Università degli Studi di Milano - Bicocca

Dipartimento di Informatica, Sistemistica e Comunicazione

Doctoral Thesis

A COMPUTATIONAL APPROACH FOR MULTI-LEVEL BIOLOGICAL COMPLEX SYSTEMS ANALYSIS

Author:
Riccardo Colombo

Supervisor:
Prof. Giancarlo Mauri
Ph.D. Coordinator:
Prof. Stefania Bandini

A thesis submitted in fulfillment of the requirements for the degree of Ph.D. in Computer Science

Academic Year 2014–2015
XXVII Series

“Spera et persevera donec transeat nox.”

A COMPUTATIONAL APPROACH FOR MULTI-LEVEL BIOLOGICAL COMPLEX SYSTEMS ANALYSIS

by Riccardo Colombo

Abstract

Due to their intrinsic nature, biological entities are universally considered as complex systems. Over years, many different computational methods pertaining to the Systems Biology field, have been devised to unravel this complexity. However, when taken alone, most of times these methods are not able to provide a deep comprehension of structural, spatial and dynamical aspects of the systems under evaluation. For this reason, approaches exploiting different levels of analysis are today a hot research topic in different areas, such as the theoretical formalization of the method, and the development of computational tools for the integration of different modeling perspectives.

In the present dissertation I developed a computational pipeline able to perform analyses exploiting, one after the other, the three main modeling frameworks for biological systems, gaining, from every level, a different type of information: i.e. identification of flux distributions and metabolic sub-phenotypes from the ensemble evolutionary FBA (a novel method inspired by Flux Balance Analysis); information on network structural properties and topological metrics from graph theory approaches; estimation of kinetic constants for mechanism-based modeling through the definition of an efficient version of the Particle Swarm Optimizer based on Fuzzy Logic. Moreover, I also redefined a network visualization strategy able to overlay flux values and topological metrics to network structure.

In order to validate the proposed pipeline I also developed a "core model" of yeast metabolism from which I identified two ensembles of flux distributions (possible solutions) in agreement with the "Crabtree-positive" and "Crabtree-negative" metabolic phenotypes. Moreover, by means of a cluster analysis, devised methods were able to define groups inside each ensemble that I identified as putative "sub-phenotypes".
Lastly, I contributed to reconstruct four reduced metabolic "core models", deriving from the Human Metabolic Atlas, and describing three tissue-specific cancer conditions and a reference state. From these models a relevant heterogeneity emerged between reference and cancer conditions in terms of metabolic flux values.

Acknowledgements

Moving from the "biological" to the "computational" field has been for me a challenge. If I succeeded, is also because of the many people that helped, supported and encouraged me during the years of my Ph.D. program. A sincere thank you goes to all of them.

First of all, I want to express my gratitude to my advisor, Professor Giancarlo Mauri. His constant supervision and advice helped me throughout the many critical steps of my work. The same gratitude goes the director of SYSBIO – Centre of Systems Biology, Professor Lilia Alberghina for her unceasing guidance and support.
A very special thank you is due to Professor Dario Pescini and to Chiara Damiani for their constant encouragement, intellectual support and assistance in solving the countless problems emerged in any possible domain.
I am also grateful to all the friends and colleagues that helped me both at DISCo and at the SYSBIO Centre, in particular: Professor Marco Vanoni, Daniela Besozzi, Marco S. Nobile, Paolo Cazzaniga, Sara Molinari, Marzia Di Filippo and Daniela Gaglio.

Moreover, I would like to thank Professor Hans V. Westerhoff who hosted me at the MCISB and provided constructive criticism and advice both in Manchester and Amsterdam. I am also profoundly grateful to Ettore Murabito for his supervision and, most of all, for his friendship in a delicate period.

Finally, my most deep gratitude goes to my parents, family, and all the friends who shared with me a long or short trait on the path of life during these last four years.

Contents

List of Figures

List of Tables

Contributions

- M. S. Nobile, G. Pasi, P. Cazzaniga, D. Besozzi, R. Colombo, and G. Mauri, "Proactive particles in swarm optimization:a self-tuning algorithm based on fuzzy logic, *Accepted to FUZZ-IEEE 2015. The 2015 IEEE International Conference on Fuzzy Systems," (Istanbul, Turkey)*, 2015.

- E. Murabito, R. Colombo, C. Wu, M. Verma, S. Rehman, J. Snoep, S.-L. Peng, N. Guan, X. Liao, and H. V. Westerhoff, "Suprabiology 2014: Promoting UK-China collaboration on systems biology and high performance computing," *Quant Biol*, pp. 1–8, 2015.

- P. Cazzaniga, C. Damiani, D. Besozzi, R. Colombo, M. S. Nobile, D. Gaglio, D. Pescini, S. Molinari, G. Mauri, L. Alberghina, and M. Vanoni, "Computational strategies for a system-level understanding of metabolism," *Metabolites*, vol. 4, no. 4, pp. 1034–87, 2014.

- C. Damiani, D. Pescini, R. Colombo, S. Molinari, L. Alberghina, M. Vanoni, and G. Mauri, "An ensemble evolutionary constraint-based approach to understand the emergence of metabolic phenotypes," *Nat Comput*, vol. 13, no. 3, pp. 321–331, 2014.

- C. Damiani, R. Colombo, S. Molinari, D. Pescini, D. Gaglio, M. Vanoni, L. Alberghina, G. Mauri, "An ensemble approach to the study of the emergence of metabolic and proliferative disorders via Flux Balance Analysis," *EPTCS*, vol. 130, pp. 92-97, 2013.

- F. Amara, R. Colombo, P. Cazzaniga, D. Pescini, A. Csikász-Nagy, M. Muzi-Falconi, D. Besozzi, and P. Plevani, "*In vivo* and *in silico* analysis of PCNA ubiquitylation in the activation of the post replication repair pathway in *S. cerevisiae*," *BMC Syst Biol*, vol. 7, no. 1, p. 24, 2013.

Chapter 1

Introduction

1.1 Complex Systems

Complexity is a common trait of many systems sorrounding our everyday life, ranging from biological to physical and sociological domains which exhibit a great variety in terms of components and relations among elements [1]. Even if we perceive complex systems under this great variety of forms, it is possible to identify some universal laws governing them. The study of these laws has been, and still is, the driving force of many scientific studies devoted to discover all the essential building blocks of modes of interaction of matter [2].

Intuitively, complexity deals with systems composed of many interconnected elements giving rise to a dynamical behavior that can not be guessed by the single components dynamics. Strikingly, in order to acquire a thorough comprehension of the system under examination it is necessary to understand both the dynamics of the elements (i.e. parts that can be described in a simple way when analyzing the behavior of the whole system) and of the system in its entirety [3]. Complex systems are commonly seen as hardly understandable systems, and this is mainly due to the fact that it is not possible to understand the whole system without analyzing each single part, and, at the same time, each part must be investigated in light of the fact that they are in relation to other parts [2].

Over years many different definitions of complex systems have been proposed in literature (see [4] for a review) ranging from quantitative and formal descriptions of complexity in specific domanins to qualitative, but rather unsatisfactory, universal statements.

Among this plethora of definitions, a particularly meaningful one has been proposed in several works [5–7] by Yaneer Bar-Yam. His definition of complexity rely on the

interdependance between "quantity of information" and scale at which the system is evaluated. A complex system is a system (physical, biological or social) that involves a number of elements, arranged in structure(s) which can exist on many scales.

In general all systems, given a constant number of components, can be represented by one of the profiles illustrated in Figure 1.1 where information is related to scale. Here the red curve represent random systems (such as the atoms in a gas) where the quantity of information is high at low scale, but when increasing the scale, this amount decreases at a fast rate. On the contrary the green curve describes the relation between information and scale for "coherent" or "ballistic" systems where the information content is not varying when moving from low to coharse grain scale (a flying cannonball could be a good example for this kind of system). Finally, the blue curve illustrates the behaviour for complex systems. In this case the information content is only slightly decreasing when scaling from fine to coharse, meaning that from each level of detail it is possible to retrieve knowledge on the structure and the dynamics of the system [5].

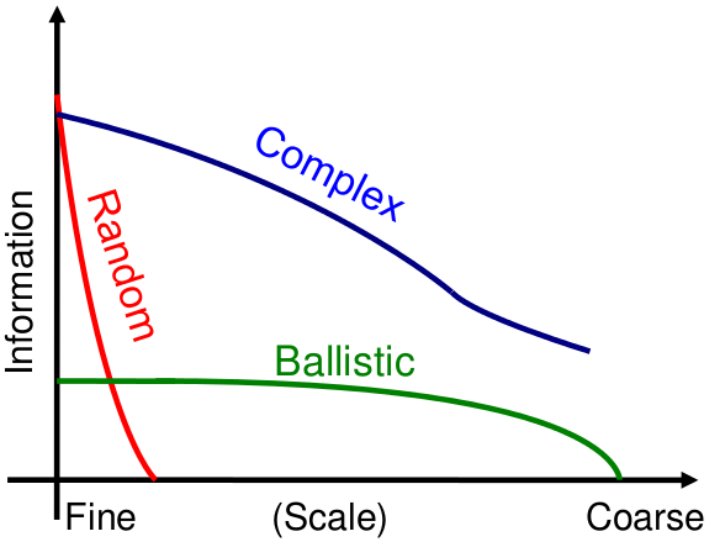


FIGURE 1.1: In this figure complexity profiles are illustrated in terms of scale at which the system is described and its content of information. Modified from [5].

This graphical representation of complexity underlines the fact that the behavior of the entire system can be interpreted in different ways depending on the scale at which it is investigated. According to this vision, complexity can be seen as the amount of information necessary to describe a system, being careful to integrate this definition with the evaluation of the level of detail in the description of the system itself.

In the context of complexity, a key concept connected to the behavior of the system is emergence [8]. This term indicates a property of the system that can not be reduced to properties belonging to single components of the system itself but that derives from the

synergistic interaction of parts at systemic level. In some ways, emergent properties can be seen as being "unpredictable" when evaluating the isolated elements of the system.

Following this definition of emergence, a reductionist approach will fail to capture the emergence of a property in a complex system. From this consideration it is clear that the best way to study these kind of systems is by means of an holistic approach investigating all the different parts in the context of the system as a whole.

It is evident that this new perspective in the study of complexity must be supported by many different tools pertaining to different scientific domains such as information and computation theory, Mathematics, applied Computer Science, Statistics, Physics and Biology. Moreover, due to the fact that complex systems can be found virtually in any scientific domain [3] and due to the increasing evidence that some of these domains are overlapping, over last decades it is arising a new discipline devoted to the investigation of universal principles governing complex systems [2].

In particular, due to the evident consequences on human health and due to the inherent structure of biological matter, a promising domain of application of studies in complexity is the investigation of biological systems ranging from population to molecular level [9]; the focus of the present work of thesis.

1.2 Complex Systems in Biology: Systems Biology

The branch of the complexity science applied to the investigation of the structure and the function of biological organisms is called Systems Biology (or sometimes Complex-Systems Biology), a research domain characterized by a deep analysis of complex interactions inside and among biological systems represented by networks and based on the holistic approach for the understanding of biological entities as "systems" [10].

The goal of Systems Biology is the integration of different disciplines, such as Biology, Chemistry, Computer Science, Mathematics and Physics as well as data that are generated in these contexts, in order to explain cellular and intercellular processes in terms of their regulation and dynamics [11].

This approach is declined in examining structure, dynamics and functioning of the system at global level, instead of investigating peculiar traits of isolated parts of a cell or of an organism [12]. This is due to the fact that many properties of "life" emerge only at system level.

In this context, the word "system" indicates not only an ensemble of components in a given configuration, but also the description of its emergent properties. The formal and

detailed description of a system is an essential step in order to acquire a full comprehension of its behavior. At the same time, it is of pivotal relevance the possibility to analyze the reactions of the system when determined stimuli or interferences are applied to the system itself.

Systems Biology adopts an approach based on the integration of the different biological knowledge on the analyzed system, and on the comprehension of how molecules interact in the network of processes giving raise to life. Due to the complexity of interactions among cellular mechanisms and due to the great number of involved components, it is impossible to intuitively and deeply understand the behavior of entire cellular networks [13].

The definition of a model is indeed a fundamental step in order to understand the process under examination: mathematical models and computer simulations have proven effective to investigate topology and dynamics of cellular processes (a significant review can be found in [11]). Moreover, the huge amount of biological data deriving from high-throughput and the impossibility to understand the system only describing interactions among molecules, justify the need for a systematic approach in modeling.

Mathematical models are able to represent biochemical systems in a realistic way under the aspect of their chemical, physical and biological behavior; they are able to integrate a great variety of empirical observations and generate new useful hypotheses.

Indeed, computational methods have a relevant advantage with respect to the traditional experimental techniques in biology, in terms of costs, saved time and increased usability. Furthermore, "experiments" that are not feasible in vivo can be realized *in silico* [14]. It is possible, for example, simulate knockouts on multiples essential genes and monitor their collective and individual effect on the cellular physiology. Clearly, these experiments can not be realized in vivo because the cell would not survive. The development of *in silico* predictive models give a great opportunity on one hand to control the system [14], on the other hand also to apply modeling techniques to test hypotheses on those components of the system that are not fully understood.

Computational analyses performed on biological systems are generally devoted to the comprehension of the following characteristics:

(1) the structure of systems through the characterization of genes, metabolism, signal transduction networks and physical structures; (2) system dynamics, i.e. their temporal evolution described through simulations (for which is essential the knowledge of chemical-physical parameters); (3) methods to control systems in order to understand the behavior of the system in response to perturbations; (4) methods to engineer desired properties

of systems, inducing, for example, the system to carry out features not present in nature (a discipline named Synthetic Biology).

To realize goals of Systems Biology it is necessary to integrate competencies and information deriving from different scientific domains such as: (1) genomics, proteomics and other molecular biology techniques, (2) computational studies in the domain of simulation, bioinformatic analyses applied to high-throughput and development of software tools, (3) technologies for wide, systematic and accurate measurements of chemical-physical parameters. This is undoubtedly a great multidisciplinary effort involving also a change of perspective in the design of studies in the different disciplines involved.

In order to define a mathematical model, a biological system has to be converted in an analogous *in silico* simplified system to facilitate the analysis, previsions, manipulation and optimization of the real system. The typical approach to build a mathematical model encompass eight phases (Figure 1.2) that defines a cycle of study [15] (or better a spiral of study) named hypothesis driven research.

The following workflow can be adopted to develop the model:

1. Data selection: the first phase of the developement requires the identification and the selection of data useful to assess if modeling goals have been reached. In this first phase is also essential to define which questions we want to answer with the study under planning: an efficient modeling process should increase the global knowledge on the model, allowing to formulate predictions on its functioning supported by experimental validation.

2. Defining the system structure and regulation through the analysis of scientific literature on the topic, and when possible, by means of de novo "wet" experiments. This phase could be particularly intricate due to the fact that the real topology and regulation mechanisms of biochemical reactions describing the network is not always clearly definable starting from literature data and additional analyses (if performed).

3. Definition of assumptions and simplifications: after having collected all the information on the biological model, the third phase is devoted to the integration of this information with admissible assumptions and simplifications to overcome the possible lack of knowledge on portions of the system. In this phase, interactions and components to be integrated in the model are also defined. In other words, during this phase it is necessary to define an adequate abstraction level. This choice will lead to the definition of models identified either as fine-grained (e.g., *mechanism-based*) or coarse-grained (e.g., *interaction-based* or *constraint-based*).

In Figure 1.3, it is presented an overview of these three main modeling approaches. This includes a list of features (quantitative *vs.* qualitative, static *vs.* dynamic, parameterized *vs.* non parameterized, single volume *vs.* compartmental, well-stirred *vs.* heterogeneous, etc.) that can be retrieved in each approach as suggested in [16]. In literature, mechanism-based approaches are used when dealing with small scale models (often defined as toy or core models), while interaction-based and constraint-based approaches, due to the limited need of parameters, are widely exploited for the analysis of large models (often defined genome-wide models). At the beginning, modeling is performed through the realization of diagrams where nodes represent components and links the interactions among them.

4. Selection of a framework for mathematical modeling: the mathematical formalism to describe the model depends on the questions to be answered with the modeling and on which methods can be applied to retrieve useful information. This step sees the conversion of the model from a diagram representation to a formal description in accordance to the selected mathematical formalism.
   Mechanism-based models (right in Figure 1.3) are widely recognized as the most powerful tool to understand biological processes due to their ability to reconstruct the dynamics of the system exploiting mainly the numerical integration of systems of differential equations. Unfortunately the difficulty to retrieve parameters limits the applicability of this approach for a large class of systems (e.g. genome-wide metabolic networks).
   Interaction-based models (left in Figure 1.3) are characterized by a network representation of biological systems and are designed to investigate qualitative and general properties (sometimes defined as "design principles"), using methods deriving from the graph theory or the topological analysis. These emerging properties are intriguingly considered transversal to different organisms [17] and can be seen as universal in the domain of life. The investigation of topological properties such as the presence of *hubs* (highly interconnected components), bottlenecks connecting hubs among them, and modularity can provide relevant information in many domains such as drug target discovery or sensitivity analysis [18].
   Constraint-based modeling framework (center in Figure 1.3) shares features both with interaction-based and mechanism based approaches. Even if constraint-based models are not able to predict the dynamics of the system, they are able to determine biological states (i.e. metabolic fluxes) thanks to the imposition of a pool of constraints (e.g. stoichiometric constraints ensuring the maintenance of the mass balance and thermodynamic constraints due to the reversibility of reactions). The outcome of this approach is often a (single) solution illustrating the flux through every reaction in the network, the so called "flux distribution".

5. Parameter estimation: following the formulation of the model, the fifth phase is committed to the definition of numerical values for the parameters of the system (e.g. molecular quantities, reaction constants, flux boundaries, etc.). The identification of these values is essential to determine if modeling outcomes are consistent with experimental observations. If these values are unknown (a quite usual case in Systems Biology), they should be identified through computational methods for parameter estimation [19].

6. Model accordance: the behavior of the model should be in accordance with experimental data, but a contrast with them indicates which further investigations should be performed. In the latter case it is necessary to identify if the contrast is generated by wrong hypotheses, simplifications, faulty structure of the model, inadequate experimental design or other factors not previously considered. The correctness of the reconstruction is evaluated against a data set used for model reconstruction in a process called cross-validation.

7. Model diagnostics: once the model has been correctly parameterized, the seventh phase is devoted to model diagnostics by performing analyses devoted to investigate model sensitivity in response to parameters variation and to determine properties such as oscillations, attractors or bistabilities.
   Lastly the sensitivity analysis shows how a parameter can influence the general behavior of the system. If the system is sensitive towards a given parameter, even small changes in its value could greatly affect the whole system. Sensitivity is related to robustness, a concept characterizing the ability of the system to remain in a solution space region where its behavior is not qualitatively varying, even if subject to a wide variation of chemical-physical parameters.

Modeling purpose In addition to the steps illustrated in this section, it is also to take into account that the goal of modeling and simulation of biological systems is (beyond the comprehension of dynamics and constitutive mechanisms of the system) the control of the system itself, governing or counteracting perturbations due to internal or external factors. The control of the system is of pivotal relevance in studies concerning the identification of therapeutic targets or the improvement in the production of determined chemical compounds involved in industrial processes.

In the present work of thesis, modeling and simulation have been exploited in order to widen knowledge and obtain insights on different biological systems such as signal transduction pathways, i.e. the Post-Replication Repair (analyzed in Chapter 2), or the energetic metabolism of yeast and mammalian cells (see Chapter 3).

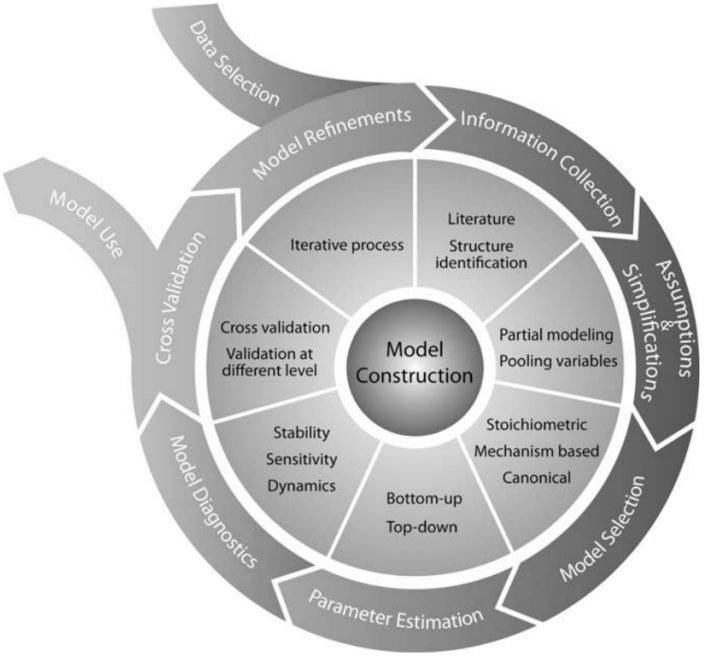


FIGURE 1.2: The "cycle" of the Systems Biology approach for a hypothesis driven research. Image from [15].

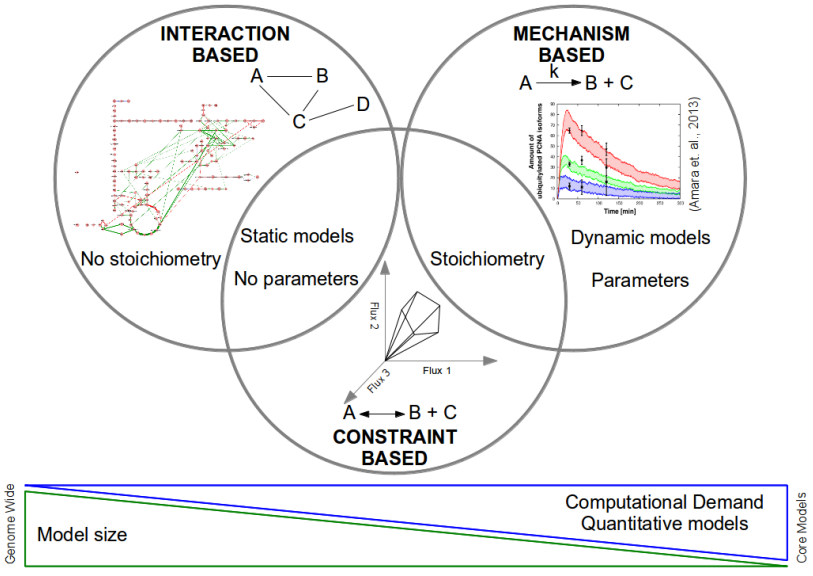


FIGURE 1.3: Schematic overview of the main modeling approaches for biological systems, together with their principal characteristics and differences. Moving from left to right: *interaction-based* approach, *constraint-based* approach and *mechanism-based* approach.

