Optimization and Ontology for Computational Systems Biology

Renato Umeton

Department of Mathematics

University of Calabria

Prof. Salvatore Di Gregorio, Advisor

Prof. Giuseppe Nicosia, Advisor

A thesis submitted for the degree of

Doctor of Philosophy in Mathematics and Informatics

23 December 2010

I would like to dedicate this thesis to
my Family (*the Umetons*)
and
to the Family that
I am going to start with Raffaella.

Acknowledgements

I would like to thank my Advisors: Professor Salvatore (Toti) Di Gregorio from University of Calabria and Professor Giuseppe Nicosia, from University of Catania; they taught me what "Research" means.

I would like to acknowledge Professor C. Forbes Dewey who boosted my learning curve while I was in his lab at Massachusetts Institute of Technology.

Special thanks go to Professor Pietro Lió from University of Cambridge UK and Professor Alessio Papini from University of Florence: their help in understanding biology has been fundamental. With respect to that, I would like to thank also Dr. Giovanni Stracquadanio from Johns Hopkins University.

I would like to thank friends and colleagues at University of Calabria, MIT and Harvard Medical School.

Microsoft Research and the Centre for Computational and Systems Biology had a role during my PhD years and I would like to thank them for that.

MIT played an important role during these years: it made me met incredible people and shaped my mind for life.

Abstract

In the context of my PhD I studied mostly problems that find their location in the bioinformatics and bioengineering fields. The artificial photosynthesis has been object of my research and problems like the efficient sequestrations of CO_2 and the optimization of the Nitrogen consumption have been taken as target. Geobacter sulfurreducens, a microorganism capable of employing biomasses to produce electrons, has been studied as well. New algorithmic approaches have been developed on both topics and new results obtained are currently under consideration for "in vitro" and "in vivo" implementations. The integration of the information coming from biological and medical resources is a problem that I tackled as well; in this case, the resulting software is currently embedded in the project Cytosolve@MIT. On a parallel track, I also studied problems that are modeled in terms of Cellular Automata, that are a computing environment that shows a straight inspiration from natural phenomena.

Contents

List of Figures

LIST OF FIGURES

List of Tables

Chapter 1

Introduction

Advances and progress in the study of living systems intrigued the human being since the very beginning of his era. Asking questions about our own origins and our own functionality is a fundamental argument for reasoning beings. For these reasons, studies in medicine and biology can be considered "privileged" as closest to this aim. In fact, both biology and medicine brought those achievements with highest impact on global population. Together with fundamental achievements in these fields, it came along the highest level of system complexity ever observed. What sorted out most of the knowledge in physics were Maxwell's equations back in 1861: through this set of partial differential equations many physical phenomena became deterministically related all of a sudden, their connections were clarified from the quantitative point of view. What is still missing in biology is the equivalent of the Maxwell's equations: we now have tons of papers about inter-system interactions, but we have no clue about what is the law model that explains all the reported characterizations from a quantitative point of view. At present, there is the feeling in the scientific community that application of computational tools to biology might be the key to sort this out.

Bioinformatics is a very young word: its first important use is back in the 80s and is related to those pioneers who started interpreting functional contributes of DNA sequences using computers. Nowadays, bioinformatics is a topic so wide that many research areas have been determined in it. From Genomics to Proteomics, and to Interactomics and Metabolomics and all of the *omics, each area is an established research field on its own [1]. Computers became more and more

important as system understanding began to grow: there are biological processes that are too vast or too complex to be directly sorted out by a researcher - on the other hand, we all know that there are tasks that are complex and time consuming for a human being but are trivial for a computer (such as traversing a big graph or checking the all of the probes in a microarray chip). The importance of computer technology applied to biology became obvious already in the study of gene regulatory networks: many cellular processes are triggered or can trigger a gene expression through promoters, activators and repressors; just the map of the regulatory regions of *Escherichia coli* promoters became a searchable database back in 1991 [2].

In the context of computer technology applied to biology, a field that is particularly fertile is the Computational Systems Biology [3]; a field where systems biology is explored with the aid of computer tools. There are no specific tools defined a priori in that: from formal algorithms to statistical approaches, from data-structures to visualizers, every computer framework that can bring our understanding further through *modeling*, is more than welcome. Modeling is the key aspect of the Computational Systems Biology: obtaining a predictive model that can quantitatively anticipate system evolution is considered a major finding. The word "model" or "runnable pathway" wraps around all of the mathematical frameworks whose evaluation can accomplish the quantitative prediction. As of today, the Computational Systems Biology field is so fertile that there are big groups of Conferences dedicated to that and many governments and big companies planned investments on that.

A parallel field that is very close to the latter is the Synthetic Biology: if Computational Systems Biology moves from the study of biological systems to the development of quantitative models, Synthetic Biology can move in the opposite direction as well. Once a researcher has a computational model that reproduces and predicts a given biological process in a reliable way, it is possible to "play" with the manner instead of experimenting on the latter, i.e., computer simulations can drive experiments. The cycle between "in vivo" and "in silico" biology is and accredited asset in this research field [4]. Where synthetic biology is concerned, it is worth mentioning the International Genetically Engineered Machine competition (iGem): in this contest, undergraduate students from world-wide

universities compete to compose build biological systems and operate them in living cells, out of standard *parts*, i.e., biological components. Having a background in computer science, I want to mention two works in Synthetic Biology that deeply grabbed my attention. The first is "Synthetic Gene Networks that Count": in this work, J.J.Collins and fellow-workers built a synthetic gene network (operated in E.coli) that counts [5]; memorizing each observed presence of a certain compound, the network remembers its exposition to the stimuli and reacts when the counter reaches a given threshold number. If we agree that a base component in computer science is the *counter*, then these researchers laid a fundamental brick in Synthetic Biology. The Ron Weiss laboratory extended this concept of programmable modules: if synthetic biology can provide the components that can be composed into modules, then these modules (such as counters, oscillators and switches), can be employed to compose *Systems* [6]. These are just two examples in the field, that cannot be exhaustively detailed in this context.

More specifically, bioengineering is the field that aims to treat biology with approaches that are typical of the engineering area: composing synthetic biology modules onto systems as we compose electrical components onto electrical circuits and devices, is an approach that belongs to this category. Bioengineering approaches have been widely adopted to boost yield, production and other outcomes in many fields such as agriculture, bioremediation and medical therapies. In the Bioengineering area, a fundamental role is played by metabolic engineering [7]. The latter is the employment of above outlined technologies to achieve a functional behavior (from cells, bacteria, etc) that is useful for human aims, through an ad-hoc tuning of the biological system metabolism. Since cellular metabolism is a very complex mechanism, its functional optimization is often mediated with the need of ensuring cell survival and strength.

In the framework of bioengineering, some problems have to be tackled with priority: indeed, recently, a committee of the U.S. National Academy of Engineering has detected fourteen "Grand Challenges for Engineering" [8], 14 areas awaiting engineering solutions in the 21st century. Two of these "Grand Challenges for Engineering" can be tackled with metabolic engineering methods: "develop carbon sequestration methods" and "manage the nitrogen cycle". The growth in emissions of carbon dioxide is a prime contributor to global warming; in fact, for the

carbon dioxide (CO_2) problem, the challenge is to develop effective and efficient systems for capturing the CO_2 and sequestering it safely away from the atmosphere. The optimized management of the nitrogen cycle is crucial for all living things. Indeed, nitrogen is an essential component of proteins and DNA/RNA. The carbon metabolism is largely influenced by the enzyme concentrations [9]; changing the natural concentration is crucial to improve the CO_2 uptake rate of a plant. The atmospheric CO_2 concentration has changed during the last 100 years more than in the past 25 million years, due to large changes in Earth environment; it seems to be reasonable that the evolutionary process cannot re-optimize the enzyme concentrations in this tight period. Even if in the bioinformatics and bioengineering era we are able to work at the enzyme level, the exhaustive search of the optimal enzyme concentrations involved in the photosynthetic metabolism, taking into account only fixed increase and decrease steps, would require testing more than 10^9 possible values. Although an in-vivo optimization is intractable, we can effectively estimate in silico the optimal concentration of the enzymes of this metabolic pathway [10]. For these reasons, the optimization of the photosynthesis has been object of my research. Chapter 3 is then focused on its study, while Chapter 2 describes the algorithms that made this study possible.

An interesting problem in Synthetic Biology concerns the integration of model information. The information coming from biomedical ontologies and runnable pathways is expanding continuously: research communities keep this process up and their advances are generally shared by means of dedicated resources published on the web. In fact, runnable pathways are shared to provide a predictive characterization of molecular processes, while biomedical ontologies detail a semantic context to the majority of those pathways [11]. Recent advances in both fields pave the way for a scalable information integration, based on aggregate knowledge repositories [12; 13], but the lack of overall standard formats impedes this progress. Having different objectives and different abstraction levels, most of these resources "speak" different languages, even if they have large superpositions of contents among each others. As a matter of fact, there is still a large chasm between today's functionality and the true ability to use ontological data to inform molecular pathways. Additionally, there is a lack of strategies for the database and ontology integration of quantitative biological sources written in

different standards (e.g., SBML [14] and CellML [15]). Chapter 4 is dedicated to these questions and describes my contribution in this important problem. This contribute can be considered particularly important in Synthetic Biology.

Chapter 2

New Optimization Algorithms

2.1 Introduction

This Chapter regards three optimization algorithms that have been object of research. The first one is AMMISCA, an evolutionary strategy that introduces a new crossover operator to find more reliable predictions in lava flow models based on Cellular Automata. Second and third algorithms are PAO and PMO2: these algorithms introduce the notion of migration in single- and multi-objective optimization, respectively. All of the algorithms proposed have in the parallelism a point of strength and all of them have been tested against a numbers of real-world problems: AMMISCA has been validated with the SCIARA-R7 model, PAO and PMO2 have been extensively stressed to assess the artificial photosynthesis, that is the object of the Chapter 3; additionally, here is presented the application of PMO2 to the Geobacter sulfurreducens, a highly-dimensional problem that is here modeled for the first time as multi-objective problem.

2.2 AMMISCA, Admissible Method for Improved Genetic Search in Cellular Automata Models

Genetic Algorithms (GAs) are widely used to incrementally reach admissible solutions for hard problems such as parameter tuning in Cellular Automata (CA) models. Here I present a genetic strategy, specifically developed for CA model calibration, exploiting the circumstance that the considered CA parameters have a physical meaning. The proposed approach has proved to be comparable and, in some cases outperforming, if compared with the standard GA proposed by Holland. As a further result, the goodness of the proposed genetic strategy opens the door to genetic tuning algorithms lacking of a standard crossover operator.

In the field of risk assessment and hazard mitigation, event simulation and predictor models have acquired a relevant position. In fact, through simulation of reliable models, risks associated with such processes can be evaluated and possibly contrastated.

Cellular Automata [16; 17] (CA) proved [18; 19; 20; 21] to be a valid choice in simulating natural phenomena such as landslides, erosion processes, lava and pyroclastic flows. They are parallel computing models, discrete in space and time, whose dynamics is determined by the application of local rules of evolution defining the CA transition function. In particular, above cited examples are based on the Di Gregorio and Serra's approach[22] for the modelling of spatially extended dynamical systems. Models based on this approach generally depend on many parameters, which must be provided with the highest possible accuracy in order to obtain satisfactory results in simulating the considered phenomenon. To do this, a parameter tuning phase through standard GA has been successfully applied in previous works[23; 24; 25; 26].

Genetic Algorithms (GAs) [27; 28] are parallel, general-purpose, search algorithms inspired by Genetics and Natural Selection. They simulate the evolution of a population of candidate solutions of a specific search problem by favoring the "survival" and the "recombination" of the best ones, in order to obtain better and better solutions. This family of algorithms has acquired an important role in all

those fields dealing with intrinsically-hard problem lacking of dedicated heuristics or ad-hoc algorithms.

Here I present the definition of AMMISCA, a genetic strategy, and its application to the parameter tuning of the SCIARA-R7 [29] CA model for lava flow simulation and forecasting. Section 2.2.1 presents the SCIARA-R7 simulation model and after that AMMISCA genetic strategy is detailed.

2.2.1 The SCIARA-R7 model

The physical behavior of lava flows can be partially described in terms of Navier-Stokes equations. Analytical solutions of these differential equations are a hopeless challenge, except for few simple, not realistic, cases. The complexity of the problem resides both in the difficulty of managing irregular ground topography and in complications of the equations, that must also be able to account for flows, exhibiting a wide diversity in their fluid-dynamical behavior due to cooling processes. An alternative approach to PDE numerical methods for Navier-Stokes [30] (or more complex) equations is offered by Cellular Automata (CA). As outlined above, CA are computational models assuming discrete space/time and easily implementable on parallel architectures. CA SCIARA-R7 for lava flows is derived from SCIARA [20] where the space is a plane, divided in hexagonal cells; each cell is characterized by a state, that specifies the mean values of physical quantities in the cell (e.g. substate altitude) and embodies a computing unit. This unit updates synchronously the substate values according to a transition function on the basis of substate values of the cell and its adjacent ones. The transition function is applied by the sequential computation of "elementary processes", that account for the phenomenon features.

From a formal point of view SCIARA-R7 is stated by the septuple $SCIARA\text{-}R7 = \langle R, L, X, S, P, \sigma, \gamma \rangle$, where

- $R = \{(x, y) | x, y \in \mathbb{N}, 0 < x < l_x, 0 < y < l_y\}$ is the set of identical hexagonal cells identified by integer co-ordinates in the finite region where the phenomenon evolves.
- $L \in R$ specifies the lava source cells (i.e. craters).

- X identifies the geometrical pattern of cells that influence the cell state change. They are, respectively, the cell itself and its adjacent cells: $X = \{(0,0),(0,1),(0,-1),(1,0),(-1,0),(-1,1),(1,-1)\}$.
- $S = Q_A \times Q_{th} \times Q_T \times {Q_O}^6$ is the set of states; more in detail, Q_A is the altitude of the cell, Q_{th} is the thickness of lava inside the cell, Q_T is the lava temperature and ${Q_O}^6$ rappresent lava outflows (6) from the central cell towards the adjacent ones.
- $P = \{p_{clock}, p_{TV}, p_{TS}, p_{chlV}, p_{chlS}, p_{adher}, p_{cool}\}$ is the set of global parameters, in which:
  - p_{clock} is the time corresponding to a CA step
  - p_{TV} is the lava temperature at vent
  - p_{TS} is the lava solidification temperature
  - p_{chlV} is the characteristic length at the vent temperature
  - p_{chlS} is the characteristic length at the solidification temperature
  - p_{adher} is the constant adherence of lava passing on a cell
  - p_{cool} is the cooling parameter
- $\sigma : Q^{6+1} \rightarrow Q$ is the deterministic state transition function, which is simultaneously applied to all cells of the CA.
- $\gamma : Q_{th} \rightarrow \mathbb{N} \times Q_{th}$ specifies the emitted lava from source cells at the CA step $t \in \mathbb{N}$.

In order to evaluate the goodness of simulations obtained with the detailed model, I have adopted the evaluation function $e_2 = \sqrt{\frac{R \cap S}{R \cup S}}$ where R and S represent the area covered by simulated and real lava flow, respectively; this evaluation function is then used to compute the fitness associated to each simulation in the genetic process.

2.2.2 AMMISCA in detail

AMMISCA, the acronym of AdMissible Method for Improved genetic Search in Cellular Automata, is a genetic strategy exploiting the circumstance that each element of the set of parameter to be tuned (P) has a physical meaning. For instance, if $parent_A$ and $parent_B$ expresses the p_{chlV} SCIARA-R7 parameter (which represents a "threshold" for lava mobility) with values 15 and 25 meters respectively, it can be erroneous to assign the next offspring to an improbable value of 50 meters (which is too *distant* from parent contributes). As anticipated above, the main, characterizing, difference between a standard Holland GA and AMMISCA regards the field which they have been designed for (Cf. Fig. 2.1). While the standard GA is a general purpose optimizer, the second one has been designed for the resolution of those problems in which parameters encoded in the individual have a physical correspondence. When this physical correspondence exists, the algorithm takes advantage of it, thanks to the different crossover strategy implemented, which strictly preserves previous obtained results.

The basic idea within the AMMISCA strategy is to go beyond the preservation of promising schemes through a different crossover, based on arithmetic average: while a one-point crossover ($ONEPT$), using a randomly selected crosspoint, can transform parent strings (e.g. AAAAA, BBBBB) into quite different strings (e.g. AABBB, BBAAA), the new crossover calculates for each parameter the average value between parent ones (as proposed in Linear crossover [31] method with $weight = 0.5$), and assigns it to next generation allele. From two parents we get only one offspring; moreover, this single individual might be too much specialized and the average-driven recombination seems to converge too much rapidly. In order to solve these problems, a sort of *anti-dimidium* is here introduced as well. In AMMISCA, as in standard GAs, there is a range for each parameter encoded in the individual, and two points inside the range rappresenting the value of the parameter introduced by parents. If we shift from a linear range to a closed one (Cf. Fig. 2.2), we obtain a circumference where minimum and maximum of the range coincide and, while the average value is assigned to the first offspring (i.e., $P_{A_{i+1}} = (P_{A_i} + P_{B_i})/2$) as the logical middle-point between parent values, the *anti-dimidium* is calculated as the point diametrically opposite to it (i.e.,

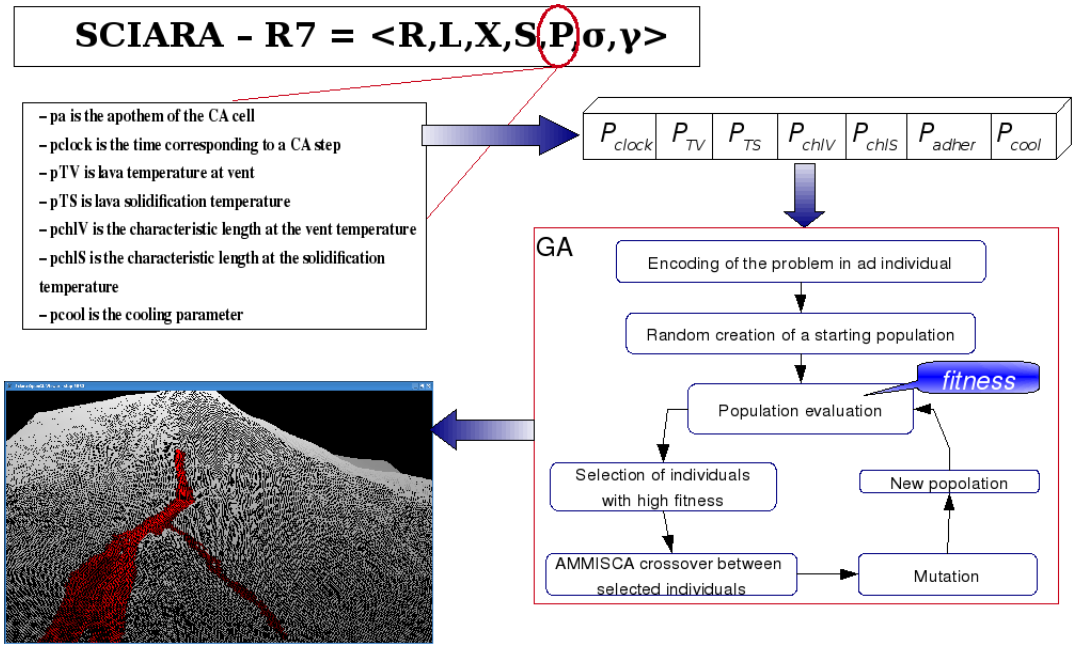
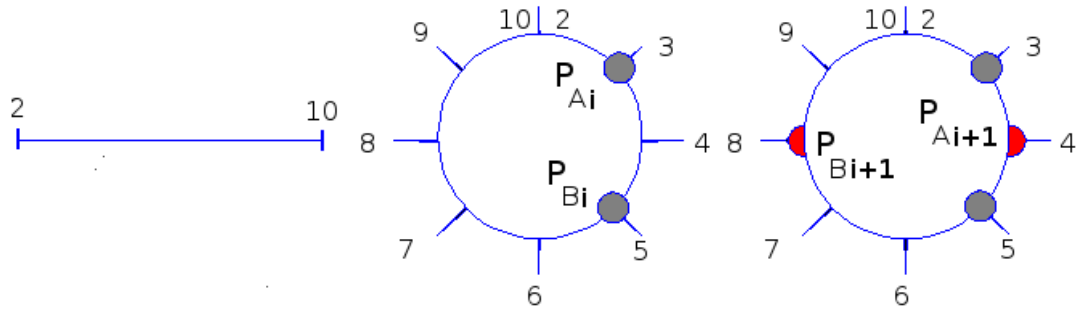


Figure 2.1: The tuning process: (1) select the part of the model that has to be tuned: the parameter set in our case; (2) encode this parameter set in the individual; (3) run the Genetic Algorithm in order to let admissible solutions evolve and recombine, favoring better solutions: in our case the fitness is evaluated through the function e_2; (4) extract the parameter set that gave the most realistic simulation; (5) adopt this set to complete the lava forecasting model.

$P_{B_{i+1}} = P_{A_{i+1}} + (P_{max} + P_{min})/2)$.

An idea, subtended by the introduction of *anti-dimidium*, concerns the following problem: some couples of parameters in SCIARA are "antagonist". This means that similar results could be obtained increasing the value of the former parameter and decreasing the value of the latter one. Hence, different clusters of good values of parameters may exist. Then, AMMISCA always suggests an "internal" ($P_{A_{i+1}}$ in Fig. 2.2) allele and an "external" ($P_{B_{i+1}}$) one: the former searches for a solution that is a specialization of the parents, while the latter explores values out of the interval defined by parent values. Finally, in the context of SCIARA clustered parameters, AMMISCA conveniently derives a new offspring by composing "internal" and "external" alleles (Cf. last line of following pseudo-code block, where alleles are exchanged with probability 0.5). Besides the application of average for model calibration as described above, I present two



(a) Linear range of a parameter p.

(b) Closed range with values of p exhibited by two individuals.

(c) Closed range with values of p assigned to the offspring off individuals in (b) according to AMMISCA "average version".

Figure 2.2: The shift from the linear range to the closed one along with average and *anti-dimidium* definition.

further variants of the algorithm which consider different offspring calculations. In particular, the first version uses the fitness associated to every parent in order to weigh their contribution and thus is labeled as a "fitness weighted average" (*FWAVG*); indeed, the more a parent is promising, the *closer* the allele will be to it. The second variant chooses a random point inside the sub-interval delimited by parents (denoted as *RWAVG* as suggested by Heuristic crossover in [32]). The pure application of the versions detailed above could result too fitness-driven and interfer with crossover function and research space inspection (the first variant, fitness weighted average), or could take longer to solve easy problems (e.g. a maximum values search in a simple cusp by means of the second variant, randomly weighted average). Then, the combination of internal ad external alleles permits to embank this problem. In order to fix all of the details given up to now, it is now presented a pseudo-code-block that states the AMMISCA crossover.

BEGIN: AMMISCA_crossover_function()
{
  crossmode = get requested crossover type *//one in* {*ONEPT, AVG, FWAVG, RWAVG*}
  for each (parameter p **in** P encoded in the individual)
    P_A = value of parameter p expressed by $parent_A$. Same for P_B.
    $P_A{}'$ = value of parameter p that will be assigned to $offspring_A$. Same for $P_B{}'$.
    $range_{min}$ and $range_{max}$ are minimum and maximum value assignable to parameter p
    if (crossmode==ONEPT) applyStandardCrossoverByHolland(P_A,P_B,$P_A{}'$,$P_B{}'$);
    else if (crossmode==AVG) $P_A{}' = (P_A + P_B)/2$;

else if (crossmode==FWAVG)
$P_A{}' = (P_A * fitness_A + P_B * fitness_B)/(fitness_A + fitness_B)$;
else if (crossmode==RWAVG) *//uses only positive random numbers*
$P_A{}' = (P_A * random_1 + P_B * random_2)/(random_1 + random_2)$;
if $(P_A{}' == (range_{min} + range_{max})/2)$
$P_B{}'$ = choose randomly, with same probability, between $range_{min}$ and $range_{min}$
else $P_B{}' = P_A{}'$+ half round of the range; *//anti-dimidium*
if $(P_B{}' > range_{max})$ $P_B{}' = P_B{}' - range_{max} + range_{min}$
if (crossmode $\neq ONEPT$) swap $P_A{}'$ and $P_B{}'$ with proability 0.5; *//alleles composition*
} **END**;

In section 2.2.3 I briefly present the main results achieved by AMMISCA applied to the calibration of the model SCIARA-R7.

2.2.3 AMMISCA Results

In order to validate the genetic strategy for a parameter tuning task, AMMISCA is used for the calibration of SCIARA-R7 model applied to the Nicolosi lava flow event which occurred at Mt Etna (Italy) in 2001.

Numbers of seeds	50

GA Setup	
Parameters num.	7
Initial number of individuals	16
Individuals replaced at each step	8
Crossover probability	1
Mutation probability	0,083
GA steps	10

Individual composition			
Parameter *(unit)*	NO. of bits	Range-min	Range-max
clock *(s)*	8	60	240
TV *(K)*	0	1323	1323
TS *(K)*	8	1023	1173
chlV *(m)*	4	0,1	5
chlS *(m)*	4	5,1	10
adher *(m)*	8	0,05	10
cool (m^3/K^3)	8	10^{-19}	10^{-12}

Figure 2.3: The GA setup for the first class of tests of AMMISCA on SCIARA-R7.

Let consider different classes of tests: first, many seeds and few GA generations are used (50 seeds and 10 generations; Cf. Fig. 2.3 for setup details), and subsequently the most promising seeds are adopted for further GA generation computation (i.e., most promising seed for 100 generations). In Table 2.1, the first test as a result of about 21000 fitness function evaluations is presented,

where the four adopted algorithms (single-point crossover and AMMISCA in its three versions) are compared by means of the contribution of the best average in the individual pool and the best individual. The most promising seed of each algorithm is further executed for ninety more generations and Fig. 2.4 displays both fitness trend and time comparison.

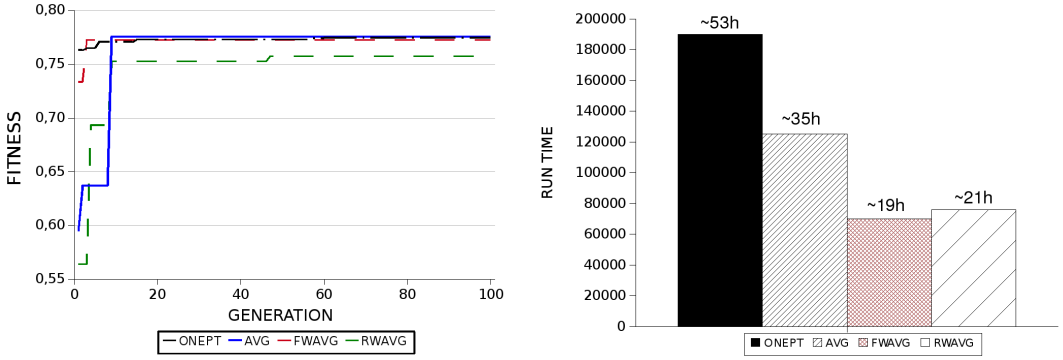
Generation	Best average fitness	Best individual
1	AVG	ONEPT
2	AVG	ONEPT
3	AVG	FWAVG
4	ONEPT	FWAVG
5	ONEPT	FWAVG
6	ONEPT	FWAVG
7	ONEPT	FWAVG
8	ONEPT	AVG
9	ONEPT	AVG
10	**ONEPT**	**AVG**

Table 2.1: First test set: generation-by-generation, for 10 generations, which algorithm gives the best results, over 50 seeds evaluation for each algorithm. ONEPT is one-point crossover; AVG is AMMISCA average version; FWAVG is fitness weighted average version; RWAVG is randomly weighted average version.

As a result, the AMMISCA strategy proves to be valid and promising, being able to outperform standard GA with single-point crossover, both in terms of obtained fitness and execution times. Table 2.1 and Fig. 2.4 indicate that AMMISCA obtains the best individual in both test cases (10 and 100 GA iterations), giving thus the most precise lava event simulation. Besides these results, AMMISCA chooses a set of individuals characterized by a high P_{clock} values leading to faster computations and lower execution times; such P_{clock} values were not taken into account by the Holland search strategy.

