete step in photosynthetic metabolism, equations for conserved quantities (i.e., nitrogen concentration) and a set of ODEs to describe the rate of concentration change in time for each metabolite (for a total of 31 differential equations). The reactions introduced in the model were categorized into equilibrium and non-equilibrium reactions; equilibrium reactions were inter-conversion between Glyceraldehyde 3-P (GAP) and Dihydroxyacetone-P (DHAP) in stroma and cytosol, xylulose-5-P (XuP5), Rib-5-P (Ri5P), ribulose-5-P (Ru5P) and Fru-6-P (F6P), Glc-6-P (G6P), and Glc-1-P (G1P). All non-equilibrium reactions were assumed to obey Michaelis-Menten kinetics, modified as necessary for the presence of inhibitors or activators.

R. spheroides, *C. reinhardtii* genome-scale metabolic networks were investigated through FBA, that is a widely used approach for studying biochemical networks. These network reconstructions contain all of the known metabolic reactions in an organism and the genes that encode each enzyme. FBA calculates the flow of metabolites through this metabolic network, thereby making it possible to predict the growth rate of an organism or the rate of production of a desired metabolite.

Chapter 6

Comparative Analysis and Design of Mitochondrial Metabolism

The bioenergetic activity of mitochondria can be thoroughly investigated by using computational methods. In particular, in this work I focus on ATP and NADH, namely the metabolites representing the production of energy in the cell. I used the BioCAD computational framework to perform an exhaustive investigation at the level of species, reactions, genes and metabolic pathways. I have considered three case studies related to the human mitochondria modeled by using different mathematical approaches, formally described with algebraic differential equations or flux balance analysis. Additionally, studies on fumarate deficiency, Ca^{2+} variation and cancer condition in mitochondria have been also conducted revealing interesting results.

6.1 Role of mitochondria in energy production and their influence in human diseases

Mitochondria are the organelles of the eukaryotic cells that play a pivotal role in the bioenergetics and regulation of many signaling pathways. They are fundamental also for the evolution of complex organisms. Specifically, mitochondria are optimized symbiotic cells useful to produce energy while simultaneously being energy-saving organelles. As a result, eukaryotic cells are able to synthesize more proteins than the prokaryotic cells (such as bacteria). Recent studies confirmed that mitochondria descend from bacteria, and indeed they lived outside the cell [80]. During the evolution, mitochondria entered in the animal and plant cells [81].

Mitochondria are important firstly for their energy productivity: they are the energy source of the cell, since they synthesize adenosine triphosphate (ATP), the chemical energy in the cell. Moreover, the mitochondrion is the site of other metabolic processes as well as carbohydrates metabolism, fatty acid oxidation and urea cycle.

The expansion of the fields of mitochondria and other mitochondrion-like organelles is mainly due to the identification of the pivotal role that mitochondria play in human disease and ageing [82], to the synergy showed by chloroplasts and mitochondria in energy output [83], and to the discovery of novel factors involved in organelle division, movement, signaling and adaptation to varying environmental conditions [84].

In the carbohydrates metabolism, the pyruvate produced from glycolysis undergoes oxidative decarboxylation to acetyl CoA, which is then oxidized in an eight-step process known as the tricarboxylic acid (TCA) cycle. The respiratory substrates NADH and FADH2 generated through the TCA cycle are then oxidized in a process coupled with ATP synthesis. Electrons are transferred from NADH and FADH2 to oxygen via enzyme complexes located on the inner mitochondrial membrane. Three of the electron carriers (complexes I, III and IV) are proton pumps, and couple the energy released by electron transfer with the translocation of protons from the matrix side to the external side of the inner mitochondrial membrane. The energy stored in the resulting proton gradient (i.e., the proton-motive force) is used to drive the synthesis of ATP via the mitochondrial enzyme ATP synthetase (complex V). Under certain conditions (e.g., fasting), acetyl CoA molecules are converted into ketones for use as an alternative energy source (fatty acid oxidation). In the urea cycle, amino acid degradation resulting in excretion of nitrogen as urea occurs partly in the mitochondrion.

Additionally, mitochondrion is also essential for several other processes, including the regulation of calcium homeostasis and other inorganic ions, cellular differentiation, cell death (apoptosis), as well as the control of the cell cycle and cell growth [85]. According to the tissue, the number, the size and the shape of mitochondrion in a cell change. For example, in the cardiac muscle, where the primary role of the heart is to pump blood that requires an intensive aerobic activity, the cells have a large number of mitochondria and a large size. Mitochondria have been also detected as responsible for several human diseases, including mitochondrial disorders, cardiac dysfunction, and type 2 diabetes [86]. The mitochondrion plays a crucial role also in cancer and in neurodegenerative disorders (as Parkinson, Alzheimer or ALS).

For these reasons, many researchers focused their attention on mitochondrion study, developing mathematical models that can simulate its metabolism, and in particular the oxidative phosphorylation. In the recent work by Bazil et al. [87], 73 algebraic differential equations are implemented to model the mitochondrial bioenergetics, including

34 biochemical reactions. Here, I used this model to in silico analyze and find the metabolites that are the most important for optimizing the energy productivity, i.e., for maximizing ATP and NADH production in matrix space. I also conducted five different studies with five different matrix calcium concentrations. As introduced above, mitochondria regulate calcium homeostasis, that is strictly linked to ATP and NADH production [88]. Additionally, calcium is an important ion inside mitochondria and its concentration is fundamental to regulate functions and acts at several levels during the ATP synthesis. The dysregulation of the mitochondrial Ca^{2+} homeostasis is involved in many pathologies. For example, an accumulation of Ca^{2+} ions in the mitochondrial space can lead to an increased generation of ROS (reactive oxygen species) that alters the permeability of the inner membrane leading the cell to apoptosis. Additionally, the dysregulation of the Ca^{2+} homeostasis is involved in neurodegenerative diseases [89]. Data from literature demonstrate that mitochondria play a crucial role in neuronal cell survival [90]. ATP metabolism, Ca^{2+} homeostasis, NAD+, NADH and ROS are key players in the cellular mechanisms, and their alteration can lead to the cell death.

The mitochondrial model was also set to simulate the cancer state. Specifically, I modified three features that have been found to vary between healthy and cancer conditions: (i) hexokinase activity, (ii) membrane potential differential, and (iii) concentration of hydrogen ions. The model includes kinetic parameters useful to mimic regulatory effects such as activation of enzymes by protein kinases. Therefore, it describes in a detailed way many features of the biochemical reaction and enzymatic action. On the other hand, this complex mathematical description introduces some limits: (i) the number of reactions is weak, therefore the complexity of the network is not captured; (ii) solving the set of DAEs requires more computational effort, and solvers can only compute approximations that may not fully agree with the real behavior of the system.

For these reasons, in this work I take also into account a mitochondrial network solved through flux balance analysis (FBA), where the system is described considering a steady state for the metabolites involved in the mitochondrial metabolism [2]. The model is composed of a set of algebraic equations and does not contain kinetic parameters. This approach permits to handle large metabolic networks (also in some cases more than 2500 reactions, 2000 metabolites and 1400 genes). The FBA mitochondrial model here considered is composed of 423 reactions (including transformation and transport reactions) and 228 metabolites. The computational time to solve the problem with FBA is highly reduced.

6.2 Computational Bioenergetics in mitochondrial metabolism

The FBA mitochondrion model [91] contains 423 reactions and 228 metabolites. By using a multi-objective optimization algorithm I maximized ATP and NADH production. In order to measure the matrix NADH productivity, I added a reaction that represents the transport of NADH from the matrix to the external environment. The aim is to find the optimal environment and the optimal metabolism for mitochondria so as to increase their bioenergetic yield.

6.2.1 Optimization of the external environment

In this experiment, the decision variables I took into account are the 73 uptake fluxes. I searched for the best values of uptake rate fluxes, which can assume a maximum value of 1000 mmolh^{-1} gDW^{-1}. The optimization finds a single Pareto point that reaches the maximum amount of ATP (1000 mmolh^{-1} gDW^{-1}), without NADH production. In another optimization experiment, I maximized ATP production and simultaneously minimized NADH production. Results are shown in Figure 6.1. We can observe that ATP production grows more rapidly than NADH consumption. I initialized the input fluxes of mitochondrial model as described in the work by Smith et al. [91]. In these conditions (before the optimization), the ATP production is equal to 139.4264 mmolh^{-1} gDW^{-1}, while NADH is totally consumed in the metabolism, and the productivity is equal to 0. After the optimization, the maximum ATP production is equal to the maximum value, 1000 mmolh^{-1} gDW^{-1} corresponding to a NADH consumption of about 211 mmolh^{-1} gDW^{-1} (Figure 6.1).

In this experiment, the optimization does not consider the limitation of substrates (as glucose or oxygen) in the biological environment, so we can consider this analysis as an asymptotic study for investigating the potentiality of mitochondria. Indeed, the optimization algorithm searches for the optimal environmental conditions without considering that the glucose and the other elements of the environment in a real cellular context, can be limited. Each uptake flux can reach the upper bound value, i.e., 1000, that is usually not feasible. Indeed in a real context, such as that of a cell, glucose availability is limited. In a second experiment, in order to include the limited availability of elements of the environment, I changed the upper bound of each uptake flux. In this way, a real context is modeled, i.e., an environment where glucose, lactate and other elements are present in a limited amount. In the original work [91], the maximum uptake rate of the fluxes was limited as follows: oxygen to 19.8, arginine to 0.0068, lysine to 0.0298, proline

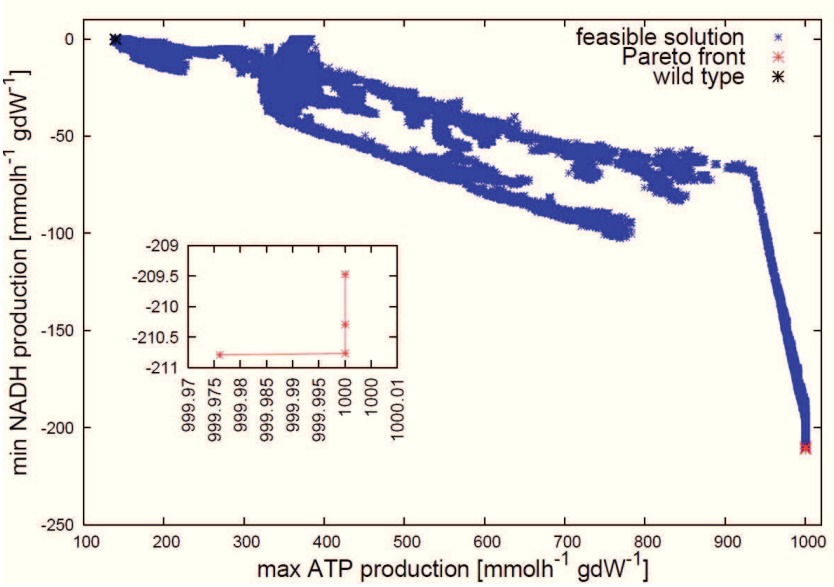


FIGURE 6.1: Maximization of ATP production and minimization of NADH production in the FBA mitochondrial model [91], carried out with 1000 individuals and halted at the 1500^{th} generation. The algorithm optimizes the uptake rate fluxes (73 exchange fluxes) to analyze the energy state of the mitochondrion. In blue the dominated feasible points, in black the wild type conditions, i.e., before optimization. The non-dominated Pareto points in red are also reported in the inset plot.

to 0.0044, aspartate to 0.1524, alpha-D-glucose to 0.9000, (R)-3-hydroxybutanoate to 0.7000, isoleucine to 0.0039, valine to 0.0106, hexadecanoic acid to 1.0000, (S)-lactate to 0.5750, HCO3- to 0.0198. These values have been validated using experimental data [91]. Therefore, in this experiment only the twelve fluxes cited above are optimized. Additionally, for each variable the domain space is constrained between 0 and +33% of the maximum uptake rate used by the authors and reported above. In this condition, NADH production does not increase. Specifically, I observed a consumption of NADH. ATP increases and, by considering only the solutions where NADH is positive, I found ATP= 185.4299 mmolh^{-1} gDW^{-1}. The optimization algorithm here used is the Non-dominated Sorting Genetic Algorithm II, also known as NSGA-II [25].

6.2.2 Optimization of the internal environment

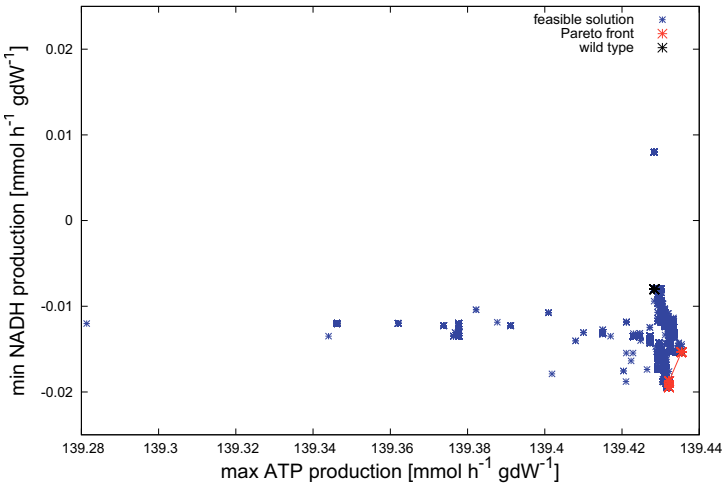
The optimization of the metabolic reactions in the FBA mitochondrial model has been performed using a genetic algorithm and a novel mutation operator. The genetic algorithm is inspired to NSGA-II with a new mutation heuristic. Since perturbing metabolic reactions (i.e., transformation fluxes) in FBA models may lead to unfeasible solutions, I have created a mutation operator that takes into account this issue. I have introduced two parameters, C and N. C is the maximum number of fluxes that can be perturbed and constrained. If the mutation results in unfeasible solutions due to the constraints

operating in the network, the procedure is repeated until a maximum number of N trials is reached, otherwise the current solution is maintained. Here, I have considered C equals to 5 and N to 10. I have performed the optimization of 229 transformation fluxes and conducted two experiments: (i) the simultaneous maximization of ATP and NADH production and (ii) the simultaneous maximization of ATP production and the minimization of NADH production. For both the experiments I have used a population of 1000 individuals. Each individual contains a vector of 229 values, and each value represents the rate of the corresponding metabolic flux. The mutation operator takes also into account the reversibility of the reactions. The algorithm has been performed until 300 generations. The results are shown in Figure 6.2. For both the experiments, the algorithm finds a set of non-dominated Pareto solutions, showed in red. Dominated and feasible solutions are shown in blue, while the point in black represents the wild type condition, i.e., the condition before the optimization.

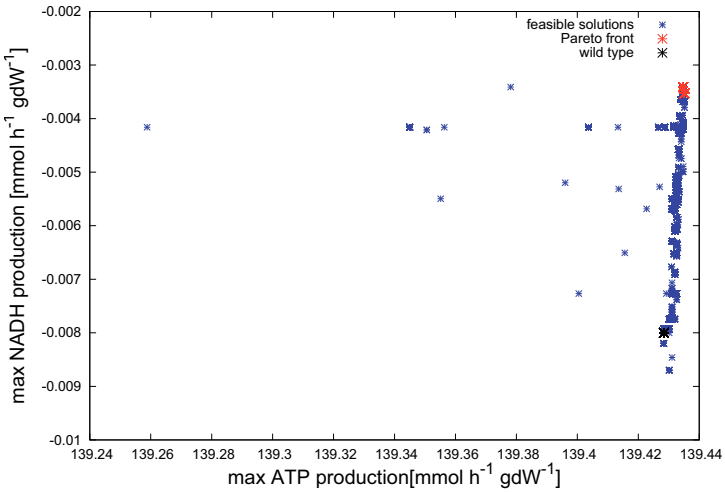
6.2.3 Identifiability Analysis to characterize mitochondrial monogenic diseases

The idea is that by performing the IA in healthy, pathological and disease conditions, we can characterize the onset of a disease by looking at the functional relations among fluxes. When taking into account a specific disease, I constrained the reaction responsible for that disease to various values and I evaluated the amount of ATP and NADH as outputs of the model. Starting from the work of Smith et al.[91] I defined a model condition as *disease status* if the ATP production is less or equal to 33% of the production under normal conditions, and *inflammation status* if the ATP production is less or equal to 66% but more than 33% of the production under normal conditions.

In this study, Identifiability Analysis (IA) is coupled with the flux balance analysis model of the mitochondrion [91]. In particular, flux balance analysis is performed to obtain optimal production of ATP and NADH while varying the flux of the fumarate hydratase, responsible for the fumarase deficiency, a monogenic disorder in the TCA cycle. I considered the fluxes of the 135 reactions occurring in the mitochondrial matrix space as functions of the fumarase flux, and inferred functionally related reactions using the alternating conditional expectation algorithm (ACE) [34]. The connection between the identifiability analysis and a constraint-based structure stems from the fact that a non-identifiable constraint involving decision variables causes them to be functionally related. Given the space of the feasible fluxes, IA has been applied in the regions with maximum ATP (healthy condition) and low ATP (pathological condition). The IA detects structural non-identifiable components of a model by fitting it repeatedly



A)



B)

FIGURE 6.2: (A) Simultaneous maximization of ATP and minimization of NADH and (B) maximization of ATP and NADH in mitochondrial metabolism in the FBA approach [91]. In blue feasible solutions and in black the wild type condition (before optimization). In red Pareto optimal solutions after optimizing internal fluxes.

to experimental data and by analyzing the estimates of each component. Details are reported in Chapter 2.

I chose the ATP and NADH productions as the objective functions to evaluate the distribution of fluxes in the network. The mitochondrial FBA model is composed of 423 reactions: 73 exchange reactions between the external environment and the mitochondrion, 135 reactions in the matrix compartment (e.g., the reactions of the Krebs cycle and beta-oxidation), and finally all the reactions that take place in the intermembrane space. Without constraining the fumarate flux, the maximum ATP production is reached when the fumarate flux is equal to 6.9721 $\text{mmolh}^{-1}\text{gDW}^{-1}$, as shown in Figure 6.3.

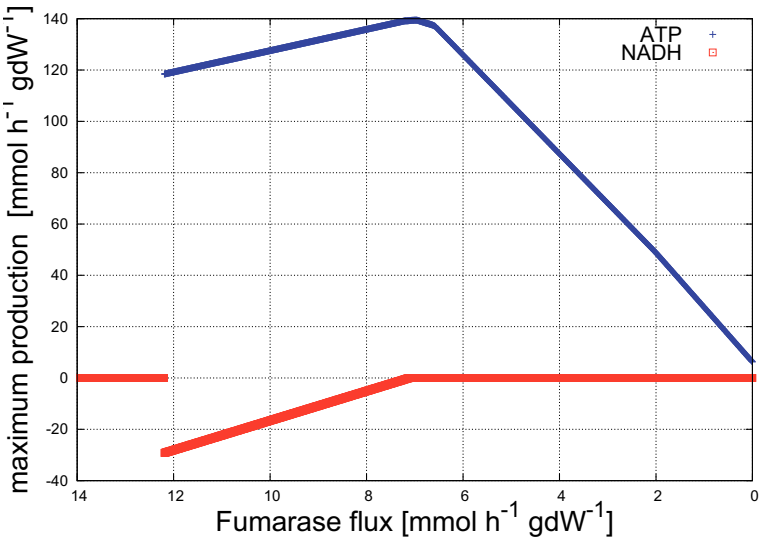


FIGURE 6.3: Maximum ATP and NADH production against different fumarate flux values.

As described by Smith et al. [91], in the fumarase deficiency conditions the ATP production is reduced until 75% of the maximum value. To evaluate the effect of the fumarate deficiency, I constrained the fumarate flux in the FBA model. I computed 2000 states of flux balance by adding constrains on the fumarase flux. In particular I constrained the reaction "Fumarate + $H_2O \longrightarrow$ (S)-Malate", firstly by imposing the knockout condition (i.e., forcing the flux of the reaction to be zero), and then increasing the flux by 0.014 $\text{mmolh}^{-1}\text{gDW}^{-1}$, until 13.986 $\text{mmolh}^{-1}\text{gDW}^{-1}$. Therefore, I obtained a 135×2000 matrix V containing all the matrix fluxes in the model corresponding to fixed fluxes of fumarase.

For the fumarase deficiency, I identified these intervals of the fumarase flux: healthy state $[10, 5.69]$, inflammation state $[5.69, 3.69[$, pathological state $[3.69, 0]$ mmolh^{-1} gDW^{-1}.

The process of repeating estimates in the matrix K is replaced by taking into account all the points given as output by the FBA run in different fumarase conditions. I adopted the Mean Optimal Transformation Approach (MOTA) [76], by fixing at 10 the maximal number of parameters allowed to enclose a functional relation. There are fluxes showing functional relations in a group of more than two elements, although in this case it is less likely to find strong relations among variables. A variable is detected in a group depending on the contribution strength of a predictor to the response. Each variable is considered once as response variable. As a result, if a functional relation is among k variables, it is tested k times, although it is unlikely that the same functional group is really detected all the k times. The r^2 column indicates how much variance of the response can be explained by the predictors. A high amount of explained variance of the response indicates a significant effect of the fixation of the predictors on the standard deviations of the response. The $cv(x) = std(x)/mean(x)$ helps to distinguish practical

identifiable from non-identifiable variables [76]. In case of practical non-identifiability, the choice of the value that needs to be fixed strongly depends on the experiments, also considering reference values in the literature. The interdependent fluxes, which are non-identifiable, may be fixed at an arbitrary value in order to improve identifiability. This would not affect the model's dynamical properties, as the variables functionally related to the fixed variable change accordingly.

I first applied the IA to detect global relations between two or more variables, allowing the fumarase flux to span all the interval $[0, 13.986]$ mmolh^{-1} gDW^{-1}. In the table summarizing the results (Table 6.1), the "flux groups" column indicates the functional relations between variables. For instance, R01361MM and R01978MM are functionally related. In other words, the response variable x_{47} is strongly related to the predictor x_{62}. In Figure 6.4 I plot the optimal transformations β found for these two reactions. We notice that the transformations are similar to each other, indicating the structural non identifiability of both variables. The functional relation between these two reactions (x_{47} and x_{62}) has been detected by the identifiability analysis applied to both reactions. This is therefore a strong relation, marked by a double asterisk in Table 6.1.

In Table 6.2, Table 6.3 and Table 6.4 have been shown the results of the IA applied to the metabolic network with various values of the fumarase flux. The (x_{28}, x_{50}) group is detected both in the healthy and in the pathological state, but not in the inflammation state. In the pathological state only four different functional relations are detected, which means that variables are mostly unrelated to one another.