1.3 Purpose and organization of the thesis

This dissertation has been conceived with the purpose to describe a novel computational pipeline for a multi-level analysis of biological complex systems focusing in particular to the investigation of metabolic networks. The goal of the pipeline is to exploit qualitatively different approaches to gain a thorough comprehension of systems under examination.

The organization of this thesis is as follows. In Chapter 2 I will introduce computational approaches for the analysis of complex biological systems (Systems Biology) and I will give an example of signal transduction pathway modeling, moreover I will illustrate the outline of the computational pipeline. In Chapter 3 I will focus on constraint-based analyses and I will provide an application of novel developed methods on a metabolic network. In Chapter 4 the focus will be on interaction-based analysis applied to metabolism and in Chapter 5 I will describe methods for mechanism-based analyses along with a novel application to estimate kinetic parameters for dynamic modeling. Lastly, in Chapter 6, along with the discussion of the realized work, I will define some perspectives and future works.

Chapter 2

Background

2.1 Computational approaches in Systems Biology

According to a well established view presented in Chapter 1, computational approaches in Systems Biology can be divided in three different classes: interaction-based; constraint-based and mechanism-based. Each of these approaches is profitably exploited to model a set of biological systems [16]. In the following, I will briefly review these different classes and I will illustrate their application to two main biological systems: signal transduction pathways and metabolic networks.

Interaction-based modeling Is the Systems Biology approach requiring less amount of information due to the fact that it considers only how components are connected among them to form patterns, and eventually the topology of the network. Biological networks can be classified in some functional categories such as signal transduction networks, metabolic networks, transcriptional regulatory networks and protein-protein interaction networks.

Static interactions in these networks are investigated through graph theory tools [20]. In the context of metabolic networks a recent study underlined the fact that the scale-free topology of these networks (few nodes having many links that are connected to many nodes having few links) is a universal property in living organisms [21].

Graph-theoretical approaches have been proven useful to analyze some additional features such as the modular structure of biological networks or to investigate centralities and lethality in protein networks. It is also to underline that recently studies are rising some doubts on the relevance of results and general principles emerging from interaction-based analyses of biological networks [22], moreover this modeling approach (and the

area of the so called network biology) can be considered still at its infancy, and next years will provide advances both in theoretical and applicative studies.

Mechanism-based approaches Are the most powerful tools in Systems Biology due to their ability of predicting cellular dynamics revealing the precise amount molecular species over time. The most widely used approach in dynamic modeling exploits numerical integration of ordinary differential equations deriving from biochemical equations represented using the canonical chemical reaction scheme:

$$A + B \xrightarrow{k_1} C + D$$

In order to simulate the system, it is essential to determine the rate constant (k_1) and the initial concentration (or numerical amount) of each reactant (A, B, C, and D).

Unfortunately some difficulties emerges when dealing with this kind of models. First of all the structure of the system could be not fully understood, then it is likely that the numerical values of many parameters are not retrievable from literature or definable with "wet" experiments. For these reasons, together with the fact that mechanism-based methods are computationally intensives, the average size of biological systems investigated is rather small, concerning mainly signal transduction pathways.

A strategy to overcome difficulties raising from mechanism-based modeling is to integrate different autonomous mechanistic sub-models in a large consensus model. However, it is not univocally defined how to integrate these modules to obtain a coherent dynamic for the global model. Even if in recent years some steps in this direction has been made [23], dynamic simulations of large-scale cell-wide models is still one of the key challenges of Systems Biology.

Constraint-based approaches As introduced in Chapter 1, an intermediate approach between network-based and mechanism-based is represented by constraint-based approaches and in particular by Flux Balance Analysis (FBA) [24] which is generally regarded as the ancestor of these methods. In FBA, metabolic networks are investigated exploiting their network topology, stoichiometric information and constraints on reversibilities and allowable fluxes. These methods can be easily applied to genome-wide networks due to the fact that they do not require kinetic parameters and are computationally less demanding than mechanistic models.

Full genome sequencing is nowadays a routine technique and complete genomes for many organisms are already publicly available. These data integrated with other sources such

as proteomics or metabolomics allow to reconstruct metabolic models containing every known metabolic reaction (and some hypothesis to fill the gaps in the network when the knowledge is not available). Some outstanding works [25, 26] lead to the reconstruction of curated genome-scale metabolic models for human and the current challenge of Systems Biology in this field is to use constraint-based methods to predict the effect of metabolic perturbations.

In this chapter I will present some examples of Systems Biology approaches for the study of signal transduction pathways (Section 2.2.1) and metabolic networks (Section 2.3) which together represent a wide part of all the Systems Biology studies. In particular, in Subsection 2.2.1.1 I will present a case study published on *BMC Systems Biology* [27] where I proposed the first mathematical modeling approach for the Post-Replication Repair pathway exploiting stochastic simulations in the context of a mechanism-based model.

In Subsection 2.3.0.2 I will introduce some examples of metabolic reconstruction and constraint-based modeling in metabolism as identified in a review [28] where I investigated the state of the art on metabolic modeling.

In the last part of the chapter (Section 2.3.1) I will illustrate the limitations of these methods for the acquisition of a deep knowledge on biochemical systems deriving from the unique use of constraint-based approaches, and I will provide the structure of a computational pipeline which is the subject of the present thesis.

2.2 Systems Biology approaches

2.2.1 Signal transduction pathways

A wide variety of environmental and internal stimuli are constantly delivered in cells. The most common classes of stimuli are hormones, temperature, light conditions, osmotic pressure, changes in concentrations of substances like glucose, ions, cofactors (e.g. cAMP) or structural modifications as in the case of DNA damage. The detection and the response to these stimuli are performed through signal transduction pathways. Typically a signaling system is composed by a receptor and a ligand that bind to constitute the signal. This binding induces a change in the activity of the receptor that, in turn, induces a propagation of the signal through a signaling cascade eventually leading to an effector that produces a response. To modulate the effect of the signaling, the cell has to tightly control levels of every component inside it.

From the molecular point of view, signaling and metabolism involve similar processes like molecular modifications (e.g. post-translactional modifications: phosphorylation, methylation, acetylation, ubiquitylation), activation or inhibition of reactions and production or degradation of substances. Nevertheless, when modeling these two classes of biological systems, some differences have to be evaluated:

1. There is an evident difference in the nature of the transferring. Indeed signaling pathways deal with information transfering and processing, while metabolism is devoted to the transport of energy and cellular building blocks.
2. Metabolism is made up of well definable classes of objects, specifically, metabolites, cofactors and enzymes that catalyze reactions. Instead, signaling pathways involve a wider variety of objects often characterized by modularity, like molecular complexes, that can assemble or disassemble to modulate the signal.
3. In metabolism, reactions convert a great amount of biological material (in the range of 10^6, 10^{12} molecules per cell), while number of molecules taking part to signaling processes are usually not exceeding 10^2, 10^4 molecules per cell.
4. Different ratio between catalysts/receptors and substrates/effectors. In the case of signaling processes the ratio of receptors and effectors is close to 1, while in the case of metabolism the ratio sees a low number of catalysts used to transform a high number of substrates (justifying the quasi steady state assumption typical of constraint-based models).

Keeping in mind these peculiarities of signal transduction pathways, it is possible to envisage a computational approach based on mechanism-based modeling that has, however, to face lacks of knowledge (limited or incomplete) on components and relations inside signaling pathways as well as effects of signaling on the whole state of the cell that impose some choices for the determination of system boundaries. Lastly it is worth to underline that the interpretation of the simulation outcomes is context- and knowledge-dependent and therefore generalizations should be carefully evaluated.

All these issues have been faced when dealing with the development and analysis of a Post-Replication Repair model of yeast.

2.2.1.1 When a mechanistic model is enough: the Post-Replication Repair model of yeast

Genomic lesions are constantly generated in living organisms due to the exposition to a great variety of damaging agents inducing DNA lesions of different nature.

In living organisms, evolution lead to the development of several systems to repair or tolerate DNA damage and counteract its negative effects on genome stability. In this subsection I will investigate the Post-Replication Repair (PRR), the pathway involved in the bypass of DNA lesions induced by sunlight exposure and UV radiation. In cells there are two different mechanisms to activate PRR via a covalent modification of the Proliferating Cell Nuclear Antigen (PCNA), a sliding clamp enclosing DNA. As described in Figure 2.1 mono-ubiquitylation leads to a mechanism called translesion synthesis (TLS), while poly-ubiquitylation induces the template switching (TS).

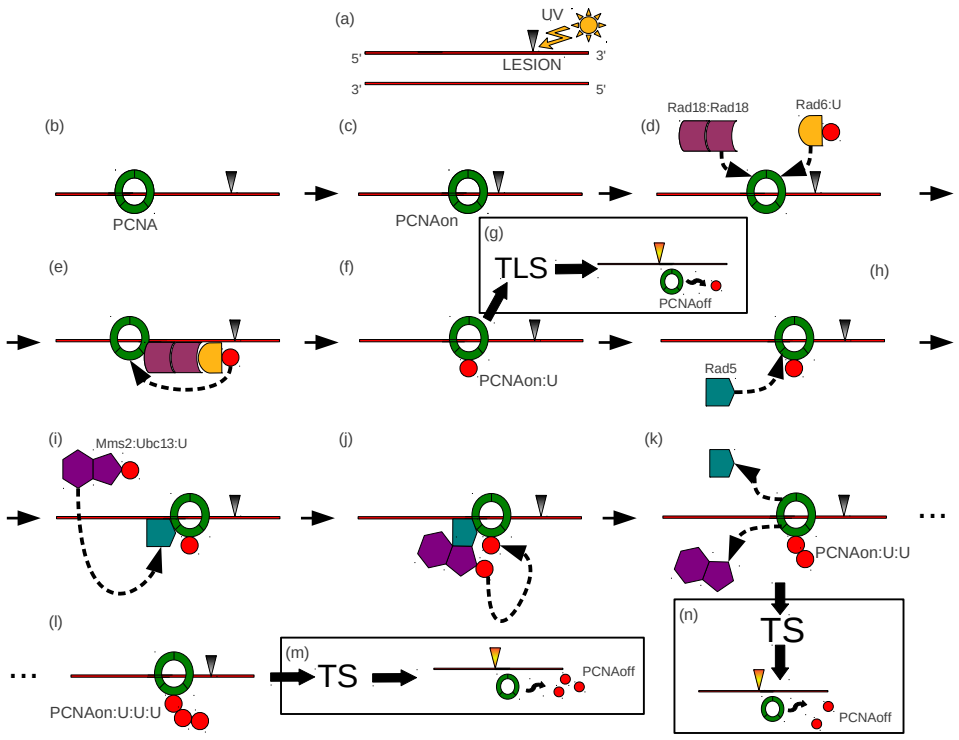


FIGURE 2.1: Graphical representation of the PRR pathway phases involved in the covalent modification of PCNA (mono- and poly-ubiquitylation) when activated by the UV-induced damage. (**a**) lesion on DNA (grey triangle) caused by the UV radiation; (**b**)(**c**) stall of the replication fork ($PCNA_{\text{on}}$); (**d**)(**e**)(**f**) PCNA is mono-ubiquitylation ($PCNA_{\text{on}}:U$) by Rad6 and Rad18; (**g**) mono-ubiquitylated PCNA can activate the Translesion DNA Synthesis sub-pathway (TLS), and hence lesion bypass and ubiquitylation signal switch-off ($PCNA_{\text{off}}$); (**h**)(**i**)(**j**) if TLS is not activated PCNA is poly-ubiquitylated by Ubc13-Mms2 and Rad5; (**k**)(**l**) poly-ubiquitylation is performed adding a single ubiquitin moiety at a time, repeating steps from (**h**) to (**k**); (**m**)(**n**) lesion bypass activated by the poly-ubiquitylated PCNA through the Template Switching sub-pathway (TS) and switch-off of the ubiquitylation signal. Figure from [27].

In [27] an approach combining *in vivo* and *in silico* studies has been used to investigate with a System Biology approach the events of PCNA ubiquitylation occurring in PRR in budding yeast cells.

The close synergy with a "wet" laboratory allowed to develop a novel *ad hoc* protocol to measure the time-course ratio between mono-, di- and tri-ubiquitylated PCNA isoforms on a single western blot[1]. Data emerging from the quantification of these western blots were used as the wet readout for PRR events in wild type and mutant *S. cerevisiae* cells exposed to acute UV radiation doses.

On the *in silico* side, a novel mechanistic model of PRR was defined on the basis of literature analysis. PCNA ubiquitylation dynamics obtained by means of stochastic simulations, evidenced a good agreement with experimental data at low UV doses (Figure 2.2), but indicated also a divergent behavior at high UV doses approximately higher than 30 J/m^2 (Figure 2.3).

This disagreement lead to the definition and realization of supplementary experiments to test the new hypothesis on the functioning of PRR.

In particular, this strategy allowed to define a UV dose for the saturation of the PRR system (leading to a stable steady state for all the analyzed PCNA isoforms) after which there is a malfunctioning in the error bypass process. Moreover, simulations shed light on an unpredicted overlap between PRR and Nucleotide Excision Repair (NER), the repair pathway known to fix UV-induced lesions during the G1 [29] and G2 [30] phases of the cell cycle. Strikingly, experimental evidences underlined that NER is required also for a proper S phase progression in response to UV irradiation.

Lasty, through a parameter sweep analysis (PSA) it was analyzed the effect of the size variation for the free ubiquitin pool (Figure 2.4), the model gave a potential explanation to the phenomenon of DNA damage sensitivity in yeast strains lacking deubiquitylating enzymes, highlighting the fact that ubiquitin concentration can affect the rate of PCNA ubiquitylation in PRR. Even if this findings suggest that the deubiquitylation of PCNA has a key role in the mechanism of ubiquitylation signal switch-off after the bypass of the damage, further *in vivo* and *in silico* investigation will be able to unveil molecular details of the preocess.

Globally, investigating for the first time the PRR pathway with a mathematical model, has been possible to determine how PRR is more complex and still far less characterized than previously thought; testifying at the same time the capabilities of the chosen Systems Biology approach.

[1]Western blot or immunoblot is a standard technique in biochemistry that allows to identify a protein of interest in a pool of proteins (deriving from a cellular extract) by means of a specific antibody.

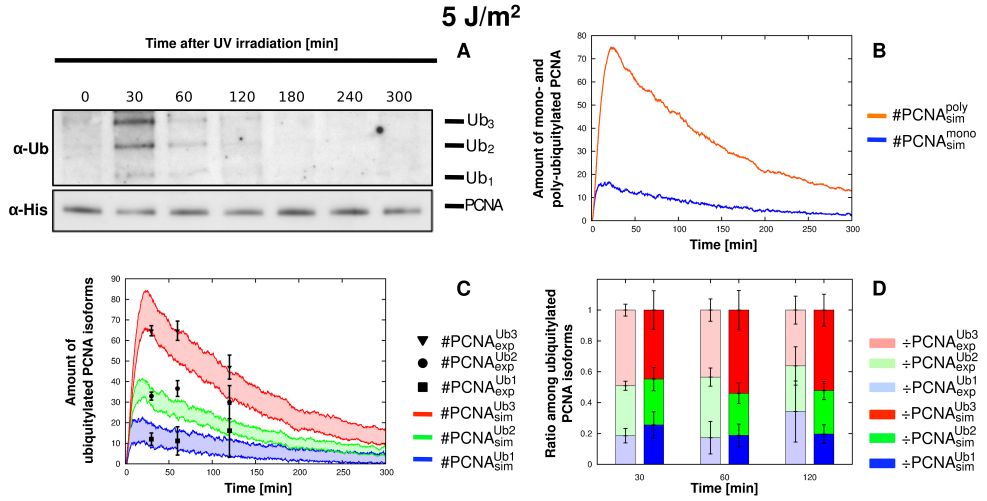


FIGURE 2.2: "Comparison between experimental and simulation results of PCNA ubiquitylation dynamics obtained on wild type yeast cells at 5 J/m^2 UV dose. The figure shows the experimental measurements on WT yeast cells irradiated at 5 J/m^2 UV dose and the comparison with the corresponding simulation results. (**A**) Representative image of a western blot showing a time-course measurement of mono-, di- and tri-ubiquitylated PCNA isoforms (top part, denoted by α-Ub) and of non modified PCNA (bottom part, denoted by α-His), sampled from 0 to 5 h after UV irradiation. The experiment was repeated 3 times. (**B**) Average dynamics of mono-ubiquitylated PCNA (blue line) and of poly-ubiquitylated PCNA (orange line), obtained from 100 independent stochastic simulations, executed starting from the same initial conditions and with an estimated number of DNA lesions equal to 1001. (**C**) Comparison between the mean dynamics of mono-, di- and tri-ubiquitylated PCNA isoforms emerging from 100 independent stochastic simulations, and the mean value of experimental data $\mu(\#\text{PCNA}^{\text{Ub}_\text{u}}_{exp})$, together with the respective standard deviation $\sigma(\#\text{PCNA}^{\text{Ub}_\text{u}}_{exp})$. Colored areas indicate the amplitude of stochastic fluctuations around the mean value $\mu(\#\text{PCNA}^{\text{Ub}_\text{u}}_{sim})$. Data are plotted by using the units representation. (**D**) Comparison between the ratio of experimental ($\div\text{PCNA}^{\text{Ub}_\text{u}}_{exp}$, left bars) and simulated ($\div\text{PCNA}^{\text{Ub}_\text{u}}_{sim}$, right bars) ubiquitylated PCNA isoforms at every sampled time point where experimental measurements yield a detectable amount of modified PCNA. Mean and standard deviation bars of both experimental and simulated ratios are plotted by using the normalized representation". Figure and caption from [27].

For what concerns mathematical methods, the the temporal evolution of the PRR pathway was simulated exploiting the stochastic algroithm tau-leaping [31] which is an efficient version of the well known stochastic simulation algorithm (SSA) [32] (see Section 5.1.2 for a more accurate description). In SSA, every reaction is executed sequentially, while in tau-leaping the speed-up is achieved through the parallel execution of several reaction steps. Simulations and analyses were performed exploiting the software

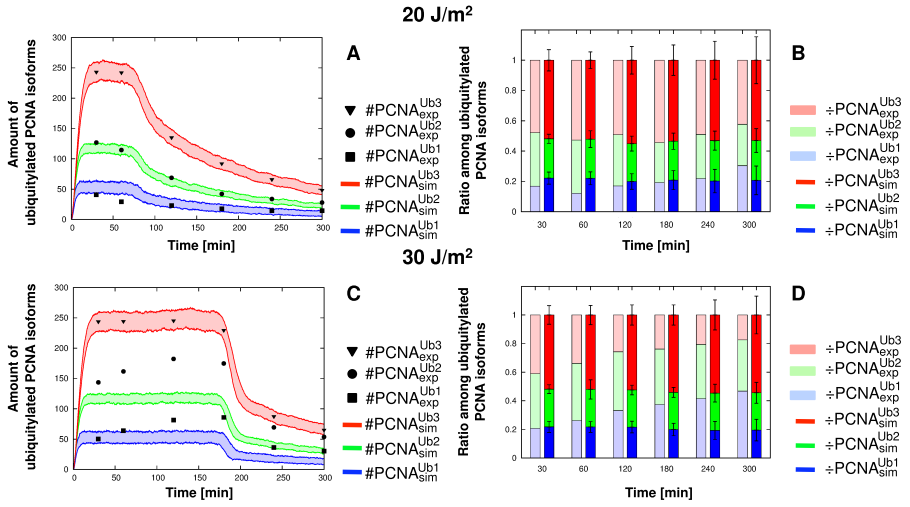


FIGURE 2.3: "Prediction of UV dose-dependent threshold and validation results on wild type yeast cells at 20 J/m^2 and 30 J/m^2 UV doses. The figure shows the experimental measurements on WT cells irradiated at 20 J/m^2 UV dose (top part) and at 30 J/m^2 UV dose (bottom part), as well as the comparison with the corresponding simulation results. As the aim of these experiments was not to carry out a precise quantification of the PCNA ubiquitylated isoforms, but only to verify the prediction of computational analysis, they were conducted with a single repetition. (**A-C**) Comparison between the value of western blot quantification $\#\text{PCNA}^{\text{Ub}_\text{u}}_{exp}$ deriving from a single experiment, and the dynamics of mono-, di- and tri-ubiquitylated PCNA isoforms $\#\text{PCNA}^{\text{Ub}_\text{u}}_{sim}$ emerging from 100 independent stochastic simulations, executed starting from the same initial conditions, with an estimated number of DNA lesions equal to 4005 (**A**) and 6007 (**B**). Colored areas indicate the amplitude of stochastic fluctuations around the mean value $\mu(\#\text{PCNA}^{\text{Ub}_\text{u}}_{sim})$. Data are plotted by using the units representation. (**B-D**) Comparison between the ratio of experimental ($\div\text{PCNA}^{\text{Ub}_\text{u}}_{exp}$, left bars) and simulated ($\div\text{PCNA}^{\text{Ub}_\text{u}}_{sim}$, right bars) ubiquitylated PCNA isoforms at every sampled time point. Mean and standard deviation bars of simulated results are plotted by using the normalized representation". Figure and caption from [27].

BioSimWare [33] which combine an efficient implementation of SSA and tau-leaping, with a handy graphical user interface.