2.2.4 Conclusions and future developments

Results can certainly be considered encouraging for the AMMISCA genetic strategy. Moreover, besides the fact that AMMISCA gives rise to the most precise lava simulation, it is interesting to note that the algorithm achieves the best solu-



(a) Fitness trend evolution of each algorithm over 100 GA generations as emerged in second test set: keep running the most promising (at 10th generation) seed for each algorithm for 90 generations more.

(b) Total time required by each algorithm over the two test sets, remarking the fact that some of them explored search zones ignored by others, at least with respect to p_{clock} parameter.

Figure 2.4: Second test set results and global time required by each algorithm to complete the two tests.

tion (in terms of fitness and required time) without a standard crossover phase as defined by Holland. Furthermore, these results route to ad-hoc tuning techniques for CA models that are similar to the analyzed one, that are CA models where the parameter set has a physical meaning.

AMMISCA can be more deeply inspected in the future, as an alternative to a standard GA algorithm. The game plan for future work is to study the AMMISCA conduct in the calibration of SCIARA model for factitious lava events [33] (the best simulation is considered as the real lava event). In fact, we can better compare standard GAs and this family of algorithms with respect to an artificial lava event so that theoretically the global optimum can be achieved during calibration. To be more precise, the referring artificial simulation can be either the simulated lava event obtained with Holland's GA or the one obtained with AMMISCA "average version" (respectively the first and the second simulation whose fitness is rappresented in Fig. 2.4). Subsequently, the second step in this validation plan would be to use AMMISCA family of algorithms to calibrate other macroscopic CA models, tuned with standard GA in the past, such as SCIDDICA [18], PYR [21] and SCAVATU [19]. Eventually, through the analysis

of AMMISCA behavior on cited models, it is possible to derive a study of fitness landscape [33] and reach a more accurate idea of the AMMISCA convergence process.

2.3 PAO: Parallel Optimization Algorithms

Another algorithm class has been designed in the context of this research is Parallel Optimization Algorithms (PAO), an optimization framework that exploits coarse-grained parallelism to let a pool of solutions exchange promising candidates 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 model outlines an optimization environment in which different niches containing different populations are evolved by different algorithms and periodically some candidate solutions migrate into another niche to spread their building block. In this archipelago approach different topologies choices can bring 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 (DE [34] and an enhanced version of CMA-ES[35]) and many archipelago topologies; its simplest topology configuration has been used to have a comprehensible comparison with the other adopted strategies and to better understand the optimization capabilities of this approach. The key difference between the enhanced version (A-CMA-ES) and the original algorithm CMA-ES, is that in the manner I introduced a set of cut-off criteria that drop unstable solutions; additionally, A-CMA-ES ensures with a constraint, a lower bound, for each enzyme concentration to be compatible with the smallest concentration observed in the natural leaf. These algorithms have been employed in the optimization of C_3 carbon metabolism: their evaluation in this context is detailed in Chapter 3.

2.3.1 PMO2: Parallel Multi-Objective Optimization

Moving beyond single optimization, another algorithm has been developed: Parallel Multi-Objective Optimization (PMO2) algorithm is an multi-objective optimization framework based on PAO that let a pool of non-dominated solutions exchange promising candidate solutions, again, in an archipelago fashion. Encapsulating the multi-objective optimization algorithms called NSGA-II[36] the framework completes with migration its multi-objective approach. NSGA-II is an elitist genetic strategy coupled with a fast non-dominated sorting procedure and a density estimation of individuals using the crowding distance; its strategy has been designed to assure an efficient approximation of the Pareto optimal set. It is important to note that this algorithm is derivative-free and, in particular, it does not make any assumption on the convexity or discontinuity of the Pareto front. Again, an island is a virtual place where a pool of candidate solutions (e.g., unfeasible, feasible and non-dominated solutions) is let evolve with a specific multi-objective optimization algorithm; communications among islands in terms of solutions evolved by potentially different algorithms (or different setting of the same optimization algorithm) are arranged through an archipelago topology. The island model outlines a multi-objective optimization environment in which different niches containing different populations (each population is a set of candidate solutions) are evolved by different algorithms and periodically some candidate solutions migrate increasing the diversity of target population.

2.3.2 PMO2 Results on Geobacter sulfurreducens

Here I present a test case in which the algorithm PMO2 is used to determine, in Geobacter sulfurreducens, the trade-off for growth versus redox properties. In the Geobacter context, I have gained the functional *desiderata* (that are fundamental for industrial processes) through the modeling of the problem in terms of a constrained multi-objective problem: goals are the maximization of both biomass and electron production.

The importance of the Geobacter sulfurreducens is well known; in fact, this is a bacterium capable of using biomasses to produce electrons to be transferred directly to an electrode; this species is a useful model for real optimization since its

genome is completely sequenced and a model of its metabolic network is available. Metabolic engineerings are surely possible. The bacterial biomass growth needs to be related to the electron transfer rate: the Geobacteraceae is a family of microorganisms known for their remarkable electron transfer capabilities which allow them to be very effective in bioremediation of contaminated environments and in harvesting electricity from waste organic matter. Bioengineering a mutant strain in order to reach faster rates in electron transport yield is highly desirable and could represent a breakthrough for massive application in biotech industry.

2.3.2.1 Maximizing Biomass and Electron Productions

Constraint-based modeling of metabolism has laid the foundation for the development of computational algorithms which allow more efficient manipulations of metabolic networks. One established approach, OptKnock, has already yield good results in suggesting gene deletion strategies leading to the overproduction of biochemicals of interest in E. Coli [37]. These increments are accomplished by dropping some redundancy in the metabolic pathways in order to eliminate reactions competing with those of interest.

Here I have optimized Geobacter sulfurreducens, modeled as an in-silico organism [38], by perturbing its 608 reaction fluxes with PMO2; additionally I ensured the constraint that steady state solutions are preferred (i.e.: $S \cdot x = 0$, where S is the stoichiometric matrix, x the perturbed flux vector and 0 is the null vector). The optimization has been designed to move towards those solutions where two crucial fluxes are maximized: Electron Production Flux and Biomass Production Flux. Five non-dominated solutions ($A - E$) are reported in Fig. 2.5 as best trade-offs. In particular, in my multi-objective constrained optimization, the solution A presents a significant slope in the constraint violation reduction: $3.4 \cdot 10^4$ is roughly 1/26.47 when compared with the initial guess solution (that showed a violation in the order of 10^6) and it keeps decreasing towards steady state solutions. To my knowledge this is the first time that a multi-objective optimization that faces both electron and biomass production is implemented for Geobacter sulfurreducens. The PMO2 approach brought a set of Pareto-optimal solutions such that: (i) an enhanced electron and biomass productions are achieved, (ii)

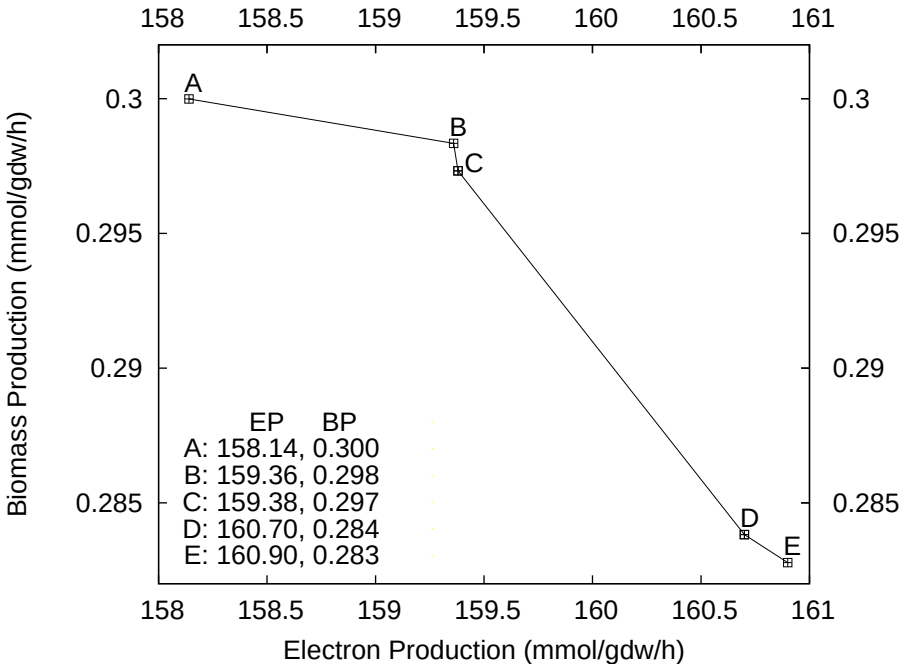


Figure 2.5: Pareto Front of Geobacter sulfurreducens: maximization of biomass production versus maximization of electron production. The units for the flux values are mmol/gDW/h (DW = dry weight).

the contraint violation is minimized by the algorithm that rewards less violating solutions, and (iii) all of the biological constraints highlighted by the Flux Balance Analysis pointed out by Cobra toolbox [39] on this pathway are intrinsically enforced because they define the search space boundaries in my algorithm. An important bound that worth mentioning is the ATP: the flux related to the latter is kept fixed at 0.45 as highlighted in [38] as best value assessed.

2.3.2.2 Geobacter conclusion

I have applied the PMO2 algorithm to the Geobacter sulfurreducens in order to stress its capabilities on a highly-dimensional problem ($\mathbb{R}^{608}$) in metabolic engineering; with respect to that I have obtained a computational model that maximizes the electron and biomass productions while preserving those bounds that ensures a biological significance. To my knowledge this is the first time that Geobacter sulfurreducens has been modeled as a multi-objective optimization

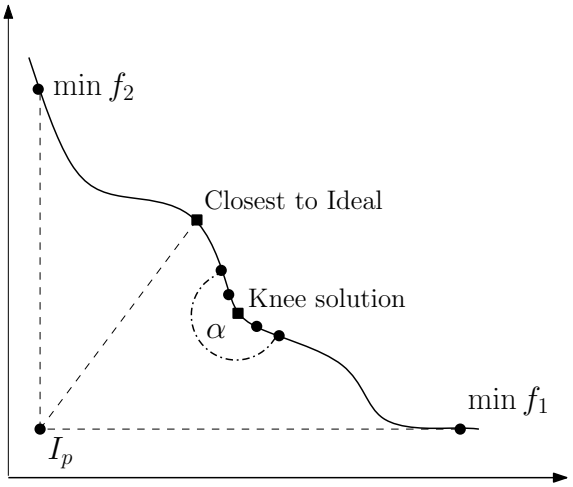


Figure 2.6: Decision making strategies. Geometrical representation of the various strategies on a bi-objective Pareto front.

problem where the search moves automatically towards steady-state solutions, contextually with biological boundaries observance and functional optimization (i.e.: biomass and electron productions).

2.3.3 Pareto Front Mining and Analysis

In addition to the success given by the practical application of the algorithm to the Geobacter sulfurreducens test case, it seems important also to specify more formal *desiderata* for a multi-objective optimization algorithm. Evaluation of these metrics on a complex real-world application is among the objects of Chapter 3.

It is worth noting that multi-objective optimization algorithms give as result a set of non-dominated solutions, instead of a single optimum (or an individual sub-optimal solution) as in single-objective optimization. In real world applications, it is useful to provide a strategy to select automatically the best trade-off solution; when the set of Pareto optimal solutions is huge, a screening strategy is mandatory. In literature, there are many trade-off selection strategies [40] typically based on the geometric notion of Pareto optimality, or heuristics based on the experimental evidence.

A natural strategy is the one that selects the Pareto optimal solution that

is closest to the ideal (Cf. Fig. 2.6) minimum of each objective. Let P a set of non-dominated solutions. The *closest-to-ideal* point is defined as:

$$x \in P : \nexists y \in P : d(y, I_p) < d(x, I_p)$$

where $d : \mathbb{R}^p \to \mathbb{R}$ is a distance metric and the ideal point is

$$I_p = \{\min f_1(x), \cdots, \min f_p(x)\}.$$

It is important to note that it is not required to know the real minimum for each objective; it is possible to use as I_p the minimum achieved for each objective by the algorithm, that is called *Pareto Relative Minimum* (PRM). Finally, the last selection criterion is the *shadow minimum* selection; according to this strategy, p points that achieves the lowest values on the k objectives considered are selected. It is always useful to select these points, since it is possible to gain more information on the best possible values achievable for each objective. The analysis of multi-objective optimization algorithms requires the definition of ad-hoc metrics; firstly, hypervolume indicator [41] is adopted. Let $X = (x_1, \cdots, x_k) \subset \mathbb{R}^k$ a k-dimensional decision vectors; the hypervolume function $V_p : \mathbb{R}^k \to \mathbb{R}$ provides the volume enclosed by the union of polytopes $p_1, \cdots, p_i, \cdots, p_k$, where p_i is formed by the intersections of the following hyperplanes arising from x_i along with the axes. In order to assess the quality of Pareto optimal sets obtained by different algorithms, it is important to compare the non-dominated solutions obtained in order to estimate which algorithm is able to cover effectively the front and which solutions are globally Pareto optimal. According to these considerations, two metrics are introduced; the *global* and *relative Pareto coverage*. Let $P_A = \cup_{i=1}^{m} P_i$ where P_i is a Pareto front; P_A is the Pareto Front defined by the union of m Pareto frontiers. Let define the *global Pareto coverage* of the $i - th$ front as follows:

$$G_p(P_i, P_A) = \frac{|x \in P_i \bigwedge x \in P_A|}{|P_A|} \tag{2.1}$$

G_p provides the percentage of Pareto optimal points of P_i belonging to P_A; it is important to note that this metric provides only a quantitative measure of the performance of the algorithm, since it strongly rewards large Pareto front.

The metric gives qualitative information if and only if the Pareto frontiers have a similar dimension. Although it is important to understand the composition of P_A, it is important to estimate how many solutions of a Pareto front are not dominated by solutions belonging to the other front considered; a solution $v \in P_i$ is called *globally Pareto optimal* if it belongs to P_A. Let P_A a global Pareto front, the *relative Pareto coverage* is defined as follows:

$$R_p(P_i, P_A) = \frac{|x \in P_i \bigwedge x \in P_A|}{|P_i|} \tag{2.2}$$

R_p measure the relative importance of the P_i front in P_A. If $R_p \to 1$, two aspects are considered; the algorithm is able to find $R_p \times |P_i|$ globally Pareto optimal solutions, or it has found $R_p \times |P_i|$ solutions in a region of the front not covered by the other methods. However, it is worth noting that algorithms that are able to generate large Pareto frontiers are important, especially in real world application, where human experts do the decision among trade-off points. For this reason, considering jointly the two metrics could effectively compare the quality of a Pareto front.

Chapter 3

Artificial Photosynthesis

3.1 The study of the C3 photosynthetic carbon metabolism

I studied the $\mathbf{C_3}$ photosynthetic carbon metabolism presented in Fig. 3.1 centering the investigation on the following four design principles.
(1) Optimization of the photosynthetic rate by modifying the partitioning of resources between the different enzymes of the $\mathbf{C_3}$ photosynthetic carbon metabolism using a constant amount of protein-nitrogen.
(2) Identify sensitive and less sensitive enzymes of the studied metabolism model.
(3) Maximize photosynthetic productivity rate through the choice of robust enzyme concentrations using a new precise definition of robustness.
(4) Modeling photosynthetic carbon metabolism as a multi-objective problem of two competing biological selection pressures: light-saturated photosynthetic rate versus total protein-nitrogen requirement.

The computational simulation of the carbon metabolism requires the definition of a set of linked ODEs to encode the relevant biochemical reactions; in my research work, I considered the model proposed by [42]. 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. The reactions introduced in the model were categorized into equilibrium and non-

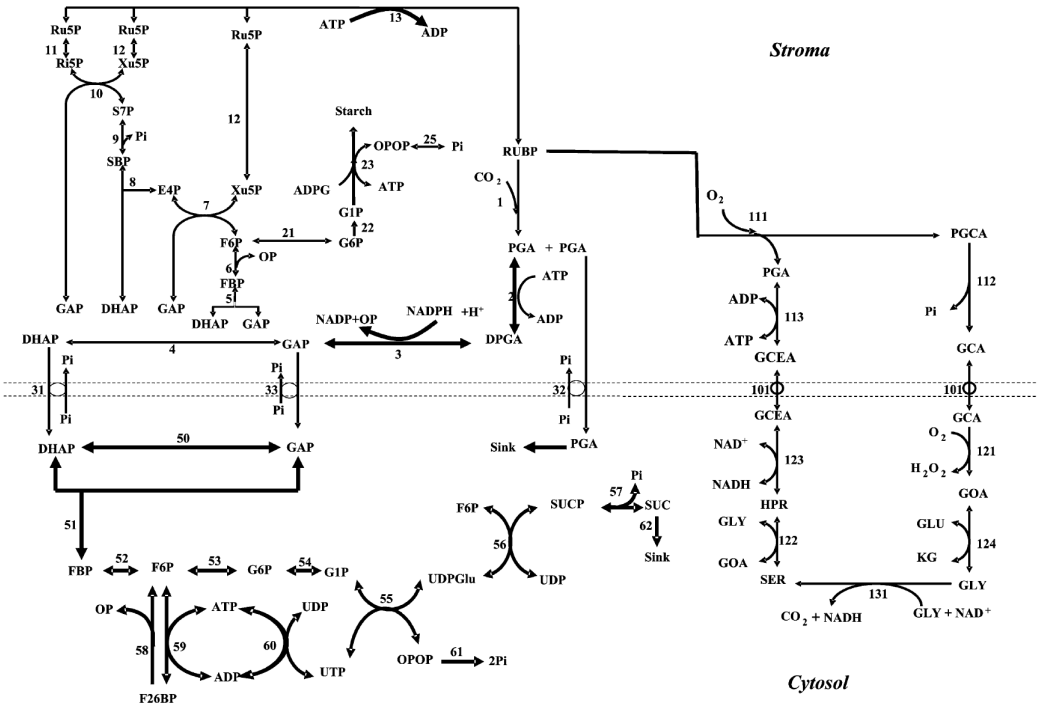


Figure 3.1: $\mathbf{C_3}$ photosynthetic carbon metabolism pathway.

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 (Cf. Appendix A1 for modeling details, and in particular section A.1.1 for metabolite nomenclature).

Main results consist in the fact that, thanks to the designed methodology PAO detailed in the Chapter 2, I have obtained an increase in photosynthetic productivity of the **135**% from 15.486 μ**mol m**$^{-2}$**s**$^{-1}$ (i.e., value measured in standard natural leaves) to **36.382** μ**mol m**$^{-2}$**s**$^{-1}$, and improving the previous best-found photosynthetic productivity value (**27.261** μ**mol m**$^{-2}$**s**$^{-1}$, **76**% of enhancement). Optimized enzyme concentrations express a maximal local robustness (**100**%) and a high global robustness (**97.2**%), satisfactory properties for a possible "in vitro" manufacturing of the optimized pathway. Morris sensitivity

analysis shows that 11 enzymes out of 23 are high sensitive enzymes, i.e., the most influential enzymes of the carbon metabolism model. Finally, I have studied the $\mathbf{C_3}$ carbon metabolism as a trade-off between the maximization of the leaf $\mathbf{CO_2}$ uptake rate and the minimization of the total protein-nitrogen concentration. This trade-off search has been carried out in six environmental scenarios: three $\mathbf{c_i}$ concentrations (referring to the estimate of $\mathbf{CO_2}$ concentration in the atmosphere characteristic of 25 million years ago, nowadays and in 2100 a.C.) and two triose-P (PGA, GAP, and DHAP): low and high export rates. Additionally, CO_2 uptake and *nitrogen consumption* are evaluated with respect to the *robustness* by means of a 3D Pareto-surface. Remarkably, the six Pareto frontiers identify the highest photosynthetic productivity rates together with the fewest protein-nitrogen usage.

3.2 Introduction

Recently, a committee of the U.S. National Academy of Engineering has detected fourteen "Grand Challenges for Engineering" [8], 14 areas awaiting engineering solutions in the 21st century. Two of these "Grand Challenges for Engineering" have been treated in my research: "develop carbon sequestration methods" and "manage the nitrogen cycle". The growth in emissions of carbon dioxide is a prime contributor to global warming, in practice, for carbon dioxide (CO_2) problem the challenge is to develop effective and efficient systems for capturing the CO_2 and sequestering it safely away from the atmosphere. The optimized management of the nitrogen cycle is crucial by all living things, in fact, nitrogen is an essential component of proteins and DNA/RNA. Indirectly, the maximization of the leaf CO_2 uptake rate and the minimization of the total protein-nitrogen concentration here obtained go in the direction to improve CO_2 capturing rate and to increase nitrogen use efficiency of natural leaf. This result has been reached thanks to specific optimization algorithms detailed in Chapter 2.

Numerous problems encountered in bioinformatics, systems biology and bioengineering can be modeled as optimization problems [43; 44] and, thus, lend themselves to the application of effective heuristic search methods and derivative-free global optimization algorithms [45]. The optimization task is conducted

with respect to a single objective function or a set of competing, conflicting, and non-commensurate objectives having nonlinear interdependence. It is necessary, hence, the usage of proper heuristics and algorithms to optimize the objective functions while satisfying several constraints.

Recently, multi-objective optimization has found important applications in a growing number of fields, for example, molecular biology, chemical engineering and biomedical engineering, and has shown to have significant benefits compared to single-objective optimization, e.g., selection of single nucleotide polymorphisms [46], protein structure prediction [47], and estimation of intracellular fluxes [48]. Here I have optimized the photosynthetic carbon metabolism in order to maximize the CO_2 uptake rate, and investigated the Pareto frontiers in the carbon metabolism in terms of photosynthetic rate versus protein-nitrogen. Using the Morris method [49], it has been evaluated the impact of enzymes on the model identifying the sensitive and insensitive enzymes. Moreover, is has been performed a new robustness analysis detecting the robust and less robust enzymes in order to keep a maximal leaf CO_2 uptake rate. Finally, robustness has been connected with multi-objective optimization. The overall framework adopted to analysis photosynthetic carbon metabolism can be used to study large-scale metabolic networks, in particular, and biomolecular systems, in general. Hopefully, the algorithms and tools designed and introduced in this study, the derivative-free global optimization algorithms, the multi-objective optimality analysis, the sensitivity and robustness analysis, although general-purpose methods, could be effective in explain key properties of many other biological systems as well.

The carbon metabolism is largely influenced by the enzyme concentrations [9]; changing the natural concentration is crucial to improve the CO_2 uptake rate of a plant. The atmospheric CO_2 concentration has changed during the last 100 years more than in the past 25 million years, due to large changes in Earth environment; it seems to be reasonable that the evolutionary process cannot re-optimize the enzyme concentrations in this tight period. Even if in the bioinformatics and bioengineering era we are able to work at the enzyme level, the exhaustive search of the optimal enzyme concentrations involved in the photosynthetic metabolism, taking into account only fixed increase and decrease steps, would require testing more than 10^9 possible values. Although an in-vivo optimization is intractable,

we can effectively estimate in silico the optimal concentration of the enzymes of this metabolic pathway [10]. For this reason, I have designed ad-hoc algorithms to optimize the enzyme concentrations in order to maximize the CO_2 uptake rate. The metabolism has been modeled as a system of ODEs, where the inputs are the enzyme concentrations and the output is the CO_2 uptake. Firstly, I maximized the CO_2 uptake rate using deterministic and stochastic optimization algorithms; I found that the designed algorithms, Advanced CMA-ES algorithm and Parallel Optimization Algorithms (i.e., A-CMA-ES and PAO, Cf. Chapter 2), are able to increase the photosynthetic rate of 135%, that is, the new best-known optimum. The Morris sensitivity analysis shows the complexity and non-linearity of the pathway; in fact Morris method unravels the insensitive and sensitive enzymes of the C_3 photosynthetic carbon metabolism model. In order to estimate the robustness of the found solutions, they have been performed both global and local robustness analysis using ad-hoc designed Monte-Carlo methods. According to which aspect or part of the dynamical system is mutated, it is possible to define four different types of robustness [50]: dynamical stability (mutation of initial conditions), constraint robustness (mutation of constraint values), parametric robustness (mutation of parameter values) and structural stability (mutation of the dynamical function). The designed robustness analysis is a parametric robustness: robustness to change of parameter values.

Finally, using the designed multi-objective optimization framework, I have discovered Pareto frontiers between two competing and conflicting objectives: the CO_2 uptake rate and the amount of protein-nitrogen. I maximized the CO_2 uptake rate while minimizing the amount of used protein-nitrogen concentration. Pareto-optimality, is used to explore the performance space of the C_3 pathway.

Pareto-optimality conditions are those in which it is impossible to make a function (target, goal, process, simulation) better off without necessarily making some else function worse off. Multi-objective optimization problems (MOP) tackle sets of competing, conflicting and non-commensurate objective functions having (strong or weak) nonlinear interdependence. MOPs generally have a set of solutions that are known as Pareto-optimal (Pareto-efficient, Pareto-surface, Pareto-front); the Pareto front, hence, represents multiple-optimized candidate solutions. The Pareto front is the solution set in which any attempt to improve

one objective function of the point on the Pareto front must result in the degradation of at least one other objective function.

3.3 The Designed Framework

In this section I outline the tools that have been adopted in the re-optimization of the photosynthetic carbon metabolism pathway, apart from those algorithms already detailed in Chapter 2, we have: sensitivity analysis, applied derivative-free optimization algorithms, and robustness analysis.

3.3.1 The method of Morris

The sensitivity analysis (SA) concerns the study of how uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model input. In particular, SA tries to identify the most influential parameters of a given model; understanding which are the most important parameters of a model could be extremely difficult since it is common to deal with non-linear, highly noise and computational expensive models. It is important to remark the differences between Robustness (RA) and SA; RA aims to evaluate which is the probability of a system to remain in a reference state under perturbations, while, SA perturbs a system in order to find which is the aspect that mainly affects its behavior and to detect the dependencies among input parameters and between input and output. SA answers the question *"which enzymes are crucial for the carbon metabolism?"* In order to perform this analysis, it has been used the Morris method, which is particularly suited when the number of uncertain parameters, called factors, is high and the model could be expensive to compute. The Morris method belongs to the class of the *one-factor-a-time* (OAT) methods [51]; OAT means that a factor is perturbed in turn while keeping all other factors fixed at their nominal value. In particular, the method varies one factor at time across a certain number of levels selected in the space of the input factors; this grid-like sampling makes the algorithm easily adaptable for discrete and continuous variables. For each variation, a factor elementary effect is computed as follows: $u_i = (Y(x_1, x_2, \ldots, x_i + \Delta x_i, \ldots, x_k) - Y(x_1, x_2, \ldots, x_i, \ldots, x_k))/\Delta x_i$ where Y is

the model, $x_1, x_2, \ldots, x_i + \Delta x_i, \ldots, x_k$ is the perturbed parameters vector and $x_1, x_2, \ldots, x_i, \ldots, x_k$ is the nominal parameters vector. For each factor, at different levels, various estimates of the elementary effect u_i are performed. In order to study the importance of the parameters, the mean μ_i and the standard deviation σ_i are computed over the elementary effects u_i of the $i-$th parameter. A high value of μ_i denotes a high linear effect for a given factor, while a high value of σ_i denotes either non-linear or non-additive behavior. The modulus version of μ_i^* has been preferred since it is better than μ_i in ranking factors in order of importance; for each enzyme are evaluated five concentrations under consideration as the nominal values of the concentrations, and successively, 20 factor levels are perturbed 10 times. Since the bounds on variables are not clearly defined, lower and upper bounds have been set at $\pm 100\%$ of the nominal value of each enzyme concentrations.