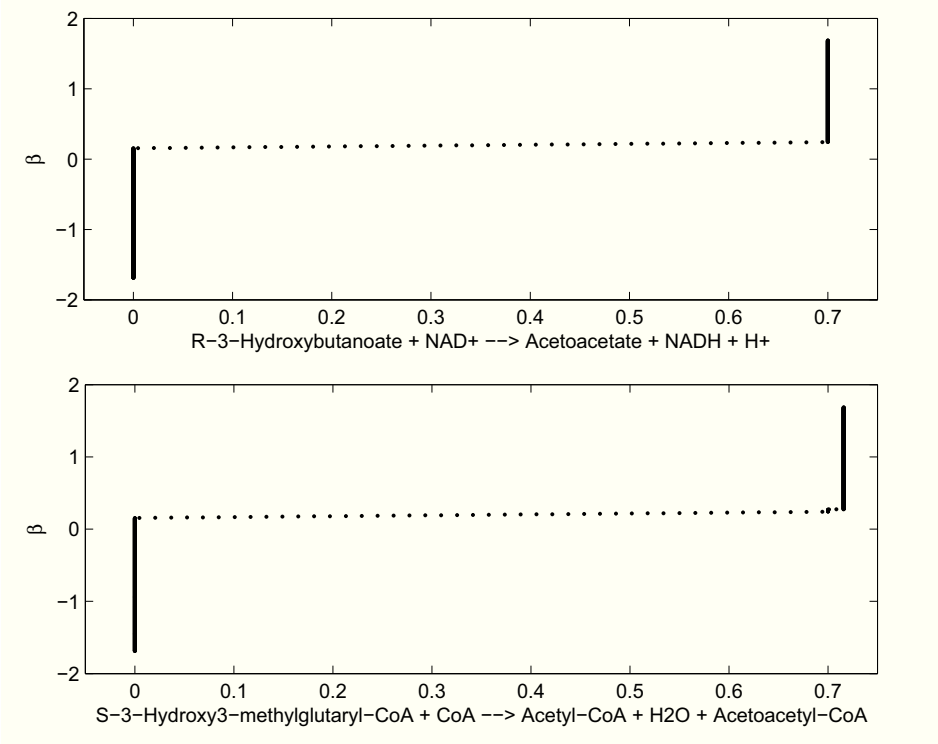


FIGURE 6.4: Optimal transformations β (y axis) found for the two fluxes R01361MM, representing the 3hydroxybutyrate dehydrogenase (top) and R01978MM, representing the hydroxy methylglutaryl–CoA synthase (bottom) (x axis) [mmolh^{-1} gDW^{-1}] in the mitochondrial FBA model [91]. This plot proves that there is a strong relation between these two fluxes, with slightly different and noisier behavior in the neighborhood of 0 and 0.7.

Variable	**Flux**	**Flux groups**	r^2	**cv**
x_1	$R00004MM$	$x_1, x_8, x_{12}, x_{27}*$	1.000	2.056
x_2	$R00014MM$	$x_2, x_{16}, x_{45}, x_{58}, x_{72}, x_{82}, x_{98}$	n.a.	0.047
x_3	$R00081MM$	x_3, x_{64}**	1.000	0.423
x_4	$R00086MM$	$x_4, x_{13}, x_{57}, x_{65}, x_{73}*$	1.000	0.404
x_5	$R00127MM$	$x_5, x_{19}, x_{87}, x_{117}, x_{118}*$	1.000	1.233
x_6	$R00157MM$	$x_6, x_{14}, x_{85}, x_{117}, x_{118}$	n.a.	1.233
x_7	$R00205MM$	$x_7, x_{21}, x_{91}, x_{109}, x_{128}$	n.a.	0.869
x_8	$R00238MM$	$x_8, x_{21}, x_{37}, x_{57}, x_{73}*$	1.000	0.884
x_9	$R00243MM$	$x_9, x_{40}, x_{75}, x_{94}, x_{131}*$	1.000	5.136
x_{10}	$R00245MM$	$x_{10}, x_{16}, x_{57}, x_{61}, x_{78}*$	1.000	0.489
x_{11}	$R00256MM$	$x_{11}, x_{68}, x_{86}, x_{90}, x_{98}, x_{109}$	n.a.	1.271
x_{12}	$R00258MM$	$x_2, x_{12}, x_{13}, x_{43}, x_{78}, x_{98}$	n.a.	0.190
x_{13}	$R00275MM$	$x_4, x_8, x_{13}, x_{52}, x_{65}*$	1.000	0.383
x_{14}	$R00330MM$	$x_8, x_{14}, x_{86}, x_{88}, x_{91}*$	0.999	3.909
x_{15}	$R00342MM$	$x_{12}, x_{15}, x_{52}, x_{66}, x_{108}, x_{134}*$	0.998	0.420
x_{16}	$R00351MM$	$x_8, x_{16}, x_{21}, x_{57}, x_{73}*$	1.000	0.556
x_{17}	$R00355MM$	$x_2, x_{17}, x_{72}, x_{82}$	0.999	0.051
x_{18}	$R00371MM$	$x_{18}, x_{91}, x_{105}, x_{128}*$	1.000	0.869
x_{19}	$R00388MM$	$x_{19}, x_{77}, x_{108}, x_{109}, x_{110}, x_{128}$	n.a.	0.304
x_{20}	$R00430MM$	$x_{14}, x_{20}, x_{134}*$	0.999	3.780
x_{21}	$R00432MM$	$x_8, x_{21}, x_{37}, x_{73}, x_{77}, x_{105}$	n.a.	0.558

Variable	Flux	Flux groups	r^2	cv
x_{22}	$R00512MM$	x_{22}, x_{129}**	0.999	2.648
x_{23}	$R00551MM$	$x_{23}, x_{88}, x_{107}, x_{111}$	1.000	0.000
x_{24}	$R00572MM$	$x_4, x_{13}, x_{24}, x_{65}, x_{73}$*	0.999	1.996
x_{25}	$R00667MM$	$x_{19}, x_{25}, x_{98}, x_{108}, x_{112}$	1.000	0.000
x_{26}	$R00705MM$	x_{26}, x_{94}*	1.000	41.821
x_{27}	$R00709MM$	$x_{12}, x_{15}, x_{27}, x_{45}, x_{57}, x_{108}$*	0.998	0.556
x_{28}	$R00713MM$	x_4, x_{12}, x_{28}*	0.981	2.246
x_{29}	$R00716MM$	$x_8, x_{29}, x_{45}, x_{90}, x_{98}$*	1.000	0.847
x_{30}	$R00740MM$	x_{30}, x_{40}*	1.000	41.821
x_{31}	$R00830MM$	$x_{31}, x_{40}, x_{75}, x_{76}, x_{94}, x_{130}, x_{131}, x_{135}$	n.a.	n.a.
x_{32}	$R00833MM$	x_{32}, x_{61}, x_{128}*	1.000	0.157
x_{33}	$R00851MM$	$x_{33}, x_{66}, x_{91}, x_{109}$*	1.000	2.648
x_{34}	$R00927MM$	x_{34}, x_{81}, x_{93}*	1.000	1.152
x_{35}	$R00941MM$	x_{35}, x_{51}**	1.000	0.867
x_{36}	$R00945MM$	$x_{19}, x_{29}, x_{36}, x_{88}, x_{105}, x_{111}$	n.a.	0.867
x_{37}	$R01082MM$	$x_{16}, x_{21}, x_{37}, x_{57}, x_{73}$*	1.000	0.578
x_{38}	$R01175MM$	$x_{38}, x_{68}, x_{98}, x_{104}, x_{108}, x_{110}$	n.a.	0.728
x_{39}	$R01177MM$	$x_{39}, x_{88}, x_{90}, x_{114}, x_{115}$*	1.000	0.841
x_{40}	$R01214MM$	$x_{31}, x_{40}, x_{75}, x_{76}, x_{94}, x_{130}, x_{131}, x_{135}$	n.a.	0.000
x_{41}	$R01218MM$	$x_8, x_{41}, x_{58}, x_{73}, x_{78}$*	1.000	1.196
x_{42}	$R01253MM$	$x_2, x_{12}, x_{13}, x_{42}, x_{58}, x_{65}, x_{82}, x_{134}$	0.999	0.000
x_{43}	$R01279MM$	$x_{43}, x_{90}, x_{104}, x_{106}, x_{108}$*	1.000	0.726
x_{44}	$R01280MM$	x_{44}, x_{53}**	1.000	1.582
x_{45}	$R01325MM$	$x_4, x_{27}, x_{45}, x_{57}, x_{134}$*	1.000	0.556
x_{46}	$R01360MM$	$x_8, x_{19}, x_{46}, x_{115}, x_{128}$*	1.000	1.112
x_{47}	$R01361MM$	x_{47}, x_{62}**	1.000	1.112
x_{48}	$R01624MM$	x_{48}, x_{97}**	1.000	1.582
x_{49}	$R01626MM$	x_{49}, x_{53}*	1.000	1.582
x_{50}	$R01648MM$	$x_{16}, x_{21}, x_{50}, x_{52}$*	0.987	2.956
x_{51}	$R01655MM$	x_{35}, x_{51}**	1.000	0.867
x_{52}	$R01700MM$	$x_{11}, x_{27}, x_{45}, x_{52}, x_{66}, x_{73}$	n.a.	0.558
x_{53}	$R01706MM$	x_{44}, x_{53}**	1.000	1.582
x_{54}	$R01799MM$	x_{24}, x_{54}*	0.999	2.648
x_{55}	$R01801MM$	$x_{43}, x_{55}, x_{117}, x_{118}$*	1.000	61.209
x_{56}	$R01859MM$	x_{56}, x_{67}*	1.000	1.152
x_{57}	$R01900MM$	$x_8, x_{16}, x_{27}, x_{45}, x_{57}$*	1.000	0.556
x_{58}	$R01923MM$	$x_{39}, x_{58}, x_{109}, x_{111}$*	1.000	0.726
x_{59}	$R01939MM$	$x_{39}, x_{45}, x_{59}, x_{79}, x_{108}, x_{112}, x_{128}$	n.a.	0.847
x_{60}	$R01940MM$	x_{60}, x_{70}*	1.000	0.798
x_{61}	$R01975MM$	$x_8, x_{61}, x_{66}, x_{78}, x_{104}$*	1.000	0.803
x_{62}	$R01978MM$	x_{47}, x_{62}**	1.000	1.112
x_{63}	$R02030MM$	$x_{15}, x_{63}, x_{105}, x_{114}, x_{116}$	n.a.	2.648
x_{64}	$R02161MM$	x_3, x_{64}**	1.000	0.423
x_{65}	$R02163MM$	$x_4, x_{13}, x_{21}, x_{65}, x_{66}$*	1.000	0.383
x_{66}	$R02164MM$	$x_8, x_{27}, x_{45}, x_{52}, x_{66}$*	1.000	0.558
x_{67}	$R02199MM$	x_{67}, x_{133}**	1.000	1.152
x_{68}	$R02241MM$	x_{24}, x_{68}*	0.999	2.648

Variable	Flux	Flux groups	r^2	cv
x_{69}	$R02313MM$	$x_{69}, x_{104}, x_{107}, x_{112}, x_{128}*$	1.000	0.847
x_{70}	$R02487MM$	$x_{19}, x_{70}, x_{90}, x_{113}*$	1.000	0.798
x_{71}	$R02529MM$	$x_{71}, x_{88}, x_{114}, x_{128}*$	0.999	0.869
x_{72}	$R02569MM$	$x_{2}, x_{12}, x_{15}, x_{37}, x_{58}, x_{65}, x_{72}, x_{82}, x_{98}$	n.a.	0.047
x_{73}	$R02570MM$	$x_{8}, x_{21}, x_{37}, x_{73}, x_{77}, x_{88}$	n.a.	0.558
x_{74}	$R02571MM$	$x_{8}, x_{74}, x_{78}, x_{109}*$	1.000	0.798
x_{75}	$R02661MM$	$x_{31}, x_{40}, x_{75}, x_{76}, x_{94}, x_{130}, x_{131}, x_{135}$	n.a.	0.000
x_{76}	$R02662MM$	$x_{31}, x_{40}, x_{75}, x_{76}, x_{94}, x_{130}, x_{131}, x_{135}$	n.a.	0.000
x_{77}	$R02765MM$	$x_{43}, x_{77}, x_{91}, x_{111}, x_{128}*$	0.999	1.152
x_{78}	$R03026MM$	$x_{21}, x_{37}, x_{52}, x_{61}, x_{62}, x_{78}$	n.a.	0.803
x_{79}	$R03102MM$	$x_{59}, x_{79}, x_{104}, x_{105}, x_{108}, x_{128}$	n.a.	0.847
x_{80}	$R03172MM$	$x_{19}, x_{39}, x_{80}, x_{98}, x_{112}*$	1.000	1.152
x_{81}	$R03174MM$	$x_{34}, x_{81}*$	1.000	1.152
x_{82}	$R03270MM$	$x_{2}, x_{16}, x_{45}, x_{58}, x_{72}, x_{82}, x_{85}$	n.a.	0.047
x_{83}	$R03314MM$	$x_{83}, x_{86}, x_{88}, x_{109}$	1.000	0.000
x_{84}	$R03381MM$	x_{84}, x_{135}	1.000	0.000
x_{85}	$R03777MM$	$x_{85}, x_{104}, x_{112}, x_{114}, x_{115}*$	1.000	0.726
x_{86}	$R03778MM$	$x_{86}, x_{90}, x_{107}, x_{114}, x_{116}*$	1.000	0.841
x_{87}	$R03857MM$	$x_{87}, x_{104}, x_{106}, x_{110}, x_{116}*$	1.000	0.726
x_{88}	$R03858MM$	$x_{88}, x_{106}, x_{107}, x_{109}, x_{113}*$	1.000	0.841
x_{89}	$R03990MM$	$x_{89}, x_{110}, x_{112}, x_{114}, x_{115}*$	1.000	0.726
x_{90}	$R03991MM$	$x_{90}, x_{107}, x_{109}, x_{111}, x_{113}*$	1.000	0.841
x_{91}	$R04170MM$	$x_{91}, x_{105}, x_{109}, x_{111}, x_{113}*$	1.000	0.841
x_{92}	$R04203MM$	x_{92}, x_{132}**	1.000	1.152
x_{93}	$R04204MM$	$x_{34}, x_{93}*$	1.000	1.152
x_{94}	$R04224MM$	$x_{31}, x_{40}, x_{75}, x_{76}, x_{94}, x_{130}, x_{131}, x_{135}$	n.a.	0.000
x_{95}	$R04355MM$	$x_{49}, x_{95}*$	1.000	1.582
x_{96}	$R04428MM$	x_{96}, x_{99}**	1.000	1.582
x_{97}	$R04430MM$	x_{48}, x_{97}**	1.000	1.582
x_{98}	$R04433MM$	$x_{98}, x_{108}, x_{112}, x_{113}*$	1.000	0.720
x_{99}	$R04533MM$	x_{96}, x_{99}**	1.000	1.582
x_{100}	$R04536MM$	$x_{49}, x_{100}*$	0.999	1.582
x_{101}	$R04537MM$	x_{101}, x_{123}**	1.000	1.582
x_{102}	$R04543MM$	$x_{102}, x_{103}*$	1.000	1.582
x_{103}	$R04544MM$	x_{103}, x_{125}**	1.000	1.582
x_{104}	$R04737MM$	$x_{104}, x_{110}, x_{112}, x_{113}, x_{115}*$	1.000	0.841
x_{105}	$R04738MM$	$x_{91}, x_{105}, x_{108}, x_{111}, x_{116}*$	1.000	0.841
x_{106}	$R04739MM$	$x_{104}, x_{106}, x_{108}, x_{112}, x_{115}*$	1.000	0.841
x_{107}	$R04740MM$	$x_{39}, x_{88}, x_{105}, x_{107}, x_{114}*$	1.000	0.841
x_{108}	$R04741MM$	$x_{104}, x_{108}, x_{110}, x_{112}, x_{115}*$	1.000	0.841
x_{109}	$R04742MM$	$x_{88}, x_{91}, x_{107}, x_{109}, x_{114}*$	1.000	0.841
x_{110}	$R04743MM$	$x_{90}, x_{106}, x_{110}, x_{112}, x_{115}*$	1.000	0.841
x_{111}	$R04744MM$	$x_{86}, x_{88}, x_{109}, x_{111}, x_{116}*$	1.000	0.841
x_{112}	$R04745MM$	$x_{91}, x_{108}, x_{110}, x_{112}, x_{115}*$	1.000	0.841
x_{113}	$R04746MM$	$x_{39}, x_{88}, x_{111}, x_{113}, x_{115}*$	1.000	0.841
x_{114}	$R04747MM$	$x_{88}, x_{91}, x_{107}, x_{111}, x_{114}*$	1.000	0.841
x_{115}	$R04748MM$	$x_{90}, x_{106}, x_{108}, x_{112}, x_{115}*$	1.000	0.841

Variable	**Flux**	**Flux groups**	r^2	**cv**
x_{116}	$R04749MM$	$x_{107}, x_{109}, x_{114}, x_{115}, x_{116}$*	1.000	0.841
x_{117}	$R04751MM$	$x_{90}, x_{110}, x_{112}, x_{115}, x_{117}$*	1.000	0.726
x_{118}	$R04754MM$	$x_{91}, x_{106}, x_{108}, x_{110}, x_{118}$*	1.000	0.726
x_{119}	$R04952MM$	x_{95}, x_{119}*	1.000	1.582
x_{120}	$R04953MM$	x_{100}, x_{120}*	1.000	1.582
x_{121}	$R04954MM$	x_{121}, x_{122}**	1.000	1.582
x_{122}	$R04956MM$	x_{121}, x_{122}**	1.000	1.582
x_{123}	$R04959MM$	x_{101}, x_{123}**	1.000	1.582
x_{124}	$R04968MM$	x_{102}, x_{124}*	1.000	1.582
x_{125}	$R04970MM$	x_{103}, x_{125}**	1.000	1.582
x_{126}	$R05064MM$	x_{40}, x_{126}	1.000	0.000
x_{127}	$R05066MM$	x_{76}, x_{127}	1.000	0.000
x_{128}	$R07162MM$	$x_{19}, x_{86}, x_{112}, x_{114}, x_{128}$*	1.000	0.304
x_{129}	$R07390MM$	x_{22}, x_{129}**	0.999	2.648
x_{130}	$R07599MM$	$x_{31}, x_{40}, x_{75}, x_{76}, x_{94}, x_{130}, x_{131}, x_{135}$	n.a.	0.000
x_{131}	$R07600MM$	$x_{31}, x_{40}, x_{75}, x_{76}, x_{94}, x_{130}, x_{131}, x_{135}$	n.a.	0.000
x_{132}	$R07603MM$	x_{92}, x_{132}**	1.000	1.152
x_{133}	$R07604MM$	x_{67}, x_{133}**	1.000	1.152
x_{134}	$R07618MM$	$x_{11}, x_{15}, x_{27}, x_{45}, x_{52}, x_{134}$	n.a.	0.424
x_{135}	$R08157MM$	$x_{31}, x_{40}, x_{75}, x_{76}, x_{94}, x_{130}, x_{131}, x_{135}$	n.a.	0.000

TABLE 6.1: Identifiability analysis applied to the FBA model of the mitochondrion with various fumarate conditions. The 135 matricial fluxes are grouped according to functional relations. r^2 indicates the amount of variance of the response explained by the predictors. A large $cv(x) = std(x)/mean(x)$ indicates that the data are scattered (practical non-identifiability). "n.a." stands for "not available", indicating that the metabolite is not significantly affected. An asterisk is added when $r^2 > 0.9$ and $cv > 0.1$, while another asterisk is added if the same functional group with $r^2 > 0.9$ and $cv > 0.1$ has been detected even if the role of response and predictors is switched, thus highlighting a strong interdependence between the variables involved.

Variable	Flux	Flux groups	r^2	cv
x_1	$R00004MM$	$x_1, x_5, x_{19}, x_{128}*$	1.000	0.539
x_2	$R00014MM$	x_2, x_{72}, x_{82}	1.000	0.029
x_3	$R00081MM$	x_3	0.999	0.027
x_4	$R00086MM$	$x_4, x_{13}, x_{37}, x_{52}, x_{65}$	1.000	0.061
x_5	$R00127MM$	x_5*	1.000	3.923
x_6	$R00157MM$	x_6*	1.000	3.922
x_7	$R00205MM$	x_7, x_{18}, x_{71}**	1.000	2.040
x_8	$R00238MM$	$x_8, x_{21}, x_{57}, x_{66}, x_{73}*$	1.000	0.302
x_9	$R00243MM$	$x_9, x_{31}, x_{76}, x_{131}, x_{135}*$	1.000	3.689
x_{10}	$R00245MM$	$x_{10}, x_{16}, x_{110}, x_{117}, x_{118}*$	1.000	0.387
x_{11}	$R00256MM$	$x_7, x_{11}, x_{18}, x_{71}*$	1.000	2.581
x_{12}	$R00258MM$	$x_{12}, x_{18}*$	0.999	0.261
x_{13}	$R00275MM$	$x_4, x_{13}, x_{57}, x_{65}$**	1.000	0.110
x_{14}	$R00330MM$	$x_{14}, x_{16}, x_{21}, x_{53}, x_{66}, x_{73}*$	1.000	2.982
x_{15}	$R00342MM$	$x_{15}, x_{45}, x_{52}, x_{66}, x_{134}*$	1.000	0.166
x_{16}	$R00351MM$	$x_8, x_{16}, x_{57}, x_{105}, x_{109}*$	1.000	0.204
x_{17}	$R00355MM$	$x_{17}, x_{60}, x_{70}, x_{74}$	0.997	0.030
x_{18}	$R00371MM$	$x_7, x_{18}*$	1.000	2.040
x_{19}	$R00388MM$	$x_{19}, x_{89}, x_{106}, x_{114}, x_{128}*$	1.000	0.373
x_{20}	$R00430MM$	$x_{14}, x_{16}, x_{20}, x_{134}*$	1.000	3.143
x_{21}	$R00432MM$	$x_8, x_{21}, x_{37}, x_{57}, x_{73}*$	1.000	0.205
x_{22}	$R00512MM$	$x_{22}*$	0.999	4.183
x_{23}	$R00551MM$	$x_{23}, x_{31}, x_{76}, x_{130}, x_{135}$	1.000	0.000
x_{24}	$R00572MM$	$x_{15}, x_{21}, x_{24}, x_{57}, x_{73}$	n.a.	1.862
x_{25}	$R00667MM$	$x_{25}, x_{31}, x_{40}, x_{94}, x_{131}$	1.000	0.000
x_{26}	$R00705MM$	$x_{26}, x_{31}*$	1.000	30.545
x_{27}	$R00709MM$	$x_{15}, x_{16}, x_{27}, x_{45}, x_{134}*$	1.000	0.204
x_{28}	$R00713MM$	x_{28}, x_{50}**	1.000	4.004
x_{29}	$R00716MM$	$x_{29}*$	0.998	1.903
x_{30}	$R00740MM$	$x_{30}, x_{131}*$	1.000	30.545
x_{31}	$R00830MM$	$x_{31}, x_{40}, x_{75}, x_{76}, x_{94}, x_{130}$	1.000	n.a.
x_{32}	$R00833MM$	$x_{15}, x_{21}, x_{32}, x_{37}, x_{73}, x_{92}*$	1.000	0.112
x_{33}	$R00851MM$	$x_{16}, x_{33}, x_{57}*$	1.000	4.642
x_{34}	$R00927MM$	$x_{34}, x_{75}, x_{76}, x_{130}*$	0.999	0.491
x_{35}	$R00941MM$	$x_{35}, x_{40}, x_{75}, x_{131}, x_{135}*$	0.999	2.024
x_{36}	$R00945MM$	$x_{36}, x_{75}, x_{76}, x_{94}, x_{130}*$	0.999	2.024
x_{37}	$R01082MM$	$x_{21}, x_{37}, x_{57}, x_{73}, x_{134}*$	1.000	0.210
x_{38}	$R01175MM$	$x_{38}, x_{104}, x_{108}, x_{110}*$	1.000	0.281
x_{39}	$R01177MM$	$x_{39}, x_{58}, x_{107}, x_{111}, x_{113}*$	1.000	0.281
x_{40}	$R01214MM$	$x_{31}, x_{40}, x_{75}, x_{76}, x_{94}, x_{130}$	1.000	0.000
x_{41}	$R01218MM$	x_{41}, x_{51}**	1.000	2.024
x_{42}	$R01253MM$	$x_{31}, x_{42}, x_{75}, x_{131}, x_{135}$	1.000	0.000
x_{43}	$R01279MM$	$x_{43}, x_{85}, x_{86}, x_{115}, x_{118}*$	1.000	0.281
x_{44}	$R01280MM$	$x_{44}, x_{49}*$	1.000	0.932
x_{45}	$R01325MM$	$x_8, x_{16}, x_{27}, x_{45}, x_{57}*$	1.000	0.204
x_{46}	$R01360MM$	$x_{46}*$	0.996	0.439
x_{47}	$R01361MM$	$x_{47}*$	1.000	0.439