In addition, a PSA was performed thanks to an *ad hoc* developed computational tool that generates a set of different initial conditions for the model and then automatically executes the corresponding stochastic simulations exploiting the tau-leaping algorithm. From the biological point of view, the PSA was used to investigate effects on the dynamics of the PRR pathway due to the variation of reaction constants, molecular amounts and the number of DNA lesions within a specified range with respect to a fixed reference

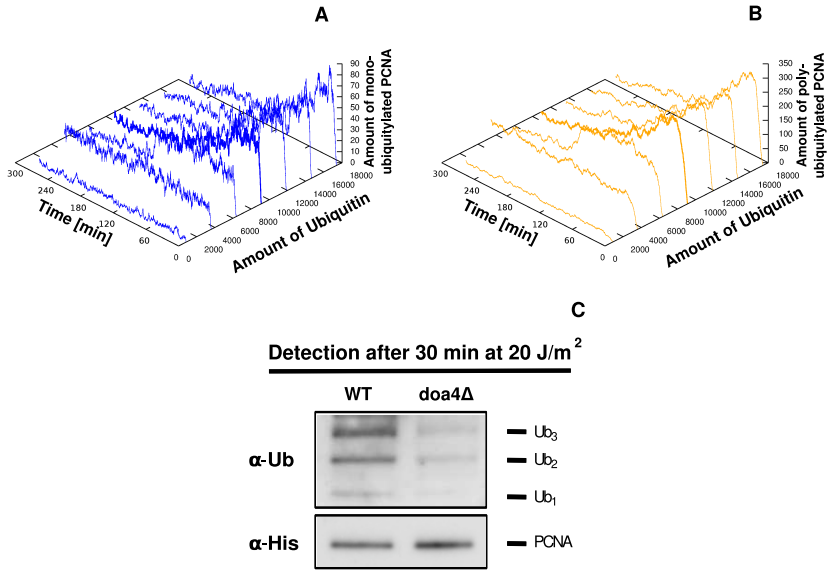


FIGURE 2.4: "Influence of free ubiquitin concentration and validation results on *doa4*Δ background yeast cells at 20 J/m^2 UV dose. The figure shows the simulated dynamics of PCNA mono-ubiquitylation (**A**) and poly-ubiquitylation (**B**) at a UV dose of 20 J/m^2, obtained from a PSA executed on the initial amount of ubiquitin, which is varied in the interval [870, 17396] molecules – mimicking the biological conditions ranging from a 10-fold reduction (corresponding to the severely impaired condition of *doa1*Δ yeast cells) to a 2-fold overexpression of the total amount of free ubiquitin in WT cells. In the plots, the thick lines correspond to the dynamics obtained with the reference value for ubiquitin amount. The simulations show that for ubiquitin amounts lower than the reference value, the amounts of mono- and poly-ubiquitylated PCNA decrease, as also observed experimentally in *doa4*Δ cells (**C**, right part). On the other hand, by increasing the ubiquitin amount occurring in the system, the dynamics show an initial peak in the amount of mono- and poly-ubiquitylated PCNA, suggesting that high amounts of ubiquitin might lead the system to a faster bypass of all lesions with respect to the physiological reference value. (**C**) Western blot showing a comparison between time-course measurements in WT yeast cells (left part) and *doa4*Δ yeast cells (right part) of the mono-, di- and tri-ubiquitylated PCNA isoforms (top part, denoted by α-Ub) and of non modified PCNA (bottom part, denoted by α-His), sampled at 30 min after UV irradiation. As the aim of these experiments was not to carry out a precise quantification of the PCNA ubiquitylated isoforms, but only to verify the prediction of computational analysis, they were conducted with a single repetition". Figure and caption from [27].

value. More precisely, the PSA varied one parameter at a time (OAT) using a linear scale sampling for molecular amounts and a logarithmic sampling scale for reaction constants. This last sampling for reaction constants was used to uniformly evaluate several orders of magnitude in order to mimic several different experimental conditions.

In the context of PRR, PSA was performed in order to assess the soundness of the

parametrization used in the model (values identified for stochastic constants and molecular amounts). In particular the PSA was performed varying parameters in the following way:

- the value of each stochastic constant was varied of 3 orders of magnitude above and 3 below the reference value manually estimated with a "trial and error" procedure;
- the value of the initial molecular amounts (at the beginning of the simulation) in the system was varied in a range between 0 and twice the reference value identified in literature, thus mimicking the biological conditions ranging from the deletion to a 2-fold overexpression of the initial species.

Besides the PSA, the PRR model was further analyzed by means of a global sensitivity analysis (SA) in order to understand how much the variation of the model input factors (molecular species amounts, kinetic constants, etc.) determines the uncertainty in the model outcome, and to identify effective control points for the dynamics of the system through the determination of which input factors cause jointly the most striking effects on the system behavior.

The SA on the values of stochastic constants was performed using a screening test called "method of the elementary effects" (EE), as described in [34–36]. This method allows to investigate how a specified model outcome changes according to a perturbation of the model input factors and this is realized by varying one input factor at a time while keeping all the others fixed.

The elementary effect can be defined as the ratio between the variation in the model output and the variation in the input factor itself. To compute global sensitivity measures, several elementary effects are estimated and averaged. These calculated measures are: the value μ^* (i.e the module of the mean of the distribution of the elementary effects), which determines the global influence of each factor on the model output, and the value σ^* (i.e. the standard deviation of the distribution of the elementary effects), which quantifies the ensemble of the factor's higher order effects.

Concerning the SA of the PRR model, I computed the sensitivity measures μ^* and σ^* by considering as input factors the set of kinetic constants associated to the model reactions and varying them over 4 orders of magnitude, 2 below and 2 above the reference value. Molecular amounts of mono- and poly-ubiquitylated PCNA isoforms (uniformly measured every 9 seconds along *in silico* simulations of the dynamics of PRR over 5 hours at the "low" 10 J/m^2 UV dose), were used as model outcomes to compute the EEs.

The strategy presented in [36] was used to sample the variation interval of each reaction constants in order to define the set of points of the parameter space used to compute the EE and hence the sensitivity measures. In particular, in this strategy the Sobol's quasi random numbers [37] allowed to obtain a radial sampling (log-scaled over the variation interval) leading to a set of a_i points, $i = 1, \ldots, 1000$, corresponding to the centers of the radial samplings. For each a_i a variation along the 25 dimensions (one for each reaction constant) of the input factor space was considered to compute the EEs, yielding a total of 1000·(25+1) different model parameterizations.

Reaction	μ^*	σ^*
1	1.2146	2.1099
4	$9.6114 \cdot 10^{-2}$	$2.2428 \cdot 10^{-1}$
12	$1.3929 \cdot 10^{-3}$	$1.1950 \cdot 10^{-2}$
18	$8.9356 \cdot 10^{-6}$	$1.1310 \cdot 10^{-4}$
23	$4.6616 \cdot 10^{-6}$	$1.6221 \cdot 10^{-5}$
9	$2.9883 \cdot 10^{-6}$	$9.2425 \cdot 10^{-6}$
25	$2.9548 \cdot 10^{-6}$	$1.4706 \cdot 10^{-5}$
17	$1.7377 \cdot 10^{-6}$	$1.5368 \cdot 10^{-5}$
20	$1.3714 \cdot 10^{-6}$	$8.2233 \cdot 10^{-6}$
21	$8.7619 \cdot 10^{-7}$	$4.4432 \cdot 10^{-6}$
22	$4.3180 \cdot 10^{-7}$	$2.9323 \cdot 10^{-6}$
2	$2.2217 \cdot 10^{-7}$	$1.8861 \cdot 10^{-6}$
7	$7.3924 \cdot 10^{-8}$	$5.0348 \cdot 10^{-7}$
13	$3.6523 \cdot 10^{-8}$	$5.4385 \cdot 10^{-7}$
8	$2.2538 \cdot 10^{-8}$	$1.2463 \cdot 10^{-7}$
15	$2.1505 \cdot 10^{-8}$	$4.7824 \cdot 10^{-7}$
24	$1.1074 \cdot 10^{-8}$	$1.2645 \cdot 10^{-7}$
14	$6.0435 \cdot 10^{-9}$	$4.8142 \cdot 10^{-8}$
19	$5.2410 \cdot 10^{-9}$	$3.4058 \cdot 10^{-8}$
10	$4.9045 \cdot 10^{-10}$	$4.4654 \cdot 10^{-9}$
16	$1.0037 \cdot 10^{-11}$	$2.0655 \cdot 10^{-10}$
3	$2.8808 \cdot 10^{-13}$	$3.5596 \cdot 10^{-12}$
6	$2.4153 \cdot 10^{-13}$	$2.4214 \cdot 10^{-12}$
5	$1.3459 \cdot 10^{-14}$	$1.2814 \cdot 10^{-13}$
11	$6.5943 \cdot 10^{-23}$	$1.7684 \cdot 10^{-21}$

Reaction	μ^*	σ^*
1	$2.4371 \cdot 10^{-1}$	$6.6924 \cdot 10^{-1}$
4	$2.0837 \cdot 10^{-2}$	$7.3272 \cdot 10^{-2}$
12	$5.9092 \cdot 10^{-3}$	$3.1596 \cdot 10^{-2}$
23	$3.5319 \cdot 10^{-5}$	$1.0420 \cdot 10^{-4}$
18	$3.9620 \cdot 10^{-6}$	$5.7120 \cdot 10^{-5}$
21	$2.3807 \cdot 10^{-6}$	$1.2263 \cdot 10^{-5}$
20	$1.5264 \cdot 10^{-6}$	$8.7254 \cdot 10^{-6}$
17	$1.0145 \cdot 10^{-6}$	$8.4656 \cdot 10^{-6}$
9	$6.6420 \cdot 10^{-7}$	$3.1911 \cdot 10^{-6}$
22	$3.0728 \cdot 10^{-7}$	$2.3997 \cdot 10^{-6}$
2	$6.5226 \cdot 10^{-8}$	$7.9826 \cdot 10^{-7}$
13	$4.0320 \cdot 10^{-8}$	$5.4867 \cdot 10^{-7}$
7	$1.7669 \cdot 10^{-8}$	$1.3968 \cdot 10^{-7}$
14	$1.5766 \cdot 10^{-8}$	$1.1217 \cdot 10^{-7}$
15	$1.2015 \cdot 10^{-8}$	$2.1648 \cdot 10^{-7}$
25	$1.0743 \cdot 10^{-8}$	$1.0942 \cdot 10^{-7}$
8	$8.3652 \cdot 10^{-9}$	$6.7295 \cdot 10^{-8}$
19	$5.4418 \cdot 10^{-9}$	$3.7516 \cdot 10^{-8}$
24	$4.8583 \cdot 10^{-9}$	$5.4275 \cdot 10^{-8}$
10	$4.8308 \cdot 10^{-10}$	$3.7435 \cdot 10^{-9}$
16	$1.6753 \cdot 10^{-11}$	$1.5816 \cdot 10^{-10}$
3	$1.4687 \cdot 10^{-13}$	$3.0104 \cdot 10^{-12}$
6	$7.6793 \cdot 10^{-14}$	$7.5556 \cdot 10^{-13}$
5	$4.3665 \cdot 10^{-15}$	$6.0235 \cdot 10^{-14}$
11	$8.1343 \cdot 10^{-23}$	$1.9211 \cdot 10^{-21}$

TABLE 2.1: Ranking of model reactions according to μ^* for the mono-ubiquitylation of PCNA (*left*) and poly-ubiquitylation of PCNA (*right*). Table from [27].

In Table 2.1 the values of μ^* and σ^* are listed for all reaction constants of mono- (left) and poly-ubiquitylation (right) outputs. The ranking of reactions is here given following decreasing values of μ^*. Same orders of magnitude of μ^*, are identified in the table by horizontal blocks.

The SA was performed exploiting the LSODA simulation algorithm [38] on a deterministic version of the PRR model equivalent to the stochastic one and derived, according to the methodology described in [32, 39], from the reactions graphically illustrated in Figure 2.1. This choice has been determined due to the huge computational time required to perform the high number of necessary stochastic simulations to perform the SA on the PRR model (about $2.6 \cdot 10^8$, assuming 10^4 independent simulations for each sample to calculate the required histogram distance [40]).

In Figure 2.5 are shown values of the measure μ^* of all reaction constants for the two model outputs – mono-ubiquitylation (*top plot*) and poly-ubiquitylation (*bottom plot*) of PCNA – and the ranking of reactions according to decreasing values of μ^*. In each plot, the inset represents the ranking on a log-scale.

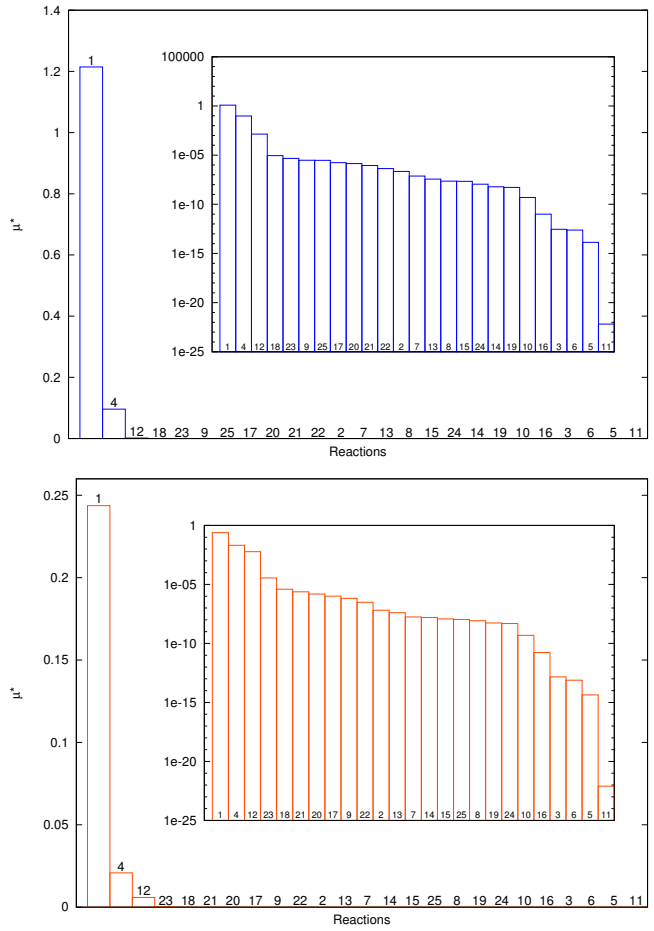


FIGURE 2.5: Sensitivity indexes μ^* and σ^* for mono- and poly-ubiquitylated isoforms. Figure from [27].

Analyzing plots Figure 2.5 and the associated tables in Table 2.1, it is possible to determine that, in response to the variation of considered input factors, the most sensitive reactions (and hence the associated rate constants) are:

- Reaction 1, which corresponds to the identification of the UV-induced lesion on DNA. The high global sensitivity of this reaction is motivated by its role in the activation of the whole pathway;
- Reaction 4, which corresponds to the loading of ubiquitin on Rad6. This is a key step for PCNA mono-ubiquitylation and, due to the stepwise mechanism of

ubiquitylation considered in the model, it also influences the downstream binding of additional moieties to PCNA (i.e., PCNA di-ubiquitylation and tri-ubiquitylation) and, therefore, it affects the whole process of PCNA ubiquitylation;

- Reaction 12, which corresponds to the loading of ubiquitin on Rad5. This is crucial for PCNA di-ubiquitylation and tri-ubiquitylation, which is less sensitive in the mono-ubiquitylation output than in the mono-ubiquitylation output since its role is downstream the first steps of the PRR pathway.

Intriguingly, the two model outcomes are only marginally influenced by other reactions.

The consistency of the PRR model was validated by comparing the outcome of stochastic simulations with the experimental measurements carried out on the wild type (WT) yeast strain at various UV doses. To this aim, by considering the western blots at each UV irradiation dose, I first quantified the values of mono-, di- and tri-ubiquitylated PCNA ratios, together with the respective mean and standard deviation of each PCNA isoform.

Then, from the outcome of stochastic simulations I derived the molecular amounts of PCNA isoforms. In particular, to tame the effect of stochastic fluctuations that are inherent in these computational analysis, I exploited the outcomes of a set of independent simulations (performed with the same initial conditions) to calculate the mean and standard deviation ($\#\mathrm{PCNA}^{\mathrm{Ub_u}}_{sim}$) of PCNA amounts.

Afterwards, since I had to compare different kinds of measurements – namely, ratios of modified PCNA derived from laboratory experiments on the one side, and molecular amounts of modified PCNA obtained from stochastic simulations on the other side – I introduced two different strategies for the graphical representation and comparison of the experimental and the computational results:

1. the first strategy, referred as "normalized representation" (NR), consists of stacked bar graphs: for each sample analyzed within the time interval of 0-5 h, the stacked bars corresponding to the normalized ratios of mono-, di- and tri-ubiquitylated PCNA isoforms obtained from stochastic simulations (denoted by $\div\mathrm{PCNA}^{\mathrm{Ub_u}}_{sim}$) are plotted side by side to the experimental bars $\div\mathrm{PCNA}^{\mathrm{Ub_u}}_{exp}$ (which, as stated above, are already expressed as ratios);

2. in the second strategy, referred as "units representation" (UR), the molecular amounts derived from stochastic simulations $\#\mathrm{PCNA}^{\mathrm{Ub_u}}_{sim}$ are compared to the western blot quantifications which, in this case, were specifically transformed into molecular quantities (denoted by $\#\mathrm{PCNA}^{\mathrm{Ub_u}}_{exp}$).

It is worth to stress the fact that the NR allows a direct comparison between the experimental and simulation results, by considering the ratio of the three ubiquitylated isoforms of PCNA with respect to the total amount of modified PCNA measured in the system. Anyway, this strategy does not give any knowledge on the actual amount of modified PCNA, and it does not allow to clearly evidence the switch-off of PCNA ubiquitylation signal as long as the DNA lesions get processed, which can be instead directly represented by using the UR.

2.3 Metabolism

A metabolic network is an abstract representation of cellular metabolism, that is the complex system of anabolic and catabolic reactions sustaining cell survival, growth and proliferation.

Due to the multiple tasks that a cell has to face, in every single instant, a great number of reactions takes place inside a cell. As a consequence the network of metabolic reactions reaches more than a thousand of reaction even for simple eucaryotic cells like yeast. Moreover due to the fact that the product of a reaction is often used as substrate of another reaction, metabolic networks are also deeply interconnected.
Formally, metabolic networks are represented exploiting directed graphs due to the fact that biochemical reactions have a directionality. In a metabolic network it is possible to attribute different meanings to nodes, and to every meaning is associated a determined type of graph [41] (see Figure 2.6):

(A) in a substrate graph, nodes represent reagents of a given reaction and links connect nodes if they take part to the same reaction;

(B) in reaction graph, nodes correspond to reactions ad a link is drawn between two reactions if a metabolite is both the product of a reaction and the substrate of another reaction, in other words reactions must share at least a common compound;

(C) in enzyme-centric graph, nodes correspond to enzymes and two enzymes establish a link if a metabolite is the product of a reaction catalyzed by the other enzyme;

(D) in substrate-enzyme bipartite graph, two types of nodes are used to represent respectively reactions and metabolites. In this last case, links connect nodes having different nature, representing both relations of substrates and products.

As stated in Section 2.1 the main computational framework for the analysis of metabolic networks is constraint-based modelling as it will also widely described in Chapter 3.

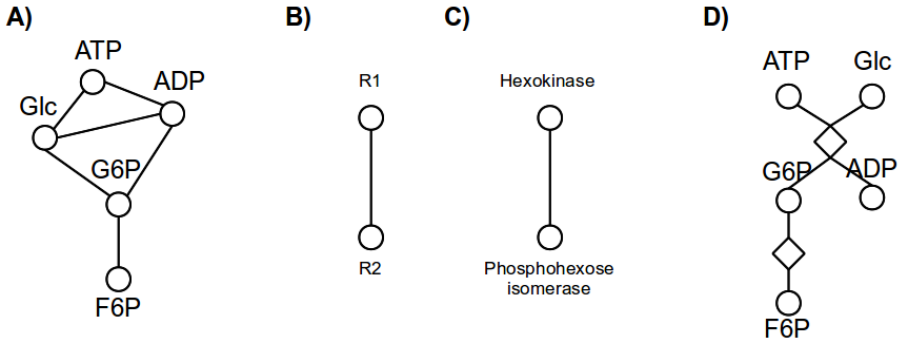


FIGURE 2.6: Metabolic network representations. A) substrate graph; B) reaction graph; C) enzyme-centric graph; D) substrate-enzyme bipartite graph.

However, the first step for this kind of analysis is the definition (or reconstruction) of a metabolic network that usually is in the form of a substrate graph [42].

2.3.0.2 Metabolic networks reconstruction

Formal representations of metabolism are typically created in a bottom-up fashion based on genomic and bibliomic data, which can possibly be integrated with data obtained from laboratory experiments. Network reconstructions can vary in size (from genome-wide networks to smaller – core – models focusing on specific metabolic pathways), and can be characterized by different levels of abstraction, according to the scope of their formulation. The generation of networks derived from top-down approaches (inference of component interactions based on high-throughput data) solicited the development of automatic reverse engineering methods, to devise a plausible network of biochemical reactions that is able to reproduce the experimental observations. Reverse engineering methods are usually applied to infer core models characterized by a small number of molecular species and biochemical reactions [43].

In Chapter 3 I will discuss in detail the process to reconstruct genome-wide and core models of metabolism with the perspective of their use for quantitative computational analyses. Nevertheless, it should be mentioned that the starting point for the development of the genome-wide models mentioned hereby are, often, metabolic network reconstructions that are intended to summarize the available experimental knowledge and may be studied with qualitative methods (e.g., the topological analysis described in Chapter 4). These manually curated maps are collected in various databases, such as EcoCyc [44], HinCyc[45] and KEGG [46].

2.3.1 Partial views from current approaches

All the three different frameworks proposed by Stelling in [16] have been applied to the modelling of metabolism. Table 2.2 gives an overview of recent works appeared in literature and dealing with the modeling of metabolism at different scales and with different goals. As it is possible to notice, FBA and ODEs are the most widely used methods. However, all these approaches give, when used on their own, only partial views on the comprehension of metabolism.

For example, constraint-based methods, and markedly FBA, are not able to determine a unique flux distribution due to the structure of the mathematical approach itself (further details will be given in Chapter 3) and, moreover, the "real" flux distribution in the cell is subject to regulatory mechanisms involving kinetic characteristics determined by enzyme expression/modification. It has also been verified that outcomes of FBA can disagree with experimental data when regulatory loops are not properly managed [47].

The most relevant point is that constraint-based models, assuming steady-state conditions, can not be used to model the dynamic behaviour of the metabolic system. This is particularly relevant because the transient of metabolic systems has an important role in the understanding of the response of cells to environmental changes. Indeed there is an increasing evidence that cellular metabolism is a highly dynamic system. A fact testified by the increasing number of publications on this topic. As an example it is worth to remember dynamics emerging from temporal variations in the concentrations of metabolic intermediates of the yeast glycolytic pathway [48, 49].

Connected to this point it is also to consider that cells may be in a suboptimal metabolic condition and this is in contrast with the fundamental assumption of FBA of the optimization towards an objective function (usually, maximization of biomass).

With regard to interaction-based approaches, even if they showed the ability to highlight relevant emergent and general properties of metabolism [17], it is to underline that these approaches showed many severe limitations. Among these, the most relevant is that universal properties are not enough to determine the dynamics and the regulation of the system. As already examined in Section 2.2.1, many aspects of metabolic networks are inherently different from other biological networks (e.g. with signaling pathways).

Finally, due to the structure of metabolic networks where both reactions and metabolites has to be drawn and investigated, several representation have been proposed (see 2.3). Strikingly, none of the them can be identified as the definitive one, suggesting the necessity of the definition/use of more advanced methods than graph theory for their analysis.

As stated in 2.1, the mechanism-based modeling of metabolism is usually seen as the end point for the understanding of cellular metabolism. However in this context, mechanistic modelling has a limited applicability due to the chronically lack of parameters such as enzyme kinetics and reliable measurements for metabolites concentrations. This is worsen by the fact that metabolic models usually encompass thousands of different metabolites and reactions making detailed kinetic modeling almost unfeasible due to high computational requirements even for some core models.