3.3.2 Derivative-Free Optimization Algorithms

As said, one of the key points of this Chapter is the CO_2 uptake optimization in the context of the carbon metabolism pathway. The optimization of the photosynthetic productivity rate has been tackled using state-of-the-art derivative-free optimization algorithms belonging to the classes of deterministic and stochastic optimizers and a new optimization framework, Parallel Optimization Algorithms (PAO). *Stochastic algorithms* taken into account are CMA-ES [35], Differential Evolution [34] and the hybrid particle swarm optimizer PPSwarm [52]. The deterministic optimizers belong to three broad sub-classes; *pattern search methods* are represented by the Hooke-Jeeves method [53], the Generalized Pattern Search [54] and the Mesh Adaptive Direct Search [55]. Finally, two *branch-and-bound algorithms* called Direct [56] and Multilevel Coordinate Search [57], together with Implicit Filtering [58] a *line-search method*, have been employed.

The ODEs system input is a partitioning of the $E = 23$ enzymes involved in the metabolic pathway; the output is an evaluation in terms of CO_2 uptake, predicting then, the photosynthetic/photo-respiratory properties of a leaf characterized by such a partitioning. This means that, abstracting the concentration of the enzymes in a vector $x = [conc_1, conc_2, \ldots, conc_E]$, the value $f(x)$ is the CO_2

uptake coming from the solution of the ODEs system when the concentration x is adopted. To solve the system of ODEs I used the ODE15S MATLAB function as proposed in [42]; this ensures an acceptable accuracy with a moderated computational cost.

In order to consider biologically meaningful concentrations, the algorithms have to look for a partitioning of the enzymes, meaning that the total amount of protein-nitrogen has to remain constant among all vectors x and equal to the amount that characterizes the vector x_0 corresponding to the enzyme concentrations measured in the natural leaf [42] (the initial concentrations). The long run comparison of the convergence processes of the algorithms reveals the presence of many local optima in the solution space; for this reason the designed algorithm, A-CMA-ES, introduces a set of cut-off criteria to CMA-ES and ensures with a constraint, a lower bound, for each enzyme concentration to be compatible with the smallest concentration observed in the natural leaf (vector x_0). Parallel Optimization Algorithm (PAO), detailed in Chapter 2, has been employed for the optimization. The PAO framework actually encloses two optimization algorithms and many archipelago topologies but its simplest configuration has been used to have a comprehensible comparison with the other adopted strategies and to better understand the optimization capabilities of this approach. The adopted configuration has two islands with 2 optimization algorithms, A-CMA-ES and DE, 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 multi-objective optimization algorithm PMO2 (detailed in Chapter 2) 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 CO_2 uptake and at the same time to minimize the total amount of protein-nitrogen needed for that. Introducing then the function $g(x) = \sum_{i=1}^{E} \frac{x[i]*WM_i}{BK_i}$, where BK_i are the catalytic number or turnover number, and WM_i the molecular weight of each enzyme respectively, the problem is now defined as finding the leaf representing the best trade-off when maximizing CO_2 uptake rate, $f(x)$, and at the same time minimizing the total amount of protein-nitrogen, $g(x)$. In other words, we are

looking for the best resulting leaf in terms of CO_2 uptake that uses the smallest amount of protein-nitrogen to gain that result. Quantitative evaluation of points obtained facing two competing and conflicting objectives is done using a Pareto front approach: non-dominated points are those solutions that are not outperformed in both objectives by other points and then represent the Pareto-optimal solutions.

3.3.3 Local and Global Robustness

The robustness is a dimensionless metric that assesses the yield of a given system, it is the property of the system itself to undergo mutations remaining in a reference state and continuing to perform its tasks in a reliable way. In biology, robustness is generally regarded as a desirable feature. The ability of a system to survive changes in the environment, and/or in the system itself, is one of the main driving forces of evolution [59]. By inspecting the photosynthesis process, it is extremely important to evaluate how the CO_2 uptake rate changes due to perturbations in the enzyme concentrations; perturbations can be caused by many factors, like bias in the synthesis process and changes in the ground elements. For instance, by mutations of the promoter sequence or on the enzyme control sites (effector binding sites) in the case of allosteric enzymes. It is then obvious the importance of seeking concentrations that maximize the CO_2 uptake rate and maintain a quasi-ideal behavior in the presence of noise. In this research, let $\Omega = \{\{p_i\}_{i=1}^{m}, \{\phi_i\}_{i=1}^{n}\}$ as a system with m parameters and n properties. *Nominal value* (N_v) is the value of a property for a given parameter set. A *trial* τ is a perturbed system generated by an α function, also called α-perturbation, such that $\tau = \alpha(\Omega, \sigma)$. The α function applies a stochastic noise σ on the reference system Ω; without loss of generality, the noise is defined by a random distribution. In order to simulate a statistically meaningful perturbation phenomenon, an ensemble, T, of perturbed systems is generated. A trial $\tau \in T$ is considered *robust* to a perturbation (mutation) of the stochastic noise σ for a given property ϕ, if the following *robustness condition* is verified:

$$\rho(\Omega, \tau, \phi, \epsilon) = \begin{cases} 1 & if \quad \mid \phi(\Omega) - \phi(\tau) \mid \leq \epsilon \\ 0 & otherwise \end{cases}$$

where Ω is the *reference system* and ϵ is a *robustness threshold*. The robustness of a system Ω is the number of *robust trials* in T (with respect to the property ϕ) over the total number of trials ($\mid T \mid$): this this measure is the *robustness* of the system. Formally, a *robustness function* Γ is defined as follows: $\Gamma(\Omega, T, \phi, \epsilon) = \frac{\sum_{\tau \in T} \rho(\Omega, \tau, \phi, \epsilon)}{|T|}$ The function Γ is a dimensionless quantity that assesses the probability that the nominal value of a property changes at most ϵ due to perturbations; high Γ values means high system robustness. Two kind of robustness analysis has been performed; the *global robustness* analysis applies a stochastic noise to each enzyme concentration; while, the *local robustness* analysis applies the noise one enzyme at time (this evaluates the *single robustness*, that is, the robustness of a single enzyme). In other words, while the global robustness analysis studies global changes of the system, the local robustness analysis studies the relative robustness of a single enzyme. The ensemble T has been generated using a Monte-Carlo algorithm; a maximum perturbation of 10% is set from the nominal value of each enzyme concentration, and the ensemble is generated as 5×10^3 trial for the global robustness analysis and 200 trials for each enzyme for the local robustness.

3.4 Experimental Results

3.4.1 Sensitivity Analysis

Sensitivity analysis perturbs a given system in order to discover which aspects primary affect its behavior, to detect the dependencies among input parameters and between input parameters and output functions. In Fig.3.2 are reported the results of the Morris sensitivity analysis on the model of the carbon metabolism. High mean values mean linear enzymatic response, while high standard deviation values assess a non-linear (or non-additive) behavior or dependencies among enzymes. Inspecting Fig. 3.2 it is possible to detect three distinct clusters, a) eleven *high sensitive enzymes* (i.e., enzymes with $\mu, \sigma > 1$), b) five *insensitive enzymes* ($\mu, \sigma < 0.1$), and c) seven *low sensitive enzymes* ($0.09 < \mu \leq 1$). Hence, the eleven *high sensitive enzymes*, Rubisco, PGA kinase, GAP dehydrogenase, FBP aldolase, FBPase, SBP aldolase, SBPase, Phosphoribulose kinase, ADPGPP, Phosphogly-

colate phosphatase, and GDC, are the most important enzymes in the studied model of the carbon metabolism.

Six enzymes of the Calvin Cycle are known to be directly regulated by light [60]; among these six are present two enzymes (PGA Kinase and GAP dehydrogenase) responsible of energy-converting reactions, which are coupled to the light reactions in the thylakoids. Rubisco, Phosphoribulose kinase, FBPase and, with somewhat lower sensitivity values, FBPase as well are controlled (and activated) by light [60].

This means that 5 out of 6 of the enzymes with the larger sensitivity values (those with the largest standard deviation in Fig. 3.2) are controlled by light. The sixth enzyme with largest sensitivity value is the SBP aldolase (third position in sensitivity value). This enzyme is not light regulated but is responsible of two different reactions of the Calvin Cycle: the aldolase controlled reactions leading to the formation of SBP and FBP (SBP aldolase and FBP aldolase are the same enzyme [61]). The fact that the same enzyme is responsible of two reactions in the same cycle can explain its substantial sensitivity. The many enzymes with large mean and standard deviation values reflect the complexity of the pathway and the non-linear interactions occurring among enzymes. For future improvements of the model it is mandatory to consider that some of the Calvin Cycle enzymes (particularly - and not surprisingly - those with higher sensitivity values) are allosteric enzymes. The use of Michaelis-Menten kinetics is, in this case, an approximation of the real situation. Moreover, it is of relevance to consider that the regulatory networks in which the Calvin Cycle enzymes are involved, go far beyond the cycle itself. For instance, the impairment of the photorespiratory enzymes (one of the aim to be achieved in order to increase photosynthetic efficiency), could cause unexpected effects on the general efficiency since photorespiration is proposed to be important for avoiding photoinhibition of photosystem II, especially in C_3 plants [62]. This implies that the variation in enzyme concentration is unlikely to be completely free (or exclusively linked to the total protein-nitrogen amount) as assumed in this model. The large variation in sensitivity of the Calvin Cycle enzymes could be linked not only to the more or less important function of the cycle itself, but also to the contemporaneous involvement of some of these enzymes in other metabolic networks and then less

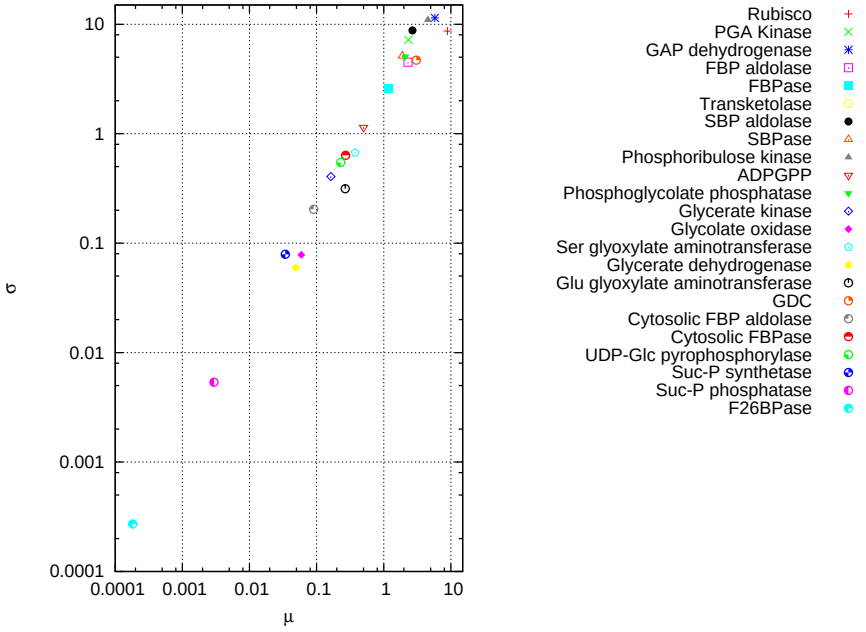


Figure 3.2: Sensitive and Insensitive Enzymes. Morris sensitivity analysis of the carbon metabolism model. For each enzyme, mean μ and standard deviation σ of the CO_2 uptake rate are reported on the x-axis and y-axis, respectively. High mean values mean linear enzymatic response, while high standard deviation values assess a non-linear behavior or dependencies among enzymes.

influenced by the Calvin Cycle selective pressures. On the contrary, enzymes with high μ value of sensitivity analysis, see Fig. 3.2, are linked to the Calvin Cycle. For instance, FBPase activity and even its mRNA expression is light regulated and hence strictly linked to photosynthesis. In order to validate the results, it has been executed a preliminary bioinformatics analysis with a BLAST [63] search on the amino acid sequences (starting from Arabidopsis genome) of the Calvin Cycle enzymes that had the most extreme sensitivity values. They have been taken into account all of the e-values calculated by BLAST as search result. The enzymes showing the highest sensitivity values, were also those with the lowest e-values in BLAST hits (corresponding to the most similar sequences found in the protein sequences database). A possible explanation of the result could be

that the amino acid sequence variation in highly sensitive enzymes is low, even in hits less related to the query sequence. Essentially, the e-value describes the random background noise. The lower the e-value, or the closer it is to zero, the more "significant" the match is (less different the sequences are). It is likely that the protein sequence is so optimized that the sequence variation is low, even in species scarcely related to the query sequence.

3.4.2 Maximal and Robust Photosynthetic Productivity

Initially, a larger family of optimization algorithms has been compared in CO_2 uptake maximization at $c_i = 270\ \mu mol\ mol^{-1}$ (reflecting the current CO_2 atmospheric concentration of 360 parts per million, ppm) and by fixing the total protein-nitrogen in the enzymes of carbon metabolism to 1 gm^{-2} of leaf area. Here, 24000 objective function evaluations are allowed as in [42]; in Fig. 3.3, I report the convergence process of the tested derivative-free optimization algorithms. It is worth noting that the EA proposed in [42] is outperformed by eight algorithms, the EA seems to stack into a local optimum after 10^4 objective function evaluations, while the designed algorithms, PAO and A-CMA-ES, achieve enhanced CO_2 uptake rates.

The most promising algorithms have been let continue the optimization process until 10^5 objective function evaluations; my PAO and A-CMA-ES algorithms found the best CO_2 uptake and they outperform H-J [53] and Differential Evolution (DE). From an optimization point of view, PAO and A-CMA-ES seem to be the most effective algorithms. The analysis of the PAO convergence shows that the algorithm rapidly reaches its best solution, and it is not able to improve it even if a large number of objective function evaluations is allowed. Surprisingly, among the three pattern search algorithms considered (H-J, GPS [54], MADS [55]), the simple H-J outperforms the other two claimed approaches. The data in Table 3.1 show the concentrations of the enzymes for the original leaf (the second column), for the optimized leaf as proposed by the evolutionary algorithm used in [42] (the third column) and four best candidates obtained by PAO and A-CMA-ES algorithms. The comparison among the robust optimized leaf (last column) and the natural leaf (second column) can help to detect the relevant enzymes

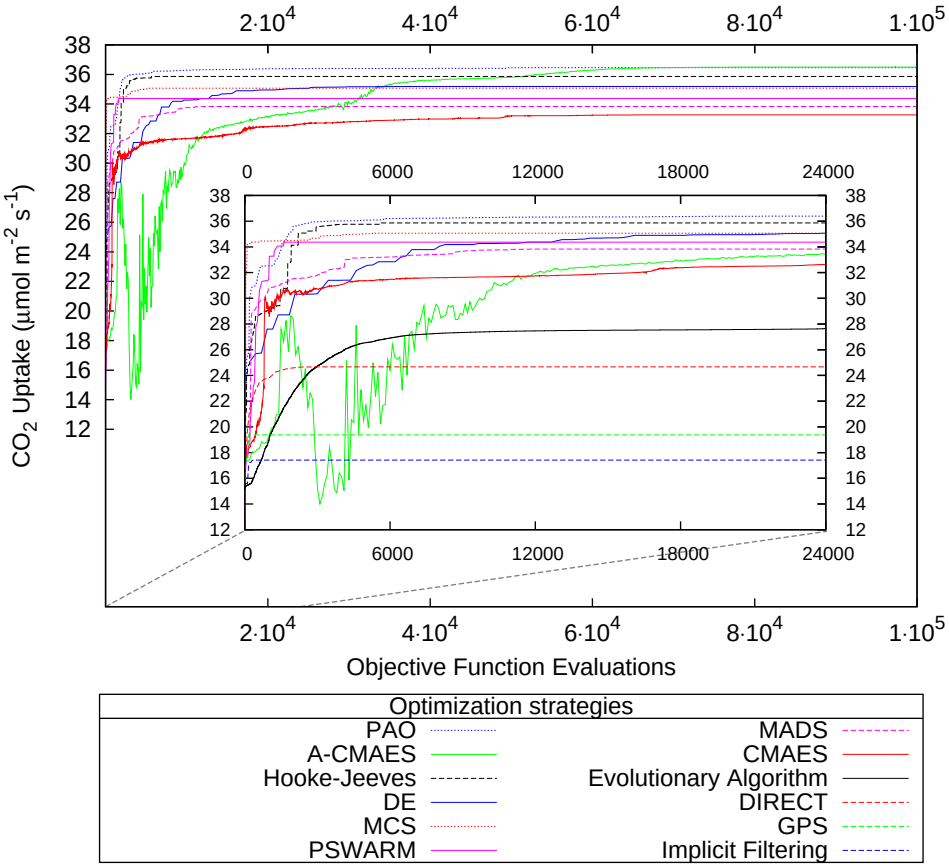


Figure 3.3: Convergence process of the derivative-free global optimization algorithms. Searching of the optimal partitioning of resources among the enzymes of carbon metabolism to maximize light-saturated photosynthetic rate (CO_2 uptake) at $c_i = 270\ \mu mol\ mol^{-1}$ (reflecting the current CO_2 atmospheric concentration). State-of-the-art optimization algorithms have been adopted and compared (in the legend from best to worst).

in order to maximize the light-saturated photosynthetic rate (see Fig. 3.4). In fact, the robust optimized leaf brings coherent relative changes with respect to the natural leaf for most of the enzymes.

In order to study the robustness of the proposed concentrations, both global and local robustness analysis have been performed; the question is "*how the gained CO_2 Uptake rate is preserved under enzyme perturbations?*"; the results are presented in Table 3.1. Two major aspects should be remarked; firstly, the concentration that achieves the maximum CO_2 uptake rate (36.495 $\mu mol\ m^{-2}s^{-1}$)

Enzyme Name	Initial Conc. $mg\ N\ m^{-1}$ *(S. Robustness %)*	Conc. $mg\ N\ m^{-1}$ found in [42] *(S. Robustness %)*	Opt. without constraints, Conc. $mg\ N\ m^{-1}$ found by A-CMA-ES *(S. Robustness %)*	Opt. with constraints, Conc. $mg\ N\ m^{-1}$ found by A-CMA-ES *(S. Robustness %)*	Opt. with constraints, Conc. $mg\ N\ m^{-1}$ found by A-CMA-ES *(S. Robustness %)*	Optimal and Robust Conc. $mg\ N\ m^{-1}$ found by PAO *(S. Robustness %)*
Rubisco	517.00 *(100)*	795.00 **(87.5)**	861.93 **(39)**	840.60 **(87)**	857.05 **(63.0)**	860.226 (100.0)
PGA kinase	12.20 *(100)*	5.06 *(100)*	3.98 **(0)**	4.90 *(100)*	4.21 *(100)*	3.989 (100.0)
GAP dehydrogenase	68.80 *(100)*	75.00 **(76.5)**	63.55 **(17)**	71.62 **(87.5)**	63.71 **(51.0)**	64.483 (100.0)
FBP aldolase	6.42 *(100)*	11.70 *(100)*	9.29 **(30.5)**	10.38 *(100)*	10.77 *(100)*	9.050 (100.0)
FBPase	25.50 *(100)*	35.90 *(100)*	27.03 **(0)**	32.07 *(100)*	31.78 *(100)*	26.889 (100.0)
Transketolase	34.90*(100)*	18.40 *(100)*	16.98 *(100)*	19.46 *(100)*	15.93 *(100)*	8.247 (100.0)
SBP aldolase	6.21*(100)*	7.43 *(100)*	5.94 **(0)**	6.95 *(100)*	5.58 *(100)*	6.661 (100.0)
SBPase	1.29 *(100)*	4.90 *(100)*	4.31 **(1)**	5.03 *(100)*	4.26 *(100)*	4.397 (100.0)
Phosphoribulose kinase	7.64 *(100)*	8.55 *(100)*	7.99 **(22.5)**	8.86 *(100)*	7.67 *(100)*	7.007 (100.0)
ADPGPP	0.49 *(100)*	4.88 *(100)*	1.22 **(0)**	2.45 *(100)*	4.75 *(100)*	0.721 (100.0)
Phosphoglycolate phos.	85.20 *(100)*	1.42 *(100)*	0.00 **(0)**	0.85 *(100)*	0.02 *(100)*	0.325 (100.0)
Glycerate kinase	6.36 *(100)*	1.31 *(100)*	0.00 *(100)*	0.03 *(100)*	0.02 *(100)*	0.005 (100.0)
Glycolate oxidase	4.77 *(100)*	1.49 *(100)*	0.00 *(100)*	1.17 *(100)*	0.02 *(100)*	0.019 (100.0)
Ser glyoxylate aminotrans.	17.30 *(100)*	3.03 *(100)*	0.00 *(100)*	0.14 *(100)*	0.02 *(100)*	0.027 (100.0)
Glycerate dehydrogenase	2.64 *(100)*	0.78 *(100)*	0.00 *(100)*	0.01*(100)*	0.02 *(100)*	0.003 (100.0)
Glu glyoxylate aminotrans.	21.80 *(100)*	4.47 *(100)*	0.00 *(100)*	0.21*(100)*	0.02 *(100)*	0.00005 (100.0)
GDC	179.00 *(100)*	18.60 *(100)*	0.00 *(100)*	1.88*(100)*	0.02 *(100)*	0.00003 (100.0)
Cytosolic FBP aldolase	0.57 *(100)*	0.28 *(100)*	2.03 **(0.5)**	0.75 *(100)*	0.89 *(100)*	2.127 (100.0)
Cytosolic FBPase	2.24 *(100)*	1.44 *(100)*	5.27 **(30.5)**	2.05 *(100)*	2.50 *(100)*	5.554 (100.0)
UDP-Glc pyrophosphorylase	0.07 *(100)*	0.07 *(100)*	0.50 **(0)**	0.56 *(100)*	0.70 *(100)*	0.531 (100.0)
Suc-P synthetase	0.20 *(100)*	0.15 *(100)*	0.03 **(30.5)**	0.09 *(100)*	0.03 **(92.5)**	0.034 (100.0)
Suc-P phosphatase	0.13 *(100)*	0.07 *(100)*	0.03 **(0)**	0.01*(100)*	0.02 *(100)*	0.031 (100.0)
F26BPase	0.02 *(100)*	0.01 *(100)*	0.00 *(100)*	0.03 *(100)*	0.02 *(100)*	0.0 (100.0)
CO_2 Uptake $\frac{\mu mol}{m^2 s}$	15.486	27.621	36.495	35.146	36.290	36.382
Local robustness %	100	76.50	0	87.0	51.0	100
Global robustness %	81.80	78.44	39.18	79.42	100.0	97.2

Table 3.1: Concentrations of the enzymes (Cf. Appendix 1 for nomenclature), and Single Robustness (S. Robustness), CO_2 Uptake, Local and Global Robustness (in the last three rows). The second and third columns report the initial concentrations of enzymes used in the simulation, (initial leaf, or natural leaf), and the optimized leaf as predicted by the evolutionary algorithm used in [42]. The last four columns show the best candidate solutions obtained by the designed PAO and A-CMA-ES algorithms. This set of candidate solutions has been obtained at $c_i = 270\ \mu mol\ mol^{-1}$ (reflecting the current CO_2 atmospheric concentration).

is extremely sensitive, and its robustness values are all below the robustness of the other solutions. In particular, by inspecting the local robustness analysis it is possible to note that many enzyme concentrations are not robust, and many of them lead to a completely unreliable pathway. By inspecting the results of local robustness analysis, it is worth noting that the Rubisco and GAP dehydrogenase

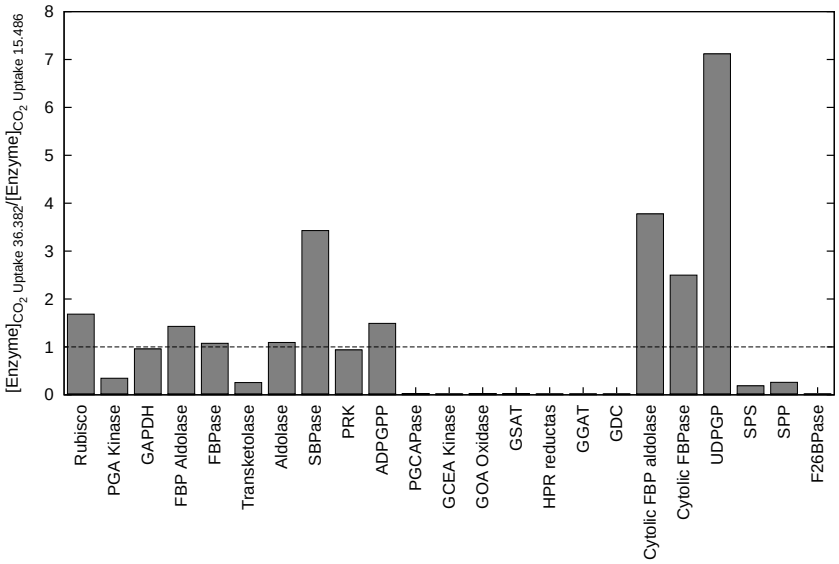


Figure 3.4: The ratio of the enzyme concentrations optimized by the PAO algorithm (36.382 $\mu mol\ m^{-2}s^{-1}$) at a $c_i = 270\ \mu mol\ mol^{-1}$ compared to the initial concentrations (15.486 $\mu mol\ m^{-2}s^{-1}$).

are the less robust enzymes for four over six candidate solutions. Using the designed optimization framework PAO I have obtained an increase in photosynthetic productivity of the 135% from 15.486 $\mu mol\ m^{-2}s^{-1}$ to 36.382 $\mu mol\ m^{-2}s^{-1}$ (last column), improving the previous best-found photosynthetic productivity value (27.261 $\mu mol\ m^{-2}s^{-1}$). Moreover, this new set of enzyme concentrations has a maximal local robustness (100%) and a high global robustness (97.2%). With respect to the initial concentration of enzymes, increases in Rubisco, FBP aldolase, SBPase, ADPGPP and a strong increases in Cytosolic FBP aldolase, Cytosolic FBPase, UDP-Glc pyrophosphorylase were required to a large increase of CO_2 uptake rate (see Fig. 3.4). Moreover, there are four enzymes, GAPDH, FBPase, SBP aldolase, and Phosphoribulose kinase, approximately maintaining the same values of the initial concentrations, while PGA kinase, Transketolase, Suc-P synthetase and Suc-P phosphatase are under-expressed; the remaining enzymes are switched off. The under- and over- expressed pattern of Fig. 3.4 is well defined, the change of concentrations of the enzymes of carbon metabolism between optimized leaf and natural leaf does not show ambiguities.

As noted in [64; 65], SBPase is a very particular enzyme: approximately 10% of increase in photosynthetic rate has been observed in transgenic plants over-expressing SBPase enzyme. It is crucial, hence, to verify if further gains could be obtained in transgenic plants if, in addition, Rubisco, FBP aldolase, ADPGPP, Cytosolic FBP aldolase, Cytosolic FBPase, and UDP-Glc pyrophosphorylase were over-expressed.

3.4.3 Multi-objective optimization of the carbon metabolism: CO2 uptake vs. Protein-Nitrogen

Pareto Optimality is one of the most fruitful and powerful approach where optimization of conflicting objectives is concerned[66; 67]. The multi-objective formulation of the re-design process poses a serious algorithmic challenge, since the defined Pareto front is not easily analyzable; for this reason, a derivative-free multi-objective optimization algorithm, PMO2, has been designed with the aim of producing a good approximation of Pareto optimal concentrations. Here I present the results of the analysis whose aim is the evaluation of the contextual maximization of the CO_2 uptake rate, while minimizing the actual amount of total nitrogen contained in the enzymes.