Variable	Flux	Flux groups	r^2	cv
x_{48}	$R01624MM$	$x_{48}, x_{97}, x_{121}*$	1.000	0.932
x_{49}	$R01626MM$	$x_{49}, x_{53}, x_{100}*$	1.000	0.932
x_{50}	$R01648MM$	x_{28}, x_{50}**	1.000	3.995
x_{51}	$R01655MM$	x_{41}, x_{51}**	1.000	2.024
x_{52}	$R01700MM$	$x_{15}, x_{27}, x_{52}, x_{73}, x_{134}*$	1.000	0.205
x_{53}	$R01706MM$	$x_{53}, x_{125}*$	0.999	0.932
x_{54}	$R01799MM$	$x_{8}, x_{21}, x_{54}, x_{73}*$	1.000	4.383
x_{55}	$R01801MM$	x_{55}	n.a.	46.807
x_{56}	$R01859MM$	$x_{56}, x_{94}*$	1.000	0.491
x_{57}	$R01900MM$	$x_{13}, x_{16}, x_{37}, x_{57}, x_{65}*$	1.000	0.204
x_{58}	$R01923MM$	$x_{58}, x_{86}, x_{90}, x_{91}, x_{105}*$	1.000	0.281
x_{59}	$R01939MM$	$x_{59}*$	0.998	1.903
x_{60}	$R01940MM$	x_{60}, x_{70}, x_{74}**	1.000	0.421
x_{61}	$R01975MM$	$x_{16}, x_{27}, x_{57}, x_{61}, x_{78}*$	1.000	0.273
x_{62}	$R01978MM$	$x_{62}*$	0.999	0.440
x_{63}	$R02030MM$	$x_{63}*$	0.999	4.179
x_{64}	$R02161MM$	x_{64}	n.a.	0.027
x_{65}	$R02163MM$	$x_{4}, x_{13}, x_{57}, x_{65}$**	1.000	0.110
x_{66}	$R02164MM$	$x_{15}, x_{27}, x_{52}, x_{66}, x_{134}*$	1.000	0.206
x_{67}	$R02199MM$	$x_{67}, x_{130}*$	1.000	0.491
x_{68}	$R02241MM$	$x_{68}*$	0.998	4.237
x_{69}	$R02313MM$	$x_{69}*$	0.999	1.903
x_{70}	$R02487MM$	x_{60}, x_{70}, x_{74}**	1.000	0.421
x_{71}	$R02529MM$	x_{7}, x_{18}, x_{71}**	1.000	2.040
x_{72}	$R02569MM$	x_{2}, x_{72}, x_{82}	1.000	0.029
x_{73}	$R02570MM$	$x_{8}, x_{21}, x_{37}, x_{52}, x_{73}*$	1.000	0.205
x_{74}	$R02571MM$	x_{60}, x_{70}, x_{74}**	1.000	0.421
x_{75}	$R02661MM$	$x_{31}, x_{40}, x_{75}, x_{76}, x_{94}, x_{130}$	1.000	0.000
x_{76}	$R02662MM$	$x_{31}, x_{40}, x_{75}, x_{76}, x_{94}, x_{130}$	1.000	0.000
x_{77}	$R02765MM$	$x_{40}, x_{76}, x_{77}, x_{94}, x_{135}*$	0.998	0.491
x_{78}	$R03026MM$	$x_{8}, x_{61}, x_{78}, x_{91}, x_{113}*$	1.000	0.273
x_{79}	$R03102MM$	$x_{31}, x_{75}, x_{79}, x_{94}*$	0.999	1.903
x_{80}	$R03172MM$	$x_{80}, x_{133}*$	1.000	0.491
x_{81}	$R03174MM$	$x_{31}, x_{81}, x_{130}, x_{135}*$	0.999	0.491
x_{82}	$R03270MM$	x_{2}, x_{72}, x_{82}	1.000	0.029
x_{83}	$R03314MM$	$x_{31}, x_{75}, x_{76}, x_{83}, x_{94}$	n.a.	0.000
x_{84}	$R03381MM$	x_{84}, x_{130}	1.000	0.000
x_{85}	$R03777MM$	$x_{39}, x_{85}, x_{115}, x_{117}, x_{118}*$	1.000	0.281
x_{86}	$R03778MM$	$x_{86}, x_{90}, x_{108}, x_{114}, x_{116}*$	1.000	0.281
x_{87}	$R03857MM$	$x_{43}, x_{87}, x_{108}, x_{111}, x_{117}*$	1.000	0.281
x_{88}	$R03858MM$	$x_{88}, x_{91}, x_{111}, x_{113}, x_{114}*$	1.000	0.281
x_{89}	$R03990MM$	$x_{85}, x_{89}, x_{106}, x_{112}, x_{115}*$	1.000	0.281
x_{90}	$R03991MM$	$x_{58}, x_{88}, x_{90}, x_{105}, x_{113}*$	1.000	0.281
x_{91}	$R04170MM$	$x_{91}, x_{108}, x_{111}, x_{113}, x_{116}*$	1.000	0.281
x_{92}	$R04203MM$	$x_{92}*$	0.999	0.491
x_{93}	$R04204MM$	$x_{93}, x_{94}*$	1.000	0.491
x_{94}	$R04224MM$	$x_{31}, x_{40}, x_{75}, x_{76}, x_{94}, x_{130}$	1.000	0.000

Variable	**Flux**	**Flux groups**	r^2	**cv**
x_{95}	$R04355MM$	$x_{49}, x_{95}*$	1.000	0.932
x_{96}	$R04428MM$	$x_{96}, x_{100}, x_{121}*$	1.000	0.932
x_{97}	$R04430MM$	$x_{97}, x_{122}*$	1.000	0.932
x_{98}	$R04433MM$	$x_{38}, x_{39}, x_{58}, x_{98}, x_{116}*$	1.000	0.277
x_{99}	$R04533MM$	x_{99}, x_{120}**	1.000	0.932
x_{100}	$R04536MM$	$x_{49}, x_{100}, x_{120}*$	1.000	0.932
x_{101}	$R04537MM$	$x_{101}, x_{121}, x_{122}*$	1.000	0.932
x_{102}	$R04543MM$	$x_{53}, x_{102}*$	0.999	0.932
x_{103}	$R04544MM$	$x_{49}, x_{103}, x_{125}*$	1.000	0.932
x_{104}	$R04737MM$	$x_{85}, x_{87}, x_{91}, x_{104}, x_{110}*$	1.000	0.281
x_{105}	$R04738MM$	$x_{86}, x_{90}, x_{105}, x_{107}, x_{109}*$	1.000	0.281
x_{106}	$R04739MM$	$x_{39}, x_{85}, x_{87}, x_{106}, x_{117}*$	1.000	0.281
x_{107}	$R04740MM$	$x_{58}, x_{86}, x_{88}, x_{107}, x_{111}*$	1.000	0.281
x_{108}	$R04741MM$	$x_{85}, x_{108}, x_{110}, x_{112}, x_{115}*$	1.000	0.281
x_{109}	$R04742MM$	$x_{105}, x_{109}, x_{111}, x_{114}, x_{116}*$	1.000	0.281
x_{110}	$R04743MM$	$x_{43}, x_{87}, x_{110}, x_{112}, x_{116}*$	1.000	0.281
x_{111}	$R04744MM$	$x_{43}, x_{91}, x_{111}, x_{113}, x_{114}*$	1.000	0.281
x_{112}	$R04745MM$	$x_{91}, x_{106}, x_{108}, x_{112}, x_{117}*$	1.000	0.281
x_{113}	$R04746MM$	$x_{39}, x_{107}, x_{109}, x_{111}, x_{113}*$	1.000	0.281
x_{114}	$R04747MM$	$x_{58}, x_{91}, x_{111}, x_{113}, x_{114}*$	1.000	0.281
x_{115}	$R04748MM$	$x_{43}, x_{85}, x_{112}, x_{115}, x_{117}*$	1.000	0.281
x_{116}	$R04749MM$	$x_{39}, x_{86}, x_{87}, x_{114}, x_{116}*$	1.000	0.281
x_{117}	$R04751MM$	$x_{85}, x_{104}, x_{106}, x_{110}, x_{117}*$	1.000	0.281
x_{118}	$R04754MM$	$x_{85}, x_{104}, x_{108}, x_{112}, x_{118}*$	1.000	0.281
x_{119}	$R04952MM$	$x_{95}, x_{119}*$	1.000	0.932
x_{120}	$R04953MM$	x_{99}, x_{120}**	1.000	0.932
x_{121}	$R04954MM$	$x_{101}, x_{121}*$	1.000	0.932
x_{122}	$R04956MM$	$x_{97}, x_{101}, x_{122}, x_{123}*$	1.000	0.932
x_{123}	$R04959MM$	$x_{121}, x_{122}, x_{123}*$	1.000	0.932
x_{124}	$R04968MM$	$x_{49}, x_{124}*$	1.000	0.932
x_{125}	$R04970MM$	$x_{103}, x_{125}*$	1.000	0.932
x_{126}	$R05064MM$	x_{94}, x_{126}	1.000	0.000
x_{127}	$R05066MM$	x_{127}, x_{130}	1.000	0.000
x_{128}	$R07162MM$	$x_{19}, x_{107}, x_{113}, x_{114}, x_{128}*$	1.000	0.373
x_{129}	$R07390MM$	$x_{129}*$	0.999	4.177
x_{130}	$R07599MM$	$x_{31}, x_{40}, x_{75}, x_{76}, x_{94}, x_{130}$	1.000	0.000
x_{131}	$R07600MM$	$x_{31}, x_{40}, x_{75}, x_{76}, x_{94}, x_{131}$	1.000	0.000
x_{132}	$R07603MM$	$x_{76}, x_{132}*$	1.000	0.491
x_{133}	$R07604MM$	$x_{31}, x_{40}, x_{76}, x_{133}*$	1.000	0.491
x_{134}	$R07618MM$	$x_{15}, x_{21}, x_{27}, x_{52}, x_{134}*$	1.000	0.168
x_{135}	$R08157MM$	$x_{31}, x_{40}, x_{75}, x_{76}, x_{94}, x_{135}$	1.000	0.000

TABLE 6.2: Identifiability Analysis results in Healthy state for FBA mitochondria model.

Variable	Flux	Flux groups	r^2	cv
x_1	$R00004MM$	$x_1, x_{44}, x_{54}*$	0.995	2.135
x_2	$R00014MM$	x_2, x_{72}, x_{82}	1.000	0.000
x_3	$R00081MM$	$x_3, x_{15}, x_{16}, x_{50}, x_{64}, x_{98}*$	1.000	0.115
x_4	$R00086MM$	$x_4, x_{13}, x_{108}, x_{117}*$	1.000	0.113
x_5	$R00127MM$	x_5*	0.912	0.267
x_6	$R00157MM$	x_6	n.a.	0.267
x_7	$R00205MM$	x_7	n.a.	0.000
x_8	$R00238MM$	$x_8, x_{50}, x_{78}, x_{91}, x_{111}, x_{114}*$	1.000	0.199
x_9	$R00243MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{30}$	1.000	n.a.
x_{10}	$R00245MM$	$x_{10}, x_{43}, x_{108}, x_{110}, x_{128}$	1.000	0.032
x_{11}	$R00256MM$	x_{11}	n.a.	0.000
x_{12}	$R00258MM$	x_{12}	n.a.	0.000
x_{13}	$R00275MM$	$x_{13}, x_{27}, x_{85}, x_{89}*$	1.000	0.104
x_{14}	$R00330MM$	$x_{14}, x_{73}, x_{91}*$	1.000	0.268
x_{15}	$R00342MM$	$x_{10}, x_{15}, x_{64}, x_{98}$	1.000	0.082
x_{16}	$R00351MM$	$x_{16}, x_{19}, x_{88}, x_{90}, x_{91}*$	1.000	0.119
x_{17}	$R00355MM$	x_{17}	0.985	0.000
x_{18}	$R00371MM$	x_{18}	0.997	0.000
x_{19}	$R00388MM$	$x_{16}, x_{19}, x_{39}, x_{107}$	1.000	0.048
x_{20}	$R00430MM$	$x_{20}*$	1.000	0.267
x_{21}	$R00432MM$	$x_8, x_{16}, x_{21}, x_{50}, x_{112}, x_{134}*$	1.000	0.119
x_{22}	$R00512MM$	x_{22}, x_{63}, x_{129}**	0.997	0.649
x_{23}	$R00551MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{30}$	1.000	0.000
x_{24}	$R00572MM$	x_{24}, x_{54}**	0.997	0.654
x_{25}	$R00667MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{30}$	1.000	0.000
x_{26}	$R00705MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{30}$	1.000	n.a.
x_{27}	$R00709MM$	$x_{10}, x_{27}, x_{38}, x_{85}*$	1.000	0.119
x_{28}	$R00713MM$	$x_{28}*$	0.952	0.466
x_{29}	$R00716MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{30}$	1.000	0.000
x_{30}	$R00740MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{30}$	1.000	n.a.
x_{31}	$R00830MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{31}$	1.000	n.a.
x_{32}	$R00833MM$	x_{32}	0.995	0.000
x_{33}	$R00851MM$	x_{33}, x_{68}**	0.965	0.691
x_{34}	$R00927MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{34}$	1.000	n.a.
x_{35}	$R00941MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{35}$	1.000	0.000
x_{36}	$R00945MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{36}$	1.000	0.000
x_{37}	$R01082MM$	$x_8, x_{37}, x_{111}, x_{114}*$	1.000	0.123
x_{38}	$R01175MM$	$x_3, x_{13}, x_{27}, x_{38}, x_{108}*$	1.000	0.217
x_{39}	$R01177MM$	$x_{10}, x_{21}, x_{39}, x_{58}*$	1.000	0.217
x_{40}	$R01214MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{40}$	1.000	0.000
x_{41}	$R01218MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{41}$	1.000	0.000
x_{42}	$R01253MM$	x_{42}	0.999	0.000
x_{43}	$R01279MM$	$x_{13}, x_{21}, x_{38}, x_{43}, x_{50}, x_{85}*$	1.000	0.217
x_{44}	$R01280MM$	$x_1, x_{24}, x_{44}*$	0.994	2.733
x_{45}	$R01325MM$	$x_{45}, x_{50}, x_{65}, x_{98}, x_{114}, x_{118}*$	1.000	0.119
x_{46}	$R01360MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{46}$	1.000	n.a.
x_{47}	$R01361MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{47}$	1.000	n.a.

Variable	Flux	Flux groups	r^2	cv
x_{48}	$R01624MM$	x_{48}, x_{97}**	0.999	5.268
x_{49}	$R01626MM$	x_{49}*	0.982	1.322
x_{50}	$R01648MM$	x_{50}, x_{104}*	0.983	1.137
x_{51}	$R01655MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{51}$	1.000	0.000
x_{52}	$R01700MM$	$x_{52}, x_{61}, x_{65}, x_{108}, x_{110}$*	1.000	0.119
x_{53}	$R01706MM$	x_{53}*	0.983	5.271
x_{54}	$R01799MM$	x_{24}, x_{54}**	0.997	0.657
x_{55}	$R01801MM$	x_{55}	n.a.	15.905
x_{56}	$R01859MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{56}$	1.000	n.a.
x_{57}	$R01900MM$	$x_{39}, x_{50}, x_{57}, x_{58}, x_{88}, x_{90}$*	1.000	0.119
x_{58}	$R01923MM$	$x_{50}, x_{58}, x_{107}, x_{109}, x_{115}, x_{128}$*	1.000	0.217
x_{59}	$R01939MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{59}$	1.000	0.000
x_{60}	$R01940MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{60}$	1.000	0.000
x_{61}	$R01975MM$	$x_{50}, x_{52}, x_{61}, x_{86}, x_{107}, x_{117}$*	1.000	0.199
x_{62}	$R01978MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{62}$	1.000	n.a.
x_{63}	$R02030MM$	x_{22}, x_{63}, x_{129}**	0.999	0.655
x_{64}	$R02161MM$	$x_{27}, x_{52}, x_{64}, x_{134}$*	1.000	0.115
x_{65}	$R02163MM$	$x_{65}, x_{117}, x_{118}, x_{134}$*	1.000	0.104
x_{66}	$R02164MM$	$x_{50}, x_{66}, x_{87}, x_{98}, x_{117}, x_{118}$*	1.000	0.119
x_{67}	$R02199MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{67}$	1.000	n.a.
x_{68}	$R02241MM$	x_{33}, x_{68}**	0.965	0.676
x_{69}	$R02313MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{69}$	1.000	0.000
x_{70}	$R02487MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{70}$	1.000	0.000
x_{71}	$R02529MM$	x_{71}	n.a.	0.000
x_{72}	$R02569MM$	x_2, x_{72}, x_{82}	1.000	0.000
x_{73}	$R02570MM$	$x_{73}, x_{107}, x_{109}, x_{111}, x_{114}$*	1.000	0.119
x_{74}	$R02571MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{74}$	1.000	0.000
x_{75}	$R02661MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{75}$	1.000	0.000
x_{76}	$R02662MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{76}$	1.000	0.000
x_{77}	$R02765MM$	x_{77}*	1.000	33.161
x_{78}	$R03026MM$	$x_{58}, x_{73}, x_{78}, x_{91}$*	1.000	0.199
x_{79}	$R03102MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{79}$	1.000	0.000
x_{80}	$R03172MM$	x_{80}	0.836	3.329
x_{81}	$R03174MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{81}$	1.000	n.a.
x_{82}	$R03270MM$	x_2, x_{72}, x_{82}	1.000	0.000
x_{83}	$R03314MM$	x_{83}	0.995	0.000
x_{84}	$R03381MM$	x_{84}	0.974	0.000
x_{85}	$R03777MM$	$x_{19}, x_{64}, x_{65}, x_{85}, x_{118}$*	1.000	0.217
x_{86}	$R03778MM$	$x_{19}, x_{50}, x_{58}, x_{78}, x_{86}, x_{104}$*	1.000	0.217
x_{87}	$R03857MM$	$x_3, x_{65}, x_{87}, x_{110}, x_{114}$*	1.000	0.217
x_{88}	$R03858MM$	$x_{39}, x_{43}, x_{73}, x_{88}, x_{114}$*	1.000	0.217
x_{89}	$R03990MM$	$x_4, x_{89}, x_{98}, x_{110}$*	1.000	0.217
x_{90}	$R03991MM$	$x_{16}, x_{90}, x_{91}, x_{109}, x_{114}$*	1.000	0.217
x_{91}	$R04170MM$	$x_{13}, x_{19}, x_{91}, x_{105}, x_{109}$*	1.000	0.217
x_{92}	$R04203MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{92}$	1.000	n.a.
x_{93}	$R04204MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{93}$	1.000	n.a.
x_{94}	$R04224MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{94}$	1.000	0.000

Variable	Flux	Flux groups	r^2	cv
x_{95}	$R04355MM$	$x_{95}*$	0.990	3.729
x_{96}	$R04428MM$	x_{96}, x_{99}**	0.999	2.360
x_{97}	$R04430MM$	x_{48}, x_{97}**	0.999	5.124
x_{98}	$R04433MM$	$x_4, x_{38}, x_{50}, x_{98}, x_{108}, x_{112}*$	1.000	0.213
x_{99}	$R04533MM$	x_{96}, x_{99}**	0.999	2.359
x_{100}	$R04536MM$	$x_{100}, x_{102}*$	0.992	2.531
x_{101}	$R04537MM$	x_{101}, x_{123}**	0.999	6.143
x_{102}	$R04543MM$	x_{102}, x_{120}**	0.992	1.294
x_{103}	$R04544MM$	x_{103}, x_{125}**	0.999	2.141
x_{104}	$R04737MM$	$x_3, x_{15}, x_{52}, x_{104}*$	1.000	0.217
x_{105}	$R04738MM$	$x_{37}, x_{50}, x_{87}, x_{90}, x_{105}, x_{111}*$	1.000	0.217
x_{106}	$R04739MM$	$x_{50}, x_{61}, x_{85}, x_{106}, x_{112}, x_{115}*$	1.000	0.217
x_{107}	$R04740MM$	$x_4, x_{39}, x_{88}, x_{107}*$	1.000	0.217
x_{108}	$R04741MM$	$x_{45}, x_{50}, x_{52}, x_{106}, x_{108}, x_{128}*$	1.000	0.217
x_{109}	$R04742MM$	$x_8, x_{50}, x_{88}, x_{90}, x_{107}, x_{109}*$	1.000	0.217
x_{110}	$R04743MM$	$x_3, x_{89}, x_{110}, x_{134}*$	1.000	0.217
x_{111}	$R04744MM$	$x_{19}, x_{50}, x_{52}, x_{58}, x_{88}, x_{111}*$	1.000	0.217
x_{112}	$R04745MM$	$x_{13}, x_{64}, x_{89}, x_{112}, x_{114}*$	1.000	0.217
x_{113}	$R04746MM$	$x_{37}, x_{50}, x_{88}, x_{98}, x_{113}, x_{114}*$	1.000	0.217
x_{114}	$R04747MM$	$x_4, x_{50}, x_{89}, x_{91}, x_{113}, x_{114}*$	1.000	0.217
x_{115}	$R04748MM$	$x_{85}, x_{89}, x_{104}, x_{115}*$	1.000	0.217
x_{116}	$R04749MM$	$x_{50}, x_{86}, x_{91}, x_{107}, x_{113}, x_{116}*$	1.000	0.217
x_{117}	$R04751MM$	$x_{10}, x_{64}, x_{87}, x_{117}*$	1.000	0.217
x_{118}	$R04754MM$	$x_4, x_{50}, x_{104}, x_{114}, x_{118}, x_{128}*$	1.000	0.217
x_{119}	$R04952MM$	$x_{119}*$	1.000	2.929
x_{120}	$R04953MM$	x_{102}, x_{120}**	0.992	2.379
x_{121}	$R04954MM$	x_{121}, x_{122}**	0.999	5.189
x_{122}	$R04956MM$	x_{121}, x_{122}**	0.999	5.189
x_{123}	$R04959MM$	x_{101}, x_{123}**	0.999	6.143
x_{124}	$R04968MM$	$x_{121}, x_{124}*$	0.979	1.594
x_{125}	$R04970MM$	x_{103}, x_{125}**	0.999	2.141
x_{126}	$R05064MM$	x_{47}, x_{126}	0.999	0.000
x_{127}	$R05066MM$	x_{31}, x_{127}	0.999	0.000
x_{128}	$R07162MM$	$x_{38}, x_{50}, x_{85}, x_{98}, x_{118}, x_{128}$	1.000	0.048
x_{129}	$R07390MM$	x_{22}, x_{63}, x_{129}**	0.998	0.655
x_{130}	$R07599MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{130}$	1.000	0.000
x_{131}	$R07600MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{131}$	1.000	0.000
x_{132}	$R07603MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{132}$	1.000	n.a.
x_{133}	$R07604MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{133}$	1.000	n.a.
x_{134}	$R07618MM$	$x_4, x_{10}, x_{61}, x_{118}, x_{134}$	1.000	0.082
x_{135}	$R08157MM$	$x_9, x_{23}, x_{25}, x_{26}, x_{29}, x_{135}$	1.000	0.000

TABLE 6.3: Identifiability Analysis results in Inflammation state in fumarase deficiency.