Pathway / aim of the model	Cell type / organ	Organism	Modeling approach & methodology	Exp. data	Reference
Glycolysis	-	*T. brucei*	CM, ODE	L	Achcar *et al.* [50]
GW metabolic network and succinic acid production	-	*S. cerevisiae*	GW, FBA	M	Agren *et al.* [51]
GW metabolic network	-	*A. niger*	GW, FBA	L	Andersen *et al.* [52]
Mitochondrial energy metabolism, Na^+/Ca^{2+} cycle, K^+ cycle	Heart, liver	*B. taurus*, *S. scrofa*, *R. norvegicus*	CM, DAE, PE, SA	L, M	Bazil *et al.* [53]
OXPHOS	Cardiomyocytes	*R. norvegicus*	CM, ODE	L	Beard [54]
Electron transport chain	Heart homogenates	*R. norvegicus*	CM, ODE, CRL	L, M	Chang *et al.* [55]
Glycolysis, OXPHOS	Not specified	Eukaryotic, *H. sapiens*	CM, Control theory	L	Cloutier *et al.* [56]
Bow-tie architecture of metabolism	Not specified	*H. sapiens*	GW, Topological analysis	L	Csete *et al.* [57]
Central metabolism	-	Yeast	CM, FBA	L	Damiani *et al.* [58]
Energy metabolism	Skeletal muscle cell	Mammal	CM, PDE	L	Dasika *et al.* [59]
Glycolysis and pentose phosphate pathway	-	*E. coli*	CM, ODE, SA	L	Degenring *et al.* [60]
Biosynthesis of valine and leucine	-	*C. glutaminicum*	CM, ODE, SDE	M	Dräger *et al.* [61]
GW metabolic network	Not specified	*H. sapiens*	GW, FBA	L	Duarte *et al.* [25]
GW metabolic network	-	*E. coli* MG1655	GW, FBA	M	Edwards and Palsson [62]
GW metabolic network	-	*H. influenzae*	GW, FBA	L	Edwards *et al.* [63]
Anabolic, catabolic, chemiosmosis pathways	-	*E. coli*	GW, Control theory	M	Federowicw *et al.* [64]
Small world behavior of metabolism	Not specified	*H. sapiens*	GW, Topological analysis	L	Fell *et al.* [65]
Cancer metabolic networks	Various (NCI-60 collection)	*H. sapiens*	Network reconstruction, FBA, gene (pair) analysis	L	Folger *et al.* [66]
GW metabolic network HepatoNet1	Hepatocytes	*H. sapiens*	GW. Network reconstruction	L	Gille *et al.* [67]
Cytochrome bc_1 complex, ROS production	Muscle, heart, liver, kidney, brain	*R. norvegicus*	CM, ODE	L	Guillaud *et al.* [68]
GW metabolic network EHMN	Not specified	*H. sapiens*	GW, Network reconstruction	L	Hao *et al.* [69]
GW metabolic network	-	*S. cerevisiae* S288c	GW, Network reconstruction, FBA	L	Heavner *et al.* [70]
GW metabolic network	-	*S. cerevisiae*	Network reconstruction	L	Herrgård *et al.* [71]
Topological properties of metabolism	-	43 different organisms	GW, Topological analysis	L	Jeong *et al.* [17]
Glycolysis, OXPHOS	-	Not specified	CM, ODE, Game theory	-	Kareva [72]
Whole-cell life cycle model	-	*M. genitalium*	GW, FBA, ODE	L, M	Karr *et al.* [23]
Glycolysis, pentose phosphate pathway	-	*T. brucei*	CM, ODE	L	Kerkhoven *et al.* [73]
Energy metabolism	Colorectal cells	*H. sapiens*	CM, FBA, EM	M	Khazaei *et al.* [74]
GW metabolic network	-	*Synechocystis* sp. PCC 6803	GW, FBA	L	Knoop *et al.* [75]
Glycolysis, gluconeogenesys, glycogen metabolism	Hepatocytes	*H. sapiens*	CM, ODE	L	König *et al.* [76]
Adenine nucleotide translocase	Heart mitochondria	*B. taurus*	CM, ODE, PE, SA	L	Metelkin *et al.* [77]
GW metabolic network	-	*Z. mays* L. subsp. *mays*	GW, Network reconstruction	L	Monaco *et al.* [78]
Xylose metabolism	-	*L. lactis IO-1*	CM, ODE, SA	M	Oshiro *et al.* [79]
GW metabolic network	-	*S. cerevisiae*	GW, Network reconstruction, FBA	L	Österlund *et al.* [80]

Continued on next page

Pathway / aim of the model	**Cell type / organ**	**Organism**	**Modeling approach & methodology**	**Exp. data**	**Reference**
GW metabolic network and succinic acid production	-	*S. cerevisiae*	GW, FBA	M	Otero *et al.* [81]
Topological properties of metabolism	-	43 different organisms, *E. coli*	GW, Topological analysis	L	Ravasz *et al.* [82]
One-carbon metabolism, trans-sulfuration pathway, synthesis of glutathione	Hepatocyte	*H. sapiens*	CM, ODE	L	Reed *et al.* [83]
Glycolysis, TCA cycle, pentose phosphate pathway, glutaminolysis, OXPHOS	HeLa cell	*H. sapiens*	CM, FBA	M	Resendis-Antonio *et al.* [84]
Modularity of metabolism	Not specified	*H. sapiens*	GW, Topological analysis	L	Resendis-Antonio *et al.* [85]
GW metabolic network	Not specified	*H. sapiens*	GW, Network reconstruction	L	Sahoo *et al.* [86]
Acetone, butanol and ethanol production	-	*C. acetobutylicum*	CM, ODE, SA	M	Shinto *et al.* [87]
Cancer metabolic networks	Various (NCI-60 collection)	*H. sapiens*	FBA	L	Shlomi *et al.* [88]
GW metabolic network	-	*S. cerevisiae*	GW, FBA	L	Simeonidis *et al.* [89]
Glycolysis	-	*S. cerevisiae*	CM, ODE	M	Teusink *et al.* [90]
GW metabolic network	Not specified	*H. sapiens*	GW, FBA	L	Thiele *et al.* [26]
Primary metabolism	-	*E. coli*	CM, ODE, EM	-	Tran *et al.* [91]
Fueling reaction network	-	*E. coli W3110*	CM, FBA	M	Varma *et al.* [92]
Reduced model of cell metabolism	-	-	CM, FBA	L	Vazquez *et al.* [93]
Small-world property of metabolism	-	*E. coli*	GW. Topological analysis	L	Wagner *et al.* [94]
GW metabolic network	-	*C. glabrata*	GW, FBA	L	Xu *et al.* [95]
Erythrocyte metabolism	Red blood cell	*H. sapiens*	Hybrid: ODE + MFA	-	Yugi *et al.* [96]
Mitochondrial energy metabolism	Various tissues	Mammal	CM, ODE	-	Yugi [97]
Modularity of metabolism	Not specified	*H. sapiens*	GW, Topological analysis	L	Zhao *et al.* [98]
ROS-induced ROS release in mitochondria network	Cardiomyocytes	*C. porcellus*	CM, ODE, PDE, RD, Finite Difference Method	M	Zhou *et al.* [99]

TABLE 2.2: Overview of some recent literature papers on the modeling and computational analysis of metabolism.
Abbreviations. CM: Core model; CRL: Chemiosmotic Rate Law; DAE: Differential Algebraic Equations; EM: Ensemble modeling; FBA: Flux Balance Analysis; GW: Genome-wide model; L: experimental data obtained from literature; M: experimental data measured with *ad hoc* experiments; MFA: Metabolic Flux Analysis; ODE: Ordinary Differential Equations; PDE: Partial Differential Equations; PE: Parameter Estimation; SA: Sensitivity Analysis; SDE: Stochastic Differential Equations. Table from [28].

Multi-level analysis to unravel complexity Due to the complex nature of biologic processes, in silico methods should consider multiple approaches to investigate systems. Multi-level analysis is today a hot research topic in different areas, such as the theoretical formalization of the method and the development of computational tools for the integration of different modeling perspectives [100]. To fulfill the need of multi-level approaches, I am developing a computational pipeline (see Figure 2.7) able to perform analyses exploiting, one after the other, three main modeling frameworks for biological systems: constraint-based analysis, interaction-based analysis and mechanism-based analysis [16].

On the whole, results emerging from the performed analyses will give a synoptic vision of the different properties of the system. In order to validate the developed method and the computational pipeline proposed, I defined a core model of the cellular metabolism in yeast.

The first step of the pipeline is a constraint-based analysis performed via FBA techniques [24] maximizing or minimizing a certain physiological aspect; crucial for this task is the optimization of an objective function achieved exploiting ensemble approaches and genetic algorithms. The second part combines results from FBA with network analysis in order to highlight emergent and general properties of the system. Finally, the last part is devoted to the retrieval of kinetic constants from fluxes and to the mechanistic simulation of the system. Hereafter I will briefly discuss the steps of my pipeline illustrated in Figure 2.7.

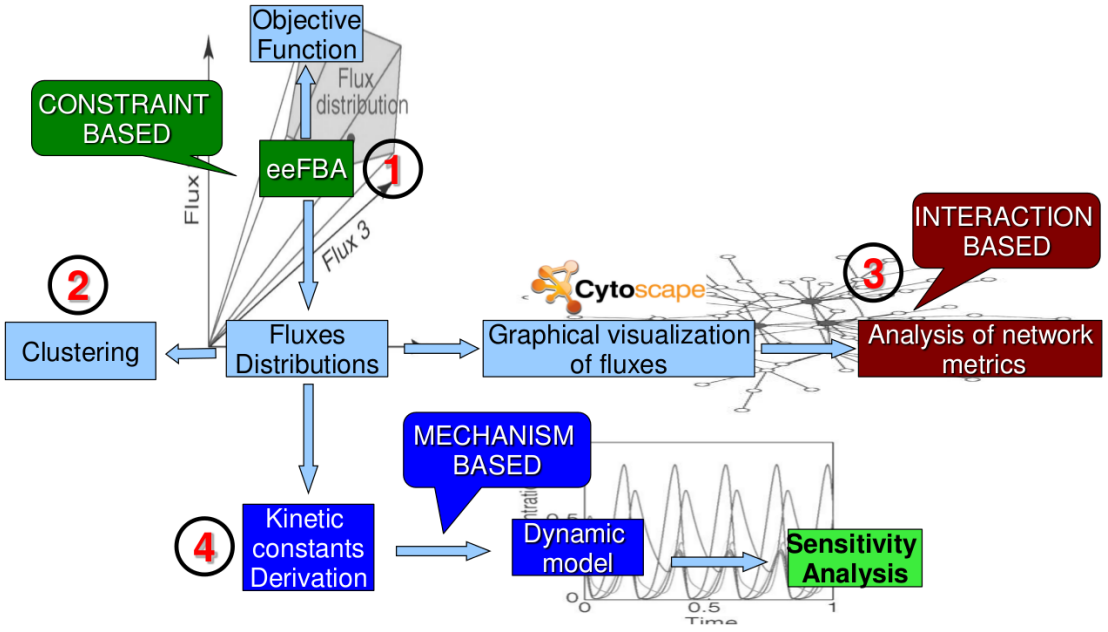


FIGURE 2.7: An overview of the computational pipeline presented in this thesis.

Overview of the computational pipeline Step (1) : Determining the objective function is pivotal in FBA; up to this date different formulations based on experimentalist sensitivity appeared in literature, but none of them can uniquely and precisely describe the behavior of cells and still it is impossible to exclude that the organism is found in a sub-optimal space.

In order to define more accurate and objective functions I defined two different automated methods not influenced by human bias. My first attempt [58] is based on Ensemble Approach [101] and Monte Carlo sampling to setup an unbiased analysis of the solutions space with the goal of identifying global network properties [102]. This approach has been used to explore the space of the possible objective functions, by randomly generating them, and then selecting those leading to the expected biological result (collection of flux distributions). In the developed "ensemble FBA" approach (eeFBA in figure 2.7) the structure of the network is never modified, while constraints are widely modified in order to mimic different cellular or environmental conditions.

My second approach for an unbiased analyses has been inspired by the observation that evolutionary algorithms have been used to optimize gene deletions in order to maximize production of a certain metabolite [103].

The adoption of an evolutionary algorithm is useful to explore the solution space in a more efficient way. In particular, the pipeline exploits a genetic algorithm (GA) [104] to evolve a population of individuals corresponding to different sampled objective functions. The population is selected accordingly to a fitness function that evaluates the distance of every single solution from the metabolic response constraint.

The use of a genetic algorithm has a double purpose: the first goal has been the identification of the solution that best matches the reference phenotype, while the second task has been the collection of an optimal individuals pool to investigate possible ensemble properties which characterize the phenotype. Up to now, the genetic algorithm has been integrated with the FBA Cobra Toolbox [105] and tested on a yeast core model.

Step (2) : from "ensemble FBA" to cluster analysis. As a second step I developed a hierarchical clustering analysis exploiting a dendrogram to illustrate how solutions obtained with the genetic algorithm cluster together. In the dendrogram, the division in two main groups corresponding to the matching and non matching solutions can be clearly observed: strikingly in this analysis there are no solutions falling into the wrong cluster. Moreover, the two ensembles (matching and non matching) can be further clustered into some well defined sub-clusters.

Step (3) , is devoted to the visualization of fluxes on the network and to network analysis. A first attempt exploits and integrates existing tools such as the CyFluxViz

[106] plugin of Cytoscape [107] to map and visualize fluxes obtained from FBA analyses. In this phase Cytoscape is also employed to analyze general (topological) properties of the network such as modules and motifs.

Step (4) : Kinetic constants derivation and dynamic modeling. Here I first evaluated the feasibility of the task implementing MetaFluxAnalysis, a LabVIEW [108] tool to determine metabolic fluxes starting from mechanistic simulations.

Then, the last phase of the pipeline has been the definition of a method to estimate kinetic constants for the yeast metabolic model starting from fluxes distributions identified after clustering solutions of the "ensemble FBA". In this context, from each cluster I derive an average flux distribution that will be the target for the estimation of kinetic constants by means of a Particle Swarm Optimizer (PSO) [109] implemented in MATLAB. In order to have a more efficient estimation of the parameters I contributed to develop a novel version of the PSO algorithm named Proactive Particles in Swarm Optimization (PPSO) [110]. The peculiarity of this algorithm is the use of Fuzzy Logic to determine the best setting for the key parameters of PSO (i.e. inertia, cognitive and social factor).

In order to validate the developed method and the computational pipeline, I defined a core model of the cellular metabolism in yeast and I am currently developing a version for the human metabolism. The choice of a core model is due to the fact that, even if genome scale networks require less assumptions on reactions to be included in the model, the interpretation and the understanding of the simulation outcomes are not always straightforward. On the other side, small scale (core) models need many more assumptions to correctly define the set of reactions, but they could shed light on design principles and emergent properties of the system under evaluation in an easier way.

Chapter 3

Constraint-based analysis

3.1 Constraint-based methods

Constraint-based analysis is the most widely used computational approach in the context of metabolic modeling. The core of this approach can be identified in the assumption that a biological system can uniquely express phenotypes that satisfy a given number of constraints, and that the metabolic system will reach a condition that can be assimilated to a steady state (referred as quasi-steady state). In other words, the setting of these boundaries, allow the identification of the solution space defining every possible functional state reachable or not reachable by the system. In this context, during the last two decades, many different approaches appeared in literature, for an exhaustive list, reader should refer to [111, 112].

In each one of these methods the first step for the analysis is the formulation of a matrix indicating the variation in units (stoichiometric coefficients) of reactants and products (metabolites, in rows) due to the application of every single reaction (represented in columns) of the system. In technical terms this is defined as the stoichiometric matrix S (see Figure 3.1).

The second key element, the quasi-steady state condition, is defined by the equation:

$$d\mathbf{x}/dt = S \cdot \mathbf{v} = \mathbf{0}, \tag{3.1}$$

where $d\mathbf{x}/dt$ are time derivatives of metabolite concentrations represented by the product of the $m \times n$ matrix S multiplied by the vector of fluxes $\mathbf{v} = (v_1, v_2, \ldots, v_n)$, where v_i is the flux of reaction i, n is the number of reactions, and m is the number of metabolites. The obtained null space corresponds to the so called *solution space* Σ.

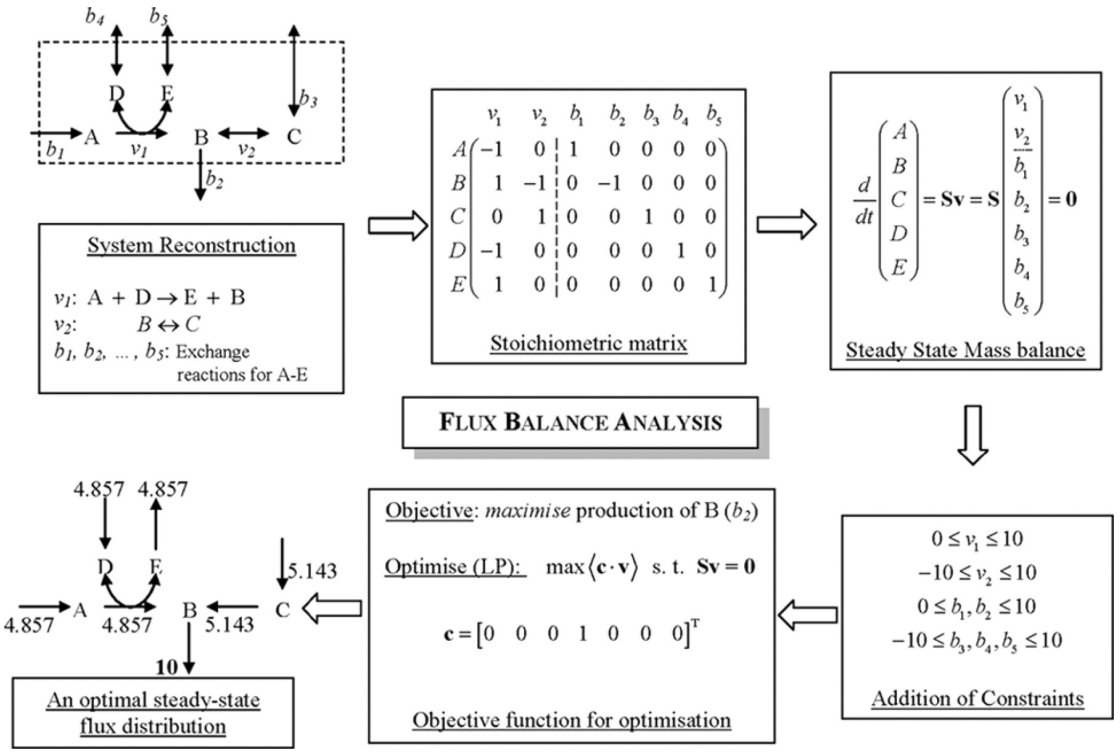


FIGURE 3.1: A scheme illustrating the main steps of constraint-based methods (FBA). Figure from [113].

Constraints due to the stoichiometric information alone, define an under-determined system because in metabolic networks the number of reactions (or equivalently fluxes) is usually greater than the number of metabolites. To further limit the solution space, it is hence necessary to incorporate additional constraints on thermodynamic properties of reactions (reversibility), and capacity constraints determined by imposing to the reactions maximum and minimum flux values determined by means of "wet" experiments. Formally $I_i = [v_i^{min}, v_i^{max}]$ define the space of acceptable fluxes $I = I_1 \times I_2 \times \ldots I_n$.

The combination of the space of acceptable fluxes and the solution space, defines the *feasible solutions space* $\Phi = I \cap \Sigma$.

3.1.1 Flux Balance Analysis (FBA) and derived methods

The ancestor of constraint-based methods can be identified in the Flux Balance Analysis (FBA) [24].

Assuming that the cell behaves optimally towards the realization of a given task (e.g. maximization of the biomass production), this method allows to compute a unique flux distribution that optimizes this specific objective in the feasible solution space. The metabolic "goal" is defined *"objective function"* (OF) and is mathematically represented

by the equation:

$$\mathbf{z} = \sum_{i=1}^{n} c_i v_i, \tag{3.2}$$

where c_i is a weight defining how much the flux v_i of reaction i contributes to the OF.

Due to the fact that all the components of the FBA are expressed in a linear form (a linear OF as described in Equation 3.2, and linear equations representing constraints), it is possible to identify this approach as a linear programming problem that can be solved using software packages (solvers) implementing the simplex algorithm. During years, however, advanced versions of FBA were developed to deal with non-linear constraints, but generally these methods are far less computational efficient when compared to the original method (a complete review of the optimization techniques for FBA can be found in [114]).

An evident limitation of FBA is caused by the use of the simplex algorithm, because, even if it is designed to find a unique optimal solution for the optimization problem; it is not able to exclude that in the system many equivalent optimal solutions can be found (and actually, this is what happens even for small networks). It is therefore important to find strategies to identify and enumerate all the different optimal flux distributions in order to evaluate the different biological meanings of the alternative optimal fluxes distributions.

Extreme Pathway Analysis [115] and Elementary Flux Modes [116] have been designed to retrieve all the different "equally optimal" fluxes distributions over the entire solution space. Unfortunately, the number of elementary flux modes grows exponentially with the number of reactions in the network, so that their enumeration becomes intractable for genome-scale networks. However, some strategies to make this computation possible have been proposed [117, 118].

Flux Variability Analysis (FVA), by constraining the objective value to be close or equal to its optimal value [119], is instead used to provide an indication of the range of variation (maximum and minimum allowed values) within each flux. Although FVA is less computational demanding with respect to Extreme Pathway Analysis and Elementary Flux Modes, it can be profitably used only for small models, and for this reason more recent versions of the method has been recently developed [120].

A relevant aspect in FBA is connected to the fact that kinetic parameters are not needed. If this on one hand allows to perform efficient simulations of otherwise hardy analyzable systems, on the other hand makes FBA unsuitable for the investigation of the system dynamics. To partially bridge this gap, methods globally referred as Dynamic Flux Balance Analysis (dFBA) have been devised to simulate dynamics changes in metabolic

networks exploiting two different strategies: Dynamic Optimization Approach (DOA) [121], makes use of a non-linear optimization over the whole evaluated time interval in order to derive flux distributions and metabolite levels, while Static Optimization Approach (SOA) [121] is performed by dividing the evaluated period into small intervals, optimizing via linear programming at the beginning of every single interval and then integrating over the entire time interval.

Looking at the applications of FBA, it is clear that the original purpose was the design of strategies to maximize the production of biochemical compounds of industrial use (e.g., biofuels [122]) in the domain of metabolic engineering; whereas later applications took advantage of the inherent properties of the FBA such as the fact that information on kinetic parameters are not needed for this kind of modeling. This great advantage respect to mechanism-based approaches lead to a renewed interest for FBA in the Systems Biology community for investigations on the physiopathological state of cellular metabolism.

The most relevant application in this context can be found for the prevision of phenotypical characteristics of microorganisms [123] where the assumption of an optimal (maximized) growth as cellular objective is particularly appropriate.