The capability of reducing the amount of nitrogen necessary to fix CO_2 in biomass is an important goal for biotechnology. Large increases in the efficiency of nitrogen usage, will be necessary to maintain or increase current food production in a sustainable manner [68]. Intensive high-yield agriculture is dependent on addition of fertilizers, especially industrially produced NH_4 and NO_3 [68]. Fig. 3.5 shows that the optimization may largely improve nitrogen usage in photosynthesis without affecting CO_2 uptake rate. Moving beyond the natural operative area (area checked in green), I found leaf configurations that expose a Pareto-optimality in the six conditions considered (three C_i atmosphere values and two triose-P export rates). The candidate highlighted as B represents a leaf with a natural CO_2 uptake ability, but employs 47% of the naturally needed protein-nitrogen. The A2 candidate is interesting as well: it needs exactly 50% of the naturally employed protein-nitrogen to gain up to 10% CO_2 uptake capacity, when compared to the natural leaf. The enzymes involved in concentration

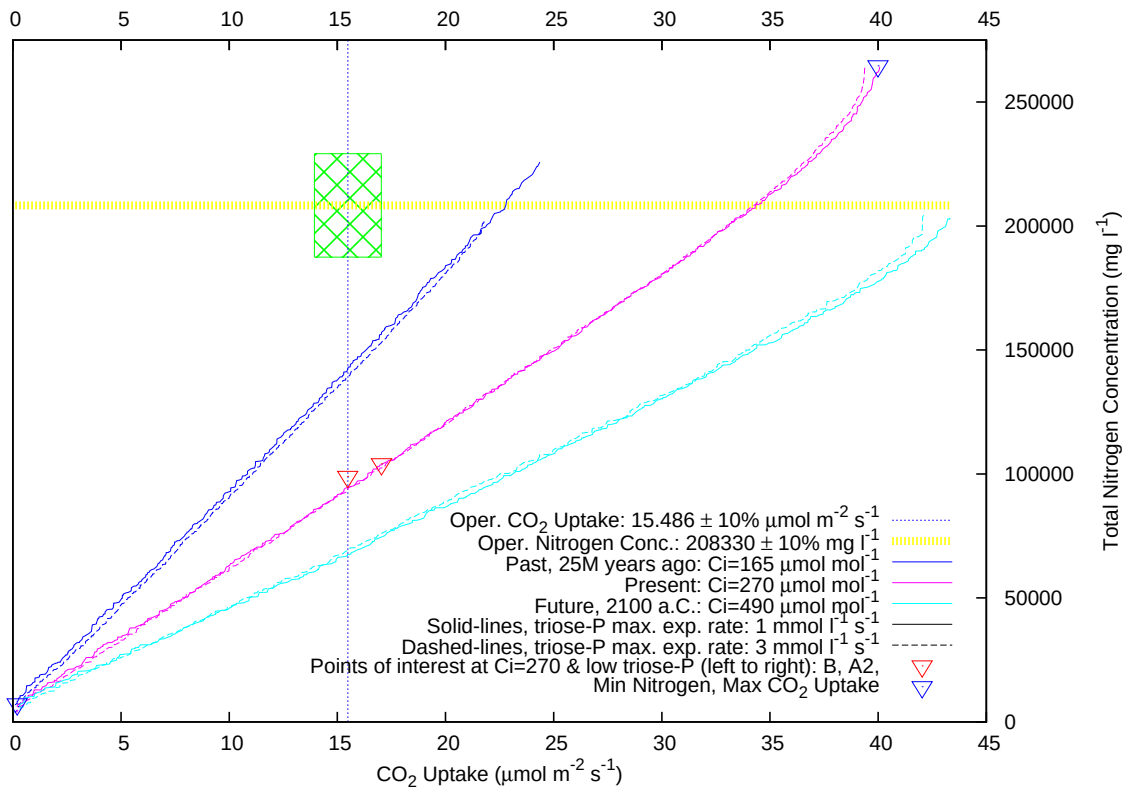


Figure 3.5: PMO2 results: multi-objective optimization of two conflicting biological pressures; leaf CO_2 uptake rate versus protein-nitrogen consumption.

variation are almost always the same: Rubisco provides nitrogen to increase the concentration of other enzymes. A slight reduction in Rubisco corresponds potentially to a large amount of protein nitrogen available for increasing concentration of the other enzymes. As a matter of fact the high concentration of Rubisco in the leaves was considered to have a possible function also as nitrogen reservoir [69].

Fig. 3.6 shows the concentration of the enzymes in the B leaf with respect to the natural concentrations. From a re-engineering point of view, the two leaves are similar; in fact, each enzyme involved shows a growth/reduction in concentration that is within the range 0.05x-2x ca. Despite this relatively small metric distance and the equal uptake rate, the biochemical effort paid by the two leaf designs is substantially different. SBPase and ADPGPP confirm their leading role in the leaf engineering. These results show that re-engineering the nitro-

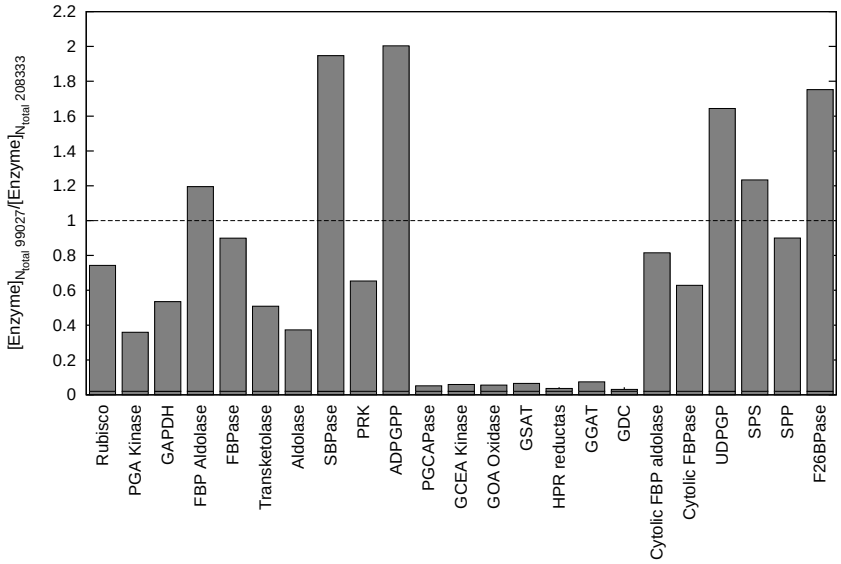


Figure 3.6: Comparison among the Pareto-optimal re-engineering candidate B (that uses a total concentration of Nitrogen equal to 99027 $mg\ l^{-1}$) and the natural leaf (whose total concentration of Nitrogen is 208333 $mg\ l^{-1}$).

gen partitioning among well determined enzymes (individuated by the detailed framework) can lead to theoretical leaves capable of reducing significantly the general amount of nitrogen without affecting the potential biomass production. It is interesting to observe that the enzymes of the photorespiration, a process acting against the general photosynthetic yield, are not kept at zero as in other models. Photorespiration has a major impact on carbon uptake, particularly under high light, high temperatures, and CO_2 or water deficits [70]. Nevertheless although the functions of photorespiration remain controversial, it is widely accepted that this pathway influences a wide range of processes from bioenergetics, photosystem II function, and carbon metabolism to nitrogen assimilation and respiration. For instance photorespiration is a major source of H_2O_2 in photosynthetic cells. Through H_2O_2 production and pyridine nucleotide interactions, photorespiration makes a key contribution to cellular redox homeostasis. Doing so, it influences multiple signaling pathways, particularly those that govern plant hormonal responses controlling growth, environmental and defense responses, and programmed cell death [70].

In summary, I modeled the $C3$ photosynthetic carbon metabolism in terms of concurrent optimization of two conflicting biological strengths: maximization of CO_2 uptake and contextual minimization of the total protein-nitrogen employed to gain that property (representative of the biochemical effort the leaf has to devote to gain that CO_2 uptake rate). I inspected the problem at three CO_2 concentrations (C_i) in the atmosphere or stroma (25M years ago environment, nowadays one, and the one predicted for the end of the century) and two triose-P (PGA, GAP, and DHAP): low and high export rates. In this context, my analysis has detected Pareto-optimal configurations in the six C_i/triose-P conditions studied. Among the others, two promising candidates for leaf re-engineering have been further inspected and compared with the natural leaf enzyme configuration. For the first time, it has been individuated a reasonably small set of key enzymes whose targeted tuning gives rise to a robust maximization of the photosynthetic rate, contextually with an efficient protein-nitrogen employment. It also interesting to note that for increasing atmospheric CO_2 it is possible to obtain a major CO_2 uptake rate with a minor protein-nitrogen concentration.

3.5 Discussion and Conclusions

Optimizing the CO_2 uptake rate is a complex task, that has been tackled by ad-hoc optimization algorithms, A-CMA-ES, PAO and PMO2; the found solution is robust and assures a gained CO_2 uptake rate of 135%. I used a multi-objective optimization approach in order to maximize the CO_2 uptake rate and minimizing the protein-nitrogen concentration; the analysis of the Pareto front shows that, for increasing CO_2 atmospheric concentrations, it is possible to obtain an improved CO_2 uptake rate with a decreasing protein-nitrogen concentration. From 1850 to 2006, fossil fuel and cement derived CO_2 emissions, released a cumulative total of $\sim$ 330 petagrams of carbon (PgC) to the atmosphere. An approximately additional 158 PgC came from land-use-change emissions, largely deforestation and wood harvest [71]. The growth rate of global average atmospheric CO_2 for 2000–2006 was 1.93 $ppmy^{-1}$ (parts per million per year) [71]. Primary production of world biomass, considering both marine and terrestrial sources, robustness an estimated global net primary production of 104.9 petagrams of carbon per year

[72], while Cellulose and Lignin, the most abundant organic resources in the world, exhibit an annual turnover rate of 4×10^{10} tonnes, or 40 petagrams [73]. My results show that the potential increase in CO_2 uptake obtainable by varying enzyme concentration of the Calvin Cycle might increase the current CO_2 uptake by 135%, hence a quantity potentially capable to counteract CO_2 emission in atmosphere by human activities. Such an increase could be obtained partly naturally by varying gene expression of the involved enzymes, or by selecting individuals that could modify the expression hence increasing their Calvin Cycle efficiency. This second mechanism would require a long time unless we consider the hypothesis of artificially modifying of DNA involved in gene expression control. This last possibility would require careful evaluation of possible risks linked to introduction in the environment of organisms capable of fast growth in a CO_2 rich atmosphere. The increase in biomass productivity and CO_2 uptake calculated by optimized enzyme partitioning might potentially counteract the current increase in atmospheric CO_2.

Photosynthesis and particularly the biochemical pathway of carbon fixation (the Calvin Cycle) has been object of many studies (for a review see for instance [74; 75; 76]) and some journals are directly entitled to this fundamental biological process. In this research I have identified key enzymes to target in order to maximize CO_2 uptake rate and minimize the protein-nitrogen in $C3$ plants. The designed methodology, including multi-objective optimization, unravelled that Rubisco, Sedoheptulosebisphosphatase (SBPase), ADP-Glc pyrophosphorylase (ADPGPP) and Fru-1,6-bisphosphate (FBP) aldolase are the most influential enzymes in carbon metabolism model where CO_2 uptake maximization is concerned. Interesting insights include the fact that the Rubisco enzyme participate with a very high concentration; additionally, some of the photorespiratory enzymes that should be almost switched off to reach the best configurations known [42] cannot be effectively switched off because they are involved in other processes carried by $C3$ plants. The pathway enzymes that lead to sucrose and starch synthesis were shown not to affect CO_2 uptake rate if maintained at their natural concentration levels. The importance of SBPase has already been pointed out by antisense transgenic plants studies [76].

3.5.1 Assessment of the quality of the results obtained thought the multi-objective optimization

The optimization performed using the PMO2 algorithm provides a large set of trade-off solutions (Cf. Appendix 1 for details on alternative solutions); in particular, 755 Pareto optimal concentrations have been found, that are the 1.83% of the total enzymes partitions explored by the algorithm.

Algorithm	Points	R_p	G_p	V_p
PMO2	**775**	**1.0**	**1.0**	**0.976**
MOEA-D	137	0	0	0.376

Table 3.2: Pareto front analysis. For each algorithm, they are reported the number of Pareto Optimal points (non-dominated points), the *relative Pareto coverage* indicator (R_p), the *global Pareto coverage* indicator (G_p), and the *hypervolume* indicator (V_p).

In order to assess the quality of the Pareto frontiers (at present C_i value of 270 $\mu mol\ mol^{-1}$ and maximal rate of triose-P (PGA, GAP, and DHAP) export of 3 $mmol\ L^{-1}\ s^{-1}$), I compare the results obtained by PMO2 and MOEA-D, another state-of-the-art evolutionary multi-objective optimization algorithm [77]. The terms of comparison are the metrics detailed in Chapter 2: Pareto Optimal points (non-dominated points), the *relative Pareto coverage* indicator (R_p), the *global Pareto coverage* indicator (G_p), and the *hypervolume* indicator (V_p). The results reported in Table 3.2 confirm the quality of the candidate solutions obtained by PMO2. Successively, from the Pareto front, they have been selected the *shadow minima* for each objective and the *closest-to-ideal* solutions; successively, they have been computed the global robustness of these concentrations. Moreover, in addition to these solutions, they have been picked 50 Pareto optimal points equally spaced on the Pareto front and their robustness have been estimated. In table 3.3, it is possible to note that the three concentrations selected by the automatic criterion are quite robust (Yield column), even if they greatly differs in terms of CO_2 Uptake rate and nitrogen concentration; this experimental evidence seems to confirm that trade-off concentrations represent robust pathway configurations despite the changes in their uptake capability and nitrogen

Selection	CO_2 Uptake	Nitrogen	Yield
Closest-to-ideal	21.213	1.270×10^5	67
Max CO_2 Uptake	39.968	2.641×10^5	65
Min Nitrogen	5.7	3.845×10^4	50
Max Yield	37.116	2.291×10^5	82

Table 3.3: Pareto Front analysis. For each Pareto optimal solution, we report the selection criterion, the CO_2 uptake rate, the nitrogen amount and the yield value.

required. However, by inspecting the Pareto front it is possible to find a new enzyme partition that achieves a slightly worse uptake rate but a remarkable increase in terms of robustness; from this analysis, it is clear that the yield is another conflicting objective and, hence, an inherent trade-off emerges.

More in detail, to inspect the relation between CO_2 uptake, Nitrogen consumption and the inherent solution robustness, it has been assessed the fitness landscape with respect to these three objectives. Figure 3.7 presents the results of this analysis by means of a 3D Pareto-surface. Despite the rugged aspect of the surface, that highlights how far from an ideal world and how real is the problem we are tackling, it is clear that Pareto relative minima are highly unstable points, while if we accept a slightly lower optimization in the functional objectives, we can obtain a significantly more reliable solution.

Finally, looking at the concentrations of the *closest-to-ideal* solutions, some more interesting results are observable; except for the GOA Oxidase, each algorithm maintains a concentration close to the natural concentrations. Remarkable increases are observable for GAP DH, GGAT, Cytolic FBP Aldolase, SPP and F26BPase enzymes. At this point, it is possible to infer that these enzymes are the best candidate for a trade-off performance leaf. Clearly, it is important to remark that modest increment of other enzymes are plausible since they have a higher molecular weight. It should be observed that even if some of the considered enzymes fall to zero in main photosynthesis models in the optimized leaf, such a low concentration could influence other important biochemical pathways. For instance photorespiration-related enzymes as Glu Glyoxylate Aminotransferase and GOA oxydase fall considerably in concentration at the optimized state. Photorespiration is by far the fastest H_2O_2 -producing system in photosynthetic cells under

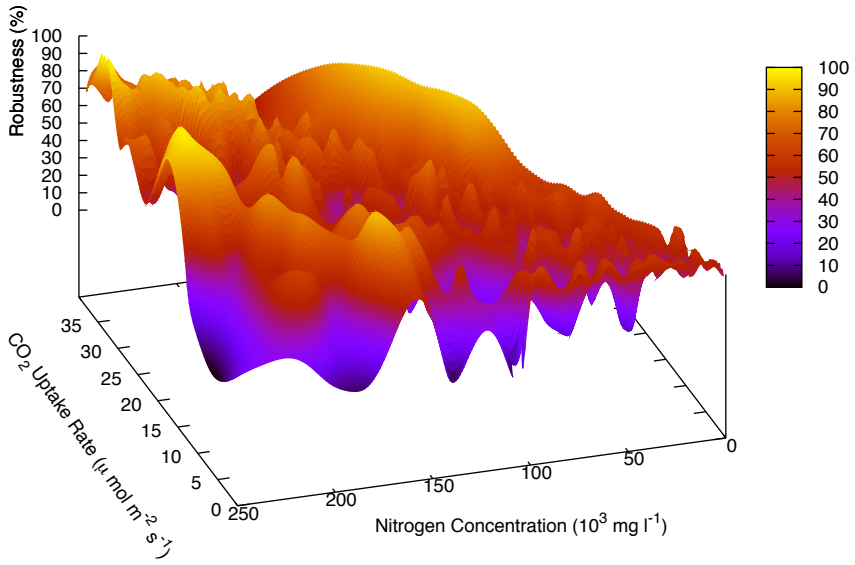


Figure 3.7: Photosynthetic Pareto surface. Robustness vs CO_2 uptake (x-axis) and Nitrogen consumption (y-axis).

many conditions [78]. H_2O_2 is an important intracellular signal [70]. Moreover the photorespiratory pathway metabolizes glycolate-2-P to Glycerate-3-P and is considered important to avoid photoinhibition of photosystem II, particularly in $C3$ plants [62]. Photorespiratory mutants of Arabidopsis with inactivation of some of the enzymes of the photorespiratoy pathway did not show negative effects at high level of external CO_2 but CO_2 fixation rates declined drastically at current atmospheric CO_2 concentration [62]. This means that models based only on the photosynthetic pathways leading to strong decrease in concentration of the photorespiratory pathway enzymes, should take into consideration that this pathway is necessary to the plant for aspects that have not been considered in current models.

From a methodological point of view, I report that the optimization method-

ologies in the systems biology framework is a thriving field of research. It has two immediate and important benefits: the improved understanding of the processes that shape the evolution of energy collecting engine at the molecular level and the improved ability to use optimization methods to predict from molecular data directions where experiments should go and drive the decision process in biotechnology.

Finally, these can be considered as points of strength: 1) as far as I know it is the first time that the overall framework, sensitivity, optimization and robustness, is used for the study of biological pathways; 2) it is the first time that local and global robustness analysis has been defined and used to study molecular entities, and 3) for the first time, the C_3 photosynthetic carbon metabolism has been characterized by CO_2 uptake rate versus protein-nitrogen Pareto frontiers which I prove to be a meaningful and effective way to address this class of bioinformatics and bioengineering problems.

The integration of optimization methods with bioinformatics is shaping at growing pace our comprehension of biological processes Optimization methodologies provide an essential tool to capture a set of assumptions and to follow them to their precise logical conclusions. They allow us to generate new hypotheses, suggest experiments, and measure crucial parameters. If the scientific progress relies on asking the right questions, the combination of optimization methods and bioinformatics will suggest more insightful questions and answers than bioinformatics techniques alone.

Explorations in Pareto front analysis suggest that its shape may reflect the amount of epistasis (where the effects of one gene are modified by one or several other genes) and pleiotropy (when a single mutation or gene affects multiple distinct phenotypic traits) in the metabolic pathway, so that simpler independent traits may generate simpler Pareto fronts. It is know that complexity and in particular fitness traits such as energy balance, growth and survival, depend on both the epistatic and pleiotropic structure of a metabolic pathway and therefore strongly influences evolutionary predictions.

Chapter 4

Biological and Medical Ontology Reasoning

4.1 The OREMP Project

The information coming from biomedical ontologies and runnable pathways is expanding continuously: research communities keep this process up and their advances are generally shared by means of dedicated resources published on the web. In fact, runnable pathways are shared to provide the characterization of molecular processes, while biomedical ontologies detail a semantic context to the majority of those pathways [11].

Recent advances in both fields pave the way for a scalable information integration based on aggregate knowledge repositories [12; 13], but the lack of overall standard formats impedes this progress. Having different objectives and different abstraction levels, most of these resources "speak" different languages.

Employing an extensible collection of interpreters, I propose a system that abstracts the information from different resources and combines them together into a common meta-format. Preserving the resource independence, the system provides an alignment service that can be used for multiple purposes. Recent examples are: 1) The new web application Cytosolve [79] uses an embedded version of this system to provide congruous parallel simulation of multiple models; 2) Using the BioModels.net database[80], a searchable dictionary of equivalent

molecular reaction paths is built. Finally, the enriched knowledge can be exported in OWL2 [81] and queried by semantically-enabled tools such as Protégé [82]. In this approach, I see a valuable tool to integrate and reason information originating from different sources, while preserving the independence of the model curation process; additionally, information sharing, integration and discovery are the primary features here provided.

4.2 Introduction

The information about molecular processes is expanding continuously and the descriptions are shared in the form of computable pathways. Biomedical ontologies are being created to provide a semantic context for the molecular species and reactions that they contain. Current advances in both topics suggest an information integration cycle based on shared knowledge-bases, but because of different languages (*i.e.*, the data formats) spoken by the data sources and different abstraction levels, there is a lack of an overall frame capable of identifying overlaps and duplications [11]. One can envision searchable biological resources, such as the Gene Ontology [83], UniProt [84], ChEBI [85], KEGG [86], Reactome [87] and BioPortal [88], defining the biological context of the pathways in a machine-readable format. Substantial effort has been devoted to the creation of ontological resources which are publicly available, but there are semantic obstacles that inhibit their combined use. On the other hand, it is desirable to inform databases of runnable pathways, such as the BioModels.net collection, the CellML repository [89] and even specialist repositories [90; 91; 92], with the information contained in the curated molecular ontologies in a manner that can be used easily. Some syntactic conversions are available among pathway data-formats [93; 94], and the state of the art for adjudication of the discrepancies between two SBML [14] models is SemanticSBML [95], which exploits machine-readable information and the user input to create a merged SBML model. Unfortunately, in the context of large-scale composite biological pathways, the merged-model approach is undesirable because it destroys the original component models and interrupts the curation process. For more than two SBML files, the tool must be run repeatedly with user-input, subjecting it to increasing human error, and suggesting that the

order in which the models are aligned matters. An alternative approach based on the use of ontologies discerns when and on which topics models are a relevant part of the large-scale context. Where bio-ontologies are concerned, the state of the art is represented by BioPortal which provides uniform access to most of the biomedical ontologies through a single user-interface and advanced tools to query over biomedical data resources. As a matter of fact, there is still a large chasm between today's functionality and the true ability to use ontological data to inform molecular pathways. Additionally, there is a lack of strategies for the database and ontology integration of quantitative biological sources written in different standards (*e.g.*, SBML and CellML [15]). What is described here is a system that creates extended ontologies out of different biochemical information sources and provides path duplication detection, sharing, integration, and knowledge discovery over heterogeneous resources. This cross-format system, I called *OREMP* (Ontology Reasoning Engine for Molecular Pathways) exports the extended ontologies in OWL2 format; the latter can be fed to Protege, where the information can be then browsed and edited at different levels of abstraction.

This framework, that I developed at Massachusetts Institute of Technology, is currently employed in the Cytosolve@MIT project [96; 97; 98].

4.3 System Architecture and Operational Workflow

Biological processes are largely modeled in terms of systems of ordinary differential equations (ODE); a forum of researchers, developers and end-users designed an encoding for these ODE systems that is based on XML: after years of discussions the result is the Systems Biology Markup Language (SBML) whose features are outlined in [14]. A runnable pathway, or simply *model*, is a set of biochemical *species* whose evolution in time is determined by the *reactions* they participate in. These reactions, as well as the species, are specified in the SBML file: the manner define by means of MathML [99] sections how species evolve. A bio-ontology is just an ontology where is formally detailed how some life-science elements relate to each others. The simplest example is the definition of a newly discovered

molecule "A" that is believed to be of type "B"; this will be formalized as

A *IsA* B.

In order to step beyond simple syntactical translation, I designed a system that merges the information from molecular pathways and curated biological ontologies into extended ontologies using a specific meta-format.

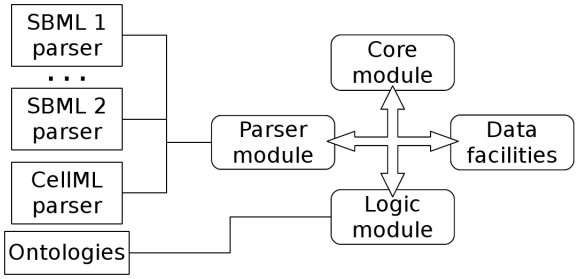


Figure 4.1: System architecture: its components are integrated to work together preserving a flexible and easily extensible architecture. Each module has different versions used on the basis of job in progress (e.g., to parse an SBML file, it will be dynamically chosen the SBML parser).

The system is composed of interchangeable and extensible components (Figure 4.1). The four components of the system are the following

- the *data access facilities*, meant to collect information about multiple pathways and existing biological databases;
- the *parser module* that can read different file formats and extracts information from those sources;
- the *core module* where knowledge from different sources can be assembled to later fill a coherent ontology;
- the *logic module* defines the conditions that identify when two biomolecular elements are in conflict, with respect to external ontologies as well;

Effectively, the information (e.g., species, reactions and references to ontologies) coming from heterogeneous resources is abstracted into our internal meta-format through these modular computational steps:

1. The *data access facility* collects information about multiple pathways and existing biological databases.

2. A *parser module* reads different file formats (i.e.: XML, RDF, SBML, CellML, etc) and extracts relevant information.

3. The *core module* assembles the knowledge, parsed from different sources, into a coherent ontology (based on our meta-format, cf. Table 4.1).

4. The *logic module* can annotate all of the species from a collection of reactions and do automated comparisons, identification of common species, and duplicate reactions.

It is worth noting that different versions of each module can in fact be used. An internal algorithm chooses the proper component implementation according to the current task (e.g., to read an SBML file, the system will invoke the SBML parser from its extensible list of parser modules). In fact, while the operational work-flow (1-4) is kept fixed, it is of note that different versions of each component may be loaded by the system. A user-configurable algorithm chooses at run-time the components that are required for the current job. This means that whenever a new modeling standard is introduced, a new parser can be connected to OREMP to interface with it as well. Similarly, different users can define different versions of the *core module*, for example, according to their understanding about how the knowledge coming from different pathways should be aggregated. This is of particular interest in domain-specific applications: according to different curators, different resources are more valuable than others and there are no gold-standards universally accepted.

A key part of this approach is the designed meta-format; around the latter the information is collated and merged together while preserving model identity. This meta-format has been designed to embed the minimalistic and quantitative MIRIAM-compliant [100] information derived from different pathways. Model annotations are preserved and extended with supplemental quantitative data to achieve a common description that can be represented as a single ontology. The structure of this ontology is presented in Table 4.1.

Entity	*has*
Annotation	type:STRING, uri:STRING, information:STRING.
Species	name:STRING, internalId:STRING, initialValue:REAL, inPathway:PATHWAY, hooks:SET_OF_ANNOTATIONS.
Kinetic reaction	internalId:STRING, kinetics:FORMULA, kineticParameters:SET_OF_PARAMETERS, inPathway:PATHWAY, reactants:SET_OF_SPECIES, catalysts:SET_OF_SPECIES, products:SET_OF_SPECIES, hooks:SET_OF_ANNOTATIONS.
Parameter	name:STRING, value:REAL.
Pathway	fullname:STRING, hooks:SET_OF_ANNOTATIONS.

Table 4.1: Main components of the minimalistic quantitative MIRIAM-compliant ontology used to abstract heterogeneous resources associated with biomolecular pathways. The format "attribute:REPRESENTATION" is used.

It is worth noting that, if we delete the link coming with the *inPathway* attribute, all of the elements abstracted in the meta-format can be disconnected from their original pathway and reasoned as if they all came from the same source. On the other hand, after this aggregate reasoning is performed, each conflict can be traced down to its source through the chain $\{Species|Kineticreaction\} \leftrightarrow Annotation \leftrightarrow Pathway$. This tunable abstraction level comes very handy when a pathway database has to be seen as a single source of information and its redundancies have to be aligned. After interpreting different formats into the internal representation (our meta-format), another computational step is taken:

6. The *logic module* computes N-order species set-set reachability of all the reactions within the loaded and aligned models.

In empirical models, as said for model repositories, the detection of duplicates is extremely important because (for instance) a duplicate reaction may lead to erroneous results. The duplicates are revealed to the user, allowing individuals to retain editorial power over their models. It also assists researchers in understanding how the resulting models of their work fit into models produced by others. The N-order reachability (duplicate reaction detection) among species sets builds a reaction composition analysis by constructing a matrix which represents a directed graph. Each vertex is a set of species and each edge is a reaction, which abstracts the overall species-set connectivity. This graph does not become a multi-graph for each set of duplicate reactions (first-order duplicate) because only one element is taken as a group representative. Through this reachability

computation, a dictionary of potentially equivalent reaction compositions is built: candidate paths of the same starting and ending sets of species, but involving alternative intermediate paths. Fig. 4.2 presents a case where first-order (N=1, R1 and R2) and N-order (R*) duplicate reaction paths overlap: the dashed arc means that is traverses more species-set apart from X and Y.