Variable	Flux	Flux groups	r^2	cv
x_1	$R00004MM$	$x_1, x_{22}, x_{33}, x_{54}, x_{112}*$	1.000	1.196
x_2	$R00014MM$	$x_2, x_{57}, x_{58}, x_{72}, x_{82}, x_{112}$	1.000	0.018
x_3	$R00081MM$	$x_3, x_{27}, x_{52}, x_{64}, x_{134}*$	1.000	0.502
x_4	$R00086MM$	$x_4, x_{61}, x_{108}, x_{115}, x_{134}*$	1.000	0.503
x_5	$R00127MM$	$x_5, x_{21}, x_{45}, x_{78}, x_{113}$	n.a.	0.658
x_6	$R00157MM$	$x_6, x_{68}*$	0.992	0.658
x_7	$R00205MM$	$x_7, x_{36}, x_{51}, x_{94}, x_{127}$	n.a.	0.000
x_8	$R00238MM$	$x_8, x_{105}, x_{109}, x_{111}, x_{116}*$	1.000	11.664
x_9	$R00243MM$	$x_9, x_{26}, x_{31}, x_{35}, x_{36}, x_{40}$	1.000	n.a.
x_{10}	$R00245MM$	$x_{10}, x_{104}, x_{110}*$	1.000	0.736
x_{11}	$R00256MM$	$x_{11}, x_{75}, x_{126}, x_{131}$	0.998	0.000
x_{12}	$R00258MM$	x_{12}, x_{42}	1.000	0.000
x_{13}	$R00275MM$	$x_{13}, x_{15}, x_{45}, x_{65}, x_{128}*$	1.000	0.476
x_{14}	$R00330MM$	$x_{14}, x_{19}, x_{39}*$	1.000	2.063
x_{15}	$R00342MM$	$x_{13}, x_{15}, x_{52}, x_{90}, x_{106}*$	1.000	0.275
x_{16}	$R00351MM$	$x_{16}, x_{57}, x_{86}, x_{109}, x_{114}*$	1.000	0.530
x_{17}	$R00355MM$	$x_1, x_{17}, x_{22}, x_{33}, x_{54}$	1.000	0.036
x_{18}	$R00371MM$	x_{18}	n.a.	0.000
x_{19}	$R00388MM$	$x_{16}, x_{19}, x_{57}, x_{106}, x_{116}*$	1.000	0.134
x_{20}	$R00430MM$	$x_{20}*$	1.000	2.055
x_{21}	$R00432MM$	$x_8, x_{21}, x_{57}, x_{105}, x_{116}*$	1.000	0.531
x_{22}	$R00512MM$	$x_1, x_{22}, x_{33}, x_{54}, x_{58}, x_{105}*$	1.000	1.196
x_{23}	$R00551MM$	$x_9, x_{23}, x_{40}, x_{69}, x_{94}$	0.999	0.000
x_{24}	$R00572MM$	$x_{15}, x_{19}, x_{24}, x_{65}, x_{128}*$	0.999	1.655
x_{25}	$R00667MM$	$x_{17}, x_{25}, x_{78}, x_{90}$	1.000	0.000
x_{26}	$R00705MM$	$x_9, x_{26}, x_{31}, x_{35}, x_{36}, x_{40}$	1.000	n.a.
x_{27}	$R00709MM$	$x_3, x_{15}, x_{27}, x_{104}, x_{115}*$	1.000	0.530
x_{28}	$R00713MM$	x_{28}, x_{50}**	0.999	4.942
x_{29}	$R00716MM$	x_{29}, x_{35}	1.000	0.000
x_{30}	$R00740MM$	$x_{30}, x_{31}*$	1.000	11.478
x_{31}	$R00830MM$	$x_9, x_{26}, x_{31}, x_{35}, x_{36}, x_{40}$	1.000	n.a.
x_{32}	$R00833MM$	x_{32}	1.000	0.000
x_{33}	$R00851MM$	$x_1, x_{22}, x_{33}, x_{54}, x_{78}*$	1.000	1.196
x_{34}	$R00927MM$	x_{34}, x_{81}, x_{93}**	1.000	824.829
x_{35}	$R00941MM$	$x_9, x_{26}, x_{31}, x_{35}, x_{36}, x_{40}$	1.000	0.000
x_{36}	$R00945MM$	$x_9, x_{26}, x_{31}, x_{35}, x_{36}, x_{40}$	1.000	0.000
x_{37}	$R01082MM$	$x_{19}, x_{37}, x_{66}, x_{73}, x_{78}*$	1.000	0.579
x_{38}	$R01175MM$	$x_{38}, x_{85}, x_{89}*$	1.000	1.401
x_{39}	$R01177MM$	$x_{13}, x_{19}, x_{21}, x_{39}, x_{114}*$	1.000	6.885
x_{40}	$R01214MM$	$x_9, x_{26}, x_{31}, x_{35}, x_{36}, x_{40}$	1.000	0.000
x_{41}	$R01218MM$	$x_{41}, x_{58}, x_{105}, x_{107}, x_{109}*$	1.000	1.056
x_{42}	$R01253MM$	x_{12}, x_{42}	1.000	0.000
x_{43}	$R01279MM$	$x_{38}, x_{43}, x_{87}*$	1.000	1.401
x_{44}	$R01280MM$	$x_{44}, x_{49}, x_{95}*$	1.000	0.624
x_{45}	$R01325MM$	$x_{13}, x_{15}, x_{45}, x_{112}, x_{128}*$	1.000	0.530
x_{46}	$R01360MM$	$x_{46}, x_{62}*$	1.000	16.699
x_{47}	$R01361MM$	x_{47}, x_{62}**	1.000	17.001

Variable	Flux	Flux groups	r^2	cv
x_{48}	$R01624MM$	$x_{48}, x_{103}*$	1.000	1.532
x_{49}	$R01626MM$	$x_{44}, x_{48}, x_{49}*$	1.000	1.574
x_{50}	$R01648MM$	x_{28}, x_{50}**	0.999	4.829
x_{51}	$R01655MM$	$x_9, x_{26}, x_{31}, x_{35}, x_{36}, x_{51}$	1.000	0.000
x_{52}	$R01700MM$	$x_{52}, x_{64}, x_{65}, x_{104}, x_{115}*$	1.000	0.533
x_{53}	$R01706MM$	$x_{53}, x_{121}*$	1.000	0.624
x_{54}	$R01799MM$	$x_1, x_{19}, x_{22}, x_{33}, x_{54}*$	1.000	1.196
x_{55}	$R01801MM$	x_{55}	n.a.	n.a.
x_{56}	$R01859MM$	$x_9, x_{26}, x_{31}, x_{35}, x_{36}, x_{56}$	1.000	n.a.
x_{57}	$R01900MM$	$x_{16}, x_{57}, x_{78}, x_{91}, x_{107}*$	1.000	0.530
x_{58}	$R01923MM$	$x_8, x_{21}, x_{58}, x_{90}, x_{115}$	n.a.	1.401
x_{59}	$R01939MM$	$x_9, x_{36}, x_{59}, x_{94}, x_{126}$	n.a.	0.000
x_{60}	$R01940MM$	$x_{31}, x_{35}, x_{36}, x_{60}, x_{76}$	n.a.	0.000
x_{61}	$R01975MM$	$x_{27}, x_{61}, x_{104}, x_{110}, x_{128}*$	1.000	11.664
x_{62}	$R01978MM$	x_{47}, x_{62}**	1.000	16.699
x_{63}	$R02030MM$	$x_4, x_{22}, x_{63}, x_{68}, x_{98}, x_{129}*$	1.000	1.196
x_{64}	$R02161MM$	$x_4, x_{45}, x_{57}, x_{64}, x_{66}*$	1.000	0.502
x_{65}	$R02163MM$	$x_{65}, x_{110}, x_{115}, x_{116}, x_{128}*$	1.000	0.476
x_{66}	$R02164MM$	$x_3, x_{27}, x_{61}, x_{66}, x_{108}*$	1.000	0.531
x_{67}	$R02199MM$	$x_{67}, x_{92}, x_{132}, x_{133}$**	1.000	n.a.
x_{68}	$R02241MM$	$x_{66}, x_{68}, x_{108}, x_{134}*$	1.000	1.196
x_{69}	$R02313MM$	x_{69}, x_{126}	1.000	0.000
x_{70}	$R02487MM$	$x_{26}, x_{40}, x_{51}, x_{70}, x_{76}$	n.a.	0.000
x_{71}	$R02529MM$	x_{71}	n.a.	0.000
x_{72}	$R02569MM$	$x_2, x_8, x_{58}, x_{72}, x_{82}, x_{91}$	1.000	0.018
x_{73}	$R02570MM$	$x_{21}, x_{73}, x_{88}, x_{107}, x_{111}*$	1.000	0.533
x_{74}	$R02571MM$	$x_{35}, x_{51}, x_{56}, x_{74}, x_{76}$	n.a.	0.000
x_{75}	$R02661MM$	$x_9, x_{26}, x_{31}, x_{35}, x_{36}, x_{75}$	1.000	0.000
x_{76}	$R02662MM$	$x_9, x_{26}, x_{31}, x_{35}, x_{36}, x_{76}$	1.000	0.000
x_{77}	$R02765MM$	$x_{77}, x_{105}, x_{107}*$	0.999	n.a.
x_{78}	$R03026MM$	$x_{16}, x_{73}, x_{78}, x_{88}, x_{111}*$	1.000	11.664
x_{79}	$R03102MM$	x_{79}, x_{131}	0.999	0.000
x_{80}	$R03172MM$	x_{80}	n.a.	5.398
x_{81}	$R03174MM$	x_{34}, x_{81}, x_{93}**	1.000	n.a.
x_{82}	$R03270MM$	$x_2, x_{72}, x_{82}, x_{98}, x_{104}, x_{112}$	1.000	0.018
x_{83}	$R03314MM$	x_{12}, x_{83}	1.000	0.000
x_{84}	$R03381MM$	$x_9, x_{26}, x_{31}, x_{35}, x_{36}, x_{84}$	1.000	0.000
x_{85}	$R03777MM$	$x_{85}, x_{117}*$	1.000	1.401
x_{86}	$R03778MM$	$x_{19}, x_{86}, x_{105}, x_{109}, x_{115}*$	1.000	6.885
x_{87}	$R03857MM$	$x_{85}, x_{87}*$	1.000	1.401
x_{88}	$R03858MM$	$x_{78}, x_{88}, x_{90}, x_{105}, x_{115}*$	1.000	6.885
x_{89}	$R03990MM$	$x_{89}, x_{118}*$	1.000	1.401
x_{90}	$R03991MM$	$x_{19}, x_{21}, x_{78}, x_{90}, x_{111}*$	1.000	6.885
x_{91}	$R04170MM$	$x_{39}, x_{86}, x_{90}, x_{91}, x_{113}*$	1.000	6.885
x_{92}	$R04203MM$	$x_{67}, x_{92}, x_{132}, x_{133}$**	1.000	n.a.
x_{93}	$R04204MM$	x_{34}, x_{81}, x_{93}**	1.000	n.a.
x_{94}	$R04224MM$	$x_9, x_{26}, x_{31}, x_{35}, x_{36}, x_{94}$	1.000	0.000

Variable	**Flux**	**Flux groups**	r^2	**cv**
x_{95}	$R04355MM$	$x_{48}, x_{95}*$	1.000	1.532
x_{96}	$R04428MM$	$x_{49}, x_{96}*$	1.000	1.532
x_{97}	$R04430MM$	$x_{97}, x_{123}*$	1.000	1.532
x_{98}	$R04433MM$	$x_{3}, x_{98}, x_{108}*$	1.000	1.233
x_{99}	$R04533MM$	$x_{99}, x_{120}*$	1.000	1.532
x_{100}	$R04536MM$	$x_{100}, x_{125}*$	1.000	1.532
x_{101}	$R04537MM$	$x_{53}, x_{101}, x_{103}*$	1.000	1.532
x_{102}	$R04543MM$	$x_{97}, x_{102}*$	1.000	0.624
x_{103}	$R04544MM$	$x_{53}, x_{103}*$	1.000	0.624
x_{104}	$R04737MM$	$x_{61}, x_{64}, x_{66}, x_{104}, x_{115}*$	1.000	6.885
x_{105}	$R04738MM$	$x_{39}, x_{78}, x_{86}, x_{105}, x_{116}*$	1.000	6.885
x_{106}	$R04739MM$	$x_{3}, x_{65}, x_{106}, x_{108}, x_{110}*$	1.000	6.885
x_{107}	$R04740MM$	$x_{39}, x_{86}, x_{107}, x_{114}, x_{116}*$	1.000	6.885
x_{108}	$R04741MM$	$x_{4}, x_{52}, x_{65}, x_{108}, x_{115}*$	1.000	6.885
x_{109}	$R04742MM$	$x_{8}, x_{91}, x_{105}, x_{109}, x_{114}*$	1.000	6.885
x_{110}	$R04743MM$	$x_{65}, x_{78}, x_{106}, x_{110}, x_{115}*$	1.000	6.885
x_{111}	$R04744MM$	$x_{73}, x_{78}, x_{105}, x_{109}, x_{111}*$	1.000	6.885
x_{112}	$R04745MM$	$x_{3}, x_{4}, x_{104}, x_{110}, x_{112}*$	1.000	6.885
x_{113}	$R04746MM$	$x_{73}, x_{90}, x_{107}, x_{113}, x_{116}*$	1.000	6.885
x_{114}	$R04747MM$	$x_{21}, x_{105}, x_{109}, x_{114}, x_{116}*$	1.000	6.885
x_{115}	$R04748MM$	$x_{61}, x_{65}, x_{66}, x_{112}, x_{115}*$	1.000	6.885
x_{116}	$R04749MM$	$x_{8}, x_{19}, x_{86}, x_{91}, x_{116}*$	1.000	6.885
x_{117}	$R04751MM$	$x_{117}, x_{118}*$	1.000	1.401
x_{118}	$R04754MM$	$x_{43}, x_{118}*$	1.000	1.401
x_{119}	$R04952MM$	$x_{119}, x_{124}*$	1.000	1.532
x_{120}	$R04953MM$	$x_{97}, x_{120}*$	1.000	1.532
x_{121}	$R04954MM$	$x_{95}, x_{96}, x_{121}*$	1.000	1.532
x_{122}	$R04956MM$	$x_{100}, x_{122}*$	1.000	1.532
x_{123}	$R04959MM$	$x_{100}, x_{102}, x_{120}, x_{122}, x_{123}*$	1.000	1.532
x_{124}	$R04968MM$	$x_{49}, x_{53}, x_{124}*$	1.000	0.624
x_{125}	$R04970MM$	$x_{120}, x_{125}*$	1.000	0.624
x_{126}	$R05064MM$	$x_{9}, x_{26}, x_{31}, x_{35}, x_{36}, x_{126}$	1.000	0.000
x_{127}	$R05066MM$	$x_{9}, x_{26}, x_{31}, x_{35}, x_{36}, x_{127}$	1.000	0.000
x_{128}	$R07162MM$	$x_{64}, x_{66}, x_{108}, x_{112}, x_{128}*$	1.000	0.134
x_{129}	$R07390MM$	$x_{3}, x_{4}, x_{63}, x_{64}, x_{129}*$	1.000	1.196
x_{130}	$R07599MM$	$x_{9}, x_{26}, x_{31}, x_{35}, x_{36}, x_{130}$	1.000	0.000
x_{131}	$R07600MM$	$x_{9}, x_{26}, x_{31}, x_{35}, x_{36}, x_{131}$	1.000	0.000
x_{132}	$R07603MM$	$x_{67}, x_{92}, x_{132}, x_{133}$**	1.000	n.a.
x_{133}	$R07604MM$	$x_{67}, x_{92}, x_{132}, x_{133}$**	1.000	n.a.
x_{134}	$R07618MM$	$x_{3}, x_{52}, x_{64}, x_{110}, x_{134}*$	1.000	0.264
x_{135}	$R08157MM$	$x_{9}, x_{26}, x_{31}, x_{35}, x_{36}, x_{135}$	1.000	0.000

TABLE 6.4: Identifiability Analysis results in Pathological state in fumarase deficiency.

6.2.4 Sensitivity and Local Robustness

For the FBA mitochondrial model I have also analyzed the sensitivity of the 135 metabolic fluxes implementing a novel perturbation that takes into account of thermodynamics constraints. Indeed, forcing a metabolic flux to a specific value in FBA approach could lead to unfeasible configuration and the FBA cannot be solved. The global sensitivity is carried out by applying a random noise on exchange fluxes. To calculate the elementary effect of the j-th metabolic flux, the v_j flux is constrained to a value obtained by adding the Δ random noise. The distribution of the elementary effect is calculated by sampling until 100 times. This perturbation considers the reversibility of the fluxes and if after perturbation the network is unfeasible, the sensitivity indexes cannot be calculated. We can see the results related to the mitochondrion network in Figure 6.5. The most sensitive reaction is the Saccharopine dehydrogenase (R001716MM), associated to the EC number 1.5.1.8, followed by ornithine aminotransferase (R00667MM) associated to the EC 2.6.1.13 and 5-aminolevulinate synthase (R00371MM) associated to EC 2.3.1.37.

In Table 6.5 I report the results of the local robustness for mitochondrial fluxes for the energy production. In particular have been shown the values obtained for the fragile metabolic fluxes, i.e., fluxes with a robustness less than 100%. In particular, in Figure 6.6 has been plotted the response in terms of ATP and NADH production after varying the metabolic fluxes whose robustness is zero. Indeed, we can see that ATP and NADH change mostly by varying values in x axis. Instead, in Figure 6.7 have been reported the most robust metabolic fluxes.

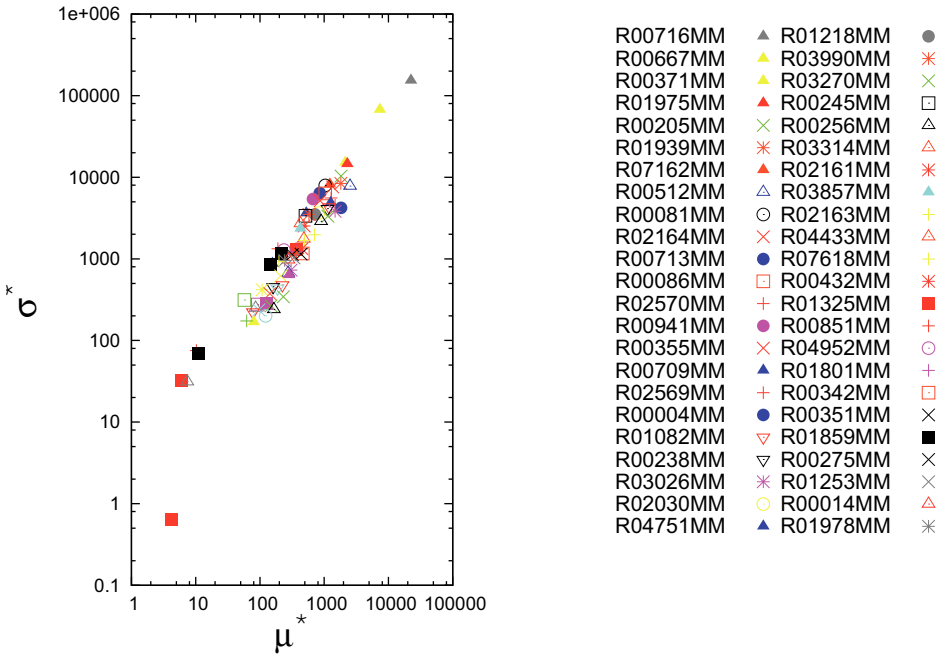


FIGURE 6.5: Sensitivity analysis results for the FBA mitochondrial model [91]. I report the two sensitivity indices for each of the 135 internal reactions. The larger is the value of the two indices, larger is the influence of the corresponding reaction on output of the model. On the key, only the most sensitive reaction IDs are reported. The labels on the key are sorted according to sensitivity ranking. For a detailed description of the key, see the work by Smith et al. [91].

	Reaction name	**LR**
1)	R00014MM Pyruvate+Thiamindiphosphate→ 2-(alpha-Hydroxyethyl)thiaminediphosphate+CO2	84.7
2)	R00081MM COMPLEX IV Oxygen+4Ferrocytochromec→4Ferricytochromec+2H2O	41.2
3)	R00086MM ATP SYNTHASE ADP+Orthophosphate+4H+→ATP+H2O	39.3
4)	R00127MM ATP+AMP→2ADP	45.8
5)	R00157MM UTP+AMP→UDP+ADP	49.4
6)	R00243MM L-Glutamate+NAD++H2O→2-Oxoglutarate+NH3+NADH+H+	99.9
7)	R00275MM 2O2.-+2H+→H2O2+Oxygen	51.7
8)	R00342MM (S)-Malate+NAD+→Oxaloacetate+NADH+H+	89.6
9)	R00351MM Citrate+CoA→Acetyl-CoA+H2O+Oxaloacetate	72.7
10)	R00430MM GTP+Pyruvate→GDP+Phosphoenolpyruvate	52.3
11)	R00432MM GTP+Succinate+CoA→GDP+Orthophosphate+Succinyl-CoA	93.5
12)	R00551MM L-Arginine+H2O→L-Ornithine+Urea	51.8
13)	R00667MM L-Ornithine+2-Oxoglutarate→L-Glutamate5-semialdehyde+L-Glutamate	49.5
14)	R00709MM Isocitrate+NAD+→2-Oxoglutarate+CO2+NADH+H+	57
15)	R00716MM N6-(L-1,3-Dicarboxypropyl)-L-lysine+NADP++H2O→ L-Lysine+2-Oxoglutarate+NADPH+H+	50.2
16)	R00833MM (R)-2-Methyl-3-oxopropanoyl-CoA→Succinyl-CoA	50.3
17)	R01082MM (S)-Malate→Fumarate+H2O	0
18)	R01175MM Butanoyl-CoA+FAD→FADH2+Crotonoyl-CoA	57.5
19)	R01177MM Acetyl-CoA+Butanoyl-CoA→CoA+3-Oxohexanoyl-CoA	0
20)	R01279MM Palmitoyl-CoA+FAD→trans-Hexadec-2-enoyl-CoA+FADH2	54.8
21)	R01325MM Citrate→cis-Aconitate+H2O	73.4
22)	R01361MM (R)-3-Hydroxybutanoate+NAD+→Acetoacetate+NADH+H+	53.3
23)	R01700MM 2-Oxoglutarate+EnzymeN6-(lipoyl)lysine→ S-succinyldihydrolipoyllysine+CO2	94.5
24)	R01900MM Isocitrate→cis-Aconitate+H2O	0
25)	R01923MM Palmitoyl-CoA+L-Carnitine→CoA+L-Palmitoylcarnitine	0
26)	R01978MM (S)-3-Hydroxy-3-methylglutaryl-CoA+CoA→Acetyl-CoA+H2O+Acetoacetyl-CoA	98.9
27)	R02161MM COMPLEX III Ubiquinol+2Ferricytochromec→Ubiquinone+2Ferrocytochromec	44.7
28)	R02163MM COMPLEX I Ubiquinone+NADH→Ubiquinol+NAD+4H+0.002O2.-	53.6
29)	R02164MM COMPLEX II Ubiquinone+Succinate→Ubiquinol+Fumarate	97.5
30)	R02313MM N6-(L-1,3-Dicarboxypropyl)-L-lysine+NAD++H2O→ L-Glutamate+L-2-Aminoadipate6-semialdehyde+NADH+H+	48
31)	R02569MM Acetyl-CoA+EnzymeN6-(dihydrolipoyl)lysine→CoA+S-acetyldihydrolipoyllysine	0
32)	R02570MM Succinyl-CoA+EnzymeN6-(dihydrolipoyl)lysine→ CoA+S-succinyldihydrolipoyllysine	0
33)	R02661MM 2-Methylpropanoyl-CoA+FAD→2-Methylprop-2-enoyl-CoA+FADH2	49.2
34)	R03026MM (S)-3-Hydroxybutanoyl-CoA→Crotonoyl-CoA+H2O	0
35)	R03102MM L-2-Aminoadipate6-semialdehyde+NAD++H2O→L-2-Aminoadipate+NADH+H+	47.6
36)	R03172MM (S)-2-Methylbutanoyl-CoA+FAD→2-Methylbut-2-enoyl-CoA+FADH2	48.5
37)	R03270MM 2-(alpha-Hydroxyethyl)thiaminediphosphate+EnzymeN6-(lipoyl)lysine→ S-acetyldihydrolipoyllysine+Thiamindiphosphate	85.3
38)	R03381MM (S)-Methylmalonatesemialdehyde+CoA+NAD+→ (R)-2-Methyl-3-oxopropanoyl-CoA+NADH+H+	49.5
39)	R03777MM Octanoyl-CoA+FAD→trans-Oct-2-enoyl-CoA+FADH2	56.1
40)	R03778MM Octanoyl-CoA+Acetyl-CoA→CoA+3-Oxodecanoyl-CoA	0
41)	R03857MM Lauroyl-CoA+FAD→2-trans-Dodecenoyl-CoA+FADH2	55.1
42)	R03858MM Lauroyl-CoA+Acetyl-CoA→CoA+3-Oxotetradecanoyl-CoA	0
43)	R03990MM Tetradecanoyl-CoA+FAD→trans-Tetradec-2-enoyl-CoA+FADH2	58.1
44)	R03991MM Tetradecanoyl-CoA+Acetyl-CoA→CoA+3-Oxopalmitoyl-CoA	0
45)	R04170MM (S)-3-Hydroxydodecanoyl-CoA→2-trans-Dodecenoyl-CoA+H2O	0
46)	R04203MM (2S,3S)-3-Hydroxy-2-methylbutanoyl-CoA+NAD+→ 2-Methylacetoacetyl-CoA+NADH+H+	49.2