The following obvious qualitative step, in terms of analyzed complexity, has been the analysis of eukcaryotic energetic metabolism through the exploitation of a OF maximizing the ATP production [93, 124] to investigate key physiological aspects of tumor metabolism such as the Warburg effect (see [93] for further reference).

A last step to improve the predictive capabilities of constraint-based methods is given by the chance to integrate regulatory and metabolic networks [125, 126], where regulation of gene expression is obtained by tuning constraints on reactions to different values according to expression levels. In the extreme case, if a gene is completely repressed, the fluxes through reactions involving the protein expressed by the given gene, will be constrained to zero.

A worthy conclusive remark to this introduction to FBA it is that, in any case, solutions obtained with constraint-based methods are only as good as the constraints used to identify them. Studies [127] highlighted the fact that identifying an appropriate OF is of pivotal importance in FBA. This rise some issues due to the fact that the identification of the "true" OF could be problematic. This is particularly true when dealing with multicellular organisms where the objective can not be unequivocally defined, but is rather a trade-off between competing tasks [128]. Lastly, even if the precise formulation the OF would be available, it would still not be possible to exclude that the organism is living in a sub-optimal state.

These concerns are gaining relevance in literature, as it happened in [129] where authors showed, following multi-objective optimization theory, that metabolism works near the Pareto-optimal surface of a space defined by competing objectives.

For the above mentioned reasons, new approaches are emerging for an unbiased analysis of the solutions space, aimed at describing global network properties [102]. In this context, I developed a novel unbiased approach called Ensemble Evolutionary FBA (eeFBA) [58] whose aim is to identify ensembles of flux distributions that comply with one or more target phenotype(s).

3.1.2 ensemble evolutionary FBA (eeFBA)

An alteration of cellular metabolic fluxes could lead to different metabolic behaviors. To understand which collection of flux distributions are compatible with the different metabolic behaviors, in [58] I proposed, in the context of constraint-based modeling, an innovative approach that focuses on the analysis of generic properties of ensembles of solutions that satisfy phenotype(s) definition.

Differently from other literature cases, in the developed approach, changes in environmental or cellular conditions are simulated by varying constraints in a systematic way, while the network structure remains fixed. In order to sample different conditions an evolutionary algorithm was adopted and for this reason it has been named evolutionary approach.

In the devised approach, another difference with respect to the classical FBA is the meaning given to the objective function $\mathbf{z}$. In eeFBA it does not represent a physiologically plausible objective (e.g. maximization of biomass) or a bioengineering design goal (e.g maximization of the production for a metabolite of interest), rather it is seen as a way to explore the feasible solutions space.

Bordel et al. already proposed in [130] an approach that maximizes a random set of objective functions to define the corners of the space of allowed solutions; in this same work, randomization has been used to study the statistical distribution of the flux values of all the reactions in a genome-scale model.
Moreover in [102] it was also proven that sampling randomly the feasible flux distributions can effectively characterize its content [102].

In the present case random sampling is instead used to search the "functional states" in "agreement" with experimentally observed phenotypes previously defined. Solutions with specified properties are quested within the feasible solutions space and for this reason a search algorithm is combined with the developed sampling method.

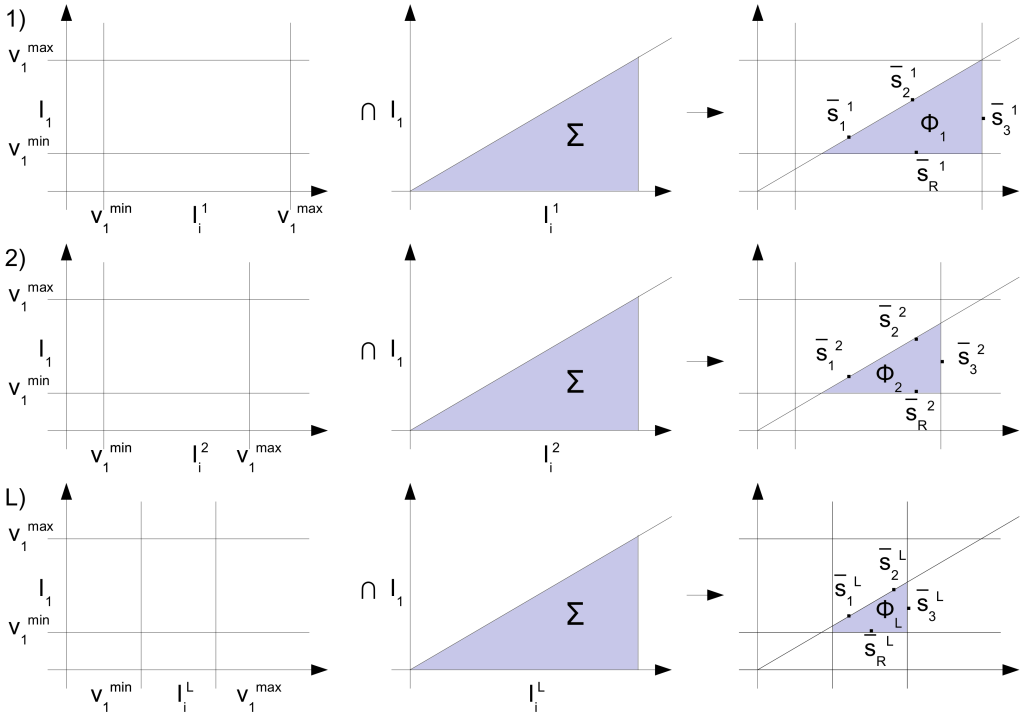


FIGURE 3.2: A graphical representation of the solution spaces according to definitions provided in Subsection 3.1.2.

The formal definition of the phenotype is expressed in terms of a metabolic response to environmental conditions variations: as an example, with this method the effects of an experimental variation in nutrient availability could be observed in terms of fluxes redistribution.

An environmental variation is mapped as variations in the boundaries of a reference flux i, i.e. $v_i \in I_i^l = [v_i^{l,min}, v_i^{l,max}]$ (a different interval for each environmental condition l with $l = 1, \cdots, L$). When it is intersected with the solution space Σ, defines the *phenotypic feasible solution space*:

$$\Phi_l = I_l \cap \Sigma \tag{3.3}$$

where $I_l = I_1 \times \cdots \times I_i^l \times \cdots \times I_n$.
The union of the L distinct Φ_l, one for each environmental condition, defines the *phenotype space*:

$$\Phi = \bigcup_l \Phi_l \tag{3.4}$$

A graphical representation for the construction of the solution spaces is given in Figure 3.2.

A metabolic response is investigated analyzing the system behavior under the L environmental conditions. Formally this is obtained assigning to the same fixed OF $\mathbf{z}_j$ in each of the distinct Φ_l a distinct optimal fluxes distribution $\hat{s}_j^l$ calculated by means of linear programming. At this step, it is then possible to define the set:

$$\mathcal{S}_j = \left\{ \hat{s}_j^l \right\}_{l=1,\cdots,L} \tag{3.5}$$

that represent the metabolic response of the system subject to the L different conditions. The exploration of Φ is realized through the definition of several, R, different OFs $\mathbf{z}_j$ and the consecutive sampling of a set of metabolic responses $\mathcal{S} = \{\mathcal{S}_j\}_{j=1,\cdots,R}$.
As stated at the beginning of this section, the OF is here used just to represent a *particular network condition* and therefore it is not intended to have any further biological meaning.

The level of complexity in the identification of a phenotype reflecting a given behavior could be raised supposing that, when increasing a flux v_i (e.g. the uptake of a nutrient, which can be simulated by increasing the boundaries of the flux), is associated an increase in another flux v_j (e.g. the secretion rate of a given metabolite).
In this case, and in in general for the eeFBA approach, *solution* will indicate a set of functional states that, besides fulfilling the imposed constraints and their variations, abides by the metabolic response definition (hereinafter also referred to as *metabolic response constraint*).

An *ensemble* of solutions is, hence, generated by different OFs representing alternative *solutions* that satisfy the specified property defined by a sufficiently loose metabolic response (e.g. the behavior in correspondence of extreme levels of nutrient uptake). The procedure to obtain such ensemble will be described in Subsection 3.1.3.
Emergent behaviors of the metabolic network under investigation can be discovered by analyzing the "typical" behavior of ensembles of *solutions* sustaining the same metabolic response in order to seek which functional states have the capability to sustain the metabolic phenotype.

It is also likely that the solutions in the same ensemble can be clustered (by means of a clustering algorithm) into different groups, suggesting that the same metabolic response is represented by different conditions of the network. An example of this application will be provided in Section 3.5.3.

In the case of multiple metabolic responses it could be of interest to compare the ensemble of solutions representative of a metabolic response α (e.g. a physiological response) against an ensemble of solutions obeying to a different response β (e.g. an altered response).

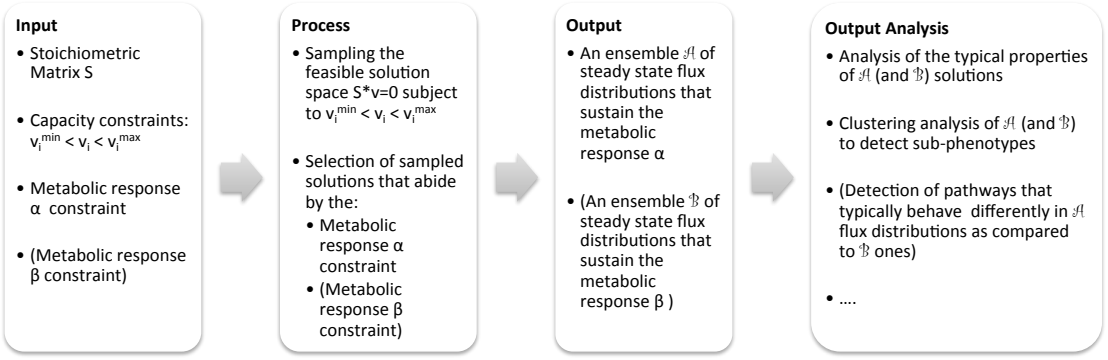


FIGURE 3.3: Workflow illustrating the eeFBA approach. Figure from [58].

This could reveal sets of fluxes with a different behavior in the two cases and that may be associated to candidate pathways responsible for the different metabolic response. In Section 3.5.3 it is shown an example exploiting a Kolmogorov-Smirnov test to identify the set of fluxes significantly different between the two ensembles.
A schematic representation of the entire workflow is provided in Figure 3.3.

3.1.3 Sampling and populating methods

Sampling the solution space In the eeFBA approach, a random set of OFs are maximized with linear programming. This is different from the commonly used method of randomly sampling the feasible solutions by means of the "hit-and-run" sampler.

In Bordel et al. [130] random OFs were generated by selecting random pairs of reactions and assigning them random weights. On the contrary, in eeFBA any number of reactions can take part in the random OF with the goal to maximize sampled solutions variability. In eeFBA number of reactions that constitutes the OF, is given by the cardinality of the set of reactions having $c_i > 0$. The fraction τ of considered reactions is drawn at random, with uniform probability in (0, 1] with the aim of obtaining an objective involving from one to all reactions with an uniform probability. To any selected reaction is then assigned a random weight c_i uniformly tossed from the interval $(0, 1]$.
An instance j of the objective functions $\mathbf{z}_j$ is defined as $\mathbf{z}_j = \sum_{i=1}^{n} c_i v_i$, where c_i takes value 0 with probability τ and takes a random value with uniform probability in $[0, 1]$ with probability $1 - \tau$.
To every $\mathbf{z}_j$ it is hence assigned an *optimal solution* $\hat{s}_j^l \in \Phi_l$ calculated with the standard FBA approach exploiting linear programming.

Populating the ensembles of solutions Within the eeFBA framework, in order to populate an ensemble of matching solutions $\mathcal{A} \subseteq \mathcal{S}$ corresponding to a phenotype of interest α, the solutions $\hat{s}_j^l$ have been characterized. In these solutions the observed redistribution of fluxes obtained when moving the boundaries of a given flux v_i while maintaining $\mathbf{z}_j$ and performing the optimization, respects the metabolic response constraint.

At this stage, with a procedure named *sampling + filtering* it is possible to sample randomly the desired number of points (i.e. $\mathbf{z}_j$) and filter them accordingly to $\mathcal{A} = \{\mathcal{S}_j \mid \mathcal{S}_j \in \mathcal{S}, A(\mathcal{S}_j) = 1\}$, where α can be converted in a Boolean fashion to the Boolean expression A (indeed the metabolic response constraint is either respected or not).

A further strategy to populate the ensemble of solutions could be the finding of the behavior closest to the metabolic response constraint. This can be done, at the price of reducing the variability within the ensembles, by means of an evolutionary algorithm. In this case a *genetic algorithm* (GA) is used to explore the phenotype space. The GA is exploited to build a set $P = \{\mathbf{z}_j\}$ of p individuals, corresponding to different sampled objective functions, and to evolve them accordingly to a fitness function $a(\mathcal{S}_j)$ quantifying the distance from the metabolic response constraints to the corresponding solutions in $\mathcal{S}_j$.

The use of the GA is motivated by two mutually exclusive reasons: (I) identify the solutions that best agree with the reference phenotype (genetic algorithms have been devised to this end); (II) enlarge the ensemble selecting the highest number of solutions to investigate the eventual presence of "global" properties characterizing the phenotype. To realize this second goal, it is possible to exploit, for each run of the GA, differently initialized populations. Another adopted strategy consisted in amplifying the internal variability of the ensemble $\mathcal{A}$ by selecting, in every run r, the best solution $\overline{\mathcal{S}_{j,r}}$ and a set of solutions "not too far from the best". In the eeFBA this is done by selecting all the individuals within a standard deviation from the best solution fitness value, $\mathcal{A} = \{\mathcal{S}_{j,r} \mid \mathcal{S}_{j,r} \in \Phi, a(\mathcal{S}_{j,r}) \leq a(\overline{\mathcal{S}_{j,r}}) + \sigma_r\}$.

3.2 Genome-wide and core models

Besides the definition of an appropriate mathematical framework to study the "metabolic system" with constraint-based approaches, it is fundamental to define an appropriate model describing it. To this end, in literature two categories of models (genome-wide models and core models, hereafter illustrated) can be identified.

Genome-wide models With the advent of high-throughput technologies [131], an unprecedented amount of quantitative data supplied by omics technologies such as genomics, transcriptomics, proteomics, metabolomics, allowed the identification of almost every component of metabolic networks. These new potentialities lead to the development of genome-wide (GW) models taking into account every single known reaction in an organism and summarizing the knowledge on the system itself.

During last ten years, a great number of GW reconstructions appeared in literature, analyzing more and more complex organisms: bacteria [75], lower [70, 95] and higher eukaryotes [25, 26], plants [78]. The most relevant goal of these GW models is, in the domain of Systems Biology, their ability to be used as "scaffolds" for computational analyses exploiting both constraint-based and network-based approaches. Moreover they can also be used as a sort of repository of the collected knowledge about metabolic pathways [132].

An automated and efficient procedure of reconstruction of a GW model for an organism of interest starts typically from the genomic sequence annotations, or from an existing model of a related organism [133] used as reference for the definition of the initial draft of the model.

The obtained draft must then undergo an accurate process of curation and refinement in order to correct errors (such as reaction gaps and wrong directionalities) due to incomplete or inaccurate annotation. This constitutes a time consuming process that needs the manual integration and control of the reactions included in the draft, even if the task can be aided by some semi-automated procedures [134–138].

An obvious challenge for Systems Biology, is the reconstruction of an accurate model of the entire human metabolism. At present the most advanced human GW model reconstruction is *Recon 2* [26] that has been curated from a previous version of the global human metabolic network, *Recon 1* [25].

Recon 1 has been defined starting from an accurate human genome sequence annotation, from which it was possible to retrieve the set of genes involved in metabolic processes and hence corresponding enzymes catalyzing reactions. This information has been subsequently integrated, in a refining and validation process, with the correct reaction stoichiometry, compartmentalization and definition of exchange reactions in order to account for the conservation of metabolic mass and charge.

Recon 2 can be seen as an extension of Recon 1 realized through the expansion of the set of metabolic reactions (7440, almost the double with respect to the previous version) supported by additional sources of metabolic information and by a further accurate process of revision and validation.

A further step in increasing the complexity of GW model is driven by the consideration that cells in different tissues are characterized by metabolic pathways that are differently active. For this reason, during last years, human GW models have been used to generate cell type-specific models. For example, as it will further described in Section 3.4.1, the INIT algorithm [139] has been used to integrate data deriving from the Human Protein Atlas [140] and other sources in order to generate cell-type specific metabolic models that will be analyzed in Section 3.4.

Several successful applications of GW models have been reported [51, 81, 141, 142], but, as already stated in several points in this thesis, some limitations emerged for their application. First of all, GW models are typically investigated through FBA methods (see Section 3.1.1) due to the fact that the dimensions of the system (thousands of metabolites and reactions) and the lack of kinetic parameters make them hardly analyzable with the computationally-demanding mechanism-based modeling techniques.

In addition, even with the classical FBA techniques it is not always possible to predict the actual metabolic flux distribution in the network, due to the presence of inconsistencies in the network itself (wrong mass and charge balance, lack of reactions or metabolites that are currently unidentified) [143], and due to the existence of thermodynamically infeasible loops. This last point should always be critically evaluated, but unfortunately the many methods devised for the complete identification and the removal of the loops [47, 144] can be profitably used only for small scale networks due to their computational cost.

Core models A complementary strategy to investigate metabolic networks, is the definition of Core models (CM), i.e. models having a lower level of complexity (see Chapter 1 for a meaningful definition) and that can be used either to investigate in detail a particular pathway (e.g., glycolysis) [90, 145], or to represent the global metabolism of a cell (see Figure 3.4 for an example) including only those pathways useful to investigate emergent properties (e.g. the presence of a metabolic phenotype).

As an example of the first goal it is worth to cite a work where Kerkhoven and co-authors [73] defined a CM evaluating exclusively the glycolysis and the pentose phosphate pathways, while the second task has been used in [84], where authors included metabolic pathways that are supposed to have a pivotal role in cancer cell growth, namely, glycolysis, TCA cycle, pentose phosphate, glutaminolysis and oxidative phosphorylation.

In the present dissertation it has been developed a yeast CM (Section 3.5 and [58]) with the aim of analyzing the design principles behind the emergence of the Crabtree effect (a metabolic behavior that is characteristic of several yeasts [146]) by means of the eeFBA

approach presented in Subsection 3.1.2. This model has been designed to include the most relevant pathways known to play a role in the emergence of the Crabtree effect but, in order to both reduce the model complexity and avoid cell-type specificities, reactions that belong to a linear cascade without branching were lumped together in a unique fictitious reaction.

Although CMs are useful to reduce the complexity of the system description and to ease the interpretation of simulation outcomes, they must be carefully exploited due to a possible undervaluation of key multi-factorial relationships between the state of the system and its response to different perturbations (e.g., in the modeling of complex diseases) [147].

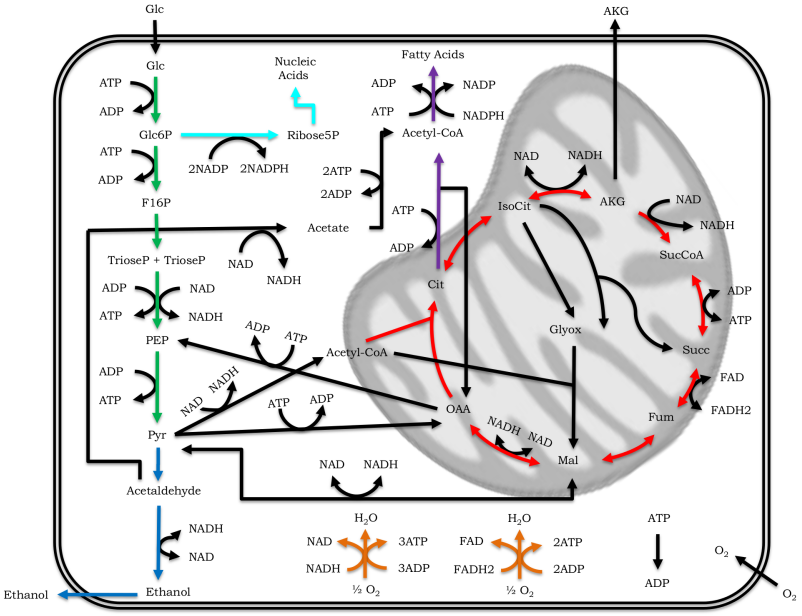


FIGURE 3.4: Yeast CM illustrating the most relevant metabolic pathways: glycolysis (green arrows), alcoholic fermentation (blue arrows), pentose phosphate pathway (light blue arrows), biosynthesis of fatty acids (violet arrows), Krebs cycle (red arrows) and OXPHOS (orange arrows). Figure from [28].

3.3 Tools for constraint-based analysis: the COBRA toolbox

In recent years many softwares have been developed to perform constraint-based analysis (a list of the most relevant is provided in Table 3.1). Among these, a sort of standard for the scientific community is the COBRA (COnstraint-Based Reconstruction and Analysis) Toolbox [149] a MATLAB environment software package for the constraint-based analyses of models.

Tool name	Purpose	Reference
BioMet Toolbox	GW models validation, FBA, probabilistic FBA, gene set analysis	[148]
Cobra Toolbox	FBA, FVA, dFBA, gap filling, network visualization	[149]
FAME	Web based FBA and FVA	[150]
FASIMU	FBA, FVA, gene deletion analysis, gap filling	[151]
OptFlux	FBA, FVA, EFM, gene deletion analysis	[152]
Pathway Tools	GW reconstruction, FBA, gap filling	[153]
Raven Toolbox	GW reconstructions, FBA, network analysis and visualization	[154]
SurreyFBA	FBA, FVA, EFM	[155]

TABLE 3.1: Main computational tools used in the modeling, simulation and analysis of metabolism. Table from [28].

COBRA can be exploited with all the systems that can be represented as a list of biochemical reactions (mainly metabolic models) and it is based on the classical aspects of constraint-based analyses, such as the determination of chemical-physical constraints, the cellular compartmentalization, the directionality of the reactions and the steady-state assumption (conservation of the global mass).

The toolbox is structured as an ensemble of tools able to perform the most part of the common analyses on metabolic networks:

- prediction of the maximal growth rate of a cell at different rates of uptake for relevant nutrients;
- description of the composition of the cellular biomass to define the OF;
- prediction of effects on growth and reduction of flux through a single reactions due to the deletion of a gene of interest (simulated exploiting information on the expression level relative to a gene "involved" in the reaction);
- determination of the optimal flux distribution towards a given OF by means of FBA techniques;
- Flux Variability Analysis (FVA), a method derived from FBA and able to analyze the optimal alternative fluxes distributions calculating, for every reaction in the system, the minimum and maximum value that the flux can assume.

A recent version (2.0) of the COBRA Toolbox includes further analysis methods such as:

- geometric FBA (dealing with the fact that different solvers may return different solutions);
- the elimination of thermodynamically infeasible loops through the Loop law;
- gap filling to replenish gaps of metabolic networks due to an incomplete knowledge on the network structure;

- tools for the metabolic engineering.