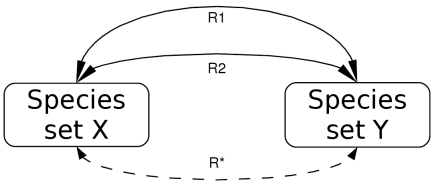


Figure 4.2: First-order and N-order reaction overlaps

The last computational step is the following:

7. The extended ontology is exported in OWL2 [81] and can be queried and edited by means of semantic tools such as Protégé [82].

From the implementation point of view, the main OREMP system functionality is written in Java, while the N-order reachability is implemented separately in Python to exploit Psyco library [101]. Additional information can be obtained using FACT++ [102] and query interface embedded in Protégé once the latter has been fed with the ontology we export.

4.4 Three Real-World Applications

Our system has been tested in three real-world applications. (i) In a simple example, we demonstrate the system's power to detect a first order duplicate reaction in the EGFR model [103] that has been factored up, but overlaps in one reaction, and the difference in quantitative results. Next application (ii) consists in the fact that Cytosolve, which is a new computational environment for parallel simulation of multiple pathways, embeds a version of the OREMP system; there it is assigned to the task of identification of common molecular species and duplicated reactions with minimal human intervention. Last application (iii) is the

combined analysis of the entire BioModels.net curated collection (currently 240 molecular pathways); OREMP has presented an aggregated view of the collection and brought to the identification of thousands of biological equivalent reaction chains, contextually a dictionary of biological building blocks has been extracted.

4.4.1 EGFR model

The combined execution of two overlapping models without detecting reaction duplication will produce an incorrect evolution of species concentrations in time. This is a concrete, quantitative effect of incorrect ontology alignment. In this example, part of a well-known EGFR (Epidermal Growth Factor Receptor) model [103] has been factored into two pieces (pathway A in Fig. 4.3 and pathway B in Fig. 4.4), containing a first order reaction pathway duplicate between the two models. The two separate model pieces are put back together and simulated simultaneously using the Cytosolve web-application, taking advantage of OREMP to inform the user about potential inconsistencies found among pathways. Without such consistency control, the evolution of the species concentration in time can lead to unpredictable values.

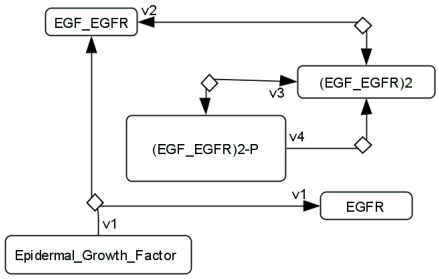


Figure 4.3: EGFR Pathway A

Fig. 4.5 presents the right parallel simulation (model A, model B) executed by Cytosolve, where our system was used to detect the conflict among the two pathways (i.e., reaction $v3$), and the user decided to zero the $v3$ rate constants in model B. Fig. 4.6 presents the same case, without accounting for the duplicated $v3$ reaction. The resulting $(EGF\_EGFR)2 - P$ and $(EGF\_EGFR)2 - PLCg$

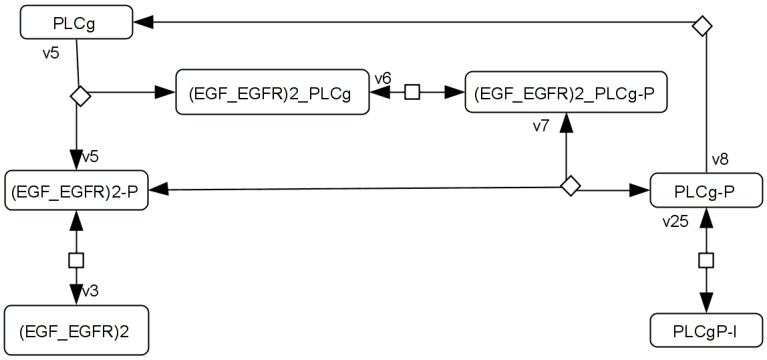


Figure 4.4: EGFR Pathway B

species concentration trends are different both in shape and magnitude, since the reaction $v3$, present in both models, led to increased species production. Note that this also triggers premature escalation of the $PLCgP-l$ concentration. The time needed by OREMP to perform this additional analysis is on the order of milliseconds.

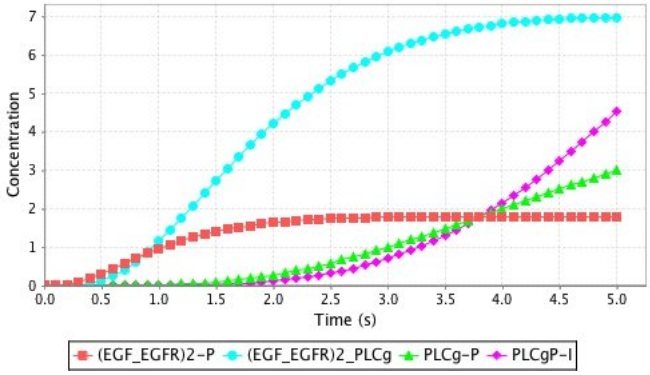


Figure 4.5: EGFR Pathway A combined with EGFR Pathway B

4.4.2 OREMP in Combining Pathways for Parallel Solution.

This system is embedded in the latest release of Cytosolve [79]. Its contribution to the integration of runnable pathways is the detection of duplicated reactions

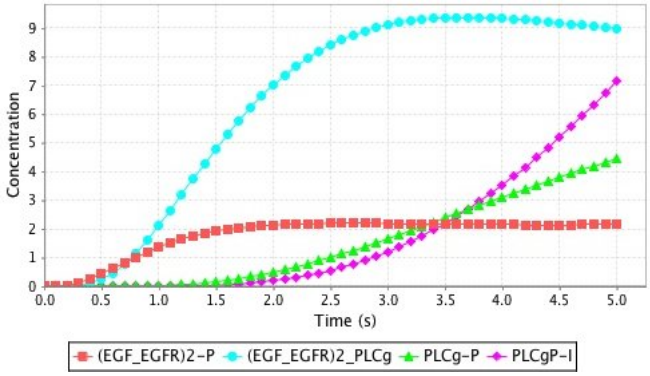


Figure 4.6: EGFR Pathway A combined with EGFR Pathway B, without accounting for the detection of the duplicate reaction

among different models. No matter the models chosen for simulation, once the species are aligned, the system identifies duplication problems in the reaction-models. From the user point of view this process is transparent: he/she receives a warning message that details the duplicated reactions and is prompted to confirm conflict elimination, and to resolve any differences in reaction kinetic rate constants. What follows is the outline of the process that starts at Cytosolve@MIT and moves from isolated pathways to their coherent parallel solution.

- Cytosolve, step 1: Multiple Simulation begins, Fig. 4.7
- Step 2: models BIOMD..1 and BIOMD..2 are selected, Fig. 4.8
- Step 3: OREMP points out the overlaps among the two models, Fig. 4.9
- Step 4: the user silences the reaction in conflict and re-uploads model 1, Fig. 4.10
- Step 5: the simulation takes place and the results are visualized, Fig. 4.11

4.4.3 OREMP in Querying Large, Independent Sources of Pathways.

The system has been tested against the entire Biomodels.net curated collection [80] that contains about 240 molecular pathways. The result of the analysis is

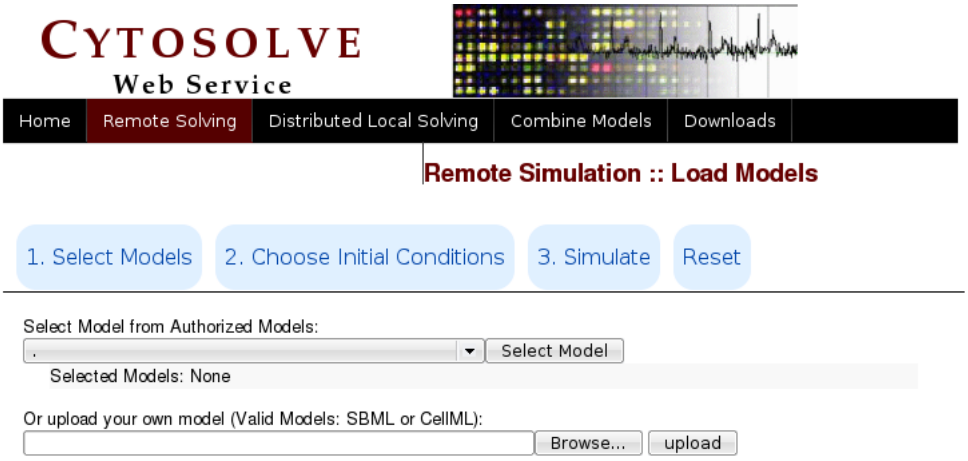


Figure 4.7: Cytosolve, step 1: Multiple Simulation begins

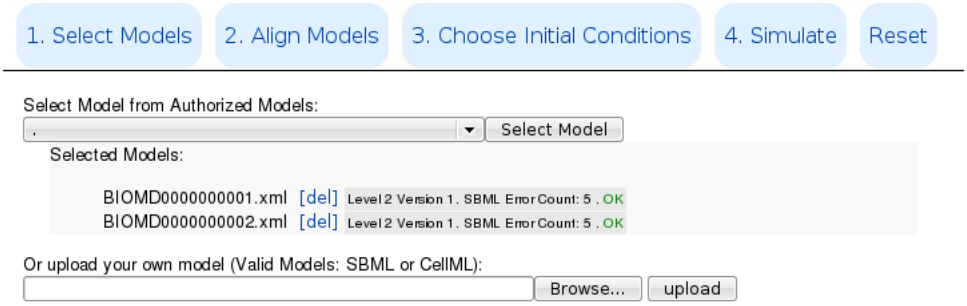


Figure 4.8: Cytosolve, step 2: models BIOMD..1 and BIOMD..2 are selected

an overall view of the database and a list of about 500 groups of overlapping reactions. This analysis took 50 seconds on a single-core 2GHz Intel CPU. The previously described knowledge-discovery-step involving N-order reachability has been taken on these resources as well. For each species configuration in the database, all alternative circuit paths have been computed. This took about 2 hours on a quad-core 2GHz AMD CPU and resulted in a dictionary of thousands "biological equivalent" circuits (*i.e.*, equivalent reaction compositions). More precisely we have obtained:

- An ordered dictionary of pathway building blocks

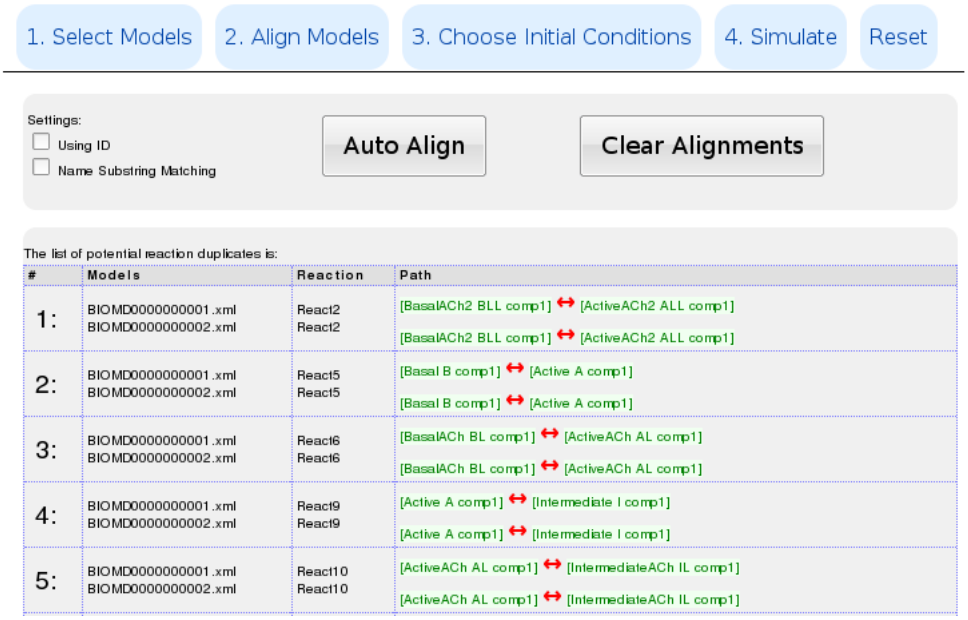
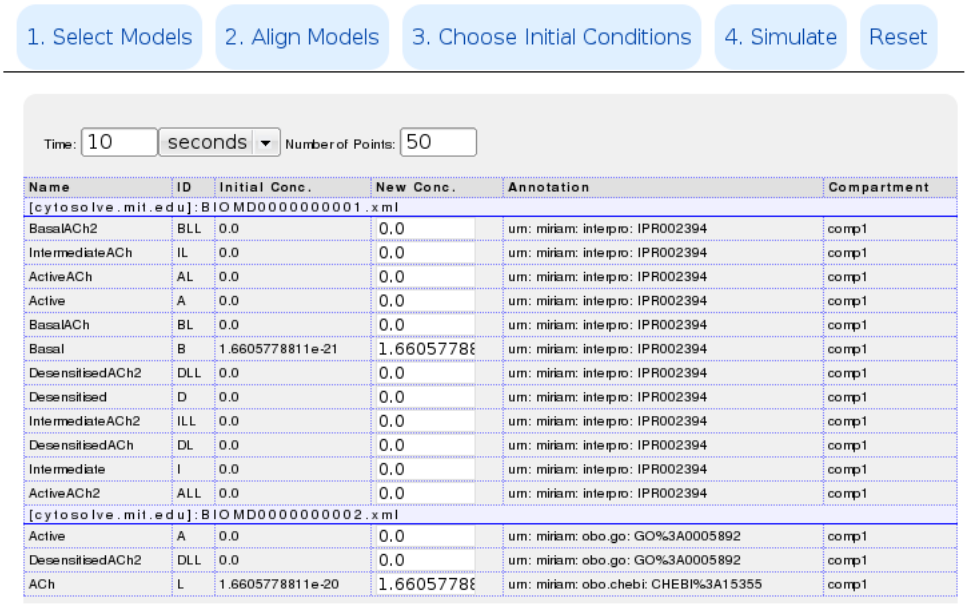


Figure 4.9: Cytosolve, step 3: OREMP points out the overlaps among the two models

- The list of equivalent reactions overall used
- All of the potentially equivalent N-order reaction compositions

With this method the observed edge/vertex ratio for the BioModels.net curated DB is 1.19, which is comparable to other biological pathway databases - HumanCyc DB [104] has a ratio of 1.01 and EcoCyc DB [105] one of 1.25 [106]. In this manner, the pathway building block dictionary obtained from the BioModels.net DB can be consulted to look up the alternative paths from one species-set vertex to another. A basic example of pathway building blocks extracted from the BioModels DB processing follows; this example includes only one species in each species-set. In the context of another EGFR model [107] (i.e.: MAP kinase cascade activated by surface and internalized EGF receptors), as detailed in biomodel 19 in [103], the system detected that the $EGF-EGFR\wedge 2-GAP-Shc$ species can directly become $EGF-EGFR\wedge 2-GAP-Shc*$ or, alternatively, the former can first become $EGF-EGFRi\wedge 2-GAP-Shc$, then $EGF-EGFRi\wedge 2-GAP-Shc*$, and finally $EGF-EGFR\wedge 2-GAP-Shc*$.



Name	ID	Initial Conc.	New Conc.	Annotation	Compartment
[cytosolve.mit.edu]:BIOMD0000000001.xml					
BasalACh2	BLL	0.0	0.0	um: miriam: interpro: IPR002394	comp1
IntermediateACh	IL	0.0	0.0	um: miriam: interpro: IPR002394	comp1
ActiveACh	AL	0.0	0.0	um: miriam: interpro: IPR002394	comp1
Active	A	0.0	0.0	um: miriam: interpro: IPR002394	comp1
BasalACh	BL	0.0	0.0	um: miriam: interpro: IPR002394	comp1
Basal	B	1.6605778811e-21	1.66057788	um: miriam: interpro: IPR002394	comp1
DesensitisedACh2	DLL	0.0	0.0	um: miriam: interpro: IPR002394	comp1
Desensitised	D	0.0	0.0	um: miriam: interpro: IPR002394	comp1
IntermediateACh2	ILL	0.0	0.0	um: miriam: interpro: IPR002394	comp1
DesensitisedACh	DL	0.0	0.0	um: miriam: interpro: IPR002394	comp1
Intermediate	I	0.0	0.0	um: miriam: interpro: IPR002394	comp1
ActiveACh2	ALL	0.0	0.0	um: miriam: interpro: IPR002394	comp1
[cytosolve.mit.edu]:BIOMD0000000002.xml					
Active	A	0.0	0.0	um: miriam: obo.go: GO%3A0005892	comp1
DesensitisedACh2	DLL	0.0	0.0	um: miriam: obo.go: GO%3A0005892	comp1
ACh	L	1.6605778811e-20	1.66057788	um: miriam: obo.chebi: CHEBI%3A15355	comp1

Figure 4.10: Cytosolve, step 4: the user silences reaction in conflict and re-uploads the model

This is just one and very simple example of our N-order analysis and the complete results about BioModels.net include a number of different species-sets in the order of 10^3. Another interesting usage example consists in asking to the system all of the possible pathways from two given species set: reading the ordered dictionary of pathway building blocks, this task can be easily achieved.

4.5 Ontologies From Pathways: Practical Advantages

From a logic point of view, the system is constructed of three layers. The bottom layer represents the original biochemical pathways, read in their primitive format (such as SBML and CellML). The second layer abstracts (through the work-flow 1-7 detailed above) the pathways into a minimalistic and quantitative meta-format (sketched in Table 4.1) that includes all the MIRIAM components. Annotations are preserved and extended with additional quantitative data to

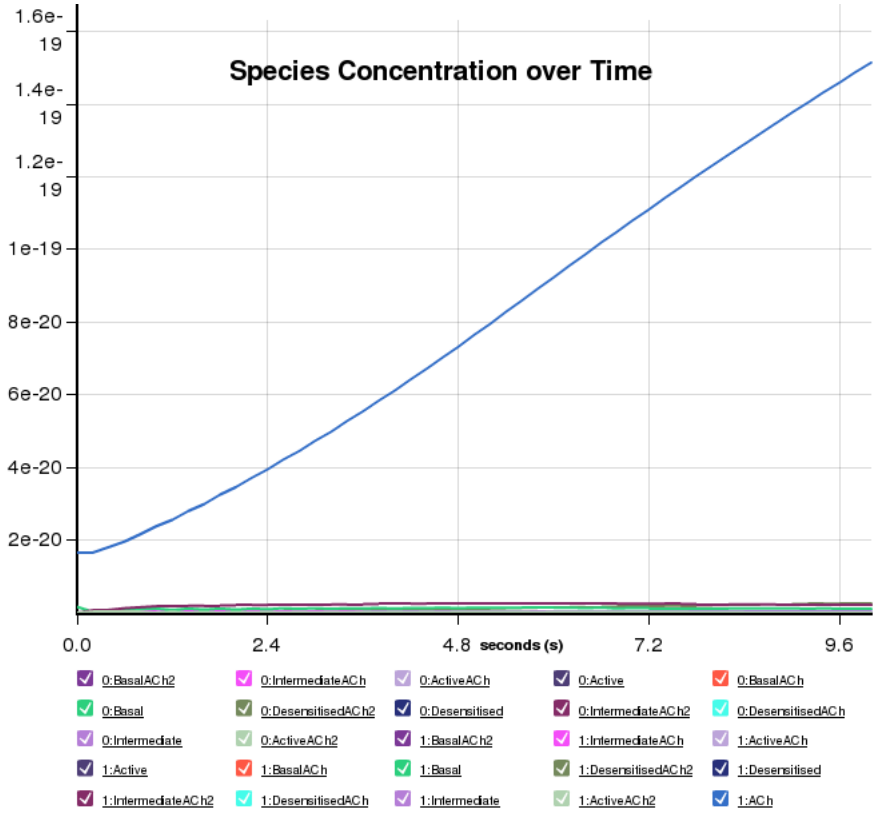


Figure 4.11: Cytosolve, step 5: the simulation takes place and the results are visualized

achieve a common description that can be represented as a single ontology. It is at this level that the extended ontology is primarily created. Entities and relations created in this manner are homogeneous in the ontological sense. This implies that several pathway collections can be combined in an ontology repository while maintaining a common semantic, meaning that the following advantages are achieved:

Sharing. Despite disparate initial data formats, the biochemical information described in each pathway is now homogeneously represented. This enables the direct reuse of componets (such as species or reactions) coming from different sources.

Integration. The system ensures a consistent merging of the resources, automatically aligning the species and showing the end-user possible duplications among reactions in the different pathways.

Knowledge discovery. Once the species alignment is done and duplicate reaction have been detected, the N-order reachability step is taken: for each reaction in each pathway the set of "alternative circuits" is computed. This means that given an arbitrary number of pathways, the system will identify all of the alternative ways to traverse from state S_0 to a state S_1 (where the states are different species configurations) within the overall set of reactions. In the last layer, all the information gathered is exported in OWL2. In the context of the Cytosolve@MIT project, I used the semantic tool, Protégé [82], to visually edit, compare, and finalize the biochemical information exported. Protégé query interface allows the user to formulate "semantically-enabled" queries that were impractical when dealing with previously heterogeneous, unaligned data repositories [108].

4.5.1 System Discussion

SemanticSBML [95] provides the state of the art tools to obtain a monolithic merged model starting from different molecular pathways. Where Cytosolve is concerned, one key component of its approach is the fact that it does not produce a monolithic model. This preserves the curation process of independent models and allows independent research laboratories to continue investigation and improvement of their own model without being forced to prematurely publish an authoritative merged resource; the independent curation process is preserved by maintaining the pathway identity, since the primitive element-pathway network is not destroyed by integration. Basically, this approach is different from SemanticSBML because it provides the user the opportunity to exploit his/her understanding to define a consistent method of knowledge integration across ontologies Another point of strength is the fact that once the system has read all of the 240 models from BioModels.net curated collection and the pathway building block dictionary is written (feeding step), the end-users can exploit this functionality to accelerate their research by taking advantage of other modelers efforts simply

by consulting this dictionary. By specifying the initial and ending set of species, modelers can use the building block dictionary to gain ideas about how other people investigated and modeled a similar problem and how cross-pathway reactions could be composed to fit their needs. The experiment detailed in section 4.4.3 provided an interesting overview of the BioModels.net collection that brought also the following achievement: from the prospective of those who curate collections of biochemical pathways, this framework can be used to find inconsistencies and redundancies within their repository since the system highlights common bricks shared among multiple models.

4.6 Conclusions

This is the first time that the information coming from different biological data sources has been aggregated into a single quantitative ontology. There, thanks to the design of the meta-format detailed, both combining operations and detail-revealing ones are allowed: OREMP application can combine several pathways, merge and combine pathway repositories, or revert to the original pathways, and inspect single-model details and query external repositories (such as UniProt and GO) referenced in pathway element annotations. The system is independent of the different file formats in which the pathways are written and contains an extensible collection of parser modules. I have selected OWL2 as export format for the extended ontologies and have adopted Protégé as default "Data Warehouse" for information storage, retrieval and reasoning. Despite its early stage, the system has been successfully employed in challenging field applications. One of the extensions to this work is an in-depth analysis on the additional constraints that conservation of mass requirements imparts to the model-merging and duplication detection problem. Secondly, Dewey Lab is now investigating ways to visualize the complex ways in which duplicate reaction paths can exist between multiple models [98].

Chapter 5

Conclusions

From an algorithmic point of view, my research brought three new algorithms: AMMISCA [109], PAO [110] and PMO2 [111]. The first one has been adopted on the tuning of an established Cellular Automata model (SCIARA [20; 29]) for the forecasting of lava flow paths. With respect to the original genetic algorithm [28], AMMISCA has given rise to the most precise lava simulation, it is interesting to note that the algorithm achieves the best solution (in terms of fitness and required time) without a standard crossover phase as defined by Holland. Additionally, it has been proved that new areas of the search space have been inspected by the new algorithm. The algorithm PAO has been designed for the optimization of a photosynthesis model: in this context the algorithm outperformed those algorithm actually considered "the state of the art" in *general purpose* optimization. Strength points of this algorithm are its distributed approach based on islands and its capability of wrapping other algorithm; this approach has exploited the concept of *migration* to combine solution building-blocks coming from different optimization niches (islands). The algorithm PMO2 algorithm has been applied to the Geobacter sulfurreducens in order to stress its capabilities in a highly-dimensional problem ($\mathbb{R}^{608}$); with respect to that I have obtained a computational model that maximizes the electron and biomass productions while preserving those bounds that ensures a biological significance. To my knowledge this is the first time that Geobacter sulfurreducens is modeled as a multi-objective optimization problem where the search moves automatically towards steady state solutions, contextually with biological boundaries observance

and functional optimization (i.e.: biomass and electron productions). Optimized configurations of the Geobacter here obtained are currently under consideration for "in vitro" and "in vivo" implementations. In fact, bioengineering a mutant strain in order to reach faster rates in electron transport yield is highly desirable and could represent a breakthrough for massive application in biotech industry.

The study of photosynthesis has been another main chapter in my research. With respect to that the designed methodology I have obtained an increase in photosynthetic productivity of the 135% from 15.486 $\mu mol\ m^{-2}s^{-1}$ (i.e., value measured in standard natural leaves) to 36.382 $\mu mol\ m^{-2}s^{-1}$, and improving the previous best-found photosynthetic productivity value [42] (27.261 $\mu mol\ m^{-2}s^{-1}$, 76% of enhancement). Optimized enzyme concentrations express a maximal local robustness (100%) and a high global robustness (97.2%), satisfactory properties for a possible "in vitro" manufacturing of the optimized pathway. Morris sensitivity analysis shows that 11 enzymes out of 23 are high sensitive enzymes, i.e., the most influential enzymes of the carbon metabolism model. Successively, I have studied the C_3 carbon metabolism as a trade-off between the maximization of the leaf CO_2 uptake rate and the minimization of the total protein-nitrogen concentration. This trade-off search has been carried out in six environmental scenarios: three c_i concentrations (referring to the estimate of CO_2 concentration in the atmosphere characteristic of 25 million years ago, nowadays and in 2100 a.C.) and two triose-P (PGA, GAP, and DHAP): low and high export rates. Additionally, CO_2 uptake and *nitrogen consumption* are evaluated with respect to the *robustness* by means of a 3D Pareto-surface. Remarkably, the six Pareto frontiers identify the highest photosynthetic productivity rates together with the fewest protein-nitrogen usage. Those leaf designs obtained in this study are currently under consideration for an "in vitro implementation" that would give rise to a CO_2 avid plant strain. The analysis of the results has shown that is possible to obtain a gain of the uptake rate while minimizing the amount of nitrogen required; the yield analysis has shown a clear propensity of remaining in a robust state of the great majority of solutions. The preliminary biological analysis of the proposed solutions provides interesting insights regarding the interactions and the behavior of the computationally designed leaves; in particular, some biological hypothesis can be inferred from the obtained results that should be linked

with the extended process of photosynthesis. The increasing CO_2 concentration requires biotechnological approaches to be tackled effectively; the aim of this part of my research has been to redesign the natural tools such that the evolutionary process can be speed up to tackle the environmental problems. Moreover, an efficient and robust plant can be considered as an innovative source of green energy, through its expected increasing of energy production due to the augmented ability of up-taking CO_2. Plants designed with the methodology I presented could truly improve life conditions Earth-wise: up-taking more and more CO_2 means counteracting what the industrial revolution brought in terms of negative consequences in the last 40 years. Indeed, in this time-span the CO_2 level moved from 280 to 380 parts per million; evidence is mounting that carbon dioxide's heat-trapping power has already started to boost average global temperatures. If carbon dioxide levels continue upward, further warming could have dire consequences, resulting from rising sea levels, agriculture disruptions, and stronger storms (e.g. hurricanes) striking more often. Say "stop burning fossil fuels" is not reasonable, because they represent 85% world's energy; then sequestration of the CO_2 in areas away from the atmosphere seems the only option, and in this context, natural tools (i.e., bioengineered plants) that can be distributed world-wide and can accomplish the uptake task efficiently can be considered a valid solution. For these reasons, this research seems fundamental and with respect to that, several competences are needed: in addition to biologists, several skills are required and, as of today, bioengineers are computer scientists are likely to be necessary as well to tackle the problem efficiently.