	Reaction name	**LR**
47)	R04224MM 2-Methylprop-2-enoyl-CoA+H2O→(S)-3-Hydroxyisobutyryl-CoA	51.4
48)	R04433MM Ubiquinone+FADH2→Ubiquinol+FAD+	55.3
49)	R04737MM (S)-3-Hydroxyhexadecanoyl-CoA+NAD+→3-Oxopalmitoyl-CoA+NADH+H+	58.8
50)	R04738MM (S)-3-Hydroxyhexadecanoyl-CoA→trans-Hexadec-2-enoyl-CoA+H2O	0
51)	R04739MM (S)-3-Hydroxytetradecanoyl-CoA+NAD+→3-Oxotetradecanoyl-CoA+NADH	56.5
52)	R04740MM (S)-3-Hydroxytetradecanoyl-CoA→trans-Tetradec-2-enoyl-CoA+H2O	0
53)	R04741MM (S)-3-Hydroxydodecanoyl-CoA+NAD+→3-Oxododecanoyl-CoA+NADH+H+	54.7
54)	R04742MM Decanoyl-CoA+Acetyl-CoA→CoA+3-Oxododecanoyl-CoA	0
55)	R04743MM (S)-Hydroxydecanoyl-CoA+NAD+→3-Oxodecanoyl-CoA+NADH+H+	56.8
56)	R04744MM (S)-Hydroxydecanoyl-CoA→trans-Dec-2-enoyl-CoA+H2O	0
57)	R04745MM (S)-Hydroxyoctanoyl-CoA+NAD+→3-Oxooctanoyl-CoA+NADH+H+	55.6
58)	R04746MM (S)-Hydroxyoctanoyl-CoA→trans-Oct-2-enoyl-CoA+H2O	0
59)	R04747MM Hexanoyl-CoA+Acetyl-CoA→CoA+3-Oxooctanoyl-CoA	0
60)	R04748MM (S)-Hydroxyhexanoyl-CoA+NAD+→3-Oxohexanoyl-CoA+NADH+H+	54.5
61)	R04749MM (S)-Hydroxyhexanoyl-CoA→trans-Hex-2-enoyl-CoA+H2O	0
62)	R04751MM Hexanoyl-CoA+FAD→trans-Hex-2-enoyl-CoA+FADH2	55.2
63)	R04754MM Decanoyl-CoA+FAD→trans-Dec-2-enoyl-CoA+FADH2	58.2
64)	R05064MM (S)-3-Hydroxyisobutyryl-CoA+H2O→CoA+(S)-3-Hydroxyisobutyrate	50.6
65)	R05066MM (S)-3-Hydroxyisobutyrate+NAD+→(S)-Methylmalonatesemialdehyde+NADH+H+	51.1
66)	R07599MM 3-Methyl-2-oxobutanoicacid+Thiamindiphosphate→ 2-Methyl-1-hydroxypropyl-ThPP+CO2	50.2
67)	R07600MM 2-Methyl-1-hydroxypropyl-ThPP+EnzymeN6-(lipoyl)lysine→ S-(2-methylpropanoyl)dihydrolipoyllysine+Thiamindiphosphate	51.3
68)	R07603MM (S)-3-Methyl-2-oxopentanoicacid+Thiamindiphosphate→ 2-Methyl-1-hydroxybutyl-ThPP+CO2	48.4
69)	R07604MM 2-Methyl-1-hydroxybutyl-ThPP+EnzymeN6-(lipoyl)lysine → S-(2-methylbutanoyl)dihydrolipoyllysine+Thiamindiphosphate	48.9
70)	R07618MM EnzymeN6-(dihydrolipoyl)lysine+NAD+→EnzymeN6-(lipoyl)lysine+NADH+H+	46.1

TABLE 6.5: Local Robustness (%) for metabolic fluxes in mitochondria. In Table the name and the ID of the metabolic reactions whose robustness value is less than 100%.

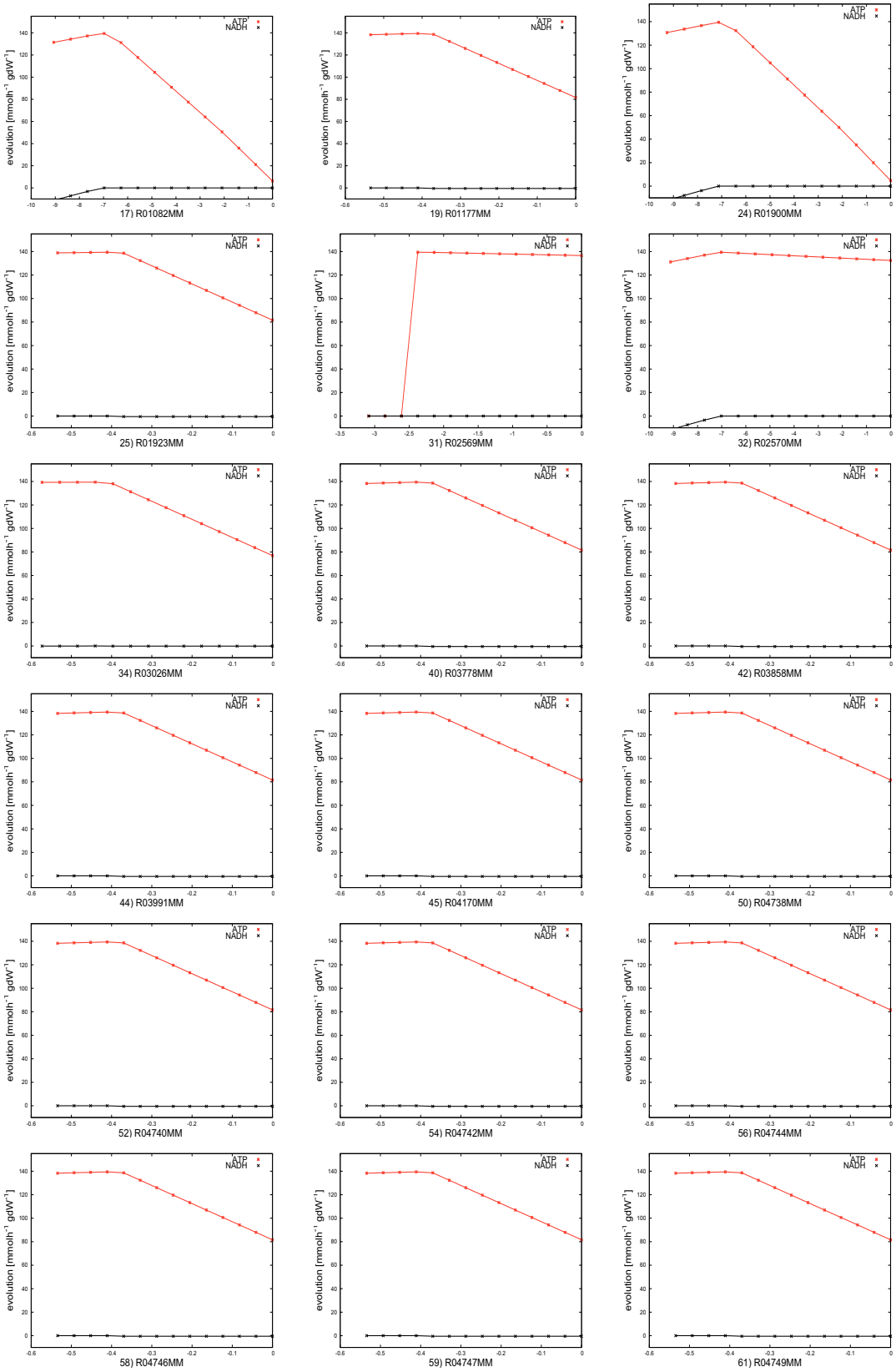


FIGURE 6.6: Response in mitochondria after perturbing metabolic fluxes, in particular in figure has been shown the ATP and NADH production for all the metabolic fluxes with a local robustness indexes equal to 0 (Table 6.5).

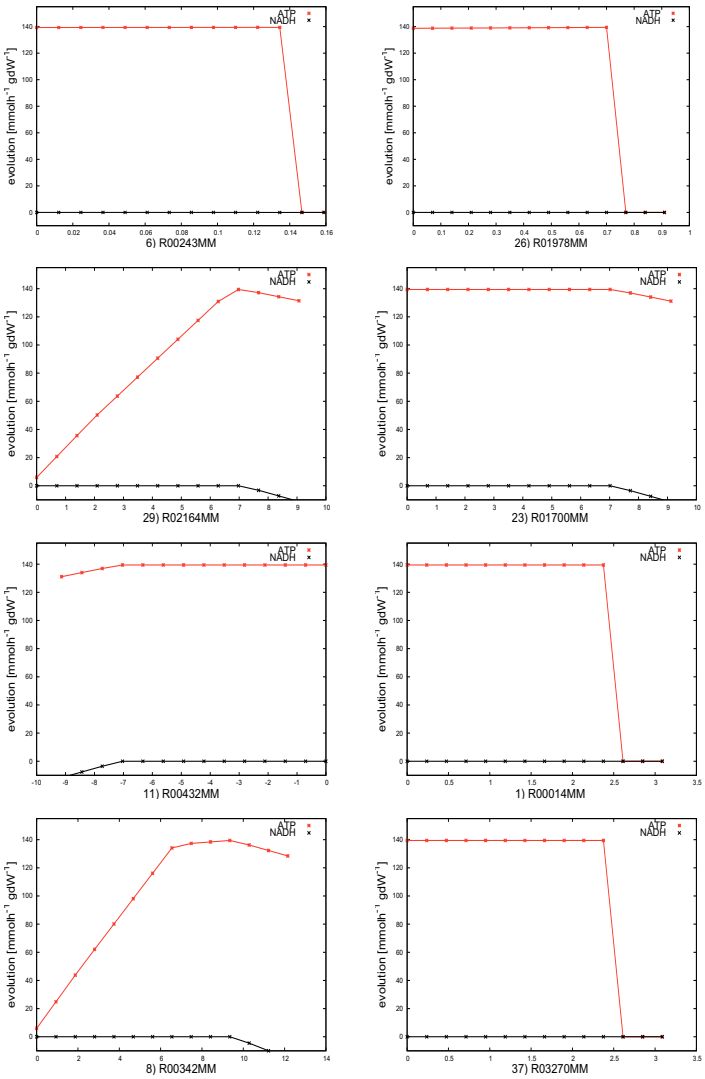


FIGURE 6.7: Response in mitochondria after perturbing metabolic fluxes, in particular in figure has been shown the ATP and NADH production for the most robust metabolic fluxes reported in Table 6.5: R00243MM=99.9%, R01978MM=98.9%, R02164MM=97.5%, R01700MM=98.8%, R00432MM=93.5%, R00014MM=84.7%, R00342MM=89.6%, R03270MM=95.3%.

6.3 Identification of Healthy and Pathological States in mitochondrial metabolism

By using NSGA-II [25], I optimized multiple energy-related objectives. The model I adopt here consists of 73 differential-algebraic equations (DAEs) to model the mitochondrial bioenergetics [87]. In particular, the model accounts for 35 biochemical reactions, including the oxidative phosphorylation, the electron transport system, the tricarboxylic

acid cycle and related reactions, the Na^+/Ca^2+ cycle and the K^+-cycle. As in the previous case studies, I maximized the production of adenosine triphosphate (ATP) and nicotinamide adenine dinucleotide (NADH).

The variable space is defined as the space of feasible initial concentrations of 50 metabolites and the population is initialized in a random way.

I simultaneously maximized ATP and NADH production varying the initial conditions (the initial metabolites contents) used to solve the DAEs system. In particular, I varied the initial conditions for 50 metabolites (the decision variables), while maintaining fixed the following: 1) matrix water volume, 2) inner membrane space water volume, 3) matrix free chloride, 4) total matrix ATP, 5) total matrix ADP, 6) total matrix GTP, 7) total matrix GDP, 8) total matrix NADH, 9) total matrix NAD, 10) total mitochondrial ubiquinol, 11) total mitochondrial ubiquinone, 12) total CO_2 matrix, 13) total O_2 matrix, 14) inter membrane space (IMS) free proton, 15) IMS free potassium, 16) IMS free magnesium, 17) IMS free calcium, 18) IMS free sodium, 19) total IMS cytochrome c2+, 20) total IMS cytochrome c3+. Before the optimization, at the fully oxidized state NADH is equal to $1.5987 \cdot 10^{-10}$ nmol/mg and ATP -0.0014 nmol/mg. After the optimization, we have the Pareto-optimal points in black shown in Figure 6.8.

If the matrix calcium content is increased from 10^{-5} to 10^{-4} nmol/mg, the ATP synthesis and NADH formation decrease (see Figure 6.8, red signs). Experiments in the figure have been conducted by incrementing calcium to 10^{-4} and $1.5 \cdot 10^{-5}$, and by reducing calcium to 10^{-6} and $10^{-5}/1.5$. These experiments can demonstrate that a perturbation in mitochondrial Ca^{2+} homeostasis has major implications for cell function at the level of ATP synthesis and NADH generation. Labels in the figure reports the maximum ATP value for each Pareto front.

For the cancer studies, I focused the analysis on mitochondrial activity and its role on cancer diseases. To simulate the cancer environment, I changed the initial condition for i) the mitochondrial membrane potential $\Delta\Phi$, ii) the total extra-mitochondrial glucose-6-phosphate content (G6P) and iii) the extra-mitochondrial free proton content (H^+). I fixed $\Delta\Phi = 2$ mV, G6P $= V_{cyt} \times 10^6$ nmol/mg and $H^+ = V_{cyt} \times 10^{-5}$ nmol/mg. Instead, in healthy state, $\Delta\Phi = 1$ mV, G6P $= V_{cyt} \times 10^{12}$ nmol/mg and $H^+ = V_{cyt} \times 10^7$ nmol/mg, where V_{cyt} is the water volume in cytosolic space. In a second step, I used the same design to minimize ATP and NADH when the mitochondrion is in a cancer condition in order to find the variables playing a crucial role to kill the cancer cell. The results of the optimization are shown in Figure 6.9. In this way, I can distinguish between healthy and pathological state. The red and green regions represent respectively the pathological and healthy state in mitochondria during the production of ATP and NADH. The purple region represents the apoptosis state of a cell in cancer conditions,

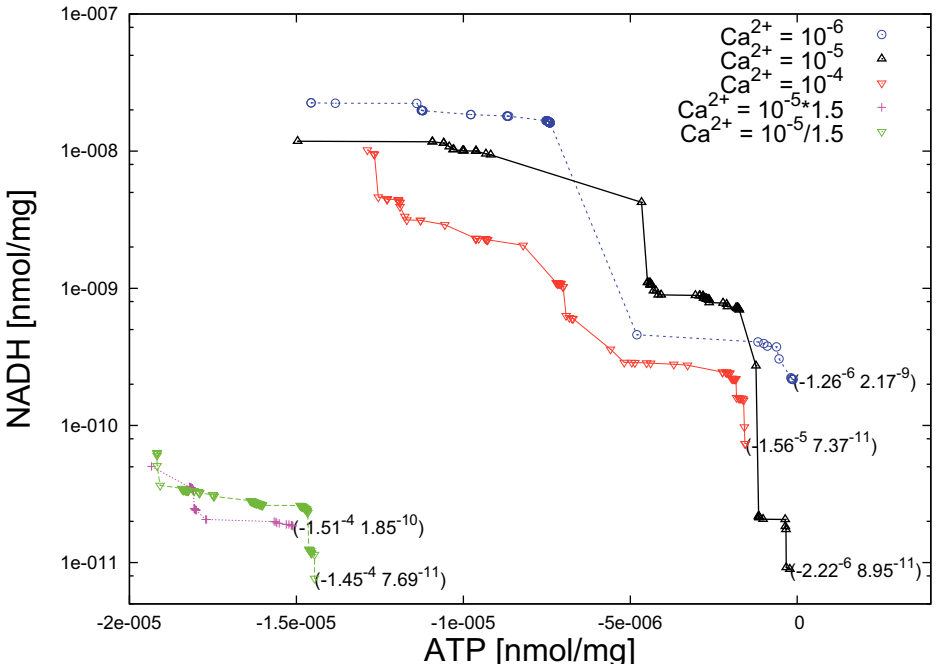


FIGURE 6.8: ATP and NADH production maximization in the mitochondrial DAEs model [87]. I varied the initial concentrations of 50 metabolites and solved the system of DAEs. I simulated five differential states based on the concentration of calcium in the matrix: the standard concentration (10^{-5} nmol/mg), two increments of the standard concentration (10^{-4} nmol/mg and $1.5 \cdot 10^{-5}$ nmol/mg) and two decrements of the standard concentration (10^{-6} nmol/mg and $10^{-5}/1.5$ nmol/mg).

i.e., when mitochondria are not able to produce bioenergy. The black line marks the Pareto optimal solutions. In the cancer experiments, the initial Ca^{2+} concentration used in all the simulations is 10^{-5} nmol/mg.

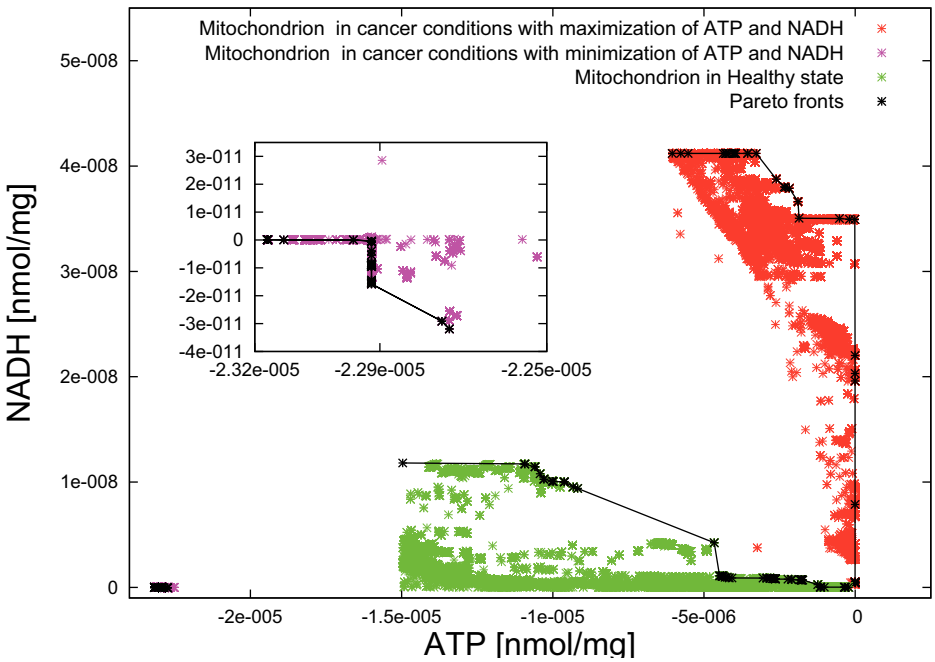


FIGURE 6.9: ATP and NADH production maximization and minimization in the mitochondrial DAEs model [87]. The regions define the states on mitochondria in different conditions. The red region represents the pathological state (cancer conditions), while the green zone the healthy state. In purple, the mitochondrial pathological state with the minimization of ATP and NADH production, i.e., during the cellular apoptosis. Pareto fronts are in black.

Chapter 7

Conclusion

After the advent of the genomic and proteomic sciences, scientists started to consider a cell as a system. From the reductionistic point of view of the 20th century we have passed to the integrative approach with the coming of bioinformatics, systems science, modeling and simulation to study how biological entities interact to form complex systems. With the Human Genome Project's completion, and with increasing amounts of expression data becoming available, growing attention is being paid to in silico biology. The term in silico biology refers to the use of computers to perform biological studies. Computations of the structures of complex biomolecules are currently routinely performed. Now, the mathematical description and computer simulation of the simultaneous action of multiple gene products is growing in importance, and in the view of many, will take center stage in biology in the coming decades [92]. With the advancements in computing power it is becoming possible to profile and model a complete biological system [4]. A big challenge was reached the last year, when the group of Markus W. Covert published the first in silico whole cell [5]. Conversely, this year is started the *Human Brain Project*, which aims to simulate the complete human brain network on supercomputers to better understand how it functions. The project is grouping many research groups of the most important European university and research centers.

With this thesis, I contribute on Systems Biology studies through the development of computational tools able to design biological systems. BioCAD software is a general purpose framework and can be seen as a black box able to analyze any type of circuit, that can be modeled with different mathematic approaches. BioCAD can be used also for other applications, such as electronic design automation for the design of electronic circuits. In this thesis, I considered biological circuits modeled through ordinary

differential equations, differential algebraic equations or by means the flux balance analysis approach. The methodology is also suitable for any simulator (e.g., SBML, Matlab, C/C++ program and NEURON). BioCAD re-designs natural biological systems for useful purposes and analyzes components and submodules. It is formed by three parts, each of them performs respectively the pre-processing analysis, optimization and post-processing analysis. The optimization is the main part and implements novel and state-of-art algorithms. It faces single- and multi-objective optimization problems, by menage continuous or discrete variable spaces. The optimization core implements also the novel Genetic Design through Multi-objective Optimization (GDMO) algorithm that finds the optimal genetic manipulation in terms of knockout in bacteria or other organisms to outperform specific biological functions. The results is a set of Pareto optimal bacterial strains. Pareto optimality has resulted to be very suited for metabolic analysis and for cross-comparing the behavior of different organisms, for example for synthetic biology targets. GDMO was also compared with previous methods, such as the most recent GDLS [11]. Performances have revealed better results and suitable genetic manipulations: for example GDMO found an acetate production that increments more than 130% in *Escherichia coli* bacterium when undergoes anaerobic conditions versus an acetate production of 91,8% obtained by GDLS. The computational time depends on the parameters of the model and of the algorithms. Time to perform FBA depends on the size of the biological circuit: for example for the *Escherichia coli* bacterium, by using a 3,4 GHz processor and 32 Gb RAM, the elapsed time is (i) 0.0629 seconds in the *i*JR904 model (931 reactions, 624 metabolites, 904 genes), (ii) 0.0759 seconds in the *i*AF1260 model (2382 reactions, 1039 metabolites, 1260 genes) and (iii) 0.1305 seconds in the *i*JO1366 model (2251 reactions, 1136 metabolites, 1366 genes). The performance of the optimization algorithm depends by the number of individuals and iterations: for the bigger *Escherichia coli* circuit, by using a population of 1000 individuals, GDMO reaches 1000 generations after four days. In the example above illustrated, the decision variables are genes, but BioCAD can menage also other elements of the metabolic circuit as well as enzymes, metabolites and reaction fluxes. At this stage can be found diverse novelty, for example the new mutation operator used in the evolutionary algorithm for optimizing the metabolic capabilities and that takes into account of constraints in flux balance approach. For instance, by analyzing the enzymes of the carbon metabolism pathway, RuBisCO has resulted the most sensitive enzyme and the UDP-glucose pyrophosphorylase (UDPGP) the most important for optimizing carbon dioxide consumption. The pre-processing step and in particular the sensitivity analysis ranks species, reactions, pathways of metabolic networks. The new Gene sets-PoSA and Fluxes-PoSA are able to find the most sensitive pathways in terms of genes and fluxes in flux balance analysis. Through the sensitivity, we can resize the set of decision variables and optimize only the most important ones. This could improve the analysis in terms of time and

computational efforts. The results revealed BioCAD overcoming previous methods and gave interesting highlights. The future directions foresee the extension of BioCAD by adding novel parts for the model checking, model order reduction, the implementation of methods for the reverse engineering of gene regulatory networks and the integration of models from data bases such as Biomodels DataBase.