The COBRA toolbox adopts the standard SBML format for the import of metabolic models, but it is also able to import from Excel spreadsheets. The software has some dependencies (sometimes giving raise to installations problems) such as (I) libSBML, an open source library for the writing, reading and manipulation of SBML models; (II) the SBML toolbox to import/export SBLM from/to MATLAB; (III) one or more supported linear programming solver such as Gurobi, CPLEX e GLPK.

The eeFBA method defined in Subsection 3.1.2, along as every script for contraint-based analysis used in this thesis, have been implemented in MATLAB exploiting the potentialities of the COBRA toolbox to manage SBML models and perform FBA.

3.4 Zooming in genome scale models, a reduction approach

In Section 3.2, it has been illustrated main advantages and disadvantages of GW and CM. In this section, starting from already existing genome-scale metabolic models, the aim is the reconstruction of manually curated core models that zoom in metabolic complexity. Here, the study has been performed not focusing on a single model for a generic cell, but on the comparison of tissue specific metabolic models describing three types of tumor and a "reference state". [1]

The choice of evaluating the relevance of the tissue specific models has been done due to the belief that the development of these kind of networks is a milestone towards the implementation of a "virtual twin", the fundamental tool in the Systems Medicine perspective [156].

3.4.1 Genome-wide models: the Human Metabolic Atlas

To approach the study of the "generic" human cell metabolism and to highlight differences, in terms of fluxes values, that can be found in cancer cell metabolism, a good starting point is the the analysis of the Human Metabolic Atlas (HMA), a database where 69 cell types and 16 cancer metabolic networks described at genome-scale level are deposited [139].

From the computational point of view these networks have been automatically generated exploiting the Integrative Network Inference for Tissues (INIT) algorithm [139] that uses as a reference the generic Human Metabolic Reaction (HMR) model [157].

[1] Disclaimer: Studies performed in Section 3.4 have been performed mainly M. Di Filippo (SYSBIO - Centre of Systems Biology, Milan - Italy) under the supervision of the author.

HMR has been built integrating information deriving from different sources representing (at the time of the publication) the state of the art for metabolic modeling: Recon 1 [25], EHMN [158], HumanCyc [159] database and KEGG [46, 160] database.

In the following step, transcriptomic, proteomic and metabolomic data have been used to select subsets of reactions that constitute the different cell or tissue specific genome-wide metabolic networks. In particular, the Human Proteome Atlas (HPA) [161] has been used for proteome data, while the Human Metabolome Database (HMDB) [162–164] has been exploited to define the metabolic pools. Then INIT "weighs" every reaction in the model accordingly to the soundness of the biological evidence and accordingly to the level of expression. As a last step, the algorithm performs an optimization in order to maximize reactions fluxes with a high weight. The process returns as output an ensemble of networks called "active" due to the fact that they should represent exclusively the portion of Human Metabolic Reaction with reactions effectively expressed in every cell/tissue type.

3.4.2 Reduction of genome-wide models

As already stated in Section 3.2, CMs are useful to reduce the complexity of the system description and to ease the interpretation of simulation outcomes. For this reason four constraint-based core metabolic models have been reconstructed and compared. Of these, one model has been obtained with a reduction of the generic HMR model while the three others have been derived from the reduction of tissue-specific genome-scale networks *iLiverCancer1715*, *iBreastCancer1771* and *iLungCancer1472* (as in the version available in the HMA database on October 2013). These models have been selected as the most harmful, among the networks generated using INIT, due to the high values of parameters such as mortality, incidence and prevalence retrieved from the last estimations of GLOBOCAN [165, 166], the online database of the International Agency for Research on Cancer.

With the reduction process, only pathways that play a pivotal role in sustaining cancer cells growth and proliferation have been selected [84, 167]: glycolysis, pentose phosphate pathway (PPP), tricarboxylic acid cycle (TCA cycle), oxidative phosphorylation, glutamine metabolism, amino acid synthesis, urea cycle, folate metabolism and palmitate synthesis.

Moreover, to obtain models that can be simulated with the FBA approach, it has been necessary to add elements such as exchange reactions (allowing metabolites to be transported between compartments of the models – cytosol, mitochondria and external environment), and "sink" reactions to define the environment around the cell, as they

introduce nutrients to be metabolized by the cell. These kind of reaction has also been introduced for those metabolites whose production is not explicitly modeled within the core models, because pertaining to non considered pathways.
Demand reactions are added for compounds that are known to be secreted by the cell in the extracellular environment. As for sink reactions, these reactions have been inserted during the curation process to allow removal of metabolites whose consumption is not entirely modeled in the network.

As discussed in Subsection 3.1.1, the fundamental step for the definition of a constraint-based model is the addition of constraints. This has been done in the four "reduced models" by means of (I) thermodynamic constraints specifying the reversibility or irreversibility of every reaction; (II) compartmentalization of metabolites; (III) environmental constraints (nutrients that can enter in the cell); (IV) capacity constraints (setting the lower bound to zero for irreversible reactions, and leaving upper bound unlimited in the allowed direction except for exchange reactions).

The fundamental assumption of FBA is the fact that metabolism behaves optimally towards a given OF. In this case, since are mainly evaluated metabolic models describing cancer cells, the "function" to be optimized has been identified in the maximization of the biomass production.

The biomass formation pseudo-reaction associated to cancer models (including all the needed metabolites, and the relative stoichiometric coefficient) has been adopted as OF for both for the reference and cancer models. In the CMs the biomass formation pseudo-reaction has been reduced from the full version to encompass only the subset of metabolites needed for biomass synthesis that are involved in the pathways here considered.
It is worth to underline that I am not neglecting the fact that normal and cancer cells exhibit different behavior in terms of growth and proliferation rate (a fact also underlined by different biomass formation pseudo-reactions in HMA models) and this choice has been done in order to highlight their different metabolic requirements.

Following the identification of some incongruities in the exploited GW models, a large amount of time has been dedicated to the manual curation of the CMs. Although the HMA has been extensively curated, the inner complexity of GW models leaves room for misinterpretations, suggesting to further check the formulation of reactions in terms of metabolites taking part to them and searching for gaps (missing metabolic reactions that wrongly interrupt a biochemical pathway), using as reference the database KEGG and the state of the art human metabolic reconstruction Recon 2 [26]. The corrections resulting from this process are listed in Table 3.2. From this table it is possible to notice that the most common error involved a wrong directionality of the reactions.

Original reactions	Revised reactions	Compartment
–	3-phospho-D-glycerate ⇒ 2-phospho-D-glycerate	Cytosol
Acetyl-CoA + H_2O + OAA ⇔ Citrate + CoA	Acetyl-CoA + H_2O + OAA ⇒ Citrate + CoA	Mitochondria
Isocitrate + NAD^+ ⇒ AKG + CO_2 + H^+ + NADH	Isocitrate + NAD^+ ⇔ AKG + CO_2 + H^+ + NADH	Mitochondria
Isocitrate + $NADP^+$ ⇒ AKG + CO_2 + H^+ + NADPH	Isocitrate + $NADP^+$ ⇔ AKG + CO_2 + H^+ + NADPH	Mitochondria
AKG + Leucine ⇒ 4-methyl-2-oxopentanoate + Glutamate	4-methyl-2-oxopentanoate + Glutamate ⇔ AKG + Leucine	Cytosol
AKG + Isoleucine ⇒ 2-oxo-3-methylvalerate + Glutamate	2-oxo-3-methylvalerate + Glutamate ⇔ AKG + Isoleucine	Cytosol

TABLE 3.2: Revised reactions after the curation phase of the core models, performed consulting KEGG database and the human metabolic reconstruction Recon 2. The first reaction is the result of a gap correction found within the glycolysis pathway, which in the starting genome-scale models, has been erroneously filled with two exchange reactions for the metabolites 3-phospho-D-glycerate and 2-phospho-D-glycerate. A revision of the directionality has been done for the other five reactions. In particular, the third and fourth reactions have been corrected within the cancer models because it is known that in tumors, unlike normal cells, the enzyme that is responsible for these reactions works mainly in the reverse direction.

	Genome-scale		Core	
Model	# reactions	# metabolites	# reactions	# metabolites
HMR database	8180	6011	274	251
Liver Cancer GW	4386	4020	257	241
Breast Cancer GW	4299	3955	244	233
Lung Cancer GW	3809	3653	236	230

TABLE 3.3: Number of reactions and metabolites for each of the genome-scale and core metabolic models.

Whereas the first reaction is an example of a gap within the glycolysis erroneously filled with two exchange reactions for the metabolites 3-phospho-D-glycerate and 2-phospho-D-glycerate, a mistake probably due to automatic gap-filling procedures performed by the INIT algorithm. In this case the correction restored the correct flow through the pathway.

The number of metabolites and reactions of the final core models, compared to their genome-wide counterparts is reported in Table 3.3. A topological analysis of the four metabolic models has been performed. The results and the analysis description can be found in Chapter 4.

The interested reader can retrieve the developed metabolic CMs from BioModels Database [168] querying for the following identifiers:
MODEL1502100000, MODEL1502100001, MODEL1502100002, MODEL1502100003.

3.4.3 Differential analysis of flux distributions

The COBRA Toolbox of MATLAB [105, 149] (see Section 3.3 for a description) has been used to perform FBA simulations in order to calculate flux distributions for each of the four CMs described in the previous section.

The four obtained flux distributions have been analyzed comparing each cancer model with the reference model.
It must be noticed that in the analysis, in order to make the three cancer models comparable among them and making the differential analysis meaningful, has been adopted a shared reference model instead of comparing each tumor against its corresponding healthy model recently published within the Human Metabolic Atlas database [169].

3.4.4 Tissue-specific cancer redistributions of metabolic flux

The accurate analysis of the metabolic flux redistribution is a pivotal step to understand mechanisms behind cancer cell growth and proliferation [170]. In this context, FBA has been used to estimate the flux of metabolites through the reactions of CMs of cancer cells and a "reference" cell (of the latter, a simplified version is illustrated in Figure 3.6 along with "active" reactions for this case). Of more interest has been the comparison of the obtained distributions that allowed to understand up- and down-regulations of fluxes in metabolic pathways. The main results of this analysis are illustrated in [171], where green and red chromatic scales highlight, respectively, the detected up-regulations and down-regulations, in terms of flux value, in the cancer condition with respect to the reference one.

A first relevant outcome is the difference in biomass production rate among the models. In particular, it is possible to notice a positive fold change of about 1.3-1.5 in favor of the cancer cells: a biologically correct outcome that confirm the ability of the model to distinguish between the the two states. Indeed, in populations of normal cells the cellular proliferation is inhibited by cell-to-cell contact, while in cancer cells this inhibition is lost [171] leading to a higher biomass production rate sustained by an extensive modification of metabolic fluxes.

Analyzing the differential values for fluxes through the biochemical pathways, it emerges that cancer cells exhibit an increased exploitation of the glycolytic metabolism, underlined by a drastic decrease of the oxidative phosphorylation (OXPHOS) and in an enhanced glycolytic activity that results in a higher level of lactate secretion, a proxy indicating high fermentative levels and a way to regenerate NAD^+ from NADH, allowing in turns an enhanced glycolysis to persist.
Moreover, lactate secretion may be able to confer another advantage to tumor cells, enhancing their invasiveness by disrupting normal tissue architecture, and promoting an acidic tumor micro-environment to evade tumor-attacking immune cells [172].
The inefficiency of glycolysis compared to OXPHOS in producing the energetic molecule ATP, reflects to an expected increase of uptake of the nutrient glucose as observed in

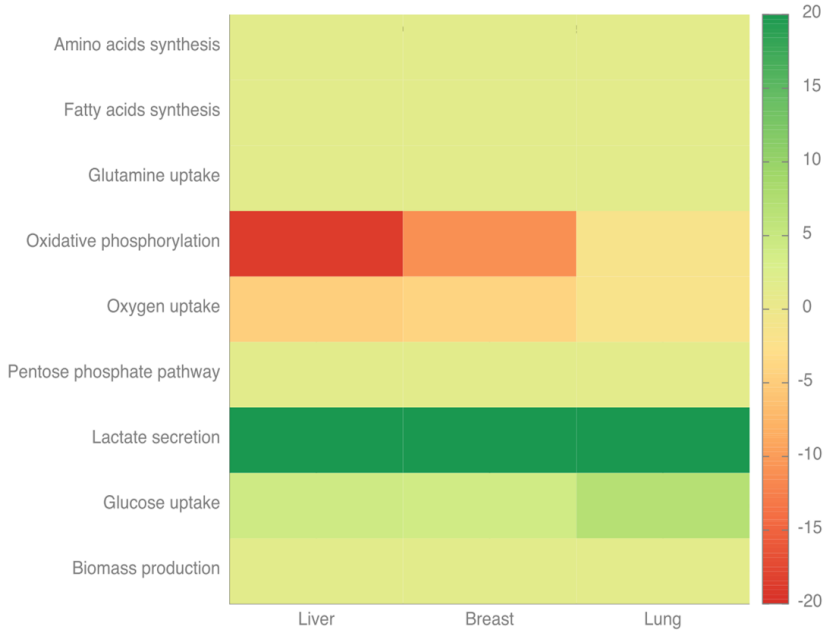


FIGURE 3.5: Main results obtained from the Flux Balance Analysis. Green and red chromatic scales highlight, respectively, the detected up-regulations and down-regulations, in terms of flux value, in the cancer cells models with respect to the reference model. From the figure it emerges that tumor cells, compared to the reference one, reprogram the metabolic pathways to satisfy their increased needs for the synthesis of macromolecular precursors essential for biomass production during tumor growth. Exploiting the FBA approach, it also emerged a heterogeneous behavior among the three investigated tumors.

cancer models (up-regulation of about 4-6 fold).
Globally these indications suggest that the CMs of cancer cells are depicting a phenotype dominated by the Warburg effect, i.e. the enhanced use of the glycolytic pathway (and lactate production), even in the presence of normal levels of oxygen [173].

The fatty acids synthesis is another fundamental pathway to sustain higher production of biomass in cancer cells. The key precursor for this synthesis is citrate (produced in the TCA cycle in the mitochondria and exported to cytosol), that is transformed in acetyl-CoA and eventually to fatty acids (simplified in the CMs with the palmitate synthesis) used as building blocks for cellular membranes. Therefore, in this view, the 1.3–1.4 fold up-regulation of ATP citrate lyase and fatty acid synthase is an expected event.

Also for glutamine metabolism has been recognized a key role in cancer growth and proliferation due to the fact that this metabolite is (along with glucose) an important source of energy, and a source of carbon for many processes such as fatty acids and aminoacids synthesis and to provide metabolites intermediates of the TCA cycle (a process called anaplerosis, that allows the TCA cycle to be a provider of biomass precursors).

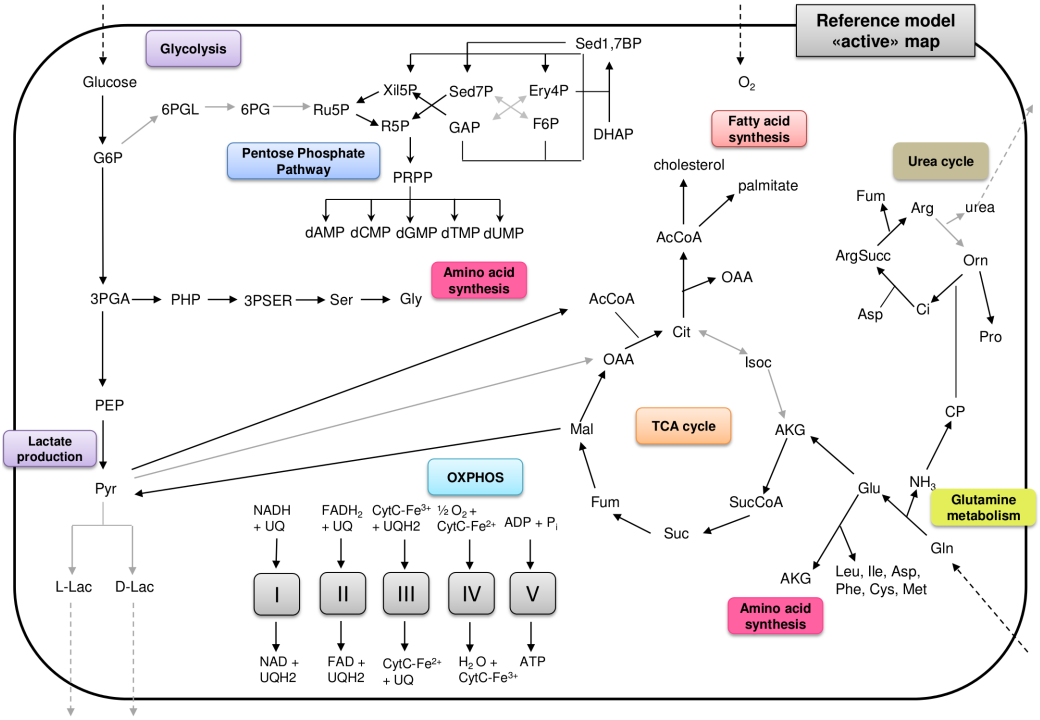


FIGURE 3.6: Schematic representation of the "active" network relative to the reference core model. This map shown, in a simplified view, the information concerning flux distribution in the reference core model within metabolic pathways under investigation in our work, namely glycolysis, pentose phosphate pathway, tricarboxylic acid cycle (TCA cycle), oxidative phosphorylation, glutamine metabolism, urea cycle, amino acid and fatty acid synthesis. The grey and black arrows correspond, respectively, to reactions having a null and a positive flux value. The direction of each black arrow is set depending on the obtained flux value in the corresponding reaction. For reasons of space, cellular compartments are not included. Abbreviations: G6P, glucose-6-phosphate; 3PGA, 3-phospho-D-glycerate; PEP, phosphoenolpyruvate; Pyr, pyruvate; L-lac, L-lactate; D-lac, D-lactate; 6PGL, glucono-1,5-lactone-6-phosphate; 6PG, 6-phospho-D- gluconate; Ru5P, ribulose-5-phosphate; R5P, ribose-5-phosphate; Xil5p, xylulose-5-phosphate; GAP, glyceraldehyde 3-phosphate; Sed7P, sedoheptulose-7-phosphate; Sed1,7BP, sedoheptulose-1,7-bisphosphate; Ery4P, erythrose-4-phosphate; F6P, fructose-6-phosphate; DHAP, dihydroxyacetone phosphate; PRPP, phosphoribosyl pyrophosphate; UQ, ubiquinone; UQH2, ubiquinol; CytC-Fe2+, ferrocytochromeC; CytC-Fe3+, ferricytochromeC; Leu, leucine; Ile, isoleucine; Asp, aspartate; Phe, phenylalanine; Cys, cysteine; Met, methionine; Gln, glutamine; Glu, glutamate; Cit, citrate; Isoc, isocitrate; AKG, α-ketoglutarate; SucCoA, succinyl-CoA; Suc, succinate; Fum, fumarate; Mal, malate; OAA, oxaloacetate; AcCoA, acetyl-CoA; CP, carbamoyl-phosphate; Orn, ornithine; Pro, proline; Ci, citrulline; ArgSucc, argininosuccinate; Arg, arginine, PHP, 3-phosphonooxypyruvate; 3PSER, 3-phosphoserine; Ser, serine; Gly, glycine.

In the developed cancer CMs the role of glutamine has been underlined by an increased glutamine uptake (about 1.4 fold). The high anaplerotic flux is a more specific indicator of cell growth with respect to the high glycolytic flux, since the latter is also stimulated by stresses not involved in biomass production [174].

Besides the increased glycolysis, another aspect underlying the Warburg effect is the presence, in cancer cells, of a down-regulated OXPHOS due to a reduced activity of all its components, and in particular due to a complete inhibition of the complex I activity.

In the analyzed models, it emerges that OXPHOS is strongly decreased in breast and liver cancer, but the lung cancer exhibit flux levels comparable to the reference case. Nevertheless, recent studies by Hooda et al. in [175], demonstated a crucial role for respiration in lung cancer cells to promote their development and growth.

The flux distributions obtained for each CM, along with model reactions, are available in Appendix A.

In conclusion, the developed CMs were able to identify different flux distributions both between reference and cancer conditions, and among the three evaluated cancer models, indicating that cell specific models should be developed and analyzed to grasp this heterogeneity. The variations among different cancer types are of great interest in the medical field for the identification of cancer type - specific drug targets in order to develop more effective treatments. Moreover, the observed heterogeneity suggest that three different sub-phenotypes (belonging to the three tissues) can be identified, all belonging to the cancer macro-phenotype.

As mentioned in Section 3.4.3, the three cancer models have been made comparable among them using the same reference model. However, an extension of the work currently in progress, is the comparison between each of our three cancer core models with their healthy tissue-specific model [169] counterparts, in order to identify specific peculiarity linked to a certain type of tissue that are not highlighted using the reference model.

Lastly, the emergence of cancer sub-phenotypes could be investigated by means of the eeFBA approach (Subsection 3.1.2), a strategy complementary to the reconstruction of distinct tissue-specific networks that makes use of a generic metabolic network to retrieve ensemble of flux distributions compatible with the cancer macro-phenotype, and that is also able to identify and describe sub-phenotypes inside the ensemble. In the next section eeFBA will be used to investigate the Crabtree effect (a phenotype exhibiting a number of common traits with the Warburg effect discussed in this section) and to identify its sub-phenotypes.

3.5 A core model of yeast to investigate the Crabtree effect

To better illustrate the eeFBA procedure defined in Subsection 3.1.2 and to investigate the emergence of the Crabtree effect [176] it has been developed a CM for the yeast metabolism consisting in 65 reactions and 37 metabolites defining the main metabolic pathways. In this simplified model, oxygen is the unique carbon source. The Crabtree is a well known phenomenon in which some yeasts (e.g. *Saccharomyces cerevisiae*) exhibit a high production of ethanol via fermentation even in presence of high external glucose concentrations (regardless of the availability of oxygen), rather than directing the production of energy towards the more efficient oxidative phosphorylation - as other yeasts do (also known as Crabtree-negative yeasts, e.g. *Kluyveromyces*) - therefore consuming less oxygen.
To demonstrate the capabilities of CMs for the investigation of the implications of the two different metabolic responses, the devised approach has been applied to the "small" model drawn in Figure 3.7, which takes into account only the main pathways and metabolites involved in the selected phenomenon.

3.5.1 Formalization of the Crabtree-positive and negative phenotypes

The Crabtree effect (CE) can be defined as the ability to repress respiration and oxidative phosphorylation as glucose concentration increases [146]. It is easily observed experimentally as the persistence of aerobic alcoholic fermentation - with a concomitant reduction in the respiration rate - when a pulse of glucose is added to the culture medium of yeasts [177, 178] or when yeasts in glucose-limited chemostats are grown at increasingly higher dilution rates, so that the flux of glucose from the medium to the cell progressively increases [179].