Another track in my research has been the semantic integration of information coming from biomedical sources. Different laboratories, distributed world-wide, are continuously testing and experimenting (in "wet laboratories") new molecules of interest for human beings or that are simply interesting from a biological point of view. Pharmaceutical companies and research groups are the main characters in this tale. It is worth noting that, just at the MIT, there are more than four research groups interested in different aspects of the same molecule: the Epidermal Growth Factor. It is obvious that different groups want to share their information without telling "too much" to general competitors. This situation is much more complicated when we move into the pharmaceutical business: com-

putational models used in the drug-design are considered extremely important. In that context a single experiment at workbench has a cost in the order of the thousands of dollars; considering these costs and the fact that each drug has to be tested against the largest set of compounds to avoid negative effects, it is obvious that whatever computational tool that can avoid an experiment and is equally reliable, can represent a significant improvement in this context. For this reason the modeling of biochemical pathways as sets of linked ODEs has been a key trend in past years. Nowadays, biochemical pathways are truly available for everyone, but there is still the lack of tools to integrate the knowledge they represent. In this context the MIT started the Cytosolve project: by means of a web-application, different researchers can connect world-wide to the same website and perform a combined simulation where all their systems of ODEs are combined and solved together. The "combining" step is transparent to the user and does not rely on the creation of a monolithic model. This means that each researcher holds his/her model identity and shares with others just "interfaces" to get a combined simulation. Having several computational pathways to combine, an important part of Cytosolve is the step in which the overlaps among models are identified. In this context I developed the OREMP library [96; 97; 98], that takes care of the model alignment task to ensure coherent distributed simulation. Additionally, in OREMP, the information coming from different biological data sources is aggregated into single quantitative ontologies. There, thanks to a specific meta-format designed, both combining operations and detail-revealing ones are allowed. OREMP application can combine several pathways, merge and combine pathway repositories, or revert to the original pathways, and inspect single-model details and query external repositories (such as UniProt[84] and GO[83]) referenced in pathway element annotations. The system is independent of the different file formats in which the pathways are written and contains an extensible collection of parser modules. OWL2 has been chosen as export format for the extended ontologies and Protégé has been adopted as default "Data Warehouse" for information storage, retrieval and reasoning. A main application of OREMP can be the data-integration and data-retrieval in the biomedical area. Feeding the system with new models published on a daily base, it is possible to build aggregated ontologies in an iterative and incremental fashion. Adopting

the system in a pharmaceutical firm would bring the integration of the knowledge coming from public repositories into the proprietary information backbone owned by the company. From the researcher/modeler point of view, OREMP could provide a semantic-aware environment in which the information would be obtained from public repositories and aggregated into structured knowledge. This means that a catalog of virtual pathway building-blocks would be available. This would significantly accelerate research in the biomedical field by boosting the modeling task and supporting knowledge sharing in this domain.

On a separate track, I studied the modeling of highway traffic in terms of Cellular Automata. The model STRATUNA has been partially re-implemented, paying attention to the coupling vehicle/driver. Main contribute on this subject is the fact that the new model STRATUNA-β4 [112; 113] gave rise to traffic forecastings whose precision varies from 88% to 99% when tested on data provided by ANAS Spa about the Italian highway A4. Additionally, I integrated a cost system that takes as input the output produced by the model. This framework connects then different highway designs to different congestion toll charges through an established cost system.

Appendix A: Artificial Photosynthesis

A.1 Modeling, Supplementary Information

The computational simulation of the carbon metabolism requires the definition of a set of linked ODEs; in my work, it is considered the model proposed by [42]. 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. 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. A general reversible reaction of the form $A + B \leftrightarrow C + D$ has been associated with the following rate equation:

$$v = V_m[A][B] - \frac{[C][D]}{k_e} M^{-1} \tag{1}$$

where M is defined as follows:

$$
\begin{aligned}
M &= K_{mA}K_{mB} \times && (2)\\
&\times \left(1 + \frac{[A]}{KmA} + \frac{[B]}{K_{mB}} + \frac{[C]}{K_{mC}} + \frac{[D]}{K_{mD}} + \right. && (3)\\
&+ \left. \frac{[A][B]}{K_{mA}K_{mB}} + \frac{[C][D]}{K_{mC}K_{mD}} \right) && (4)
\end{aligned}
$$

following the standard kinetic equation for a reversible reaction with two substrates and two products, where $[A], [B], [C], [D]$ represent the metabolite concentrations and $K_{mA}, K_{mB}, K_{mC}, K_{mD}$ are the Michaelis-Menten constants for the metabolites A, B, C, D, while k_e is the equilibrium constant of this reaction and V_m the maximum rate of reaction. For a general non-reversible reaction $A + B \rightarrow C + D$, the generalized rate equation was:

$$v = V_m \frac{[A][B]}{([A] + K_{mA])([B]+K_{mB})} \tag{5}$$

Contrariwise, the presence of a competitive inhibitor (E) changes the apparent Michaelis-Menten constant of the corresponding substrate; in this case, a non-reversible reaction $A + B \rightarrow C + D$ has the following reaction rate:

$$v = V_m \frac{[A][B]}{\left([A] + K_{mA}\left(1 + \frac{[E]}{K_l}\right)\right)([B] + K_{mB})} \tag{6}$$

where K_l is the inhibition constant. These generic equations were used to describe the enzyme catalyzed steps of the Calvin cycle, starch synthesis, triose-P export, Suc synthesis and the PCOP. For the Rubisco enzyme, a different equation has been adopted to correlate the rate of carboxylation and oxygenation to total Rubisco concentration (R_t). The solution of this equation can be approximated to:

$$v_c = W_c \min\{1, [R_t]/[E_t]\} \tag{7}$$

where W_c is calculated as:

$$W_c = \frac{V_{Cmax}|CO_2|}{|CO_2| + K_{M11}\left(1 + \frac{|O_2|}{K_{M12}}\right)} \tag{8}$$

where V_{Cmax} represents the maximum rate of Rubisco carboxylation, K_{M11} the Michaelis-Menten constant of CO_2 and K_{M12} the Michaelis-Menten constant estimating O_2. The model uses a large number of constants and parameters and no consistent set of them are available for any specie of plant; in order to face this problem, these parameters were picked from literature. The model assumed that the total protein-nitrogen in the enzymes is 1 $g\ m^{-2}$; the mass nitrogen in each enzyme, in a 1 m^2 leaf area, was computed based on the number of active sites, catalytic rate per active site, molecular mass of each enzyme, and the ratios between V_m of different enzymes. Mole of each protein is then calculated based on the molecular mass and the mass of each protein, i.e., the total concentration of the adenylate nucleotides ([CA]) in the chloroplast stroma (that is, the sum of [ATP] and [ADP]) was assumed to remain constant. The V_m for each enzyme was then calculated based on the amount of each enzyme and the volume of the compartment that it occupies in 1 m^2 leaf area. The total concentration of the adenylate nucleotides ([CA]) in the chloroplast stroma, the sum of [ATP] and [ADP], was assumed to remain constant. Similarly, the sum of [NADPH] and [NADP] in the chloroplast stroma ([CN]) was assumed constant. The export of PGA, GAP or DHAP from the chloroplast to the cytosol is associated with a counterimport of the phosphate, mediated by a phosphate translocator. Consequently, the total concentration of phosphate in the stroma ([CP]) is assumed constant. Finally, a set of ODEs encodes the rates of changes in concentration of the metabolite, that is represented by the difference between the rates of reactions generating the metabolites and the rates of the reactions consuming the metabolites. It is clear that the volume of the chloroplast stroma can be different from the cytosol one in a typical higher plant cell; in this scenario, it has been assumed a $1:1$ ratio in calculating the concentrations of the two compartments.

A.1.1 Enzyme nomenclature reference

Here is reported the complete reference to each of the enzyme used, together with abbreviations and unique *EC* number.

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

Table 1: Enzyme abbreviations [114] used in the text or used in the tables or figures are here listed.

A.1.2 Alternative leaves

Many other leaf designs have been studied in addition to those detailed above; here I report more about alernative solutions. Fig. 1 reports the changes in the concentrations of Carbon-metabolism enzymes with respect to their natural values when three alternative strategic leaf designs are considered. *Maximal* CO_2 *Uptake* (Top plot), *Minimal Nitrogen Consumption* (Middle plot), and *Closest-to-ideal solution* (Bottom plot). The maximal rate of triose-P (PGA, GAP, and DHAP) export is kept fixed to the value of 1 $mmol\ L^{-1}\ s^{-1}$ and the C_i has value 270 $\mu mol\ mol^{-1}$ to reflect nowadays condition.

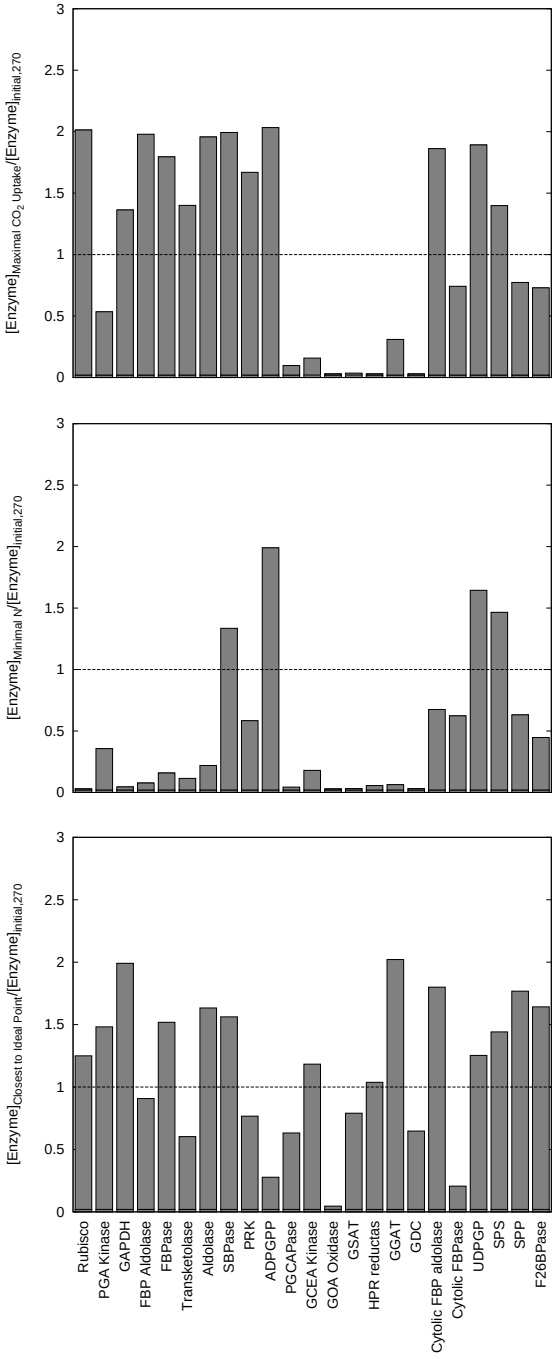


Figure 1: Alternative leaf designs.

APPENDIX A: ARTIFICIAL PHOTOSYNTHESIS

In Fig. 2 are reported those leaves obtained when the optimization is carried out in an alternative scenario: maximal rate of triose-P (PGA, GAP, and DHAP) is 3 $mmol\ L^{-1}\ s^{-1}$. Top plot shows the comparison among optimized enzyme concentrations at a $C_i = 270\ \mu mol\ mol^{-1}$ (i.e., nowadays concentration of CO_2 in the atmosphere) and the natural leaf. Middle plot reports, enzyme-wise, the changes among the leaf optimized for 2100 a.C. environment ($C_i = 490\ \mu mol\ mol^{-1}$) and the one optimized for nowadays conditions. Instead of future, bottom plot reports the w.r.t. the leaf design optimized for $C_i = 165\ \mu mol\ mol^{-1}$ (i.e., concentration estimated to be in place 25M years ago)

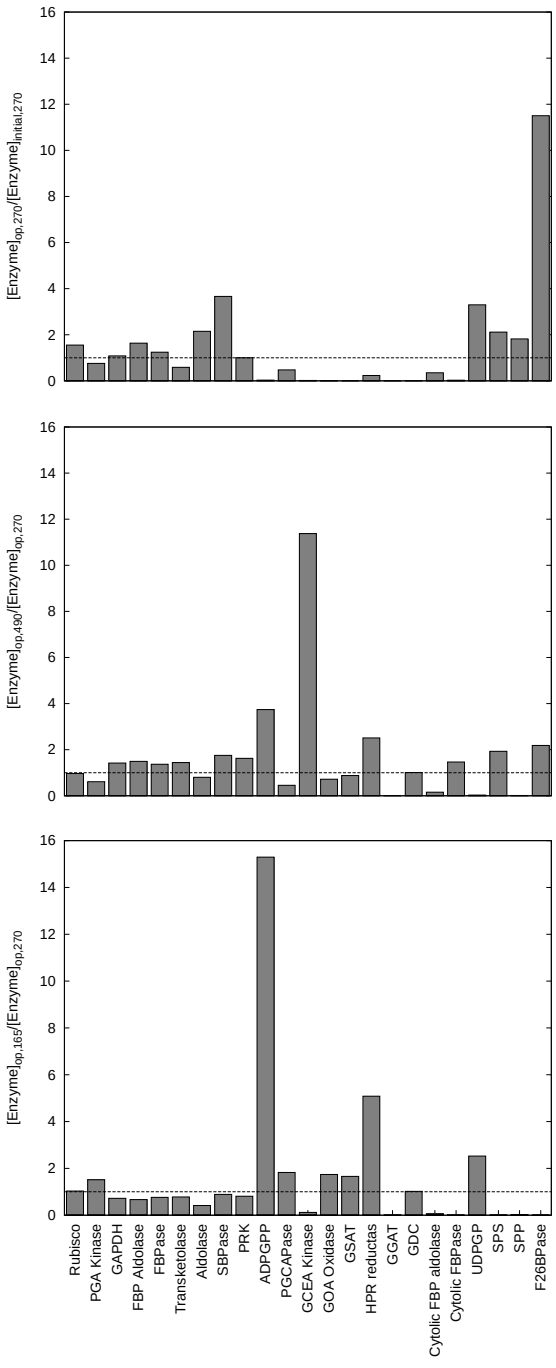


Figure 2: Change in optimized enzyme concentrations with respect to different atmospheric CO_2. Maximal rate of triose-P (PGA, GAP, and DHAP) is 3 $mmol\ L^{-1}\ s^{-1}$.

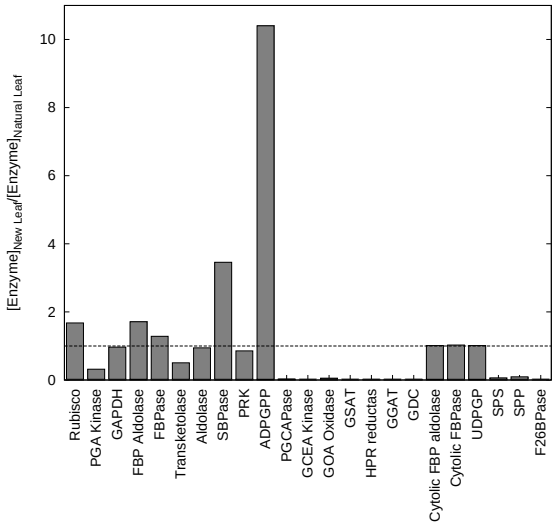


Figure 3: Optimized leaf when Cytosolic FBP aldolase, Cytosolic FBPase and UDPGP are kept at their natural value.

Fig. 3 reports the changes in the concentrations of Carbon-metabolism enzymes with respect to their natural values when three metabolites are kept constant: Cytosolic FBP aldolase, Cytosolic FBPase, UDPGP. The maximal rate of triose-P (PGA, GAP, and DHAP) export is kept fixed to 1 $mmol\ L^{-1}\ s^{-1}$ and the C_i has value 270 $\mu mol\ mol^{-1}$, reflecting nowadays condition

APPENDIX A: ARTIFICIAL PHOTOSYNTHESIS

Fig. 4 reports those leaves optimized for the environment in place 25M years ago: Minimal Nitrogen Consumption (Top plot) and Maximal CO_2 Uptake (Bottom plot) are compared to the natural leaf.

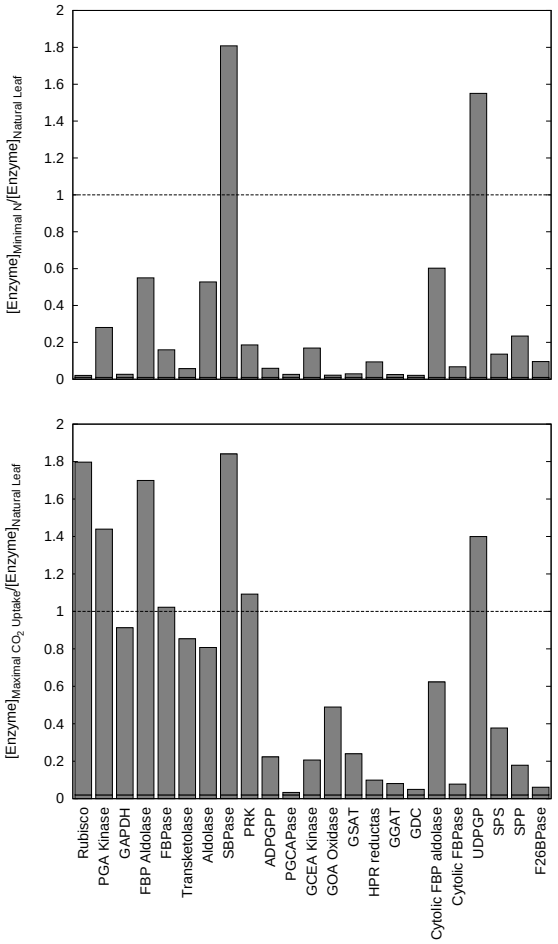


Figure 4: Alternative leaves obtained when the maximal rate of triose-P (PGA, GAP, and DHAP) export is kept fixed to the value of 1 $mmol\ L^{-1}\ s^{-1}$ and the C_i has value 165 $\mu mol\ mol^{-1}$ to reflect 25M years ago environment.

APPENDIX A: ARTIFICIAL PHOTOSYNTHESIS

Fig. 5 reports those leaves optimized for the environment predicted for the end of the century: the figure reports changes in the concentrations of Carbon-metabolism enzymes with respect to their natural values when two alternative strategic leaf designs are considered: Minimal Nitrogen Consumption (Top plot) and Maximal CO_2 Uptake (Bottom plot). The maximal rate of triose-P (PGA, GAP, and DHAP) export is kept fixed to the value of 1 $mmol\ L^{-1}\ s^{-1}$ and the C_i has value 490 $\mu mol\ mol^{-1}$.

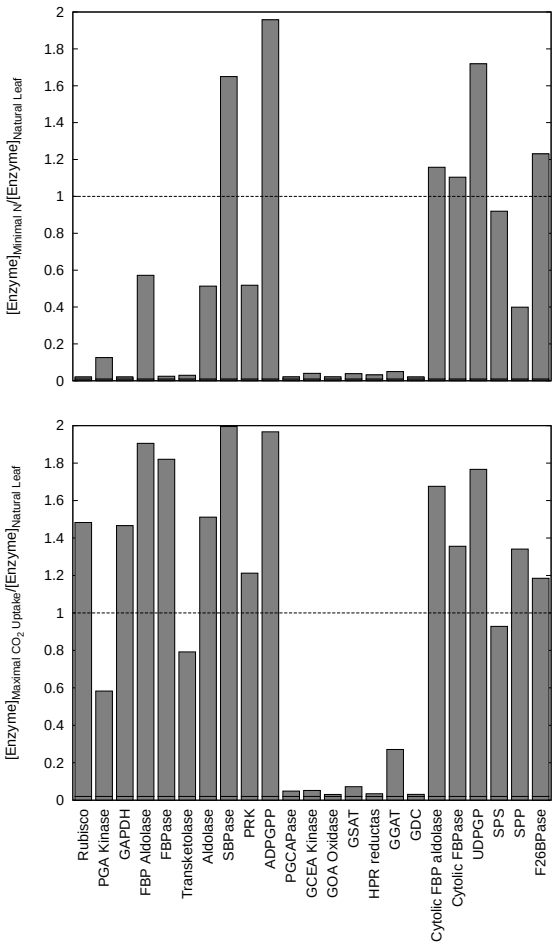


Figure 5: Alternative leaves obtained when the maximal rate of triose-P (PGA, GAP, and DHAP) export is kept fixed to the value of 1 $mmol\ L^{-1}\ s^{-1}$ and the C_i has value 490 $\mu mol\ mol^{-1}$ to reflect the 2100 a.C. environment.

Coming figures present how the system behaves when: only six enzymes are

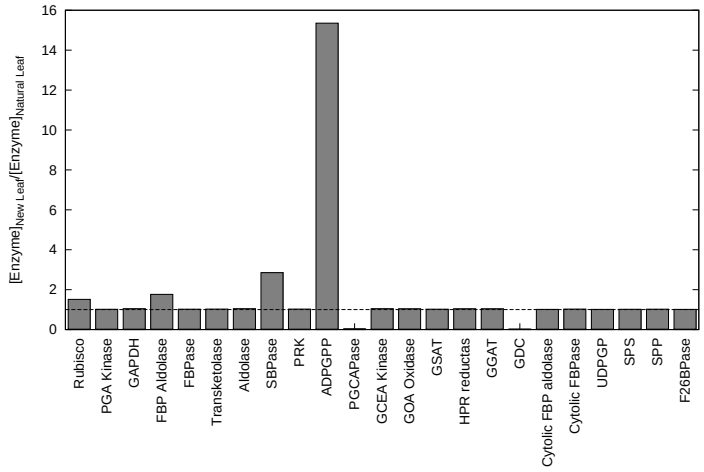


Figure 6: Optimization of CO_2 uptake rate perturbing 6 enzymes only (Rubisco, FBP aldolase, SBPase, ADPGPP, Phosphoglycolate phos., and GDC) while the remaining 19 enzymes are maintaining to their initial concentrations. For the 6 enzymes we defined the following constraint: the concentration must be $\geq$ 0.02 $mg\ N\ m^{-1}$. *Rubisco, FBP aldolase, SBPase, ADPGPP are overexpressed*, while *Phosphoglycolate phos., and GDC are quasi switched off*. This configuration obtains CO_2 uptake rate of 32.89 μ mol $m^{-2}\ s^{-1}$, it wastes about 3.492 μ mol $m^{-2}\ s^{-1}$ of CO_2 uptake rate but it uses *only 6 enzymes*.

varied from their natural concentration (Fig. 6), the Rubisco is kept fixed (Fig. 7), or only six enzymes are varied and one of them - Rubisco - can change with bounds of $\pm 15\%$ (Fig. 8).

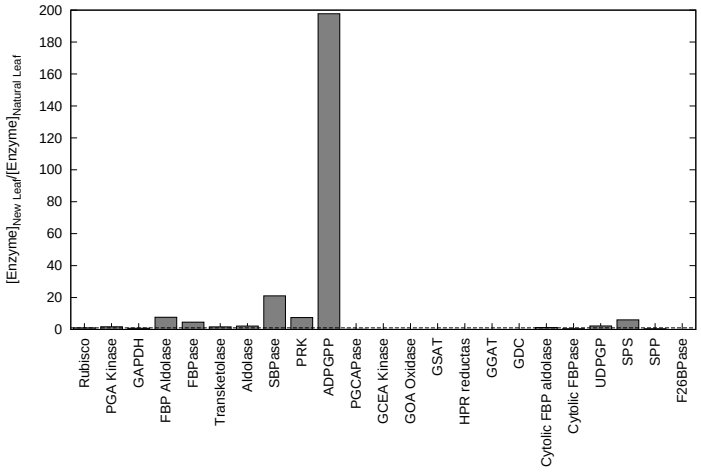


Figure 7: Optimization of CO_2 uptake rate perturbing 24 enzymes while the Rubisco is maintaining to its initial concentration. This configuration obtains CO_2 uptake rate of 22.26 μ mol m^{-2} s^{-1}. This leaf points out the centrality of Rubisco in the optimization process.

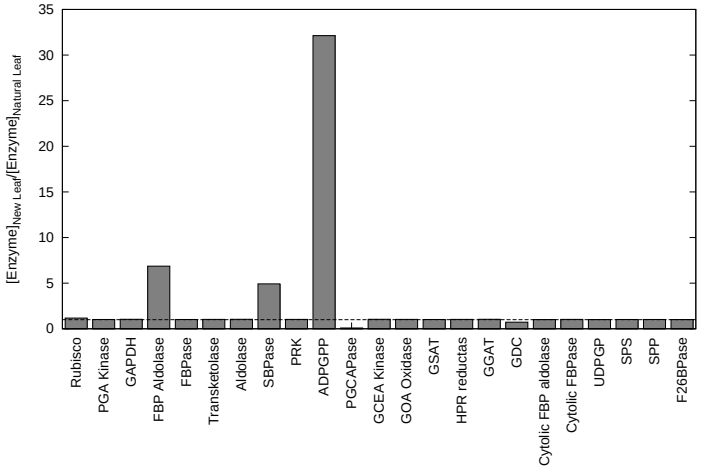


Figure 8: Optimization of CO_2 uptake rate perturbing 6 enzymes only (Rubisco, FBP aldolase, SBPase, ADPGPP, Phosphoglycolate phos., and GDC) while the remaining 19 enzymes are maintaining to their initial concentrations. In this optimization the Rubisco is allowed to increase up to 15%; this constraint has been inserted in order to have more feasible biotechnological results. *FBP aldolase, SBPase, and ADPGPP are overexpressed*, while *Phosphoglycolate phos.* is switched off and *GDC* is close to its initial value. This configuration obtains CO_2 uptake rate of 25.246 μ mol m^{-2} s^{-1}.

Appendix B: Highway Traffic

B.1 A Cellular Automata model for highway traffic simulations

Contextually with bioinformatics and bioengineering topics, I explored more engineering problems as well. Highway traffic is one of them: its evolution is regulated by parallel and acentric interactions among vehicles. In this Appendix is reported STRATUNA, a model for highway traffic forecasting, together with a cost system, directly fed by simulation data.

Cellular Automata are an established formal support for modelling traffic. STRATUNA is a Cellular Automata model for simulating two/three lanes highway traffic. It is based on an extensive specification of the driver response to the surrounding conditions. The model is deterministic with regard to driver behavior, even if values of parameters ruling the reactivity level of the drivers are assigned stochastically. Probability distribution functions were deduced by field data and applied to vehicular flow generation (vehicle types, driver desired speed, entrance-exit gates). A partial implementation of STRATUNA has been performed and applied to Italian highway A4 from Venice to Trieste. Simulations have been compared with available field data with results that may be considered positive. Fair results in flow forecasting lead to the implementation of an established cost system in which simulation directly provides cost forecasting in terms of congestion toll.

B.2 Introduction

Cellular Automata (CA) are a computational paradigm for modelling high complexity systems [115] which evolve mostly according to the local interactions of their constituent parts (acentrism property). Intuitively a CA can be seen as a d-dimensional space, partitioned into cells of uniform size, each embedding a computational device, the elementary automaton (EA), whose output corresponds to its state. Input for each EA is given by states of EA in neighboring cells, where neighboring conditions are determined by a pattern invariant in the time and equal for each cell. EA are in an arbitrary state at first (initial conditions), subsequently CA evolves by changing simultaneously states to all of the EA at equal discrete time steps, according to the EA transition function (parallelism property).

CA were used for modelling highway traffic [116] because of acentric and parallel characteristics of such a phenomenon. As a matter of fact, when highway structural features are fixed and there are no external interferences out of the vehicular interactions (normal conditions), the traffic evolution emerges by the mutual influences among vehicles in driver sight range.