Appendix A

Supplementary Figures and Tables

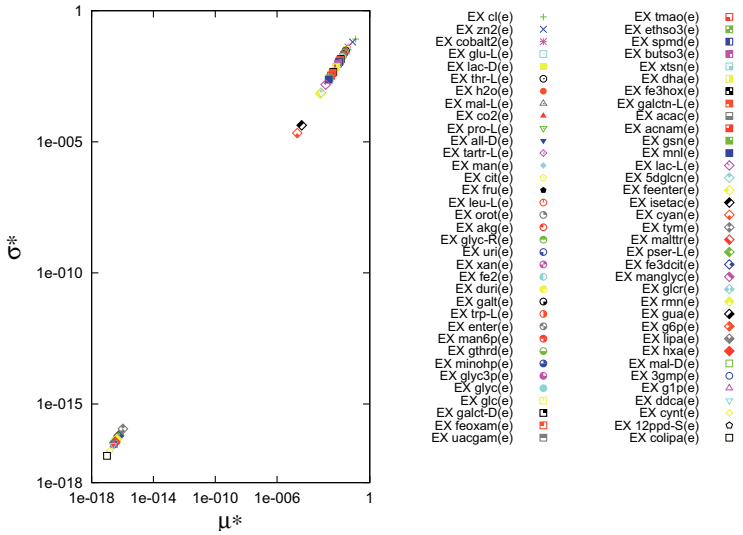


FIGURE A.1: Species-oriented Sensitivity Analysis in *i*AF1260 *E. coli*. I investigate the input fluxes of the model (299 nutrients) and evaluate their sensitivity with respect to all the fluxes of the model using the Morris method [28]. I find that only 70 fluxes (reported in the key) are influent, while the other fluxes have sensitivity indices equal to zero.

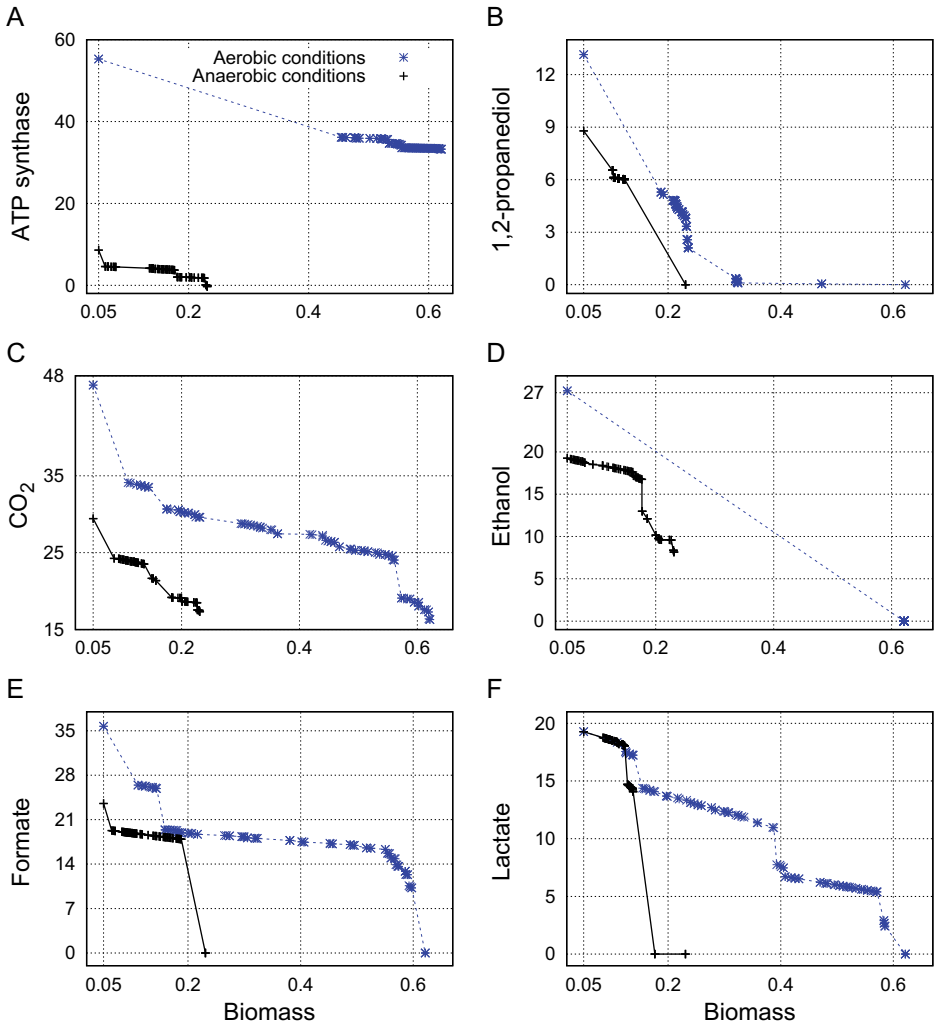


FIGURE A.2: Pareto fronts for six experiments using GDMO in *i*AF1260 *E. coli*. I simultaneously maximize the biomass formation [h^{-1}] against ATP synthase rate (A), 1,2-propanediol production rate (B), CO_2 (C), ethanol (D), formate (E), and lactate (F) production rates [$mmolh^{-1}$ gDW^{-1}]. In blue, the aerobic condition with O_2=10 $mmolh^{-1}$ gDW^{-1}, while in black, the anaerobic condition.

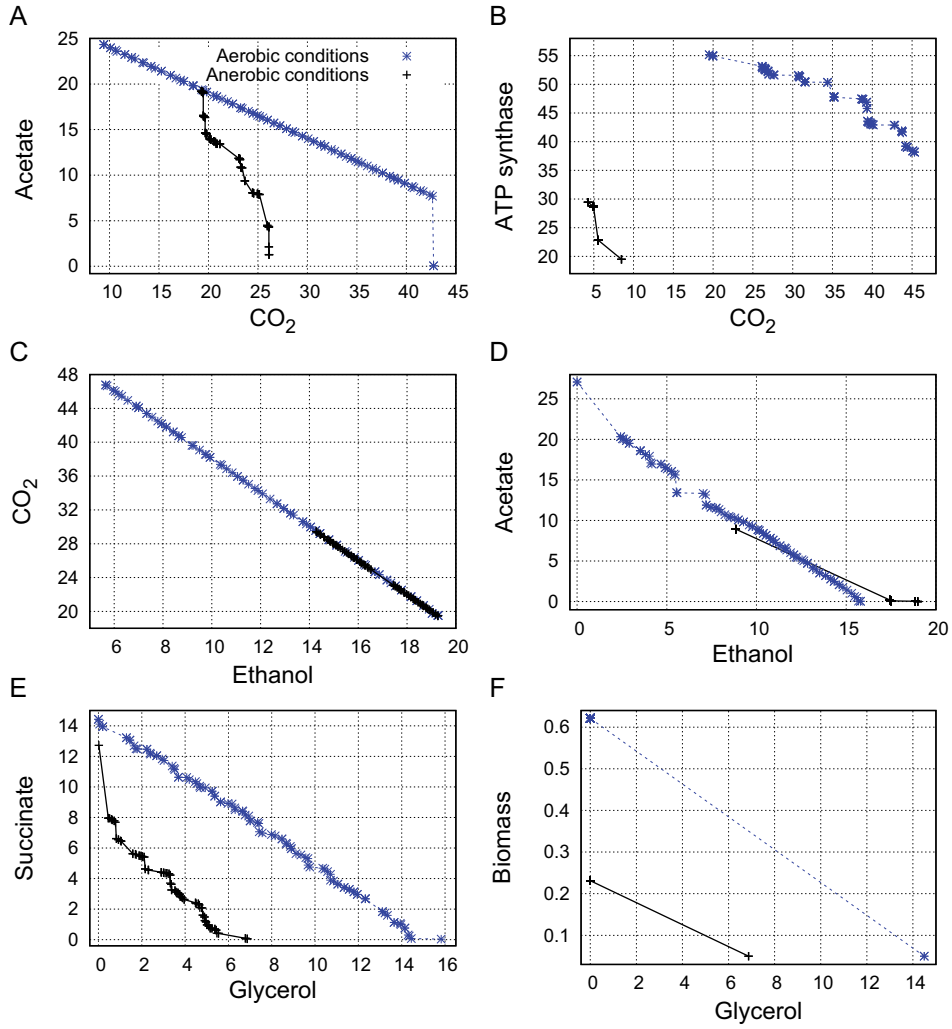


FIGURE A.3: Pareto fronts for six experiments using GDMO in *i*AF1260 *E. coli*. I simultaneously maximize the production rates of acetate and CO_2 (A), ATP synthase rate and CO_2 production rate (B), ethanol and CO_2 production rates (C), acetate and ethanol production rates (D), succinate and glycerol production rates (E), glycerol production rate and biomass formation (F) [h^{-1}]. All flux rates are expressed in [mmolh^{-1} gDW^{-1}], except for biomass formation, that is expressed in [h^{-1}]. The maximizations are performed in anaerobic (black signs) and aerobic (blue signs, O_2=10 mmolh^{-1} gDW^{-1}) conditions.

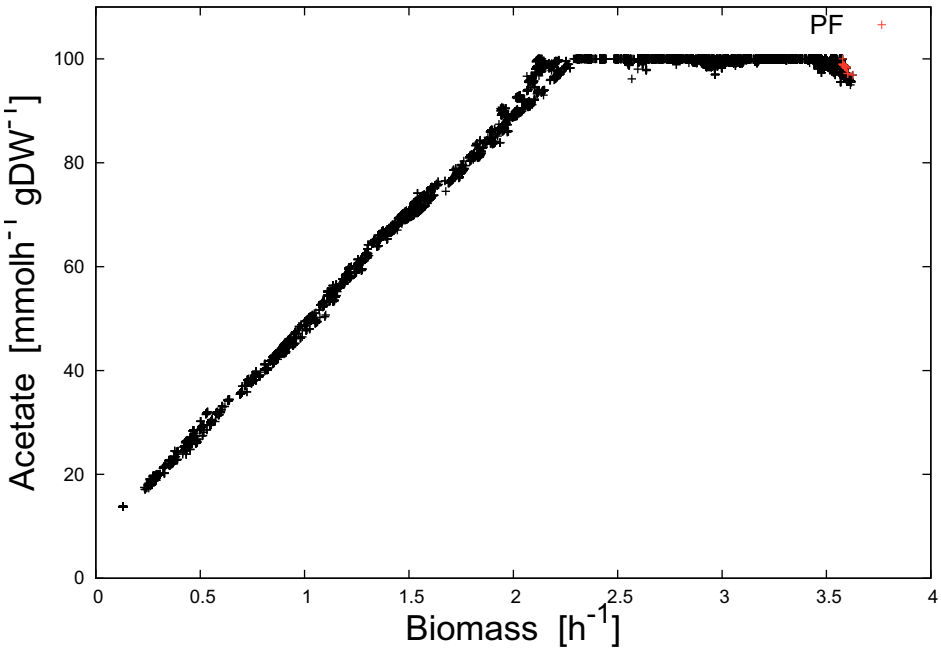


FIGURE A.4: Flux design in *i*AF1260 *E. coli* for the maximization of the acetate production. By means of GDMO I find the optimal strain A1 reported in Table 3.1. In a second optimization, I search for the optimal nutrients in order to maximize the acetate production [mmolh^{-1} gDW^{-1}] and biomass formation [h^{-1}]. In red Pareto front and in black feasible points.

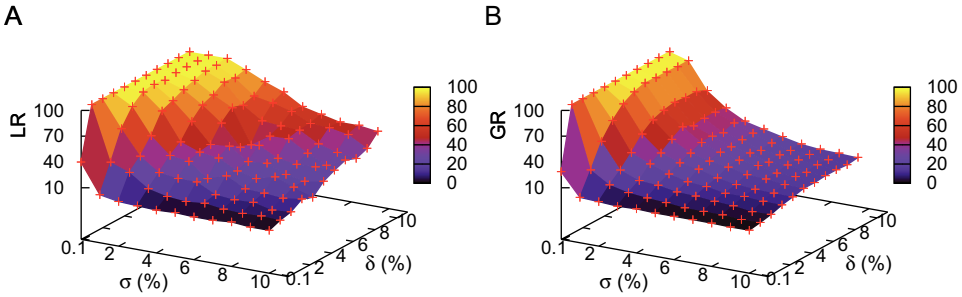


FIGURE A.5: Local (A) and Global (B) Robustness variation versus σ and δ values. These results are calculated for the choice of μ and σ. σ is proportional to the perturbed parameters/fluxes, while δ is proportional to the metrics (in this case acetate and biomass productions). For the local value, I report the minimum value found (this value is associated with the glucose uptake rate). Since I are interested in the behavior of a *i*AF1260 *E. coli* strain when subjected to small perturbations, and since the behavior is acceptable when the deviation from the original value is as small as possible, I choose the values of ϵ equal to 1% of the metrics and σ equal to 1% of the perturbed variables.

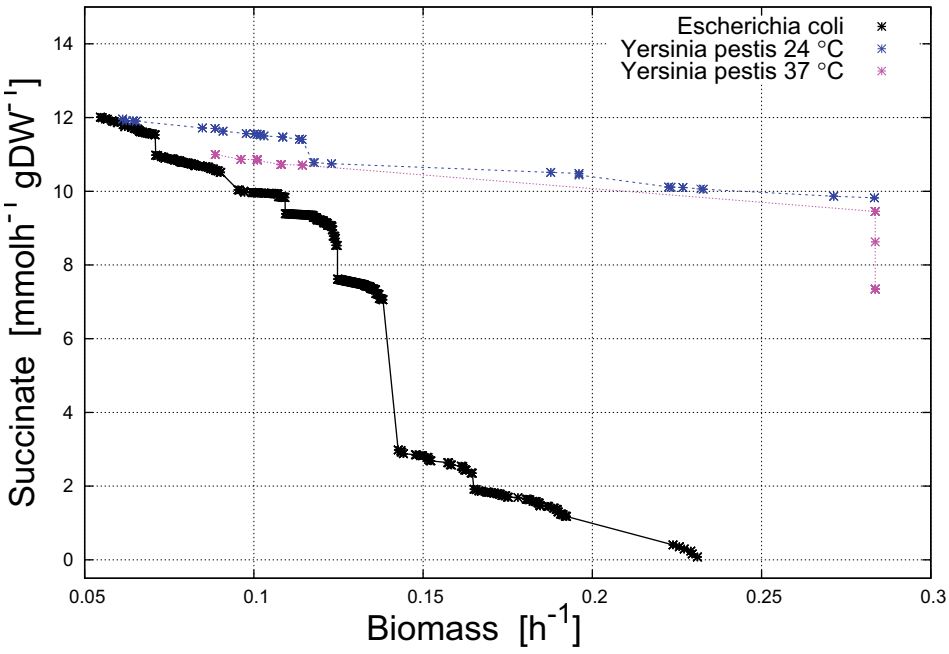


FIGURE A.6: Pareto fronts obtained by optimizing the succinate production [mmolh^{-1} gDW^{-1}] and the biomass formation [h^{-1}] using GDMO in two organisms: *E. coli* [47] and *Y. pestis* [51]. The algorithm is not applicable for *M. barkeri* [53] and *G. sulfurreducens* [52]. Indeed, the *M. barkeri* reconstruction does not provide succinate, and for *G. sulfurreducens* I obtain the maximum yield in wild type, probably because of a different metabolic mechanism and a low number of reactions in the model.

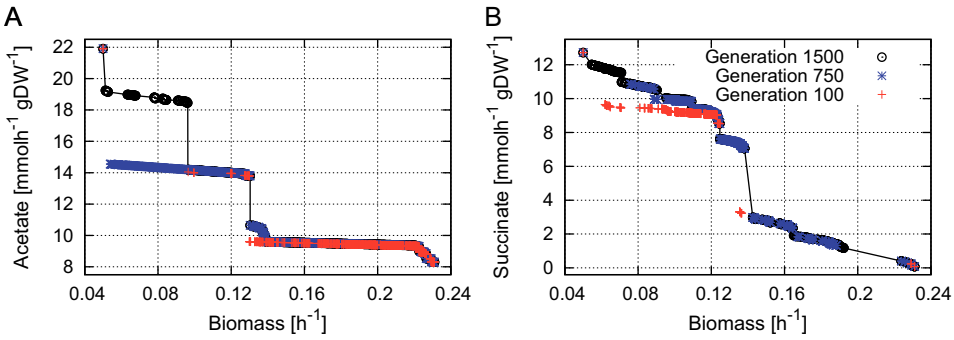


FIGURE A.7: Convergence process of GDMO for acetate (A) and succinate (B) production in *i*AF1260 *E. coli*. I report the Pareto fronts obtained during the initial phase of the method (generation 100, in red), in an intermediate phase (generation 750, in blue), and the final result (generation 1500, in black). The experiments have been conducted using a population of 1000 individuals.

TABLE A.1: Results of the Global Robustness (GR), Local Robustness (LR) and the normalized volume of the robust parameters (R) related to acetate and succinate optimization in *i*JO1366 *E. coli*. Points have been selected from Pareto fronts of Figure 3.7-A-C. "W" indicates the wild type configuration

Strains of *i*JO1366 *E. coli*	GR (%)	LR (%)	R	Acetate ($mmolh^{-1} \cdot gDW^{-1}$)	Biomass (h^{-1})	KC
W	22.88	17.29	1.36	4.446	1.033	0
A	12.81	15.03	1.29	19.790	0.016	19
B	25.56	38.35	1.79	10.644	0.702	8
C	54.61	56.39	1.90	16.208	0.252	1
Strains of *i*JO1366 *E. coli*	GR (%)	LR (%)	R	Succinate ($mmolh^{-1} \cdot gDW^{-1}$)	Biomass (h^{-1})	KC
W	4.46	0	1	0	1.033	0
A	94.58	98.49	1.23	1.072	1.028	3

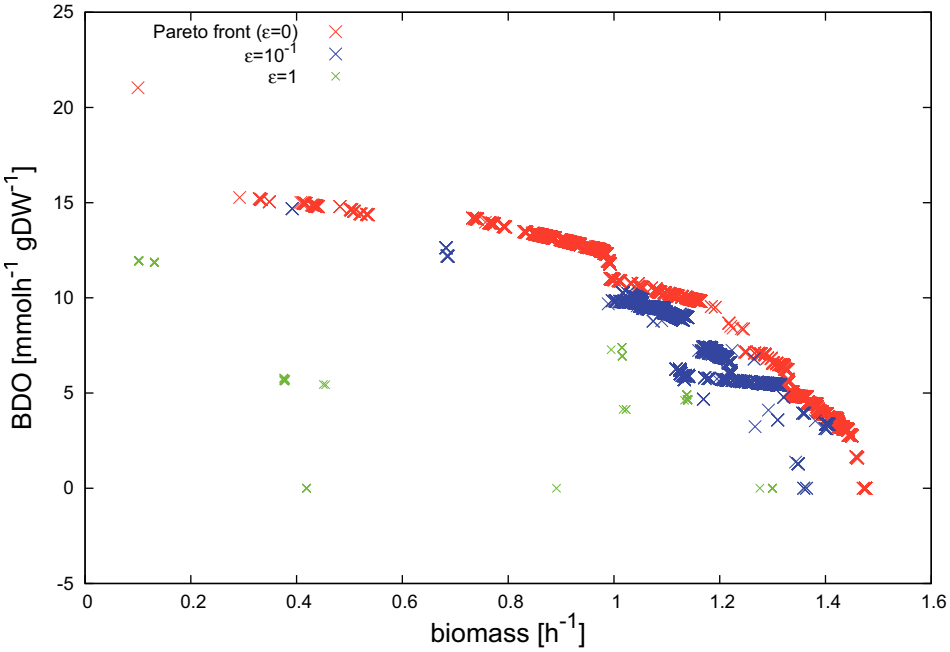
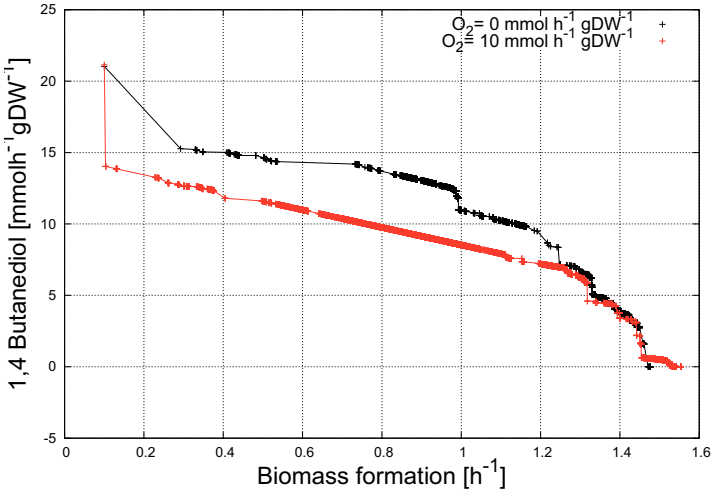
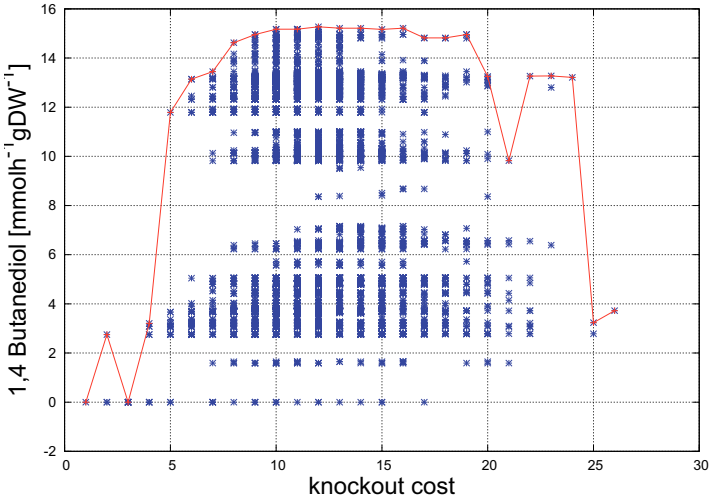


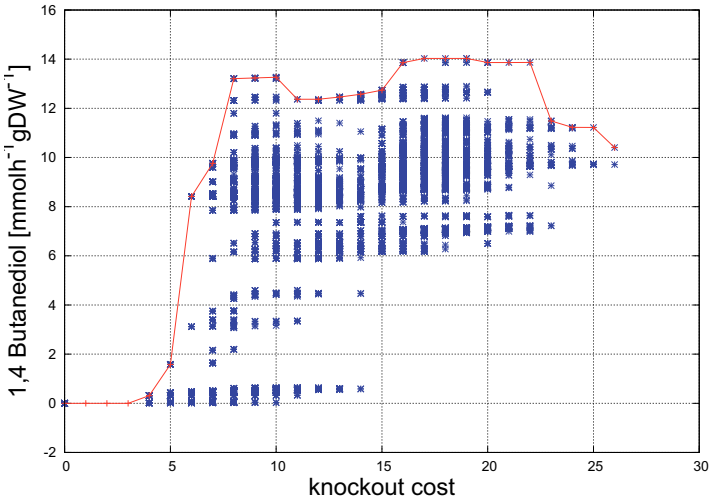
FIGURE A.8: ϵ-dominance and Pareto front in the original *i*JR904 *E. coli* obtained for maximizing BDO production.