To study the CE exploiting the eeFBA it is necessary to give a formal definition to the metabolic response constraints corresponding to Crabtree-positive ($C\oplus$) and Crabtree-negative ($C\ominus$) phenotypes.

To do this, two fluxes whose activity traditionally define the CE has been evaluated: the first one is the oxygen uptake v_o, a proxy the oxidative phosphorylation, and the second one is ethanol secretion v_e, a proxy for the fermentative metabolism. In our approach, these two fluxes are considered in function of the glucose uptake v_g due to an enhanced difference between $C\oplus$ and $C\ominus$ yeasts in the kinetics of glucose uptake, as observed in [178, 180].

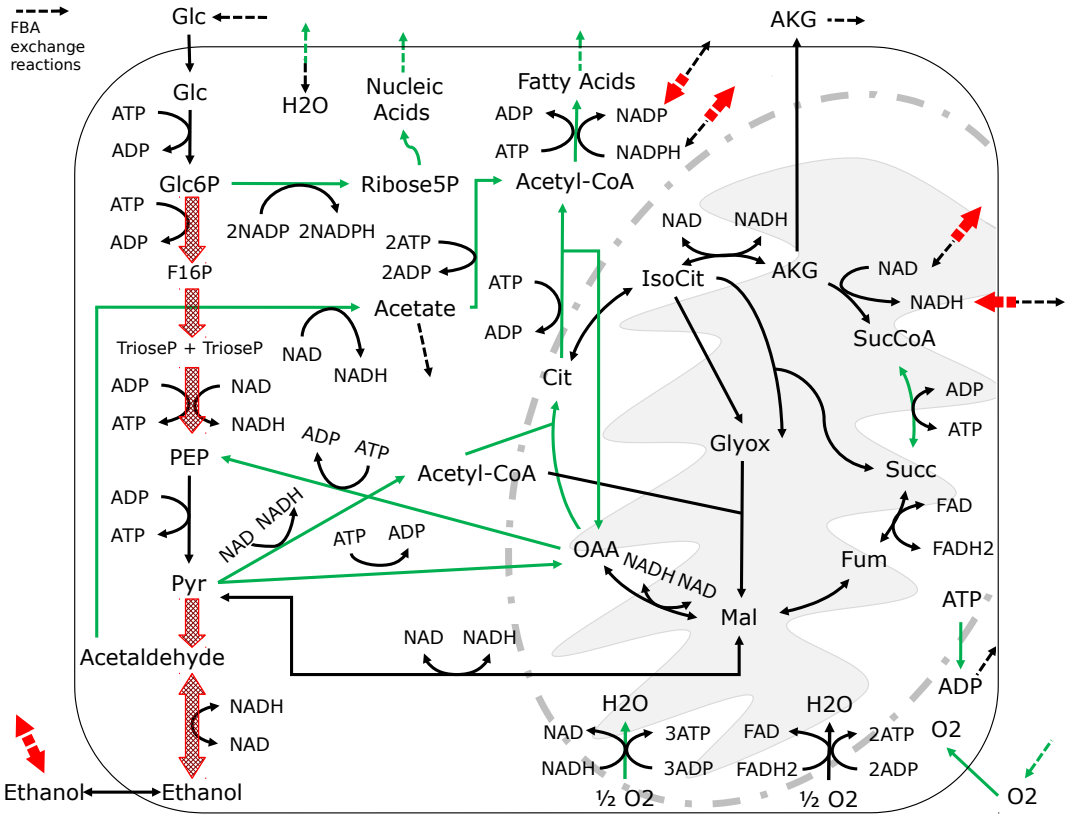


FIGURE 3.7: Wiring diagram and differential fluxes in the CM of yeast. Dashed lines represent exchange reactions for external metabolites, dead-end metabolites and redox cofactors. Colored arrows indicate differentially expressed pathways in C⊕ and C⊖ . Red thick arrows are fluxes "higher" in the C⊕ case when compared to the C⊖ case; green thin arrows illustrate the opposite case. Figure from [58].

In particular, the uptake of this nutrient can be described exploiting a series of glucose uptake indicating the set of glucose values at which each *in-silico* experiment is run and defined as:

$$\{v_g^i \mid \forall\, i < j \quad v_g^i < v_g^j\}_{i,j=1,\dots,L} \tag{3.6}$$

Moreover, to obtain a description more adherent to the actual biological situation, in both phenotypes it must imposed that a null glucose uptake, $v_g^1 = 0$, corresponds to oxygen uptake and ethanol secretion null fluxes, $v_e(v_g^1) = v_o(v_g^1) = 0$.

As stated before, a peculiar characteristic of the C⊕ yeast can be recognized in the fact that the ratio of aerobic glycolysis over respiration grows proportionally to the glucose uptake. Or equivalently it is possible to state that at maximal values of glucose uptake the ethanol secretion flux must overcome the oxygen uptake flux. It is also necessary to exclude those solutions in which respiration is not simultaneously employed with fermentation, meaning that, the oxygen consumption and ethanol secretion fluxes

must be at least once non null in the evaluated interval. All the previously enumerated constraints can be summarized in the logical expressions of Equation 3.7 used to filter the solutions relative to the C⊕ :

$$\Big(\sum_{l=1}^{L} v_e(v_g^l) > 0\Big) \wedge \Big(\sum_{l=1}^{L} v_o(v_g^l) > 0\Big) \wedge \Big(v_e(v_g^1) = 0\Big) \wedge \Big(v_o(v_g^1) = 0\Big) \wedge \Big(v_e(v_g^L) - v_o(v_g^L) > 0\Big) \quad (3.7)$$

In this application, the filtering process has been applied, in order to explore relationship between fermentation and respiration, only at the extreme levels of glucose uptake to favor the emergence of the widest set of behaviors that satisfy Equation 3.7.
A similar reasoning has been applied for the C⊖ phenotype where a null or a specific fermentation level has not been imposed; on the contrary, it has been verified that the oxygen consumption does not grow faster than ethanol production as a function of glucose uptake (i.e. is not C⊕), and of course that oxygen uptake must increase as function of glucose uptake (i.e. for maximal values of glucose uptake it has to be greater than zero). Under these premises, it is possible to define the C⊖ metabolic response Boolean constraint:

$$\Big(v_e(v_g^1) = 0\Big) \wedge \Big(v_o(v_g^1) = 0\Big) \wedge \Big(v_o(v_g^1) - v_o(v_g^L) < 0\Big) \quad (3.8)$$

To populate the two ensembles corresponding to the phenotypes of interest, following the sampling of the space Φ, the retrieved solutions were firstly filtered exploiting the Expression 3.7 to populate the C⊕ ensemble and then the filter defined in Expression 3.8 was used on the remaining solutions to populate the C⊖ ensemble.
Accordingly to this framework, it will be selected with the same probability both solutions where the ratio of aerobic glycolysis over respiration is slightly growing as a function of glucose uptake will and a solution in which this rate significantly increases.

If instead it is preferable to define an ensemble of C⊕ solutions where the response constraint is maintained as much as possible (e.g. the increase in the aerobic glycolysis/respiration is maximum), the search algorithm must be biased, for example, by means of a fitness function able to do it.

To achieve this goal, in [58] I defined for the C⊕ case a fitness function $a(\mathcal{S}_j)$ to be minimized:

$$\Big(v_e(v_g^1) + v_o(v_g^1) + \frac{v_o(v_g^L)}{v_e(v_g^L)}\Big) \quad \frac{(V_{\text{MAX}}\ L)^2}{\sum_{l=1}^{L} v_e(v_g^l) \quad \sum_{l=1}^{L} v_o(v_g^l)} \quad (3.9)$$

Qualitatively speaking, in the heuristic Expression 3.9, the sum terms $v_e(v_g^1) + v_o(v_g^1)$ requires that neither respiration nor fermentation should be observed when glucose uptake is null, whereas the ratio $\frac{v_o(v_g^L)}{v_e(v_g^L)}$ is intended to amplify the fact that the ethanol secretion should overcome the oxygen uptake at the highest glucose intake level. If $v_o(v_g^L) = v_e(v_g^L) = 0$, the second term of Expression 3.9 leads to an indefinite form that has been solved by setting the fraction value to infinity to penalize those phenotypes with no ethanol production at high glucose.

The last term $\frac{(V_{\text{MAX}}\ L)^2}{\sum_{l=1}^{L} v_e(v_g^l)\quad \sum_{l=1}^{L} v_o(v_g^l)}$ has been introduced to impose a greater weight to solutions having a not null glucose kinetics of the two main fluxes, with V_{MAX} being the higher bound value for all the fluxes so that $\sum_{l=1}^{L} v_o(v_g^l) \leq LV_{\text{MAX}}$.

Analogously in the C$\ominus$ case, the fitness function has been defined in the following terms:

$$v_e(v_g^1) + v_o(v_g^1) + \frac{v_e(v_g^L)}{v_o(v_g^L)} + \frac{v_o(v_g^1)}{v_o(v_g^L)} \tag{3.10}$$

In this case, the third term of Expression 3.10 describes the "constraint" by which in the C$\ominus$ phenotype the oxygen uptake should overcome the ethanol secretion at high glucose.

For the C$\ominus$ phenotype, the indefinite form has been solved imposing the fraction value to infinity to penalize those solutions with no oxygen production at high glucose levels. Eventually, the last term has been included to favor those solutions which increase respiration as glucose uptake increases, and the indefinite form has been fixed to infinity if $v_o(v_g^1) = v_o(v_g^L) = 0$.

3.5.2 Tuning the Genetic Algorithm in eeFBA

The GA used with the eeFBA has been implemented in MATLAB in order to integrate it with the other scripts governing the sampling procedure and the execution of the FBA optimization. In this context, it is necessary to illustrate the parametrization used to tune the GA (see [181] for a general introduction on GA and [103] for an application to metabolism), keeping in mind that the goal of modeling the CE was to give a proof of concept for the effectiveness of the eeFBA method and not the identification of the most effective set of the GA parameters. In all the performed runs of the GA

- the initial population has been set to 100 individuals;
- the tournament selection has been of size 4;

- a single point crossover has been used: a random crossover point h is drawn then for the child vector $\mathbf{c}$, then c_i is equal to the c_i of the first parent for all $i < h$ or to the c_i of the second parent otherwise;
- a uniform random mutation has been applied: each entry in the vector $\mathbf{c}$ can be replaced by another uniform random number in [0,1], with a probability of 0.05.

3.5.3 Results emerging with the eeFBA approach

Crabtree-positive and Crabtree-negative ensembles The striking aspect emerging from the eeFBA applied to the investigation of the Crabtree effect is a significative difference in population composition for the ensembles $\mathcal{A}_{\oplus}$ (solutions relative to the C$\oplus$ phenotype), and the ensemble $\mathcal{A}_{\ominus}$, (solutions relative to the C$\ominus$ phenotype). This non intuitive difference will be analyzed in the next lines.

By means of the filtering+sampling algorithm 42448 random OFs have been obtained; among these the filter returned an ensemble of 933 C$\oplus$ solutions (of which 930 are unique solutions, i.e. at least the value of one flux across the different values of glucose uptake is not identical) and en ensemble of 4746 C$\ominus$ solutions (of which 4471 are unique). From an easy comparison of these values it emerges that it is more than 5 times likely to observe the metabolic response typical of the Crabtree-negative yeasts than of observing the CE (estimated as 11% against a 2% respectively) when performing unbiased sampling of the phenotype space. Moreover the fact that the 87% of cases the metabolic response observed can not be ascribed neither to C$\oplus$ nor to C$\ominus$, suggests that the devised constraint-based method is effective in excluding biologically implausible solutions from the phenotype space.
For what concerns the GA, I performed 1000 runs, and from these runs 7455 individuals have been selected according to the C$\oplus$ fitness function (of these 397 are unique solutions), where the C$\ominus$ fitness functions resulted in 11181 (of which 884 are unique) individuals. After checking $\mathcal{A}_{\oplus}$ and $\mathcal{A}_{\ominus}$ with the Boolean filter to discard the false positives, only the population of $\mathcal{A}_{\ominus}$ was reduced to 867 individuals, highlighting the effectiveness of the chosen heuristic.

Perhaps the most significant differences emerged when comparing the properties of the ensembles obtained with the sampling+filtering and GA algorithms. The clustering analysis has been performed twice by means of a hierarchical clustering [182], each one using data deriving from the GA or from the sampling+filtering. For each of these two cases, the analysis has been performed by merging the C$\oplus$ and C$\ominus$ solutions. The sampling+filtering case returned a dendrogram (not shown) that does not allow us to

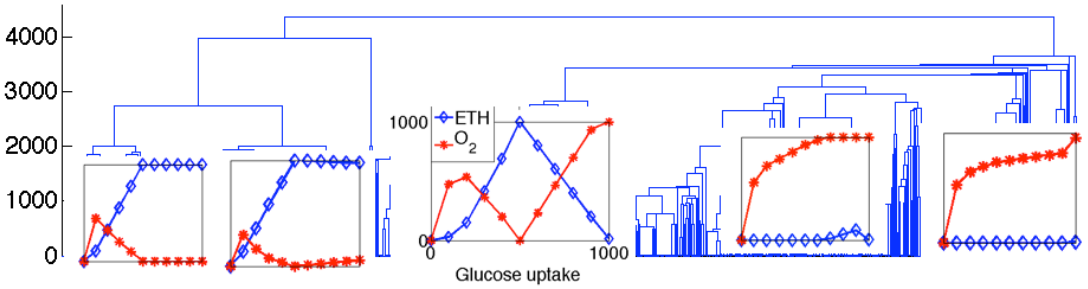


FIGURE 3.8: Dendrogram representing the hierarchical clustering of the solutions (C⊕ and C⊖ merged together) obtained with the GA. In the inserted plots: average of oxygen uptake and ethanol secretion fluxes as a function of glucose uptake for the main clusters. Figure from [58].

identify well separated clusters of solutions, not even those corresponding to the sets of C⊕ and C⊖ solutions.

Instead, a clear distinction between two major groups can be clearly observed in the dendrogram obtained with the GA (shown in Figure 3.8). In this case, the main branch on the left are the set of C⊕ solutions, while on the right the main branch corresponds to the set of C⊖ solutions. Strikingly, no solutions fall into the cluster representative of the wrong ensemble, and even more relevant, some sub-clusters can also be clearly identified within the two ensembles. The main sub-clusters (10 in total) can be identified cutting the histogram at the value 2000 on the Y axis (an arbitrarily chosen distance) and selecting sub-clusters composed of more than 10 individuals. In these sub-clusters the existence of distinct behaviors compliant with the same metabolic response definition (either C⊕ or C⊖) clearly emerged from the analysis of the average behavior of the fluxes representative of respiration and fermentation: representative behaviors of the 5 major sub-clusters are depicted in the plots attached to the dendrogram in Figure 3.8.

Moreover the complete flux profile has been analyzed for those sub-clusters that are most characterizing of the C⊕ and C⊖ phenotypes. In particular, in Figure 3.9 is represented a heat map for all the fluxes of Crabtree-positive (top) and Crabtree-negative (bottom) solutions. The most different fluxes between the two phenotypes are labeled 1, 2 and 3.

Globally, these areas show in the heat map a marked difference in fluxes value, confirming that a rearrangement of a subset of fluxes is associated with Crabtree-positive and Crabtree-negative solutions/phenotypes (as also elucidated in 3.5.3). Within some clusters of the C⊖ phenotype is also possible to identify smaller subset of reactions (grey rectangles) underlining a higher plasticity.

Crabtree-positive vs Crabtree-negative average behavior: differentially expressed pathways At a more biochemical level the differences between the C⊖ e C⊕

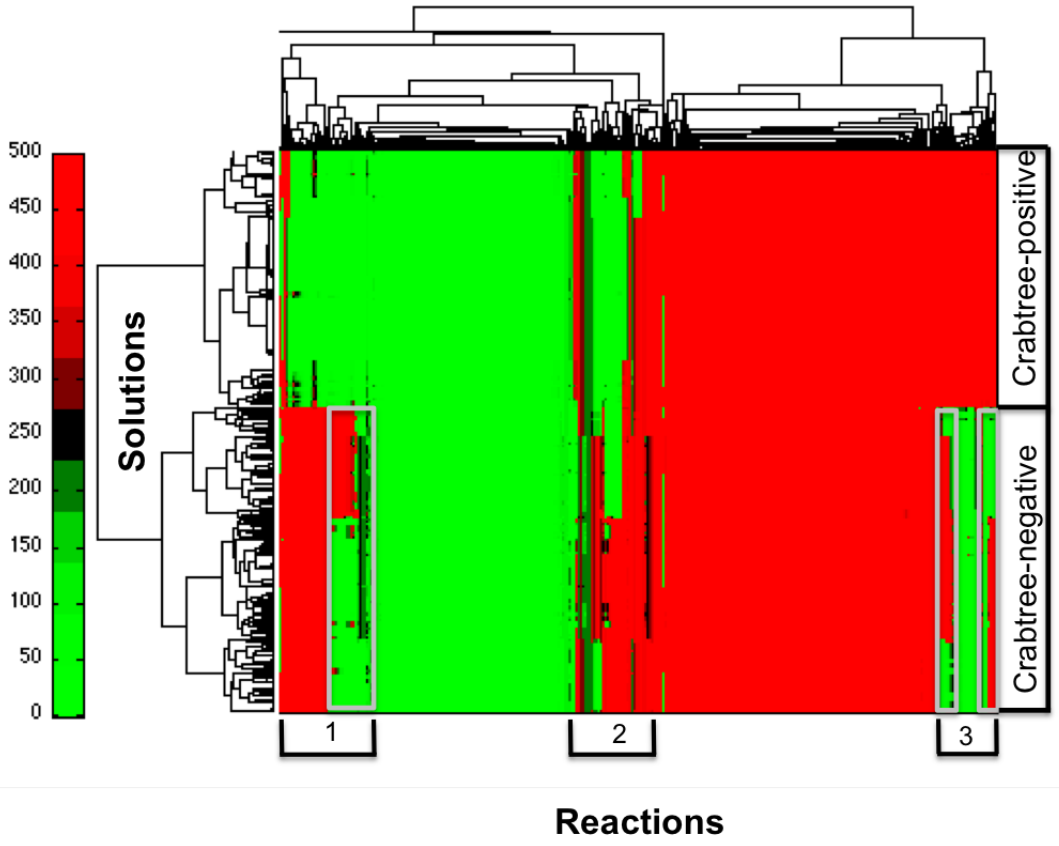


FIGURE 3.9: Heatmap illustrating the flux profile of the identified clusters (in the heatmap are not shown clusters exhibiting a behavior similar to that of the central cluster in Figure 3.8). In the heatmap each row represent a solution (i.e. a flux distribution), while columns illustrate the value of the fluxes for all the reactions of the model at the 11 evaluated levels of simulated glucose. Adjacent fluxes are the most similar, regardless of the relative level of glucose. Numbers at the bottom indicate most different fluxes between C⊕ and C⊖ solutions. Grey rectangles highlight regions with a higher plasticity. Figure from [58].

ensembles has been investigated limiting the analysis to the solutions belonging to the clusters that better characterize the C⊕ and C⊕ phenotypes. To do this, out of the 10 obtained clusters, 3 were discarded because exhibiting an hybrid behavior close to those of solutions not belonging neither to C⊕ , nor to C⊖ (i.e. similar to the small plot in the middle of Figure 3.8).

In this context a Kolmogorov-Smirnov test has been applied to identify fluxes that significantly distinguish the two ensembles. The statistical test has been performed on all the 10 non-null levels of glucose: the fluxes that are significantly different for at least 9 out of 10 levels of glucose are marked in Figure 3.7. In particular red arrows indicate that for a given reaction, fluxes are higher in the C⊕ case that in the C⊖ one; on the contrary a green arrow indicate that higher fluxes are found in the C⊖ case. Black

arrows belong to reactions whose Kolmogorov - Smirnov cumulative value is under the threshold of 9.

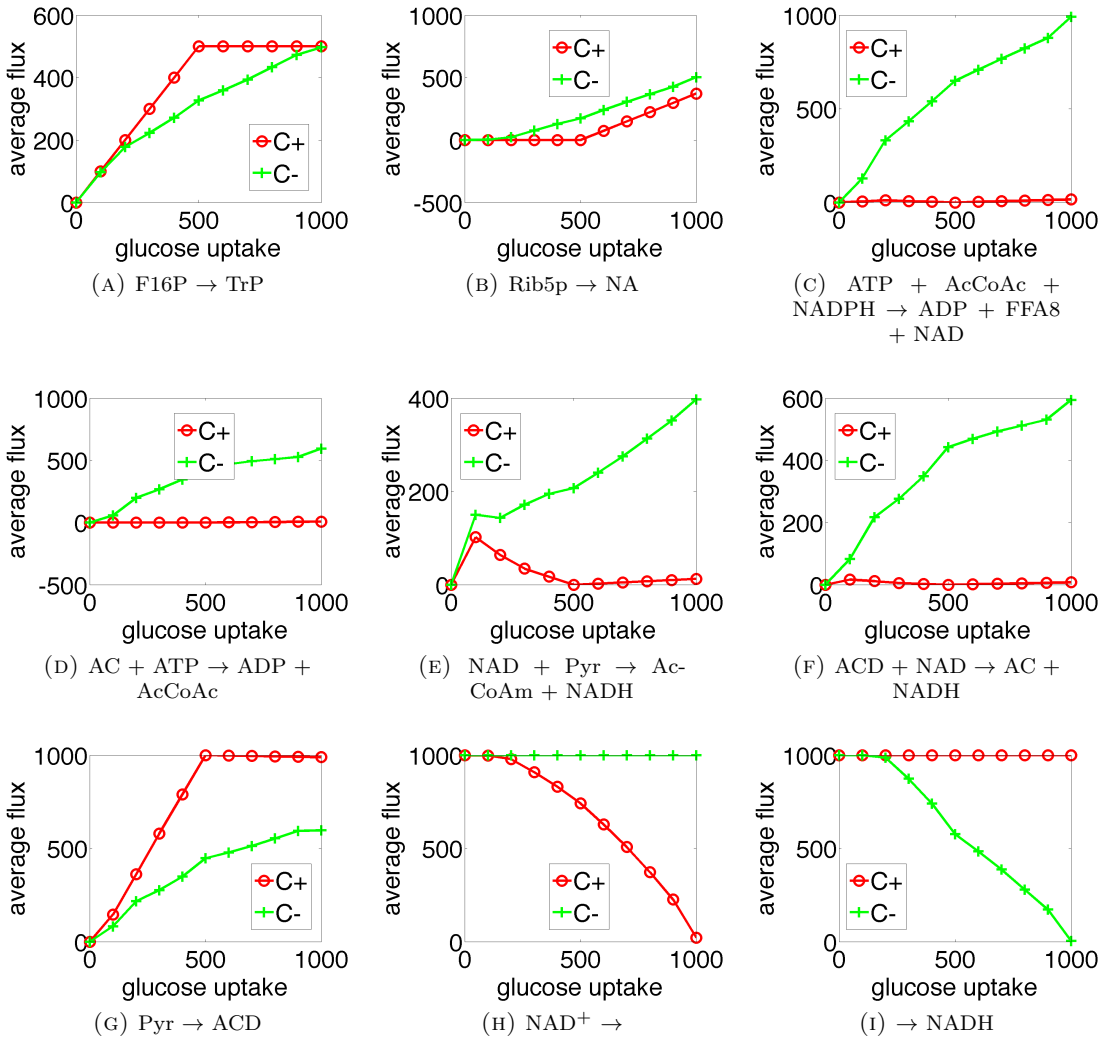
Significative fluxes emerging from the Kolmogorov - Smirnov test confirmed that the eeFBA approach is able to describe fluxes (evaluated as a function of glucose uptake) typical of the Crabtree effect and of the Crabtree-negative phenotype (Figure 3.10). This is underlined by the differences in reactions ascribable to fermentation and respiration. In particular, in accordance to experimental results [178, 180], it emerges that the glycolytic pathway is enhanced in the Crabtree phenotype; a fact confirmed by higher values for fluxes belonging to glycolysis (e.g. F16P $\rightarrow$ Trp in Figure 3.10a) in the case of C$\oplus$ solutions with respect to fluxes distributions classified under the C$\ominus$ phenotype. Moreover, the expected phenotype is sustained by other fluxes besides glycolysis:

- pyruvate $\rightarrow$ acetylCoA (pyruvate carboxylase, Figure 3.10e) is enhanced in C$\ominus$ phenotype [178, 180] as it is a bridge towards the TCA cycle and the respiratory metabolism of yeast;
- pyruvate $\rightarrow$ acetaldehyde (pyruvate decarboxylase, Figure 3.10g) is enhanced in C$\oplus$ phenotype [178, 180] as it leads to ethanol production, final step of the fermentative metabolism;
- acetaldehyde $\rightarrow$ acetate (acetaldehyde dehydrogenase, Figure 3.10f) and acetate $\rightarrow$ acetylCoA (acetylCoA synthetase, Figure 3.10d), redirecting C towards the TCA cycle, have higher values in the case of C$\ominus$ [178, 180];
- acetylCoA $\rightarrow$ fatty acids (Figure 3.10c) and ribose 5-phosphate $\rightarrow$ nucleic acids (Figure 3.10b) [180]: according to [178] show higher fluxes in C$\ominus$ cells because of a greater production of biomass in comparison to C$\oplus$ cells [183].