The main CA models of highway traffic [117; 118; 119; 120] may be considered "simple" in terms of external stimuli to the driver and corresponding reactions, but they are able to reproduce the basic three different phases of traffic flow (i.e., free flow, wide moving jams and synchronized flow) by simulations to be compared with data (usually collected automatically by stationary inductive loops on highways).

STRATUNA (Simulation of highway TRAffic TUNed-up by cellular Automata), is a new CA model for highway traffic with the aim of describing more accurately driver surrounding conditions and responses. I referred to a previous CA model [121; 122], that was enough satisfying in the past, but now it is dated for the different technological situations (e.g., the classification of vehicles on the base of pure acceleration, deceleration features is no more realistic). Reference data for deducing STRATUNA parameters and for real-simulated event comparison are the timed highway entrance-exit data, that are comprehensive of the vehicle type.

Next section outlines the STRATUNA model, while the transition function is described in the third section. Implementation of the model is discussed together with simulation results and comparison with real event in the fourth section. The cost system is detailed in fifth section. Conclusions are reported at the end of this appendix.

B.3 The STRATUNA general model

STRATUNA is based on a "macroscopic" extension of CA definition [115], involving "substates" and "external influences". The set of "state values" is specified by the Cartesian product of sets of "substate values". Each substate represents a cell feature and, in turn, a substate could be specified by sub-substates and so on. Vehicular flows at tollgates and weather conditions are external influences, generated by dataset or probabilistic functions according to field data and are applied before the CA transition function.

Only one-way highway traffic is modelled by STRATUNA (complete highway is obtained by a trivial duplication). One-dimension is sufficient, because a cell is a highway segment, 5m long, whose specifications (substates) encloses width, slope and curvature in addition to features of possible pairs vehicle-driver. The STRATUNA time step, the driver minimum reaction time, may range from 0.5s to 1s (CA clock).

An 8-tuple defines $STRATUNA = \langle R, E, X, P, S, \mu, \gamma, \tau \rangle$, where:

- $R = \{x | x \in \mathbb{N}, 1 \leq x \leq n\}$ is the set of n cells, forming the highway.
- $E \subset R$ is the set of entrance-exit cells in R, where vehicles are generated and annihilated.
- $X = \langle -b, -b+1, ..., 0, 1, ..., f \rangle$ defines the EA neighboring, i.e the forward (f) cells and backward (b) cells in the driver sight, when visibility is maximum (no cloud, sunlight etc.).
- $P = \{length, width, clock, lanes\}$ is the set of global parameters, where $length$ is the cell length, $width$ is the cell width, $clock$ is the CA clock,

lanes is the number of highway lanes (1, 2 .. from right to left), that includes an additional lane 0, representing from time to time the entrance, exit, emergency lane.

- $S = Static \times Dynamic \times (Vehicle \times Driver)^{lanes}$ specifies the high level EA substates, that are clustered in typologies, i.e statical and dynamical features of highway segment corresponding to the cell, vehicle and driver features (there are at most as many pairs vehicle-driver as lanes). Such substates are detailed in the Table 2.

- $\mu : \mathbb{N} \times R \rightarrow Dynamic$ is the "weather evolution" function, that determines *Dynamic* values for each step $s \in \mathbb{N}$ and each cell $c \in R$.

- $\gamma : \mathbb{N} \times E \rightarrow Vehicle \times Driver$ is the vehicle-driver pair *normal* generation function for each step $s \in \mathbb{N}$ and each cell $c \in E$.

- $\tau : S^{b+1+f} \rightarrow S$ is the EA transition function. The visibility reduction to b' backward cells and to f' forward cells involves that cells out of range will be considered without information.

Substate	Sub-substates hierarchy
Static	*CellNO, Slope, CurvatureRadius,*
	SurfaceType, SpeedLimit, Lane1SpeedLimit
Dynamic	*BackwardVisibility, ForwardVisibility, Temperature,*
	SurfaceWetness, WindDirection, WindSpeed
Vehicle	*Type, Length, MaxSpeed, MaxAccelerat., MaxDecelerat. ;*
	CurrentSpeed,CurrentAcceleration, Xposition, Yposition,
	Indicator, StopLights, WarningSignal
Driver	*Origin, Destination, DesiredSpeed,*
	PerceptionLevel, Reactivity, Aggressiveness

Table 2: Substates and related sub-substates.

B.4 The STRATUNA transition function

An overview of the transition function will be here given with the aim of exposing the leading ideas and the adopted choices concerning STRATUNA, together with a better specification of the mentioned substates and sub-substates.

A vehicle is specified by constant and variable (during all the simulation) values of sub-substates. Constant properties are *Type* (motorcycle, car, bus / lorries / vans, semitrailers / articulated), *Length*, $MaxSpeed$, $MaxAcceleration$, $MaxDeceleration$. The main mechanism of the traffic evolution is related to the determination of the new values of the variable sub-substates of $Vehicle$, i.e., $Xposition$ and $Yposition$ (they individuate the cell co-ordinates x, y of the middle point in the vehicle front) $CurrentSpeed$, $CurrentAcceleration$, *Indicator* (with values: null, left, right, hazard lights) *StopLights* (on, off), $WarningSignal$ (on, off).

Note that the vehicle space location is not identified by a sequence of full cells as in other CA models [116] , but it is more accurate because portions of cell and positions between two lanes can be considered occupied. *Indicator* and $WarningSignal$ sub-substates in the simulation hold a larger role than indicator and a generic warning signal in the real events. When a real driver wants to change lane, not always he uses the indicator, but drivers around detect such a manoeuvre from his behavior (e.g., a short beginning moving toward the new lane before to decide overtaking). Of course simulation doesn't account for these particular situations, but this problem doesn't exist, a driver in the simulation communicates his intention to change lane always by the indicator. Sub-substate $WarningSignal$ is activated when driver wants to signal that he needs the lane immediately ahead of his vehicle to be free. This situation corresponds in the real word to different actions or their combination, e.g., sounding the horn, blinking high-beam lights, reducing "roughly" the distance with vehicle ahead and so on. Through such sub-substates, *Indicator*, *StopLights*, $WarningSignal$ a communication protocol could be started between vehicles.

The single vehicle V moving involves two computations, i.e., the objective determination of the future positions of vehicles "around V" and the subjective V driver reaction. The former one is related to the objective situation and

forecasts all the spectrum of possible motions of all the vehicles, that can potentially interact with V, i.e., the vehicles in the same cells, where V extends more the next vehicles ahead and behind such cells for each lane in the range of the neighborhood.

In first instance, some *Static* and *Dynamic* sub-substates determine highway conditions (e.g., highway $surface\_slipperiness$ is computed by $SurfaceType$, $SurfaceWetness$ and $Temperature$); subsequently, they are related to the $Vehicle$ sub-substates in order to determine the temporary variable max_speed that guarantees security with reference only to the conditions of highway segment represented by cell. It accounts for the vehicle stability, speed reduction by limited visibility and speed limits in the lane, occupied by the vehicle. If max_speed is smaller than $DesiredSpeed$, $desired\_speed = max\_speed$ otherwise $desired\_speed = DesiredSpeed$. $Slope$ and $surface\_slipperiness$ determine the temporary variables $max\_acceleration$ and $max\_deceleration$, correction to sub-substates $Max-Acceleration$ and $MaxDeceleration$.

The next computation step determines "objectively" the "free zones" for V, i.e. all the zones in the different lanes, that cannot be occupied by the vehicles around V, considering the range of the speed potential variations and the lane change possibility, that is always signalled by $Indicator$. Note that the possible deceleration is computed on the value of $max\_deceleration$ in the case of active $StopLights$, otherwise a smaller value is considered, because deceleration could be only obtained by shift into a lower gear or by relaxing the accelerator.

The last computation step involves the driver subjectivity. First of all, the cell number corresponding to vehicle position $CellNO$ is compared with the cell number of $Destination$ in order to evaluate if the exit is so close to force approaching lane 1 (if in other lanes) or continuing in lane 1 slowing down opportunely to the ramp speed limit.

The driver aims in the other cases to reach/maintain the $desired\_speed$; different options are perceived available, each one is constituted by actions (Fig. 9) involving costs (e.g. the cost of the gap between the new value of $CurrentSpeed$ and the $desired\_speed$). The driver chooses the option, among all the possible ones, with minimal sum of the costs.

All is based on a driver subjective perception and evaluation of an objec-

tive situation by sub-substates *PerceptionLevel*, *Reactivity*, *Aggressiveness*. *PerceptionLevel* concerns the perception of the free zones; their widths are reduced or (a little bit) increased by a percentage before to compute on their new values the various possibilities to reach free zones in security conditions, considering the variable values of $Vehicle$ sub-substates more max_speed, $max\_acceleration$ and $max\_deceleration$.

Reactivity is a collection of constants for determining costs by means of function of the same type expressed in Fig. 9. Examples are "remaining in a takeover lane", "staying far from *desired_speed*", "breaking significantly", "starting a takeover protocol".

Aggressiveness forces the deadlocks, that could be generated by a cautious *PerceptionLevel*, e.g. when the entrance manoeuvre is prohibited in a busy highway, because free zones are very much reduced in the perception phase. The stop condition increases at each step the *Aggressiveness* value, it implies a proportional increase of the percentage value of *PerceptionLevel* from negative values to positive ones until the free zone in a lane remains shorter than the distance between two consecutive vehicles, where the entrance could be performed. *Aggressiveness* value comes back to zero when stop condition ends.

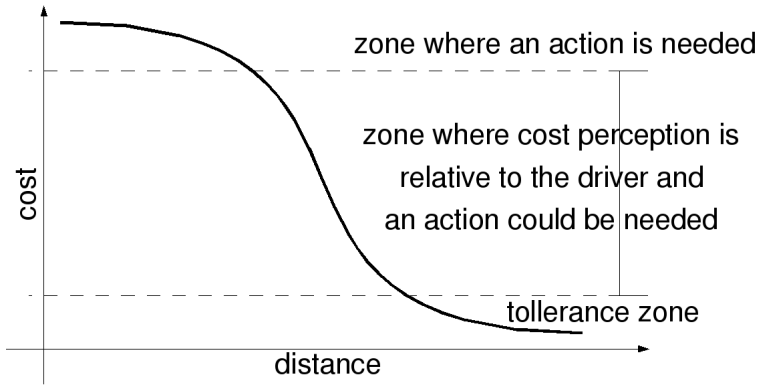


Figure 9: The function that connects the distance from front vehicle with a cost.

B.5 STRATUNA implementation

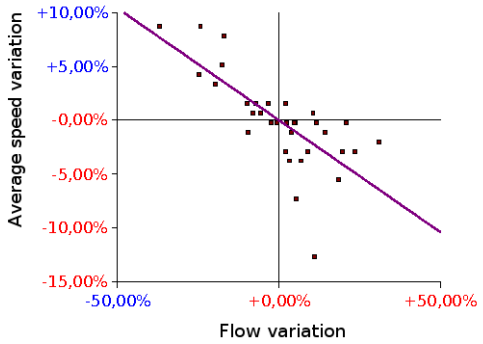
At present, STRATUNA has been partially implemented in a simplified form in order to perform a preliminary validation. The implemented model is the $\beta 4$ version: $STRATUNA_{\beta 4} = \langle R, E, X', P, S', \gamma_{\beta 4}, \tau_{\beta 4} \rangle$.

The function μ disappeared, because no weather evolution is considered, but only constant average conditions. Therefore $X' = \langle -r, -r+1, ..., 0, 1, ..., r \rangle$ substitutes X where r is a radius, accounting for the average visibility of an average driver and *Dynamic* substate is no more considered. *Indicator* lacks of hazard lights value, *PerceptionLevel* value is always 1, behavior involving *Aggressiveness* was not implemented and *Reactivity* is considered only for "staying far from *desired_speed*".

The generation function $\gamma_{\beta 4}$, was tailored for the traffic of Italian highway A4, characterized (in the area covered by data) by two lanes and twelve entrances/exits. Data are composed by around 1 milion of tolltickets, they are related to 5 non-contiguous weeks and grouped in five categories, depending on vehicle number of axles (it is reducible to our vehicle classification). Due to problems of time synchronization among tollgates, these datasets have to be considered partial and incomplete. For these reasons, a data cleaning step was mandatory for the following infrequent situations: (i) missed tickets: transits without entrance or starting time; (ii) transits across two or more days; (iii) transits that end before they begin; (iv) vehicles too fast to be true: exceeding 200 km/h as average speed. Afterwards, the average speed was related to the total flow for each of the 34 days.

The result of this quantitative study is summarized in the following chart: each day is rappresented as a dot; a shift over x-axis and y-axis is a variation respectively of "total flow" and "average speed" from their averaged values over all of the days (Fig. 10a).

DesiredSpeed distribution (Fig. 10b) according to the vehicle *Type* are easily deduced by highway data in free flow conditions for vehicles covering short distance in highway. The probability to park in the rest and services areas is minimal in short distance cases. Parking in the rest and services areas cannot be detected by data and causes errors; they justify the slightly higher values of average speed



(a) Daily flow and speed fluctuation from the average

Vehicle type	Desired speed	Flow share
I	122.80 km/h	93.4%
II	112.77 km/h	4.6%
III	113.29 km/h	0.5%
IV	102.61 km/h	0.1%
V	93.90 km/h	1.4%

(b) Share and desired speed for each type of vehicle in selected case of freeflow

Figure 10: Daily and selected data.

obtained in the simulated cases, in comparison to the same values of corresponding real events. Finally a statistical sampling treatment was performed to select meaningful subsets. After scaling flow values and vehicle generation rate, some validation sets were designed. Each set provides a number of vehicles (each one specified by the couple $\langle Origin, Destination \rangle$) and the average real speed ($\overline{rS}$) over all its vehicles and over all the event. Being 95% of real traffic, generated vehicles are all cars. Validation sets concern conditions from freeflow to congestion situation. In order to give a recapitulation of salient characteristics of the implemented transition function, a pseudo-code block is here presented. It is worth noting these remarks: (i) "return" ends the evolution of the single EA at each evolution step; (ii) functions starting in lowercase are actions enqueued to be performed in further steps; (iii) underlined functions represent the beginning of a synchronized protocol (e.g., actions in consecutive steps of takeover-protocol are: control a freezone on the left, light on the left indicator, start changing $Yposition$, and so on).

BEGIN: TransitionFunction()
FindNeighbours(); ComputeSpeedLimits();
ComputeTargetSpeed(); DefineFreeZones();
AssignTheCost_PM_WhereAFreeZoneIsReduced();
if(ManouvreInProgress())
  continueTheManouvre(); **return**;

APPENDIX B: HIGHWAY TRAFFIC

```
if(myLane==0) //I'm on a ramp
  if(IWantToGetIn())
    if(TheRampEnded())
      if(ICanEnter())
        enter(); return;
      else
        if(IHaveSpaceProblemsForward())
          slowDown(); return;
        else followTheQueue(); return;
    else //the ramp is not ended yet
      if(IHaveSpaceProblemsForward())
        followTheQueue(); return;
      else keepConstantSpeed(); return;
  else //I want to get out
    if(TheRampEnded()) deleteVehicle(); return;
    else
      if(IHaveSpaceProblemsForward())
        followTheQueue(); return;
      else keepConstantSpeed(); return;
  //end lane==0
else if(myLane==1)
  if(MyDestinationIsNear()) slowDown();
  if(MyDestinationIsHere()) goInLowerLane();
else //myLane==2 or more
  if(ICanGoInLowerLane())
    if(GoingInLowerLaneIsForcedOrConvenient())
      goInLowerLane();
  else //I cannot go in lower lane
    if(MyDestinationIsNear())
      slowDown(); goInLowerLane();
if(!IHaveSpaceProblemsForward()) //every lane
  if(TakeoverIsPossibleAndMyDestinationIsFar())
    if(TakeOverIsDesired()) takeover();
    else followTheQueue();
  else followTheQueue();
else //I have space problems forward
  if(TheTakeoverIsForced()) takeover();
```

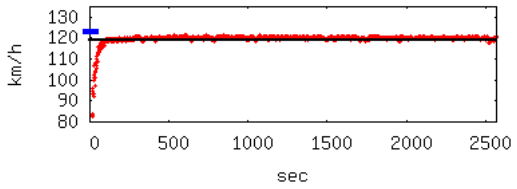
return;

END;

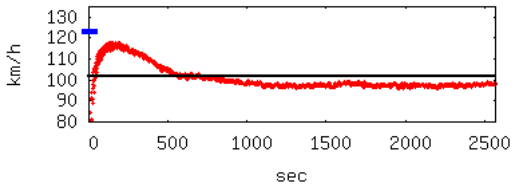
B.5.1 Results of simulations with STRATUNA B4

Here I report five significant simulations for typical highway conditions: freeflow (Fig. 11a), moderated-flow next to congestion (Fig. 11b and Fig. 11c) and locally congested situations (Fig. 11d). In addition to $\overline{rS}$ (represented in figures as a line) I consider step-by-step average simulated speed ($\overline{sS}$, represented in figures as fluctuating curves) and average simulated desired speed ($\overline{sDS}$, represented in figures as an invariant notch, Cf. Fig. 10b). Simulation conditions contemplate, at the beginning, for all of the cases, an empty highway, fed at each entrance with vehicles according to appropriate generation rate. Initially, average speed is low, because generated vehicles start from null speed. After this very first phase, $\overline{sS}$ increases since vehicles can tend to their *DesiredSpeed* value, until the small number of vehicles in the highway permits free flow conditions (i.e., when simulation time $<$ 500s). To provide a goodness measurement, simulations reported are accompanied with two error quantification: e_1 and e_2. The first one measures the average relative error (over all CA steps) between $\overline{sS}$ and $\overline{rS}$; the second one is the same as first but calculated after 500 seconds of simulated time in order to skip the initial phases of model evolution.

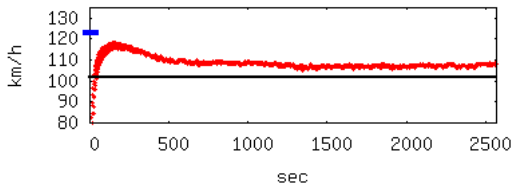
In the freeflow case, $\overline{sS}$ matches $\overline{rS}$ during the whole simulation, remaining slightly higher than field data, with very short oscillations (Fig. 11a). In the moderated flow case (Fig. 11b), after the same initial phase, $\overline{sS}$ became definitely lower than $\overline{rS}$ with moderate oscillations. Such a behavior is not correct, also if its error rate is low: the cars in the simulation must be faster than corresponding real cars, because they don't waste time to park in the rest and services areas. This problem depends clearly on the driver subjective evaluation, that came out too much cautiously because the partial implementation of transition function reduced the moving potentiality (reaction rigidity). A possible solution could be a shorter time step, that is equivalent to a more rapid reactivity. The utilized time steps have been 1s, the standard average reaction time of the average driver. Simulation was repeated with time step 0.75s, obtaining a more realistic result



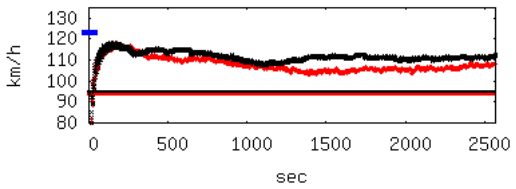
(a) Freeflow: *clock*=1s; err_1=1.29%; err_2=0.86%.



(b) Moderated flow: *clock*=1s; err_1=4.87%; err_2=3.44%.



(c) Moderated flow (tuned reactivity value): *clock*=0.75s; err_1=6.47%; err_2=5.83%.



(d) Locally congested situations. Lighter (darker) curve: *clock*=1s; err_1=14.79% (17.27%); err_2=13.94% (17.12%).

Figure 11: Average speed fluctuation in selected case study.

(Fig. 11c).

After this, two simulations, where the implementation performance is lower than previous simulations, are reported in Fig. 11d. Both account for the same particular real situation, when a largest vehicle flow occurs only from one entrance; both run on the same model specifications and feed function. $\overline{sS}$ became quickly significantly higher than $\overline{rS}$. This means that the reaction rigidity of the driver was rewarded by a higher speed in this particular case because the entrance filtering creates synchronization.

Classical patterns of highway traffic (moving jams and synchronized flow) have been observed in the simulations of congested traffic, but the lack of data collected automatically by stationary inductive loops (single vehicle data[116]) does not permit a serious comparison.

B.6 Cost system for congestion toll

Theories on congestion pricing have been under research since the 1920's and there are numerous references in literature about methods to estimate the costs for operating a car (fuel costs, maintenance, etc) in addition to the costs that each individual traveler imposes on other travelers due to the fact that each car increases the congestion of the highway. Road pricing has been implemented in various countries worldwide in order to reduce the traffic congestion problems in urban roads and highways. Here I propose an established cost system in which the simulation model can guide to business advantages. Assuming all vehicles are only cars, the principle of congestion pricing [123] provides a direct curve of correlation between traffic volume and its costs. In fact, every motorist making a trip introduces personal expenses in terms of private marginal costs, MC, (that are operating car costs plus the value of time spent in the highway) and takes a social cost (whose average will be denoted as AC). The difference between MC and AC represents the cost that a driver induced on his road neighbors [124]: if c is the hourly average generalized travel cost (as above, it is composed by car operating costs plus value of travel time) and is supposed to be invariable, $dist$ is the covered distance (assumed to be 1 km in the second part of Eq. (9)), $V(q)$ is a function of the flow q and represents the speed of vehicles, then AC, with respect to a certain flow value q, is given by:

$$AC(q) = c\frac{dist}{V(q)} = \frac{c}{V(q)} \tag{9}$$

Thus the total cost $T(q)$ of those vehicles is simply $T(q) = qAC(q) = (qc)/V(q)$. This means that for each new vehicle joining the flow q, we have the following marginal cost for the community:

$$MC(q) = \frac{d}{dq}T(q) = \frac{V(q)c - qc\frac{d}{dq}(V(q))}{V(q)^2} = AC(q) - \frac{qc}{V(q)^2}\frac{d}{dq}(V(q)) \tag{10}$$

Assuming that MC increases much more rapidly than AC when congestion begins (i.e. a flow $q > q'$), the difference between these two values is the considered

money that motorists have to pay if we want to charge the cost they are imposing to the society. This means that the "congestion toll" r is given by:

$$r = MC(q') - AC(q') = \frac{qc}{V(q)^2}\frac{d}{dq}(V(q)) \tag{11}$$

This quantity could be equal to zero when there is no congestion (i.e. flow $q <= q'$), increases when the flow increases and subsequently decreases when $V(q)$ increases. Now I introduce a model that is widely used and empirically verified over several highway models to establish the correlation between the flow and the speed of vehicles composing it: the Drake model [125]. Let q_0 be the maximum flow capacity (vehicles per hour per lane), V_0 the corresponding speed at maximum flow capacity and V_f the speed in free flow condition, then in the framework of Drake model, q is given by:

$$q = V(q)\frac{q_0}{V_0}\sqrt[\delta]{\delta\ ln(V_f/V(q))} \tag{12}$$

The speed-flow relationship given by Eq. (12) where δ is a parameter equal to 2 [125], can be used inside Eqs. (9-11) to estimate the congestion toll when the flow is higher than q' and the Drake model is a good approximation. As a result the congestion toll is given by:

$$r = \frac{c}{V(q)}\frac{ln(V_f) - ln(V(q))}{ln(V(q)) - ln(V_0)} \tag{13}$$

Assuming that European euro/km rates [126] are also valid for Italy, we can take cost values reported in Table 3 as input and then derive the value of $c = 1.08 euro/km$. Moreover, in order to resolve Eq. (13), values for V_0 and V_f are needed; while the value of speed at free flow can be considered as the one presented in Fig. 10b ($V_f = 122.8 km/h$), the inference of a proper value for V_0 needs more attention. The evaluation of a realistic V_0 value is where our STRATUNA model can help and, in fact, leads to cost forecasting through speed forecasting.

In fact, our model has the expressively needed for speed forecasting and has exhibited a predicting reliability for different flow volumes even in its partially implemented version (detailed above). Therefore, it can be used, together with

	euro/km	euro/hour
Petrol	0.112844	0.31090068
Tires	0.00806588	0.0222226
Service labour costs	0.02184835	0.06019521
Replacement parts	0.01472219	0.04056164
Parking and tolls	0.01409571	0.03883562
Standing charges	0.21981479	0.60561986
Total	**0.39139093**	**1.07833562**

Table 3: Total of Standing Charges and Running Costs, Assuming 15000 km per Year

the cost system object of this section, to foresee how different highway designs influence the speed at maximum capacity. This enables a straightforward calculation of the corresponding income for the highway owners and for the society. I now present the curves of AC and MC, as stated by Eqs. (9-10), with the aim of fixing the cost system.

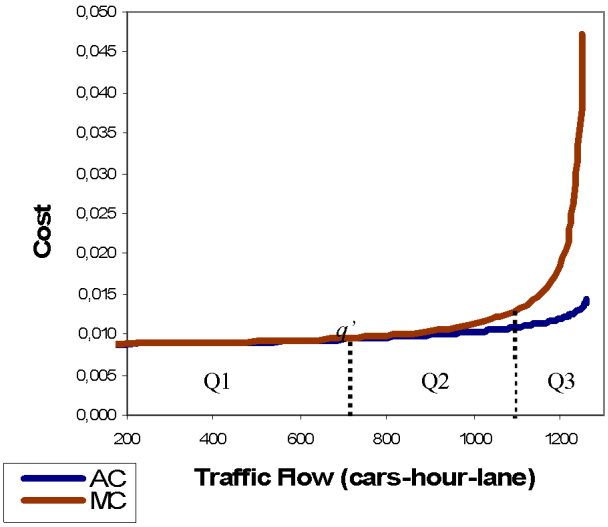


Figure 12: Cost of AC and MC in Relation to the Flow q.

Up to a traffic volume of about 680 cars per hour per lane, the private cost of a motorist (MC) is, in fact, identical to the one that he imposes to others (AC). This, presented in Fig. 12 as $Q1$, can be traced back to the free flow condition; same tracing is possible from $Q2$ ad $Q3$ (Cf. Fig. 12) to moderated flow and traffic jams. For quantity of cars $q > q'$ we have AC costs that increase more rapidly

than MC: first linearly and then polynomially. This increasing cost, induced to others with heavier flow, can be represented by Fig. 13: more cars means slower speed, that means more breaking/accelerating, low gears usage, higher petrol consumption and so on.

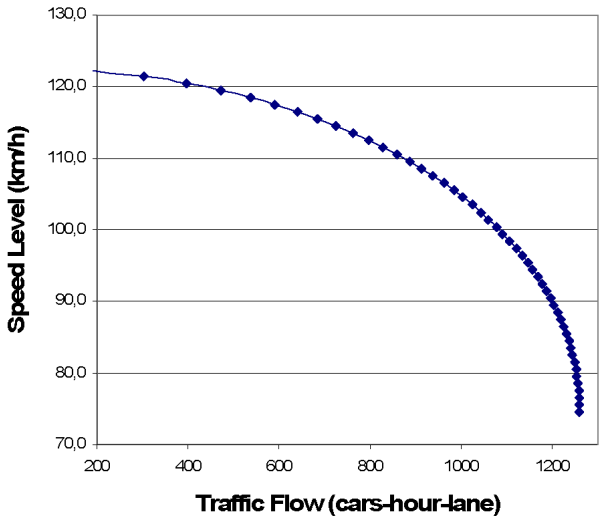


Figure 13: Speed-Flow Chart.

Now that the cost system has been satisfactory detailed, I propose in Fig. 14 the congestion toll (euro/km) evolution, in relation with the V_0 value deduced by our model.

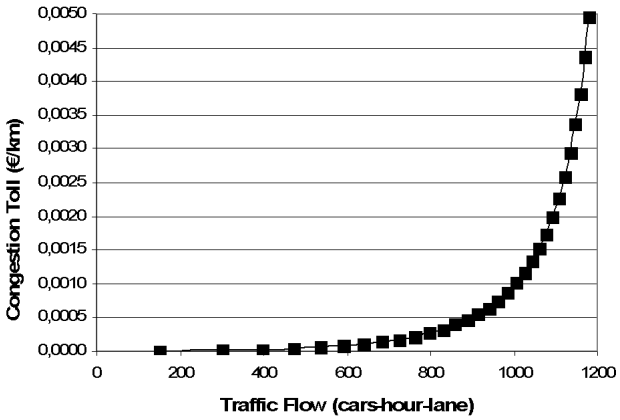


Figure 14: Congestion Toll with respect to Traffic Flow.