A)

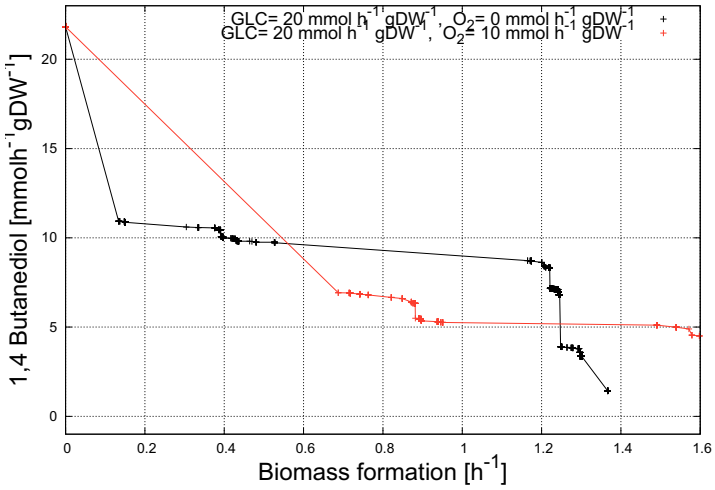


B)

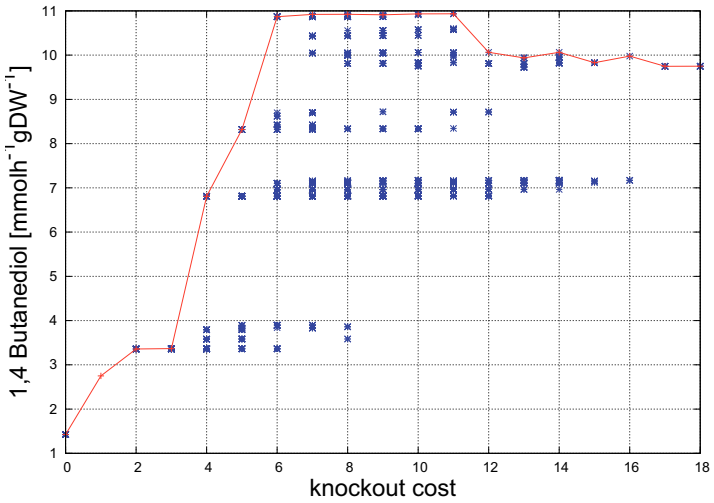


C)

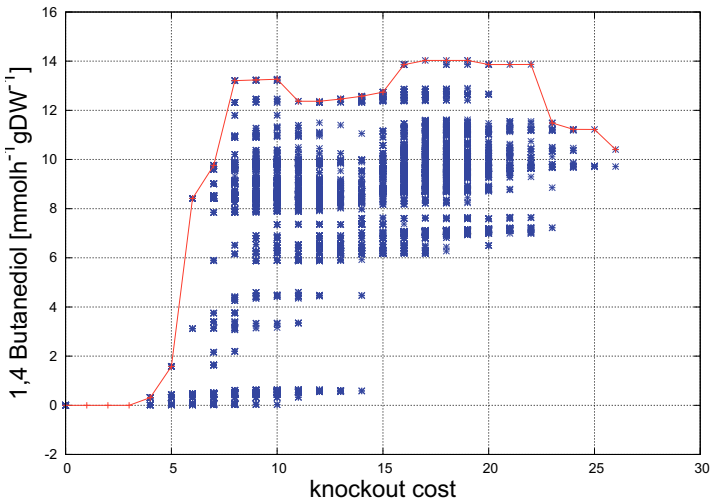
FIGURE A.9: (A) Genetic strategies optimization to maximize 1-4 butanediol (BDO) and biomass in the original *i*JR904 *E. coli* in anaerobic/aerobic conditions and glucose feed equal to 20 $mmolh^{-1}$ gDW^{-1} by using C=10, I=1000 gen=3000. Knockout cost versus BDO production for genetic strategies optimization with C=10 in anaerobic (B) and aerobic conditions (C).



A)



B)



C)

FIGURE A.10: (A) Genetic strategies optimization to maximize 1-4 butanediol (BDO) and biomass in *i*JO1366 *E. coli* in anaerobic/aerobic conditions and glucose feed equal to 20 mmolh^{-1} gDW^{-1} by using C=10, I=1000 gen=3000. Knockout cost versus BDO production for genetic strategies optimization with C=10 in anaerobic (B) and aerobic conditions (C).

TABLE A.2: Knocked out gene sets occurrences for maximizing BDO production in *i*JO1366 *E. coli* [50] with respect to 726 Pareto strains (front in black, C=50, Figure 4.5).

Gene sets ID	*Occurrences*	*Frequency*
((b0902 and b0903) or (b0902 and b3114) or (b3951 and b3952) or ((b0902 and b0903) and b2579))	450	0.619%
(b3640 or b2251)	190	0.26%
b3029	179	0.246%
b3844	128	0.176%
b3867	125	0.172%
b2529	121	0.167%
((b4079 and (b2481 and b2482 and b2483 and b2484 and b2485 and b2486 and b2487 and b2488 and b2489 and b2490)) or (b4079 and (b2719 and b2720 and b2721 and b2722 and b2723 and b2724)))	118	0.163%
b0221	111	0.153%
(b2528 and b2529)	102	0.140%
b3236	102	0.140%
b0243	96	0.132%
(b3417 or b3428)	91	0.126%
(b3161 or b3709 or b1473 or b0112)	89	0.123%
b1855	85	0.117%
(b1602 and b1603)	81	0.112%
(b2530 and b2529)	79	0.109%
b2508	77	0.106%
b0511	76	0.105%

TABLE A.3: Knockout strategies obtained through GDMO for maximizing BDO production by using C=50 in the *i*JO1366 *E. coli* [mmolh^{-1} gDW^{-1}].

Str	BDO	Biom	kcost	Genes	Pathways	Reactions
A_1	10.8692	0.15085				
	662.67%	-88.96%	6	((b0902 and b0903) or (b0902 and b3114) or...	Pyruvate Metab.	PFL
				...(b3951 and b3952) or ((b0902 and b0903) and b2579))		
				b3029	Oxidative Phosphorylation	QMO2
					Oxidative Phosphorylation	QMO3
				b3919	Glycolysis/Gluconeogenesis	TPI
					Transport, Inner Membrane	GLYCLTt2r
					Transport, Inner Membrane	L-LACt2
				b2492, b0904	Transport, Inner Membrane	FORt
A_2	8.3188	1.2205				
	483.71%	-10.70%	5	((b0902 and b0903) or (b0902 and b3114) or ...	Pyruvate Metab.	PFL
				...(b3951 and b3952) or ((b0902 and b0903) and b2579))		
				(b1602 and b1603)	Oxidative Phosphorylation	THD2pp
A_3	6.8094	1.2457				
	377.80%	-8.85%	4	((b0902 and b0903) or (b0902 and b3114) or ...	Pyruvate Metab.	PFL
				...(b3951 and b3952) or ((b0902 and b0903) and b2579))		
A_4	3.3581	1.3011	2			
	135.63%	-4.80%				
				((b4079 and (b2481 and b2482 and b2483 and b2484 and ...	Pyruvate Metab.	FHL
				... b2485 and b2486 and b2487 and b2488 and b2489 ...		
				... and b2490)) or (b4079 and (b2719 and b2720 and ...		
				... b2721 and b2722 and b2723 and b2724)))		

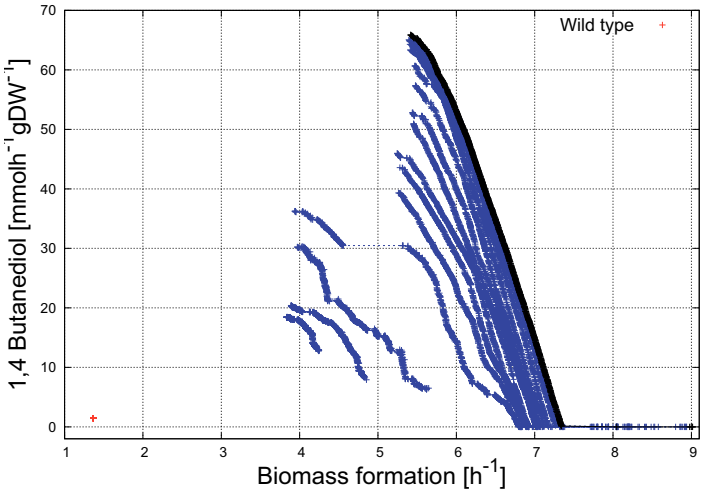


FIGURE A.11: Environmental optimization to maximize Biomass versus 1-4 butanediol synthetic production in *i*JR904 *E. coli*, in anaerobic conditions and maintaining glucose feed fixed to 20 mmolh^{-1} gDW^{-1}. In blues, I show the feasible solutions found each 100 iterations of the algorithm.

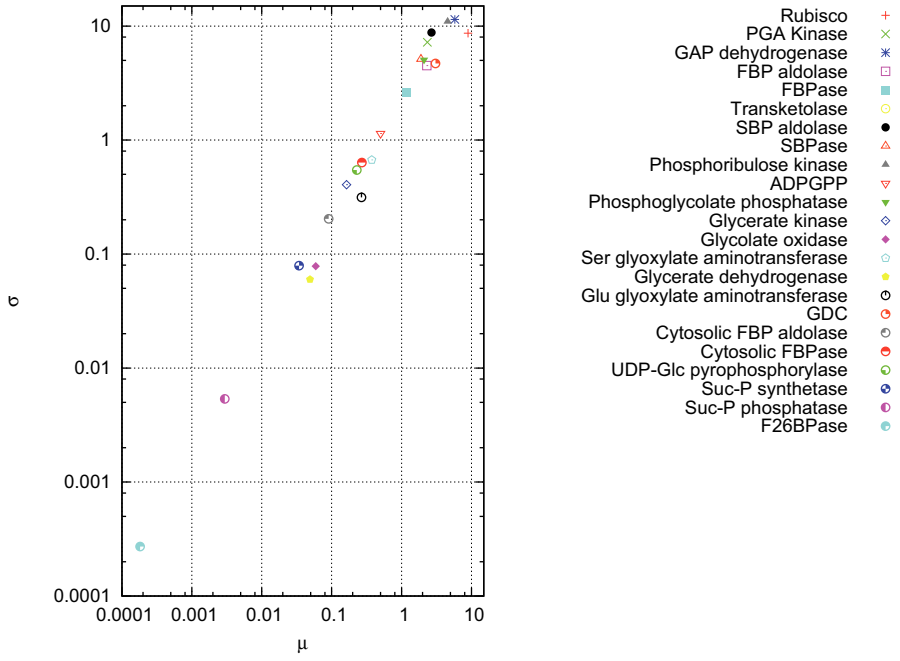


FIGURE A.12: Sensitive enzymes in photosynthetic carbon metabolism obtained by the Morris method [28].

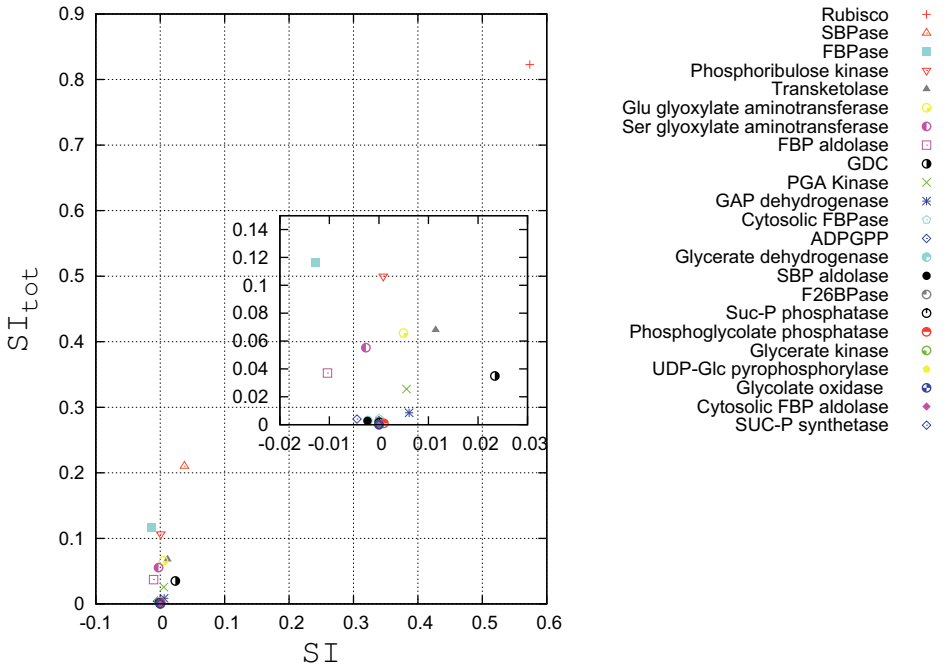


FIGURE A.13: Sensitive enzymes in photosynthetic carbon metabolism obtained by the Sobol' method. SI and SI_{tot} are the sensitive indexes calculated by the Sobol' method [29], and are comparable with the indexes μ^* and σ^* found by the Morris method.

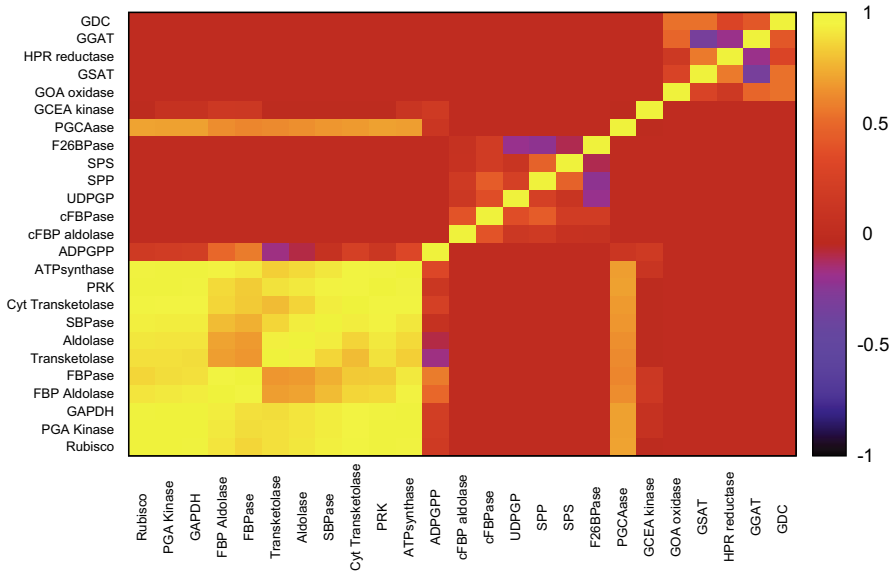


FIGURE A.14: Enzyme correlation matrix in the photosynthetic carbon metabolism. The correlation matrix measures the interdependence between the parameters and gives an idea of the compensation effects of change in the parameter values on the model output. We perform the derivative-based Global Sensitivity Measures (DGSM) with the toolbox SensSB [27]. The matrix shows that RuBisCO has a high correlation with the following enzymes: PGA kinase, GAPDH, cyt-Transketolase, PRK.The enzyme PGA kinase shows a high correlation with GAPDH and PRK. GAPDH is correlated with PRK and FBA aldolase is correlated with FBPase.

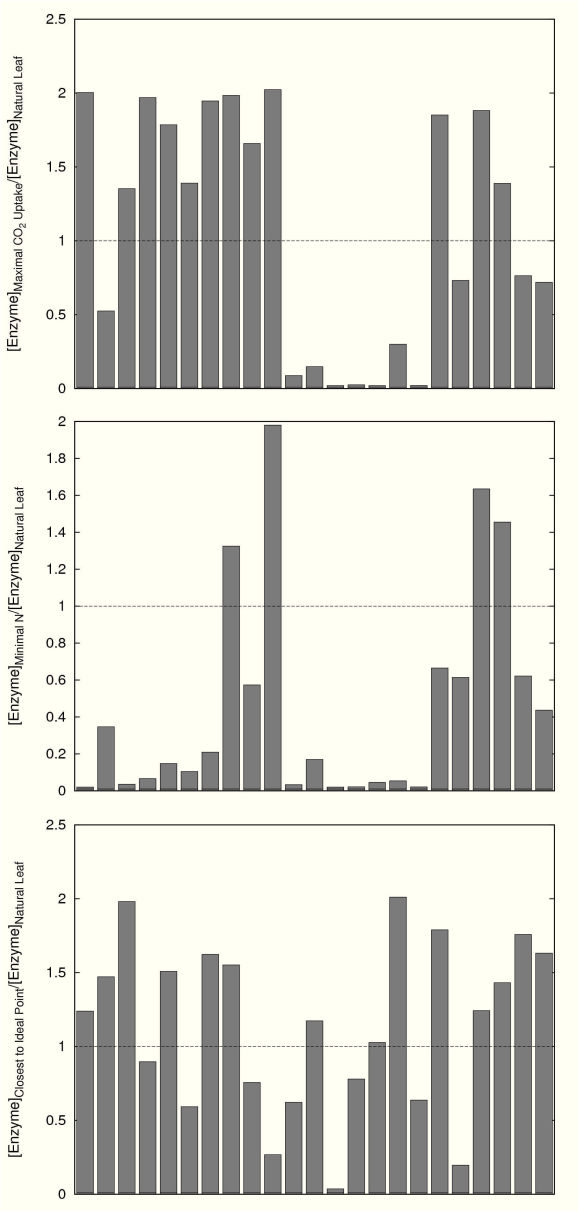


FIGURE A.15: Changes in the concentrations of carbon metabolism enzymes with respect to their natural values when three alternative strategic leaf designs are considered: (i) maximal CO_2 uptake: CO_2 uptake rate is 39.968 (Top plot); (ii) minimal nitrogen consumption: CO_2 uptake rate is 5.7 (middle plot); (iii) closest-to-ideal solution: CO_2 uptake rate is 21.213 (bottom plot). The maximal rate of triose-P (PGA, GAP, and DHAP) export is kept fixed at 1 mmol L^{-1} s^{-1} and the c_i is 270 μmol mol^{-1} to reflect nowadays condition.

Enzyme Name	Sensitive Enzymes	Fragile Enzymes	Important for maximizing CO_2 uptake rate	Light control. Energy convert	Best solution
RuBisCO	X	X	X	X	860.226 (100)
PGA kinase	X			X	3.989 (100)
GAP DH	X	X		X	64.483 (100)
FBP aldolase	X		X		9.05 (100)
FBPase	X			X	26.889 (100)
Transketolase					8.247 (100)
SBP aldolase	X				6.661 (100)
SBPase	X		X	X	4.397 (100)
PRK	X			X	7.007 (100)
ADPGPP	X		X		0.721 (100)
PGCA Pase	X				0.325 (100)
Glycerate kinase					0.005 (100)
Glycolate oxidase					0.019 (100)
GSAT					0.027 (100)
Glycer. dehyd.					0.003 (100)
GGAT					0.00005 (100)
GDC	X				0.00003 (100)
Cyt. FBP ald.					2.127 (100)
Cyt. FBPase					5.554 (100)
UDPGPP					0.531 (100)
SPS					0.034 (100)
SPP					0.031 (100)
F26BPase					0.00 (100)
CO_2 Uptake $\frac{\mu\mathbf{mol}}{\mathbf{m^2 s}}$					36.382
(Local R. %, Global R. %)					(<u>100</u>, 97.2)

TABLE A.4: Sensitivity and Fragility in photosynthetic carbon metabolism. Eleven enzymes resulted sensitive and two of them fragile (RuBisCO and GAP dehydrogenase). Six of the sensitive enzymes were coincident with the known light controlled enzymes of the cycles. Both fragile enzymes were light controlled. A first conclusion is that most sensitive enzymes are key enzymes that can strongly influence the CO_2 uptake with slight concentration variation. The fact that these enzymes are mostly light controlled confirms the strict control of light availability on the Calvin Cycle.

Enzyme Name	Initial Conc. $mg\ N\ m^{-1}$ (the natural leaf)	Optimal Conc. of Var4-1 Sensitive Enz. $mg\ N\ m^{-1}$	Optimal Conc. of Var4-2 Sensitive Enz. $mg\ N\ m^{-1}$	Optimal Conc. of Var4-3 Sensitive Enz. $mg\ N\ m^{-1}$
RuBisCO	517.00 (100)	*517.00* **(99.5)**	*517.00* (100)	*517.00* **(98.5)**
PGA kinase	12.20 (100)	*12.20* (100)	*12.20* (100)	*12.20* (100)
GAP DH	68.80 (100)	*68.80* (100)	*68.80* (100)	*68.80* (100)
FBP aldolase	6.42 (100)	14.76 (100)	10.40 (100)	*6.42* (100)
FBPase	25.50 (100)	*25.50* (100)	*25.50* (100)	*25.50* (100)
Transketolase	34.90 (100)	*34.90* (100)	*34.90* (100)	*34.90* (100)
SBP aldolase	6.21 (100)	*6.21* (100)	*6.21* (100)	*6.21* (100)
SBPase	1.29 (100)	198.05 (100)	*1.29* **(70)**	91.13 (100)
PRK	7.64 (100)	*7.64* (100)	*7.64* (100)	*7.64* (100)
ADPGPP	0.49 (100)	*0.49* (100)	43.52 (100)	46.52 (100)
PGCA Pase	85.20 (100)	63.30 (100)	220.93 (100)	131.79 (100)
Glycerate kinase	6.36 (100)	*6.36* (100)	*6.36* (100)	*6.36* (100)
Glycolate oxidase	4.77 (100)	*4.77* (100)	*4.77* (100)	*4.77* (100)
GSAT	17.30 (100)	*17.30* (100)	*17.30* (100)	*17.30* (100)
Glycer. dehyd.	2.64 (100)	*2.64* (100)	*2.64* (100)	*2.64* (100)
GGAT	21.80 (100)	*21.80* (100)	*21.80* (100)	*21.80* (100)
GDC	179.00 (100)	0.02 (100)	0.49 (100)	22.19 (100)
Cyt. FBP ald.	0.57 (100)	*0.57* (100)	*0.57* (100)	*0.57* (100)
Cyt. FBPase	2.24 (100)	*2.24* (100)	*2.24* (100)	*2.24* (100)
UDPGPP	0.07 (100)	*0.07* (100)	*0.07* (100)	*0.07* (100)
SPS	0.20 (100)	*0.20* (100)	*0.20* (100)	*0.20* (100)
SPP	0.13 (100)	*0.13* (100)	*0.13* (100)	*0.13* (100)
F26BPase	0.02 (100)	*0.02* (100)	*0.02* (100)	*0.02* (100)
CO_2 Uptake $\frac{\mu mol}{m^2 s}$	15.486	22.420	20.626	22.156
(Local R. %, Global R. %)	(100, 81.80)	(99.5, 91.8)	(70, 69.4)	(98.5, 92.9)

TABLE A.5: Optimization and Robustness in photosynthetic carbon metabolism. Concentrations of the enzymes, individual robustness, CO_2 uptake rate (at $c_i = 270\ \mu mol\ mol^{-1}$, reflecting current CO_2 atmospheric concentration), global and local robustness values. The second column reports touchstone concentrations used in simulations, i.e. the initial/natural leaf (modeled by Zhu et al.[60]). Columns 3-5 present enzyme values obtained as result of simulations (varying three different set of 4 sensitive enzymes), and the robustness values associated with each leaf engineering.