Above results show clearly that, through a simulation model, the test of dif-

Flow type	Free flow	Moderated flow	Congestion
Corresponding min. flow	0	681	1101
Corresponding max. flow	680	1100	1260
Min. congestion toll	+0.0	+0.0001	+0.0023
Max. congestion toll	+0.0	+0.0023	+0.0339

Table 4: Congestion Toll and Different Traffic Flows. Flows are measured in *cars per hour per lane*, while tolls are reported in *euro per car per km*.

ferent highway designs is possible and then, to each design, is linkable a simulated V_0 value, leading to the appropriate congestion toll. In other words, through the simulation of different highway design, differentiated V_0 values follow; then, the optimal congestion cost is derivable from it by means of the reported congestion toll system. As a result, I report in Table 4 the congestion toll that the price system of the simulated and analyzed highway could implement in relation to free flow, moderated flow and traffic jams.

B.7 Conclusions

These results of the reduced version $\beta 4$ of the STRATUNA model are very encouraging, considering that discrepancies between statistics deduced by real data and simulations are in part justified by unavoidable inaccuracies in the available real data and by imprecision introduced by parking in the rest and services areas. This is an interesting starting point in order to implement the full model. An important problem will be to tune some values of variables concerning the driver subjective behavior to solve problems of congested situations. The implemented model, used together with an established cost system, guides the interesting problem of the appraisal of the right price for a toll ticket. Indeed, the simulator shows the ability of associating to a simulated highway a value of average speed at maximum capacity. Thanks to this value, it is possible to establish a congestion toll mechanism. This mechanism, widely used worldwide, gives to motorists the perception of the costs they are imposing to other travelling and non-travelling people. The CA approach demonstrates its validity and leads to interesting emerging phenomena, both from the traffic forecasting and from an

economical point of view; in the latter, STRATUNA gives a feedback that connects different highway designs to different congestion toll charges through an established cost system. Accessing to other types of data concerning highway traffic would be important for the approach completeness.

References

[1] Ramsden, J.: Bioinformatics: An Introduction. 2nd edn. Springer Publishing Company, Incorporated (2009) 1

[2] Collado-Vides, J., Magasanik, B., Gralla, J.D.: Control site location and transcriptional regulation in escherichia coli. Microbiol. Mol. Biol. Rev. **55**(3) (September 1991) 371–394 2

[3] Kitano, H.: Computational systems biology. Nature **420**(6912) (November 2002) 206–210 2

[4] Ventura, B.D., Lemerle, C., Michalodimitrakis, K., Serrano, L.: From in vivo to in silico biology and back. Nature **443**(7111) (October 2006) 527–533 2

[5] Friedland, A.E., Lu, T.K., Wang, X., Shi, D., Church, G., Collins, J.J.: Synthetic gene networks that count. Science (New York, N.Y.) **324**(5931) (May 2009) 1199–1202 3

[6] Purnick, P.E.M., Weiss, R.: The second wave of synthetic biology: from modules to systems. Nature Reviews. Molecular Cell Biology **10**(6) (June 2009) 410–422 PMID: 19461664. 3

[7] Stephanopoulos, G.N., Aristidou, A.A., Nielsen, J.: Metabolic Engineering : Principles and Methodologies. Academic Press (October 1998) 3

[8] U. S. National Academy of Engineering: Engineering's grand challenges Details at http://www.engineeringchallenges.org/. 3, 25

[9] Barber, J.: Photosystem II: the engine of life. Quarterly Reviews of Biophysics **36**(01) (2003) 71–89 4, 26

[10] Nedbal, L., Červenỳ, J., Rascher, U., Schmidt, H.: E-photosynthesis: a comprehensive modeling approach to understand chlorophyll fluorescence transients and other complex dynamic features of photosynthesis in fluctuating light. Photosynthesis Research **93**(1) (2007) 223–234 4, 27

[11] Bauer-Mehren, A., Furlong, L.I., Sanz, F.: Pathway databases and tools for their exploitation: benefits, current limitations and challenges. Molecular Systems Biology **5** (2009) 290–303 4, 48, 49

[12] Dao, N., McCormick, P.J., Dewey, C.F.: The human physiome as an information environment. Annals of Biomedical Engineering **28**(8) (2000) 1032–1042 4, 48

[13] Hunter, P., Borg, T.: Integration from proteins to organs: the physiome project. Nature Reviews. Molecular Cell Biology **4**(3) (2003) 237–243 4, 48

[14] Hucka, M., Finney, A., Sauro, H., Bolouri, H., Doyle, J., Kitano, H., and the rest of the SBML forum: The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics **19**(4) (2003) 524–531 5, 49, 50

[15] Lloyd, C., Halstead, M., Nielsen, P.: CellML: its future, present and past. Progress in Biophysics and Molecular Biology **85**(2-3) (2004) 433–450 5, 50

[16] von Neumann, J.: Theory of Self Reproducing Automata. University of Illinois Press (1966) 7

[17] Wolfram, S.: A New Kind of Science. Wolfram Media (January 2002) 7

[18] Di Gregorio, S., Rongo, R., Siciliano, C., Sorriso-Valvo, M., Spataro, W.: Mount ontake landslide simulation by the cellular automata model sciddica-3. Physics and Chemistry of the Earth, Part A **24** (1999) 97–100 7, 15

[19] D'Ambrosio, D., Di Gregorio, S., Gabriele, S., Gaudio, R.: A cellular automata model for soil erosion by water. Physics and Chemistry of the Earth, Part B **26** (2001) 33–40 7, 15

[20] Crisci, G.M., Di Gregorio, S., Rongo, R., Spataro, W.: The simulation model sciara: the 1991 and 2001 at mount etna. Journal of Vulcanogy and Geothermal Research **132** (2004) 253–267 7, 8, 64

[21] Crisci, G.M., Di Gregorio, S., Rongo, R., Spataro, W.: Pyr: a cellular automata model for pyroclastic ows and application to the 1991 mt. pinatubo eruption. Future Generation Computer Systems **21** (2005) 1019–1032 7, 15

[22] Di Gregorio, S., Serra, R.: An empirical method for modelling and simulating some complex macroscopic phenomena by cellular automata. Future Generation Computer Systems **16** (1999) 259–271 7

[23] D'Ambrosio, D., Di Gregorio, S., Iovine, G.: Simulating debris ows through a hexagonal cellular automata model: Sciddica s3-hex. Natural Hazards and Earth System Sciences **3** (2003) 545–559 7

[24] Atkinson, P.M., Foody, G.M., Darby, S., Wu, F. In: Brains versus Brawn - Comparative Strategies for the Calibration of a Cellular Automata-based urban growth model. CRC Press (2004) 7

[25] Straatman, B., White, R., Engelen, G.: Towards an automatic calibration procedure for constrained cellular automata. Computers, Environment and Urban Systems **28** (2004) 149–170 7

[26] D'Ambrosio, D., Spataro, W.: Parallel evolutionary modelling of geological processes. Parallel Computing **33** (2007) 186–212 7

[27] Holland, J.H. In: Nonlinear environments permitting effcient adaptation. New York: Academic (1967) 7

[28] Holland, J.H.: Adaptation in Natural and Articial Systems. University of Michigan Press, Ann Arbor (1975) 7, 64

[29] Avolio, M., Crisci, G.M., Di Gregorio, S., Rongo, R., Umeton, R.: Introduction of more physical features in the cellular automata model for lava flows sciara: preliminary results regarding the viscosity (2007) Asia and Oceania Geosciences Society. 8, 64

[30] Lamb, H.: Hydrodynamics. Cambridge University Press (1879) 8

[31] Wright, A.H.: Genetic algorithms for real parameter optimization. In Rawlins, G.J., ed.: Foundations of genetic algorithms. Morgan Kaufmann, San Mateo, CA (1991) 205–218 10

[32] Wright, A.H.: Genetic algorithms for real parameter optimization. In Rawlins, G.J., ed.: Foundations of Genetic Algorithms, First Workshop on the Foundations of Genetic Algorithms and Classier Systems. Morgan Kaufmann, San Mateo, CA (1990) 205–218 12

[33] D'Ambrosio, D., Spataro, W., Iovine, G.: Parallel genetic algorithms for optimising cellular automata models of natural complex phenomena: an application to debris-ows. Computers and Geosciences **32** (2006) 861–875 15, 16

[34] Storn, R., Price, K.: Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. Journal of Global Optimization **11**(4) (1997) 341–359 16, 29

[35] Hansen, N., Ostermeier, A.: Completely derandomized self-adaptation in evolution strategies. Evolutionary Computation **9**(2) (2001) 159–195 16, 29

[36] Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation **6**(2) (2002) 182–197 17

[37] Burgard, A.P., Pharkya, P., Maranas, C.D.: Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnology and Bioengineering **84**(6) (December 2003) 647–657 18

[38] Mahadevan, R., Bond, D.R., Butler, J.E., Esteve-Nuez, A., Coppi, M.V., Palsson, B.O., Schilling, C.H., Lovley, D.R.: Characterization of metabolism in the Fe(III)-Reducing organism geobacter sulfurreducens by Constraint-Based modeling. Applied and Environmental Microbiology **72**(2) (February 2006) 1558–1568 18, 19

[39] Becker, S.A., Feist, A.M., Mo, M.L., Hannum, G., Palsson, B.O., Herrgard, M.J.: Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox. Nat. Protocols **2**(3) (March 2007) 727–738 19

[40] Cutello, V., Narzisi, G., Nicosia, G.: A multi-objective evolutionary approach to the protein structure prediction problem. Journal of the Royal Society Interface **3**(6) (2006) 139–151 20

[41] Zitzler, E., Brockhoff, D., Thiele, L.: The hypervolume indicator revisited: On the design of pareto-compliant indicators via weighted integration. Lecture Notes in Computer Science **4403** (2007) 862 21

[42] Zhu, X.G., de Sturler, E., Long, S.P.: 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** (2007) 513–526 23, 30, 35, 37, 43, 65, 69

[43] Floudas, C.A., Pardalos, P.M., eds.: Optimization in Computational Chemistry and Molecular Biology: Local and Global Approaches. Kluwer Academic Publishers (2000) 25

[44] Dasika, M.S., Maranas, C.D.: Optcircuit: An optimization based method for computational design of genetic circuits. BMC Systems Biology **2** (2008) 24 25

[45] Conn, A.R., Scheinberg, K., Vicente, L.N.: Introduction to Derivative-Free Optimization. SIAM (2009) 25

[46] Hubley, R.M., Zitzler, E., Roach, J.C.: Evolutionary algorithms for the selection of single nucleotide polymorphisms. BMC Bioinformatics **4:30** (2003) 26

[47] Cutello, V., Narzisi, G., Nicosia, G.: A multi-objective evolutionary approach to the protein structure prediction problem. Journal of the Royal Society Interface **3**(6) (2006) 139–151 26

[48] Nagrath, D., Avila-Elchiver, M., Berthiaume, F., Tilles, A.W., Messac, A., Yarmush, M.L.: Integrated energy and flux balance based multiobjective framework for large-scale metabolic networks. Annals of Biomedical Engineering **35**(6) (2007) 863–885 26

[49] Morris, M.: Factorial sampling plans for preliminary computational experiments. Technometrics **33**(2) (1991) 161–174 26

[50] Gunawardena, J.: Models in systems biology: the parameter problem and the meanings of robustness. In: Elements of Computational Systems Biology. John Wiley and Sons (2010) 21–43 27

[51] Saltelli, A., Tarantola, S., Campolongo, F.: Sensitivity analysis in practice: a guide to assessing scientific models. John Wiley & Sons Inc (2004) 28

[52] Vaz, A., Vicente, L.: A particle swarm pattern search method for bound constrained global optimization. Journal of Global Optimization **39**(2) (2007) 197–219 29

[53] Hooke, R., Jeeves, T.A.: "Direct Search" solution of numerical and statistical problems. Journal of ACM **8**(2) (1961) 212–229 29, 35

[54] Lewis, R., Torczon, V.: Pattern search algorithms for bound constrained minimization. SIAM Journal on Optimization **9**(4) (1999) 1082–1099 29, 35

[55] Audet, C., Dennis, J.E.: Mesh adaptive direct search algorithms for constrained optimization. SIAM Journal on Optimization **17**(1) (2007) 188–217 29, 35

[56] Jones, D.R., Perttunen, C.D., Stuckman, B.E.: Lipschitzian optimization without the lipschitz constant. Journal of Optimization Theory and Applications **79**(1) (1993) 157–181 29

[57] Huyer, W., Neumaier, A.: Global optimization by multilevel coordinate search. Journal of Global Optimization **14**(4) (1999) 331–355 29

[58] Gilmore, P., Kelley, C.T.: An implicit filtering algorithm for optimization of functions with many local minima. SIAM Journal on Optimization **5**(2) (1995) 269–285 29

[59] Jen, E.: Robust Design: A Repertoire of Biological, Ecological, and Engineering Case Studies. Oxford University Press (2005) 31

[60] Mohr, H., Schopfer, P.: Plant Physiology. Springer Verlag (1995) 33

[61] Raines, C.A.: The calvin cycle revisited. Photosynthesis Research **75** (2003) 1–10 33

[62] Takahashi, S., Bauwe, H., Badger, M.: Impairment of the photorespiratory pathway accelerates photoinhibition of photosystem II by suppression of repair but not acceleration of damage processes in Arabidopsis. Plant Physiology **144**(1) (2007) 487–494 33, 46

[63] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research **25**(17) (1997) 3389–3402 34

[64] Lefebvre, S., Lawson, T., Zakhleniuk, O.V., Lloyd, J.C., Raines, C.A.: Increased sedoheptulose-1,7-bisphosphatase activity in transgenic tobacco plants stimulates photosynthesis and growth from an early stage in development. Plant Physiology **138**(451–460) (2005) 39

[65] Tamoi, M., Nagaoka, M., Miyagawa, Y., Shigeoka, S.: Contribution of fructose-1,6-bisphosphatase and sedoheptulose-1,7-bisphosphatase to the photosynthetic rate and carbon flow in the calvin cycle in transgenic plants. Plant Cell Physiology **47**(380–390) (2006) 39

[66] Sheng, W., Xiao, L., Mao, Z.: Soft error optimization of standard cell circuits based on gate sizing and multi-objective genetic algorithm. In: Proceedings of the 46th Annual Design Automation Conference. DAC '09, New York, NY, USA, ACM (2009) 502–507 39

[67] McConaghy, T., Palmers, P., Gielen, G., Steyaert, M.: Simultaneous multi-topology multi-objective sizing across thousands of analog circuit topologies. In: Proceedings of the 44th annual Design Automation Conference. DAC '07, New York, NY, USA, ACM (2007) 944–947 39

[68] Tilman, D., Cassman, K.G., Matson, P.A., Naylor, R., Polasky, S.: Agricultural sustainability and intensive production practices. Nature **418** (2002) 671–677 39

[69] Hirel, B., Gouis, J.L., Ney, B., Gallais, A.: The challenge of improving nitrogen use efficiency in crop plants: towards a more central role for genetic variability and quantitative genetics within integrated approaches. J. Exp. Bot. **58**(9) (2007) 2369–2387 40

[70] Foyer, C.H., Bloom, A.J., Queval, G., Noctor, G.: Photorespiratory metabolism: genes, mutants, energetics, and redox signaling. Annual Review of Plant Biology **60** (2009) 455–484 41, 46

[71] Canadell, J.G., Quéré, C.L., Raupach, M.R., Field, C.B., Buitenhuis, E.T., Ciais, P., Conway, T.J., Gillett, N.P., Houghton, R.A., Marland, G.: Contributions to accelerating atmospheric CO_2 growth from economic activity, carbon intensity, and efficiency of natural sinks. Proceedings of the National Academy of Sciences **104**(47) (2007) 18866–18870 42

[72] Field, C.B., Behrenfeld, M.J., Randerson, J.T., Falkowski, P.: Primary production of the biosphere: integrating terrestrial and oceanic components. Science **281**(5374) (1998) 237–240 43

[73] Coughlan, M.P.: The properties of fungal and bacterial cellulases with comment on their production and application. Biotechnology & Genetic Engineering Reviews **3** (1985) 39–109 43

[74] Woodrow, I.E., Berry, J.A.: Enzymic regulation of photosynthetic co2 fixation in c3 plants. Annu Rev Plant Physiol Plant Mol Biol **39** (1988) 533–594 43

[75] Geiger, D.R., Servaites, J.C.: Diurnal regulation of photosynthetic carbon metabolism in c3 plants. Annu Rev Plant Phys Plant Mol Biol **45** (1994) 235–256 43

[76] Raines, C.A.: The Calvin cycle revisited. Photosynthesis research **75**(1) (2003) 1–10 43

[77] Zhang, Q., Li, H.: Moea/d: A multiobjective evolutionary algorithm based on decomposition. Evolutionary Computation, IEEE Transactions on **11**(6) (2007) 712–731 44

[78] Noctor, G., Veljovic-Jovanovic, S.D., Driscoll, S., Novitskaya, L., Foyer, C.H.: Drought and oxidative load in the leaves of c3 plants: a predominant role for photorespiration? Ann. Bot. **89** (2002) 841–50 46

[79] Ayyadurai, V.A.S., Dewey, C.F.: CytoSolve: a scalable computational method for dynamic integration of multiple molecular pathway models. Cellular and Molecular Bioengineering **3**(4) (2010) In press 48, 56

[80] Le Novère, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J., Hucka, M.: BioModels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Research **34**(Database issue) (2006) D689–691 48, 57

[81] W3C OWL Working Group: OWL 2 Web Ontology Language: Document Overview. W3C Recommendation (27 October 2009) Available at http://www.w3.org/TR/owl2-overview/. 49, 54

[82] Noy, N.F., Sintek, M., Decker, S., Crubezy, M., Fergerson, R.W., Musen, M.A.: Creating semantic web contents with protege-2000. Intelligent Systems, IEEE [see also IEEE Intelligent Systems and Their Applications] **16**(2) (2001) 60–71 49, 54, 62

[83] Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., Sherlock, G.: Gene ontology: tool for the unification of biology. the gene ontology consortium. Nature Genetics **25**(1) (2000) 25–29 49, 67

[84] The UniProt Consortium: The universal protein resource (UniProt). Nucleic Acids Research **35**(Database issue) (2007) D193–197 49, 67

[85] Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcantara, R., Darsow, M., Guedj, M., Ashburner, M.: ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Research **36**(suppl_1) (2008) D344–350 49

[86] Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., Yamanishi, Y.: KEGG for linking genomes to life and the environment. Nucleic acids research **36**(Database issue) (January 2008) 484, D480 49

[87] Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., D'Eustachio, P.: Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Research **37**(Database issue) (January 2009) D619–622 PMID: 18981052. 49

[88] Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M., Chute, C.G., Musen, M.A.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research **37**(suppl_2) (2009) W170–173 49

[89] Lloyd, C.M., Lawson, J.R., Hunter, P.J., Nielsen, P.F.: The CellML model repository. Bioinformatics **24**(18) (2008) 2122–2123 49

[90] Hines, M.L., Morse, T., Migliore, M., Carnevale, N.T., Shepherd, G.M.: ModelDB: a database to support computational neuroscience. Journal of Computational Neuroscience **17**(1) (2004) 7–11 49

[91] Moraru, I.I., Schaff, J.C., Slepchenko, B.M., Blinov, M., Morgan, F., Lakshminarayana, A., Gao, F., Li, Y., Loew, L.M.: The virtual cell modeling and simulation software environment. IET Systems Biology **2**(5) (2008) 352–362 49

[92] The Virtual Cell Website: Virtual cell repository. http://www.nrcam.uchc.edu/vcell_models/published_models.html 49

[93] Schilstra, M.J., Li, L., Matthews, J., Finney, A., Hucka, M., Le Novère, N.: CellML2SBML: conversion of CellML into SBML. Bioinformatics **22**(8) (2006) 1018–1020 49

[94] European Bioinformatics Institute: SBML converters. http://www.ebi.ac.uk/compneur-srv/sbml/converters/SBMLConverters.html 49

[95] Krause, F., Uhlendorf, J., Lubitz, T., Schulz, M., Klipp, E., Liebermeister, W.: Annotation and merging of SBML models with semanticSBML. Bioinformatics (2009) btp642 49, 62

[96] Umeton, R., Yankama, B., Nicosia, G., Dewey, Jr., C.: Oremp: Ontology reasoning engine for molecular pathways. In D'Aquin, M., Castro, A.G., Lange, C., Viljanen, K., eds.: Proceedings of ORES 2010, 1st International Workshop on Ontology Repositories and Editors for the Semantic Web - satellite conference in ESWC 2010, 7th Extended Semantic Web Conference, May 30, 2010, Heraklion, Crete, Greece. Volume 596 of CEUR Workshop Proceedings., CEUR-WS.org (2010) 26–30 50, 67

[97] Umeton, R., Yankama, B., Nicosia, G., Dewey, C.: A cross-format framework for consistent information integration among molecular pathways and ontologies. In Magjarevic, R., Lim, C.T., Goh, J.C.H., eds.: Proceedings of WCB 2010, 6th World Congress on Biomechanics, August 1 - 6, 2010, Singapore. Volume 31 of IFMBE Proceedings., Springer Berlin Heidelberg (2010) 1595–1598 50, 67

[98] Yankama, B., Umeton, R., Ayyadurai, S., Dewey, C.: Editing and aligning complex molecular pathways using 3D models. In: Proceedings of the BMES 2010, Biomedical Engineering Society, October 6 - 9, 2010, Austin, TX 50, 63, 67

[99] W3C Math Working Group: Mathematical Markup Language (MathML): Overview. W3C Recommendation (21 October 2010) Available at http://www.w3.org/TR/MathML/. 50

[100] Le Novère, N., Finney, A., Hucka, M., Bhalla, U.S., Campagne, F., Collado-Vides, J., Crampin, E.J., Halstead, M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J.L., Spence, H.D., Wanner, B.L.: Minimum information requested in the annotation of biochemical models (MIRIAM). Nature Biotechnology **23**(12) (2005) 1509–1515 52

[101] Rigo, A.: Representation-based just-in-time specialization and the psyco prototype for python. In: Proceedings of the 2004 ACM SIGPLAN symposium on Partial evaluation and semantics-based program manipulation - PEPM '04, Verona, Italy (2004) 15–26 54

[102] Tsarkov, D., Horrocks, I.: FaCT++ description logic reasoner: System description. In: Proceedings of the International Joint Conference on Automated Reasoning (IJCAR 2006). Volume 4130 of Lecture Notes in Artificial Intelligence., Springer (2006) 292–297 54

[103] European Bioinformatics Institute: Biomodel 19 Available at http://www.ebi.ac.uk/biomodels-main/BIOMD0000000019. 54, 55, 59

[104] SRI International: Humancyc pathways db Available at http://www.humancyc.org. 59

[105] SRI International: Ecocyc pathway db Available at http://www.ecocyc.org. 59

[106] Wang, H., He2, H., Yang, J., Yu, P.S., Yu, J.X.: Dual labeling: Answering graph reachability queries in constant time. In: Data Engineering, Interna-

tional Conference on. Volume 0., Los Alamitos, CA, USA, IEEE Computer Society (2006) 75 59

[107] Schoeberl, B., Eichler-Jonsson, C., Gilles, E.D., Mller, G.: Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. Nature Biotechnology **20**(4) (April 2002) 370–375 PMID: 11923843. 59

[108] Massachusetts Institute of Technology: Oremp project Details at http://cytosolve.mit.edu/oremp. 62

[109] Umeton, R., Di Gregorio, S.: Admissible method for improved genetic search in cellular automata model AMMISCA: a strategy in genetic calibration - preliminary results. In Serra, R., Villani, M., Poli, I., eds.: Proceedings of WIVACE 2008, 5th Italian Workshop on Artificial Life and Evolutionary Computation, September 8 - 10, 2008, Venice, Italy, World Scientific (2008) 99–108 64

[110] Stracquadanio, G., Umeton, R., Papini, A., Liò, P., Nicosia, G.: Analysis and optimization of c3 photosynthetic carbon metabolism. In Rigoutsos, I., Floudas, C.A., eds.: Proceedings of BIBE 2010, 10th IEEE International Conference on Bioinformatics and Bioengineering, May 31 - June 3, 2010, Philadelphia, PA, USA, IEEE Computer Society (2010) 44–51 64

[111] Stracquadanio, G., Umeton, R., Papini, A., Liò, P., Nicosia, G.: Key enzymes for the optimization of co2 uptake rate and nitrogen concentration in the c3 photosynthetic carbon metabolism. In Fava, F., Nicotra, F., eds.: Proceedings of IBS 2010, 14th International Biotechnology Symposium and Exhibition, September 14 - 18, 2010, Rimini, Italy. Journal of Biotechnology, Elsevier (2010) To appear 64

[112] Di Gregorio, S., Umeton, R., Bicocchi, A., Evangelisti, A., Gonzalez, M.: Highway traffic model based on cellular automata: Preliminary simulation results with congestion pricing considerations. In Castilla-Rodriguez, I., Longo, F., eds.: Proceedings of EMSS 2008, 20th European Modeling and

Simulation Symposium, September 17 - 19, 2008, Campora S.G., CS, Italy, Inder Science (2008) 665–674 68

[113] Di Gregorio, S., Umeton, R., Bicocchi, A., Evangelisti, A.: A cellular automata model for highway traffic with preliminary results. In Serra, R., Villani, M., Poli, I., eds.: Proceedings of WIVACE 2008, 5th Italian Workshop on Artificial Life and Evolutionary Computation, September 8 - 10, 2008, Venice, Italy, World Scientific (2008) 235–244 68

[114] Enzyme Nomenclature Number: E.C. database Available at http://www.chem.qmul.ac.uk/iubmb/enzyme/. 72

[115] Di Gregorio, S., Serra, R.: An empirical method for modelling and simulating some complex macroscopic phenomena by cellular automata. Future Generation Computer Systems **16** (1999) 259–271 83, 84

[116] Schadschneider, A.: Cellular automata models of highway traffic. Physica A **372** (2006) 142–150 83, 86, 93

[117] Nagel, K., Schreckenberg, M. Journal of Physics, Part I **2** (1992) 2221 83

[118] Wolf, D.E.: Cellular automata for traffic simulation. Physica A **263** (1999) 438–451 83

[119] Knospe, W., Santen, L., Schadschneider, A., Schreckenberg, M.: Towards a realistic microscopic description of highway traffic. J. Phys. A: Math. Gen. **33** (2000) 477–L485 83

[120] Lárraga, M., del Ríob, J., lcaza L., A.: Cellular automata for one-lane traffic flow modeling. Transportation Research Part C **13** (2005) 63–74 83

[121] Di Gregorio, S., Festa, D.C. In: Cellular Automata for Freeway Traffic. Volume 5. (1981) 133–136 83

[122] Di Gregorio, S., Festa, D., Rongo, R., Spataro, W., Spezzano, G., Talia, D. In: A microscopic freeway traffic simulator on a highly parallel system. Volume 11 of Advances in Parallel Computing. (1996) 69–76 83

[123] Pigou, A.: The Economics of Welfare. MacMillan, London (1920) 94

[124] Li, M.Z.F.: The role of speed-flow relationship in congestion pricing implementation with an application to singapore. Transportation Research Part B: Methodological **36**(8) (2002) 731 – 754 94

[125] Drake, J., Schofer, J., May, A.: A statistical analysis of speed-density hypotheses. in vehicular traffic science. In Edie, L.C., Herman, R., Rothery, R., eds.: Proceedings of the Third International Symposium on the Theory of Traffic Flow, Elsevier Science Publishing Company (2006) 112–117 95

[126] Automobile Association Limited: Running costs for petrol cars Available at http://www.theaa.com/allaboutcars/advice/advice_rcosts_petrol_table.jsp. 95