TABLE A.6: Enzyme abbreviations defined in the text or used in the tables and figures for the photosynthetic carbon metabolism are listed below.

Rubisco	ribulose bisphosphate carboxylase = Ribulose-1,5-bisphosphate carboxylase/oxygenase	EC 4.1.1.39
PGA Kinase	phosphoglycerate kinase = 3-Phosphoglycerate kinase	EC 2.7.2.3
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase = GAP dehydrogenase	EC 1.2.1.12
Phosphoribulose kinase	Ribulose-5-phosphate kinase=PRK	EC 2.7.1.19
FBP aldolase	FBP Fructose 1,6bisphosphate aldolase	EC 4.1.2.13
FBPase	FBP Fructose 1,6bisphosphate phosphatase	EC 3.1.3.11
Transketolase	Transketolase	EC 2.2.1.1
SBP aldolase	Sedoheptulosebisphosphate aldolase	EC 4.1.2.13 (see FBP aldolase)
SBPase	Sedoheptulosebisphosphatase	EC 3.1.3.37
ADPGPP	ADP glucose pyrophosphorylase	EC 2.7.7.27
Cytosolic FBP Aldolase	Fructose 1,6bisphosphate aldolase	EC see the chloroplast isoform
Cytosolic FBP	Cytosolic FBP ase 6 Fructose 1,6bisphosphate phosphatase	EC see the chloroplast isoform
UDP-Glc pyrophosphorylase	UDPGP = UDP glucose pyrophosphorylase	EC 2.7.7.9
Suc-P synthetase	SPS Sucrose phosphate synthetase	EC 2.4.1.14
Suc-P phosphatase	SPP Sucrose phosphate phosphatase	EC 3.1.3.24
F26BPase	Fructose 2,6bisphosphatase	EC 3.1.3.46
Phosphoglycolate phosphatase	PGCA phosphatase	EC 3.1.3.18
Glycerate kinase	GCEA kinase	EC 2.7.1.31
Glycolate oxydase	Glycollate GCA oxydase	EC 1.1.1.79
Ser Glyoxylate aminotransferase	Glyoxylate:serine aminotransferase = GSAT	EC 2.6.1.45
Glycerate dehydrogenase	GCEA dehydrogenase	EC 1.1.1.29
Glu glyoxylate aminotransferase	GGAT = Glutamate:Glyoxylate aminotransferase	EC 2.6.1.44
GDC	Glycine decarboxylase = Gly decarboxylase	EC 1.4.4.2

Abbreviations

MOO	**M**ulti **O**bjective **O**optimization
GDMO	**G**enetic **D**esign through **M**ulti-objective **O**ptmization
GDLS	**G**enetic **D**esign through **L**ocal **S**earch
SA	**S**ensitivity **A**nalysis
PoSA	**P**athway-**o**riented **S**ensitivity **A**nalysis
SoSA	**S**pecies-**o**riented **S**ensitivity **A**nalysis
RoSA	**R**eaction-**o**riented **S**ensitivity **A**nalysis
RA	**R**obustness **A**nalysis
GR	**G**lobal **R**obustness
LR	**L**ocal **R**obustness
R	**G**local **R**obustness
PoRA	**P**athway-**o**riented **R**obustness **A**nalysis
FBA	**F**lux **B**alance **A**nalysis
ODE	**O**rdinary **D**ifferential **E**quation
DAE	**D**ifferential **A**lgebraic **E**quation
PDE	**P**artial **D**ifferential **E**quation
NSGA II	**N**on-dominated **S**orting **G**enetic **A**lgorithm 2
BioCAD	**Bio**logical **C**omputer-**a**ided **D**esign
GPR	**G**ene-**P**rotein-**R**eaction
EDA	**E**lectronic **D**esign **A**utomation
IA	**I**dentifiability **A**nalysis
glpk	**G**NU **L**inear **P**rogramming **K**it
ATP	**A**denosine **t**ri**p**hosphate
NADH	Nicotinamide adenine dinucleotide
KEGG	**K**yoto **E**ncyclopedia **G**enes and **G**enomes

MILP	**M**ixed **I**nteger **L**inear and **P**rogramming
mmolh^{-1} **gdW**$^{-1}$	millimoles per gram dry cell weight per hour
GP	**G**ene-**P**athway
RP	**R**eaction-**P**athway
ACE	**A**lternating **C**onditional **E**xpectation
EE	**E**lementary **E**ffect
PCA	**P**rincipal **C**omponent **A**nalysis
VMR	**V**ariance-to**M**ean **R**atio
BDO	1,4-**B**utanediol
$\mathbf{c}_i$	**c**arbonate **i**ons

Bibliography

[1] Bernhard O. Palsson. *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press, 1 edition, 2006.

[2] Jeffrey D Orth, Ines Thiele, and Bernhard O. Palsson. What is flux balance analysis? *Nature Biotechnology*, 2010.

[3] Trey Ideker, Timothy Galitski, and Leroy Hood. A new approach to decoding life: Systems Biology. *Annual Review of Genomics and Human Genetics*, 2(1):343–372, 2001.

[4] Lindsay Edwards and Ines Thiele. Applying systems biology methods to the study of human physiology in extreme environments. *Extreme Physiology & Medicine*, 2 (1):8, 2013.

[5] Jonathan R. Karr, Jayodita C. Sanghvi, Derek N. Macklin, Miriam V. Gutschow, Jared M. Jacobs, Benjamin Bolival, Nacyra Assad-Garcia, John I. Glass, and Markus W. Covert. A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell*, 150(2):389–401, 2012.

[6] Hiroaki Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.

[7] Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[8] Dan M. Bolser, Pierre-Yves Chibon, Nicolas Palopoli, Sungsam Gong, Daniel Jacob, Victoria Dominguez Del Angel, Dan Swan, Sebastian Bassi, Virginia González, Prashanth Suravajhala, Seungwoo Hwang, Paolo Romano, Rob Edwards, Bryan Bishop, John Eargle, Timur Shtatland, Nicholas J. Provart, Dave Clements, Daniel P. Renfro, Daeui Bhak, and Jong Bhak. Metabase – the wiki-database of biological databases. *Nucleic Acids Research*, 2011.

[9] A. Sangiovanni-Vincentelli. Eda meets biology! the bumpy road ahead [perscetives]. *Design Test of Computers, IEEE*, 29(3):49–50, 2012.

[10] Ines Thiele and Bernhard O. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 5(1):93–121, 2010.

[11] Desmond S. Lun, Graham Rockwell, Nicholas J. Guido, Michael Baym, Jonathan A. Kelner, Bonnie Berger, James E. Galagan, and George M. Church. Large-scale identification of genetic design strategies using local search. *Molecular Systems Biology*, 5, 2009.

[12] A.P. Burgard, P. Pharkya, and C.D. Maranas. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering*, 84(6):647–657, 2003.

[13] Miguel Rocha, Paulo Maia, Rui Mendes, Jose Pinto, Eugenio Ferreira, Jens Nielsen, Kiran Patil, and Isabel Rocha. Natural computation meta-heuristics for the in silico optimization of microbial strains. *BMC Bioinformatics*, 9(1):499, 2008.

[14] Kiran Patil, Isabel Rocha, Jochen Forster, and Jens Nielsen. Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics*, 6(1):308, 2005.

[15] Jole Costanza, Giovanni Carapezza, Claudio Angione, Pietro Lió, and Giuseppe Nicosia. Robust design of microbial strains. *Bioinformatics*, 28(23):3097–3104, 2012.

[16] Adam M. Feist, Christopher S. Henry, Jennifer L. Reed, Markus Krummenacker, Andrew R. Joyce, Peter D. Karp, Linda J. Broadbelt, Vassily Hatzimanikatis, and Bernhard Ø Palsson. A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. *Molecular Systems Biology*, 3(121):291–301, 2007.

[17] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, , the rest of the SBML Forum:, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novére, L. M. Loew, D. Lucio, P. Mendes,

E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[18] José Oscar Sendin, Antonio Alonso, and Julio Banga. Multi-objective optimization of biological networks for prediction of intracellular fluxes. In Juan Corchado, Juan De Paz, Miguel Rocha, and Florentino Fernández Riverola, editors, *IWPACBB 2008*, Advances in Soft Computing, pages 197–205. Springer Berlin / Heidelberg, 2009.

[19] Robert Schuetz, Nicola Zamboni, Mattia Zampieri, Matthias Heinemann, and Uwe Sauer. Multidimensional optimality of microbial metabolism. *Science*, 336(6081): 601–604, 2012.

[20] Adam M. Feist and Bernhard O. Palsson. The biomass objective function. *Current opinion in microbiology*, 13(3):344–349, 2010.

[21] Amit Varma and Bernhard O. Palsson. Parametric sensitivity of stoichiometric flux balance models applied to wild-type escherichia coli metabolism. *Biotechnology and Bioengineering*, 45(1):69–79, 1995.

[22] Jennifer L. Reed, Iman Famili, Ines Thiele, and Bernhard O. Palsson. Towards multidimensional genome annotation. *Nature reviews. Genetics*, 7(2):130–141, 2006.

[23] Carlos A. Coello. An updated survey of ga-based multiobjective optimization techniques. *ACM Comput. Surv.*, 32(2):109–143, 2000.

[24] K. Deb. *Multi-Objective Optimization Using Evolutionary Algorithms*. Wiley paperback series. Wiley, 2009.

[25] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *Evolutionary Computation, IEEE Transactions on*, 6 (2):182–197, 2002.

[26] Manoj Kumar, Mohammad Husian, Naveen Upreti, and Deepti Gupta. Genetic algorithm: Review and application. *International J. of Information Technology and Knowledge Management*.

[27] Maria Rodriguez-Fernandez and Julio R. Banga. Senssb: a software toolbox for the development and sensitivity analysis of systems biology models. *Bioinformatics*, 26 (13):1675–1676, 2010.

[28] M.D. Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174, 1991.

[29] I. Sobol. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55(1-3):271–280, 2001.

[30] Giovanni Stracquadanio, Renato Umeton, Alessio Papini, Pietro Liò, and Giuseppe Nicosia. Analysis and optimization of c3 photosynthetic carbon metabolism. In Isidore Rigoutsos and Christodoulos A. Floudas, editors, *IEEE BIBE, Philadelphia, PA, USA, May 31-June 3*, pages 44–51. IEEE Computer Society, 2010.

[31] Hong-Xuan Zhang and John Goutsias. A comparison of approximation techniques for variance-based sensitivity analysis of biochemical reaction systems. *BMC Bioinformatics*, 11(246), 2010.

[32] F. Campolongo, J. Cariboni, and W. Schoutens. The importance of jumps in pricing european options. *Reliability Engineering and System Safety*, 91(10), 2006.

[33] M. Laumanns, L. Thiele, K. Deb, and E. Zitzler. Combining convergence and diversity in evolutionary multiobjective optimization. *Evolutionary Computation*, 10(3):263–282, 2002.

[34] L. Breiman and J.H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391): 580–598, 1985.

[35] Duncan S Callaway, MEJ Newman, Steven H Strogatz, and Duncan J Watts. Network robustness and fragility: Percolation on random graphs. *Physical Review Letters*, 85(25):5468–5471, 2000.

[36] Guy Shinar, Uri Alon, and Martin Feinberg. Sensitivity and robustness in chemical reaction networks. *SIAM Journal of Applied Mathematics*, 69(4):977–998, 2009.

[37] Marc Hafner, Heinz Koeppl, Martin Hasler, and Andreas Wagner. "glocal" robustness analysis and model discrimination for circadian oscillators. *PLoS Computational Biology*, 5(10), 2009.

[38] R. Donaldson and D. Gilbert. A model checking approach to the parameter estimation of biochemical pathways. In Monika Heiner and Adelinde Uhrmacher, editors, *Computational Methods in Systems Biology*, volume 5307 of *Lecture Notes in Computer Science*, pages 269–287. Springer Berlin / Heidelberg, 2008.

[39] H. Lodhi and D. Gilbert. Bootstrapping parameter estimation in dynamic systems. In Tapio Elomaa, Jaakko Hollmen, and Heikki Mannila, editors, *Discovery Science*, volume 6926 of *Lecture Notes in Computer Science*, pages 194–208. Springer Berlin/Heidelberg, 2011.

[40] R. Umeton, G. Stracquadanio, A. Sorathiya, A. Papini, P. Lio, and G. Nicosia. Design of robust metabolic pathways. In *Design Automation Conference (DAC), 2011 48th*, pages 747 –752, 2011.

[41] G Nicosia, S. Rinaudo, and E. Sciacca. An evolutionary algorithm-based approach to robust analog circuit design using constrained multi-objective optimization. *Knowledge-Based Systems*, 21(3):175 – 183, 2008. The 27th SGAI International Conference on Artificial Intelligence.

[42] G. Stracquadanio and G. Nicosia. Computational energy-based redesign of robust proteins. *Computers & chemical engineering*, 35(3):464–473, 2011.

[43] Hal Alper, Kohei Miyaoku, and Gregory Stephanopoulos. Construction of lycopene-overproducing E. coli strains by combining systematic and combinatorial gene knockout targets. *Nature Biotechnology*, 23(5):612–616, 2005.

[44] Laura R. Jarboe, Xueli Zhang, Xuan Wang, Jonathan C. Moore, K. T. Shanmugam, and Lonnie O. Ingram. Metabolic engineering for production of biorenewable fuels and chemicals: Contributions of synthetic biology. *Journal of Biomedicine and Biotechnology*, 2010:1–18, 2010.

[45] Shota Atsumi, Tung-Yun Y. Wu, Eva-Maria M. Eckl, Sarah D. Hawkins, Thomas Buelter, and James C. Liao. Engineering the isobutanol biosynthetic pathway in Escherichia coli by comparison of three aldehyde reductase/alcohol dehydrogenase genes. *Applied microbiology and biotechnology*, 85(3):651–657, 2010.

[46] Vincenzo Cutello, Giuseppe Narzisi, and Giuseppe. Nicosia. A multi-objective evolutionary approach to the protein structure prediction problem. *Journal of the Royal Society Interface*, 3(6):139–151, 2006.

[47] Adam M. Feist, Christopher S. Henry, Jennifer L. Reed, Markus Krummenacker, Andrew R. Joyce, Peter D. Karp, Linda J. Broadbelt, Vassily Hatzimanikatis, and Bernhard Ø Palsson. A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. *Molecular Systems Biology*, 3(121):291–301, 2007.

[48] Anthony P. Burgard, Shankar Vaidyaraman, and Costas D. Maranas. Minimal reaction sets for escherichia coli metabolism under different growth requirements and uptake environments. *Biotechnology Progress*, 17(5):791–797, 2001.

[49] R. Mahadevan and C.H. Schilling. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering*, 5(4), 2003.

[50] Jeffrey D. Orth, Tom M. Conrad, Jessica Na, Joshua A. Lerman, Hojung Nam, Adam M. Feist, and Bernhard O. Palsson. A comprehensive genome-scale reconstruction of Escherichia coli metabolism-2011. *Molecular Systems Biology*, 7:53+, 2011.

[51] Pep Charusanti et al. An experimentally-supported genome-scale metabolic network reconstruction for *Yersinia pestis* CO92. *BMC Systems Biology*, 5(1):163, 2011.

[52] Jun Sun, Bahareh Sayyar, Jessica E. Butler, Priti Pharkya, Tom R. Fahland, Iman Famili, Christophe H. Schilling, Derek R. Lovley, and Radhakrishnan Mahadevan. Genome-scale constraint-based modeling of *Geobacter metallireducens*. *BMC Systems Biology*, 3(1), 2009.

[53] Adam M. Feist, Johannes C. Scholten, Bernhard O. Palsson, Fred J. Brockman, and Trey Ideker. Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. *Molecular Systems Biology*, 2, 2006.

[54] Claudio Angione, Giovanni Carapezza, Jole Costanza, Pietro Lio', and Giuseppe Nicosia. Computing with metabolic machines. In Andrei Voronkov, editor, *Turing-100*, volume 10 of *EPiC Series*, pages 1–15, 2012.

[55] J.D. Keasling. Manufacturing molecules through metabolic engineering. *Science*, 330(6009):1355 – 1358, 2010.

[56] H. Alper et al. Identifying gene targets for the metabolic engineering of lycopene biosynthesis in escherichia coli. *Metabolic Engineering*, 7(3):155 – 164, 2005.

[57] H. S. Fong et al. In silico design and adaptive evolution of escherichia coli for production of lactic acid. *Biotechnology and Bioengineering*, 91(5):643–648, 2005.

[58] H. R. Yim et al. Metabolic engineering of escherichia coli for direct production of 1,4-butanediol. *Nature Chemical Biology*, 7(7):445–452, 2011.

[59] Jennifer Reed, Thuy Vo, Christophe Schilling, and Bernhard Palsson. An expanded genome-scale model of escherichia coli k-12 (ijr904 gsm/gpr). *Genome Biology*, 4 (9):R54, 2003.

[60] X.G. Zhu, E. De Sturler, and S.P. Long. Optimizing the distribution of resources between enzymes of carbon metabolism can dramatically increase photosynthetic rate: a numerical simulation using an evolutionary algorithm. *Plant Physiology*, 145(2):513–526, 2007.

[61] S. Imam, S. Yilmaz, U. Sohmen, A.S. Gorzalski, J.L. Reed, D.R. Noguera, and T.J. Donohue. irsp1095: A genome-scale reconstruction of the rhodobacter sphaeroides metabolic network. *BMC systems biology*, 5(1):116, 2011.

[62] R.L. Chang, L. Ghamsari, A. Manichaikul, E.F.Y. Hom, S. Balaji, W. Fu, Y. Shen, T. Hao, B.O. Palsson, and K. Salehi-Ashtiani. Metabolic network reconstruction of chlamydomonas offers insight into light-driven algal metabolism. *Molecular systems biology*, 7(1), 2011.

[63] A.M. Smith, S.C. Zeeman, and S.M. Smith. Starch degradation. *Annual Review of Plant Biology*, 56:73–98, 2005.

[64] C.N. Hunter, F. Daldal, and M.C. Thurnauer. *The purple phototrophic bacteria*, volume 28. Springer Verlag, 2008.

[65] H. Kitano. Biological robustness. *Nature Reviews Genetics*, 5(11):826–837, 2004.

[66] H. Kitano. Towards a theory of biological robustness. *Molecular systems biology*, 3 (1), 2007.

[67] M. Jose, Y. Hu, R. Majumdar, and L. He. Rewiring for robustness. In *Design Automation Conference (DAC), 2010 47th ACM/IEEE*, pages 469–474. IEEE, 2010.

[68] M. Welch, A. Villalobos, C. Gustafsson, and J. Minshull. You're one in a googol: optimizing genes for protein expression. *Journal of The Royal Society Interface*, 6 (Suppl 4):S467–S476, 2009.

[69] F.S. Chapin, E.D. Schulze, and H.A. Mooney. The ecology and economics of storage in plants. *Annual review of ecology and systematics*, 21:423–447, 1990.

[70] P. Millard. The accumulation and storage of nitrogen by herbaceous plants. *Plant, Cell & Environment*, 11(1):1–8, 2006.

[71] A. Bar-Even, E. Noor, N.E. Lewis, and R. Milo. Design and analysis of synthetic carbon fixation pathways. *Proceedings of the National Academy of Sciences*, 107 (19):8889–8894, 2010.

[72] C.A. Raines. Transgenic approaches to manipulate the environmental responses of the c3 carbon fixation cycle. *Plant, cell & environment*, 29(3):331–339, 2006.

[73] N. Sun, L. Ma, D. Pan, H. Zhao, and X.W. Deng. Evaluation of light regulatory potential of calvin cycle steps based on large-scale gene expression profiling data. *Plant molecular biology*, 53(4):467–478, 2003.

[74] P. Singh, D. Kaloudas, and C.A. Raines. Expression analysis of the arabidopsis cp12 gene family suggests novel roles for these proteins in roots and floral tissues. *Journal of experimental botany*, 59(14):3975–3985, 2008.

[75] IA Paponov, S. Lebedinskai, and EI Koshkin. Growth analysis of solution culture-grown winter rye, wheat and triticale at different relative rates of nitrogen supply. *Annals of botany*, 84(4):467–473, 1999.

[76] S. Hengl, C. Kreutz, J. Timmer, and T. Maiwald. Data-based identifiability analysis of non-linear dynamical models. *Bioinformatics*, 23(19):2612–2618, 2007.

[77] R. Storn and K. Price. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *J. of Global Optimization*, 11(4): 341–359, 1997.

[78] R. Umeton, G. Stracquadanio, A. Sorathiya, P. Liò, A. Papini, and G. Nicosia. Design of robust metabolic pathways. In *Proceedings of the 48th Design Automation Conference*, pages 747–752. ACM, 2011.

[79] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–195, 2001.

[80] J Cameron Thrash, Alex Boyd, Megan J Huggett, Jana Grote, Paul Carini, Ryan J Yoder, Barbara Robbertse, Joseph W Spatafora, Michael S Rappé, and Stephen J Giovannoni. Phylogenomic evidence for a common ancestor of mitochondria and the sar11 clade. *Scientific reports*, 1, 2011.

[81] Pietro Lió and Nick Goldman. Modeling mitochondrial protein evolution using structural information. *Journal of molecular evolution*, 54(4):519–529, 2002.

[82] S. Raha, B.H. Robinson, et al. Mitochondria, oxygen free radicals, disease and ageing. *Trends in biochemical sciences*, 25(10):502, 2000.

[83] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, et al. Energy conversion: mitochondria and chloroplasts. In *Molecular Biology of the Cell*. Garland Science, 2002.

[84] T. Kuroiwa, H. Kuroiwa, A. Sakai, H. Takahashi, K. Toda, and R. Itoh. The division apparatus of plastids and mitochondria. *International review of cytology*, 181:1–41, 1998.

[85] Pietro Lió. Phylogenetic and structural analysis of mitochondrial complex i proteins. *Gene*, 345(1):55–64, 2005.

[86] U. Sengupta, S. Ukil, N. Dimitrova, and S. Agrawal. Expression-based network biology identifies alteration in key regulatory pathways of type 2 diabetes and associated risk/complications. *PloS one*, 4(12):e8100, 2009.

[87] J.N. Bazil, G.T. Buzzard, and A.E. Rundell. Modeling mitochondrial bioenergetics with integrated volume dynamics. *PLoS computational biology*, 6(1):e1000632, 2010.

[88] Ulrich Mühlenhoff, Nadine Richhardt, Jana Gerber, and Roland Lill. Characterization of iron-sulfur protein assembly in isolated mitochondria a requirement for atp, nadh, and reduced iron. *Journal of Biological Chemistry*, 277(33):29810–29816, 2002.

[89] Beal M. Flint. Mitochondria in neurodegeneration: Bioenergetic function in cell life and death. *Cerebral Blood Flow & Metabolism*, 19:231–245, 1999.

[90] Gillian M Borthwick, Margaret A Johnson, Paul G Ince, Pamela J Shaw, and Douglass M Turnbull. Mitochondrial enzyme activity in amyotrophic lateral sclerosis: implications for the role of mitochondria in neuronal cell death. *Annals of neurology*, 46(5):787–790, 1999.

[91] A.C. Smith and A.J. Robinson. A metabolic model of the mitochondrion and its use in modelling diseases of the tricarboxylic acid cycle. *BMC systems biology*, 5 (1):102, 2011.

[92] Bernhard Palsson. The challenges of in silico biology. *Nature Biotechnology*, 18 (11):1147–1150, 2000.