Intriguingly, the eeFBA approach has revealed also less expected differences in fluxes related to redox cofactors such as NAD^+ and NADH (Figure 3.10h and 3.10i). In the case of C$\oplus$ solutions, analyses suggest an enhanced used of NADH and a surplus of NAD^+ (with respect to the C$\ominus$ solutions) due to the presence of a higher uptake of NADH and a higher outflow of NAD^+.

In conclusion it is worth to underline some relevant features emerged comparing the Crabtree-positive and Crabtree-negative phenotypes of yeasts by means of the eeFBA.

- The sampled set of OFs has been divided, by means of a evolutionary algorithm, in two distinct ensembles of solutions (i.e., specific distributions of metabolic fluxes), that represent the Crabtree-positive yeasts and the Crabtree-negative yeasts.



(A) F16P → TrP (B) Rib5p → NA (C) ATP + AcCoAc + NADPH → ADP + FFA8 + NAD

(D) AC + ATP → ADP + AcCoAc (E) NAD + Pyr → AcCoAm + NADH (F) ACD + NAD → AC + NADH

(G) Pyr → ACD (H) NAD^+ → (I) → NADH

FIGURE 3.10: The average flux as a function of glucose uptake of a specific reaction for both the C⊕ and the C⊖ ensembles. The high variability in each ensemble is due to the fact that the whole ensemble has been evaluated instead that characterizing the distinct sub-phenotypes that can be observed in the heat map (Figure 3.9). Figure from [58].

- In eeFBA, a genetic algorithm has been used for the first time to identify functional states that are in agreement with experimentally observed phenotypes. Albeit, evolutionary algorithms have been previously employed within the context of FBA, e.g. to rapidly identify gene deletion strategies to optimize a desired phenotypic objective function [103] or for nonlinear optimization problems [184].

- A cluster analysis has been exploited to further characterize the ensembles. The procedure (implemented in MATLAB) has been able to divide the ensemble into subsets highlighted with grey rectangles in in the heat map reported in Figure 3.9. These regions showing high plasticity may represent distinct sub-phenotypes sustaining the same macro-phenotype (see [185]). This fact suggest that the developed

method could be profitably used for the characterization of metabolic plasticity in cases where a macro-phenotype, such as cancer, may be compatible with different metabolic configurations [146, 186–188], or metabolic rearrangements emerging from the modification of evolutionary, physio-pathological or developmental constraints.

- The "performance" of the system in terms of produced biomass has not been the target of the investigations due to the fact that even poorly performing phenotypes can be relevant in the search for the "design principles" of the phenomenon under study. Moreover, it has been shown, at least for energetic metabolism in mammals [189], that an optimal phenotype under a given selective condition, may be suboptimal or even inefficient under a different condition. As well in yeast, the same fitness level, in static environmental conditions, has been identified in strains showing different abilities to respond to changes in carbon source. In the case of a change in the carbon source (i.e. a dynamic alteration of the environment), the fitness of the strains may change of a significative value [190].

- All the analyses has been done exploiting a simple model representing the main pathways of yeast energetic metabolism and thereby confirming the potentialities of CMs.

Chapter 4

Interaction-based analysis

The second step of the pipeline presented in this thesis (see Figure 2.7) is devoted to the investigation of metabolic networks exploiting interaction-based analysis, a framework that makes extensive use of graph theoretical techniques and concepts from literature on random and scale-free networks.
Before illustrating the computational approaches applied to the models investigated in Chapter 3, I will introduce, in the following section, some basic concept of graph theory and a short description of network classification.

4.1 Elements of graph theory

The analysis of the structural properties of a network (i.e. its topology) is made possible thanks to the standard notions in graph theory [191, 192], of which I will introduce some key concepts.

A graph G is defined as a pair of elements V and E

$$G =< V, E >$$

where:

- V is the ensemble of graph components called nodes: $V = \{v_1, v_2, \ldots, v_n\}$;
- E is the ensemble of links defining interactions among nodes: $E = \{e_1, e_2, \ldots, e_m\}$; each element e_h of the ensemble E, with $h = 1, \ldots, m$, corresponds to a pair of nodes $e_h = (v_i, v_j)$ where $v_i, v_j \in V$.

A graph is said to be directed or oriented, if each link belonging to it is defined by an ordered pair of nodes, more formally $\forall(v_1, v_2) \in V$, $e = (v_1, v_2)$, $e' = (v_2, v_1) \Rightarrow e \neq e'$. Metabolic networks belong to this class of graphs where the orientation represent the directionality of the mass flow through a certain reaction. If instead links do not show any directionality, the graph can be defined as "indirected". This is the case, for example, of protein-protein interaction networks, where it is obvious that two proteins establish a mutual relation. Therefore, the choice to use a directed or indirected graph is only dictated by the nature of the biological system under evaluation [191].

Another relevant aspect for the study of biological networks is the definition of a subgraph that can be seen as a portion of G. The relevance of these structures is due to the fact that they are able to define substructures such as motifs and modules that help to understand the structure of the network under examination reducing its complexity.

Node degree In graph theory, the most significant characteristic of a node is its degree k [192].
For every node v_i in the network, its degree k_i can be calculated as the number of nodes connected to it through a link $k_i = |K_i|$ being $K_i = \{e_1 \in E \,|\, e_j = (v_h, v_k)\, v_h, v_k \in V, h \neq k$ then $h = i$ or $k = i\}$. In the case of a directed graph (e.g. metabolic networks), it is possible to assign to every v_i an *in-degree* and an *out-degree*:

- the *in-coming degree* of v_i, k_i^{in}, is the number of nodes forming a link that ends in v_i. $k_i^{in} = |K_i^{in}|$ being $K_i^{in} = \{e_1 \in E \,|\, e_j = (v_h, v_k)\, k = 1, \ldots, m\, j = 1, \ldots, m\}$;
- the *out-coming degree* of v_i, k_i^{out}, is the number of nodes forming a link that starts in v_i. $k_i^{out} = |K_i^{out}|$ being $K_i^{out} = \{e_1 \in E \,|\, e_j = (v_h, v_k)\, k = 1, \ldots, m\, j = 1, \ldots, m\}$

.

Moreover, for an indirected graph having n nodes and m links, it is possible to calculate the *average degree* as $< k >= \frac{2m}{n}$ [191].

Bipartite graphs A bipartite graph V is a network in which nodes can be divided into two separated ensembles V_1 and V_2, such that $V = V_1 \cup V_2$ and $V_1 \cap V_2 = \emptyset$. In these graph, each node of V_1 establish a link with a node of V_2 [192].

As already discussed in Section 2.3, bipartite graphs are a common representation for metabolic networks where it is possible to identify two ensembles of nodes: metabolites and reactions [192].

Degree distribution When the degree of every node of the network is calculated, it is possible to define the degree distribution $P(k)$ [191, 192] as the probability that a given node of the network assumes the degree k:

$$P(k) = \frac{N(k)}{n}$$

where $N(k)$ is the number of nodes having degree k. Since $P(k)$ is a probability:

$$\sum_{k=1}^{\infty} P(k) = 1,$$

this means that the sum of the degree distributions, calculated for every value of k that can be identified in the network, has always value 1. The relevance of the degree distribution in network analysis is due to the fact that it allows to discriminate different network topologies such as random, scale-free and hierarchical (see Section 4.2 for more details).

The node degree can greatly vary in many real networks (including biological networks); it is possible to identify isolated nodes having $k = 0$ and hubs having high value of k and hence an elevated number of connections. Often the degree distribution function is represented with a logarithmic graph where it is easier to identify hubs possibly existent in the network.

Clustering coefficient The clustering coefficient, C_i [191, 192] is the fraction of existing links among the nodes connected to node v_i of the network and indicates the extension of the connections of these last nodes among them. This is defined as follows:

$$C_i = \frac{2N_i}{k_i(k_i - 1)}$$

where k_i is the degree of the node v_i, N_i indicates the number of links among all the nodes connected to v_i, and $k_i(k_i - 1)/2$ being the total number of possible links among all the nodes connected to node v_i. C_i can assume only values in the interval [0,1], indeed:

- $C_i = 0$ if all nodes connected to the node v_i are not connected among them;
- $C_i = 1$ if all nodes connected to the node v_i are connected among them.

The clustering coefficient is a measure for the local density of the network, since the higher is the value C_i, the more nodes connected to the node v_i are interconnected among them, and higher is the probability that a clique will be formed [1].

[1] A clique is a subgraph where for each possible pair of nodes there is a link connecting them.

4.2 Network topologies

Thanks to the global measures of the graph given by the functions $P(k)$ e $C(k)$ described in Section 4.1, it is possible to identify three different network topologies: (I) random networks, (II) scale-free (or scale invariant) networks, (III) hierarchical networks.

4.2.1 Random networks

In random networks, links among the n nodes are randomly defined and every pair of nodes have the same probability p to be connected. Steps to reconstruct a random network are hereafter listed:

1. initialization of n isolated nodes;
2. selection of a pair of nodes and generation of a random number in the interval [0,1]: if the number is higher than the probability p, the pair of nodes will be connected by means of a link, otherwise the two nodes will remain disconnected;
3. step 2 is repeated for each of the $n(n-1)/2$ pair of nodes.

In this way it is possible to generate a statistically homogeneous network, where the most part of the nodes has the same degree of the average degree $< k >$ of the whole network [191, 192]: in a random graph the degree distribution $P(k)$ follows a Poissonian distribution, and there are few nodes having a markedly high or low degree.
The clustering coefficient $C(k)$ in a random network, is a constant function independent from the nodes' degree, a fact that underlines the lack of modularity in this kind of networks.

A further relevant property of random networks is the small-world effect: the average length of paths [2] is proportional to the logarithm of the number of nodes in the graph (i.e. $l \sim logN$) [191]. The small-world property is shared by all the complex networks and indicates that, regardless the dimension of the network, every pair of nodes can be connected exploiting a contained number of links; a property that allows a fast transmission of information through the network.

4.2.2 Scale-free networks

A great number of biological networks exhibit the so called scale-free network topology. This kind of topology can be obtained starting from a random network and modifying

[2] The average length of paths is defined as the average shortest-path between two nodes.

it by adding new nodes that will preferentially connect to nodes already having a high degree. This process is defined preferential attachment (or rich get richer) [191, 192], and constitutes an hypothesis to explain the formation of hubs: indeed, if a node has already established many links, it will attract the formation of new links with a higher probability [191].

Differently from random networks, scale-free networks are not statistically homogeneous, but are characterized by the presence of few hubs and of many nodes with few connections. Scale-free networks can be recognized by means of an exponential degree distribution $P(k) = k^{-\gamma}$, also defined power-law. In general, the γ coefficient assumes a value in the interval [2, 3], suggesting that these networks are characterized by the ultra small-world property, where the average length of paths is $l \sim log(logN)$ [191]. The ultra small-world property has been identified in metabolic networks, where short paths of 3 or 4 reactions connect the most part of metabolites pairs, meaning that the local perturbation (removal of a metabolic node) could rapidly reach the entire network.

For what concerns the clustering coefficient, random and scale-free networks share the same behavior (i.e. $C(k)$ is a constant function and consequently they do not exhibit modularity).

Lastly, scale-free networks are robust towards the random removal of a node (random attack). However, if nodes are selectively removed, and if a removed node is a hub, scale-free networks exhibit a high vulnerability resulting in a dramatic disconnection of the graph.

4.2.3 Hierarchical networks

Hierarchical networks integrate scale-free topology and presence of local clusters. This is underlined by the fact that, as in the case of scale-free networks, the degree distribution is a power-law with $\gamma \cong 2.2$ [191, 192].

Instead, the modularity of hierarchical networks can be deduced by the fact that the $C(k)$ is no longer a constant function but proportional to $1/k$, meaning that given a node, other nodes connected to it will tend to form a clique. The identification of clusters can be seen as the observation of subnetworks such as modules and motifs.

A module (or cluster) is a subgraph that can be functionally separated by the rest of the network and that is involved in a determined biological function showing a high intra-connection among nodes and a low inter-connection with the rest of the network. A motif is a small subgraph having a well defined structure that can be found also in

the global structure of the network with a frequency markedly higher with respect to the number of occurrences obtainable in a random network of the same size.

From the biological point of view, modularity seems to be a key aspect to sustain many different functions. This fact is confirmed by the finding that the average clustering coefficient $< C >$ is significantly higher in biological networks with respect to random network of equal size and degree distribution [191].

In Section 4.4.1 I will apply graph theory tools investigate topological properties of GW metabolic networks previously analyzed by means of constraint-based methods (see Chapter 3).

4.3 Network and fluxes visualization

Due to the pervasive nature of networks in the domain of biological complex systems, during last years a strong need of tools able to represent them in an effective way has emerged in the modeling community. To satisfy this need, visual representations of biological networks have been extensively used to give an immediate representation of the complexity beyond the systems and to sum-up some relevant information [193]. The recourse to visualization strategies has been motivated by the fact that our brain has acquired, through evolution, a remarkable capability to process visual information in order to identify patterns (e.g. biochemical pathways) and other relevant topological features (e.g. hubs) [194].

The "visual complexity" of these graphical representations range over various orders of magnitude spanning from the description of a single pathway (signal transduction, metabolic pathway, interaction pool of a protein), to the representation of the interactions involving every single component of cellular systems. A pioneering example in this sense is the *Biochemical Pathways Poster* developed by Michal in 1993 [195] that provided an organized representation of the cellular metabolism still used as a reference in many laboratories.

However, the later development of high-throughput -omics technologies has imposed a change of paradigm for the definition of these representations. Indeed nowadays, due to huge amount of data to be described, it is no longer possible to apply the manual curation and refinement, and this reason motivated the thrive of softwares able to automatically generate network visualizations.

Even if the huge effort has produced outstanding results (see [196] for an extensive review), in this domain some challenges are still open.

A first challenge is related to the scalability of these methods: since most of the softwares make use of standard visualization packages, the most common layout of the network is often a very uninformative "hair ball" [193]. The definition of this layout is mainly due to the lack of knowledge on the inner organization of the network (e.g. cellular localization of elements, molecular functions, structure of proteic complexes, etc.), but also to the difficulty of representing the system in a way expected by the user (generally a biologist having a *forma mentis* dictated by the classical biochemistry textbooks). A second challenge is connected to the retrieval of desired information and to the network navigation for the exploration of the "surroundings" of a given element; an activity that could generate insights to direct the investigation of the system.
A last challenge can be identified in the enrichment of the visualization by adding further information (e.g. attributes from external sources and database), and maintaining a good readability of the relevant information.

In the following sections I will illustrate the visualization approaches that have been used in this thesis to represent metabolic networks and flux distributions.

4.3.1 PAINT4NET

The Paint4Net toolbox [197] (an extension of the COBRA Toolbox (see Section 3.3) in MATLAB), has been designed to offer a visualization tool for metabolites and reactions in metabolic networks.

Paint4Net exploits two visualization approaches. The first approach has been devised to represent large networks, in this case the information about nodes and links is shown only when the cursor is moved over one of these elements (Figure 4.1); while the second approach is instead more suited to represent subnetworks or few metabolic pathways since it shows every element label (Figure 4.1). Moreover, the latter approach allows to map fluxes values on links representing reactions, making it a useful tool to have an overview of flux distributions.

Additional features of the software allow to retrieve a list of metabolites (including the dead-end metabolites), exclude selected metabolites from the representation and find subnetworks detached from the main component.

Relying on the COBRA Toolbox of MATLAB, Paint4Net is able to deal with models in SBML format [198], and in this context it is able to take advantage of the Bioinformatics toolbox of MATLAB to generate a layout for the network under investigation.

Unfortunately, Paint4Net has many limitations: (I) the layout of the network is automatically generated and often it is particularly hard to decipher by the user, (II) links

are assigned in a disorganized way exploiting curved connectors that limit the readability of the model (see Figure 4.1 for an example), (III) the size of the visualized network is limited at about a thousand reactions, whereas GW networks entail many thousands of reactions.

For these reasons it has been preferred to use a more powerful and reliable tool (see next section), even if it is not usable under the MATLAB environment.

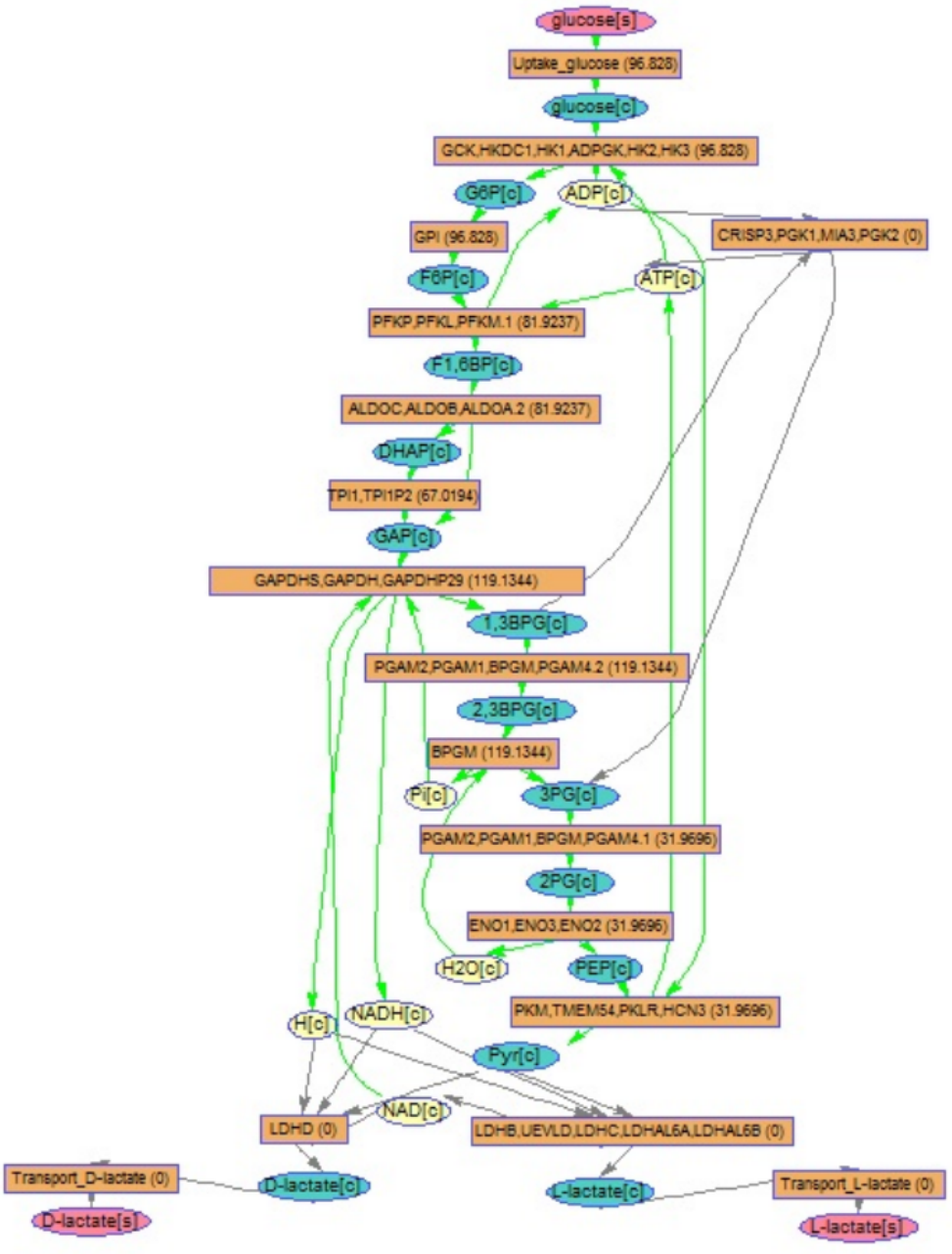


FIGURE 4.1: An example of metabolic map for the glycolysis obtained with Paint4Net.

4.3.2 Network analysis: Cytoscape

Cytoscape [107, 199], is a constantly maintained and developed bioinformatic software to visualize networks of molecular interaction, biological pathways and in general is able to represent biological networks of various nature (as well as non biological ones). The software is commonly regarded as one of the most complete (in terms of functionalities) and reliable (see Figure 4.2 for a comparison with the state of the art) environments for network analysis and visualization.

Feature	CY	GM	VA	OS	CD	AR	IN	GG	PI	PR	BL	PA
Free for academic use	✓	✓	✓	✓	✓				✓	✓	✓	✓
Free for commercial use	✓	✓	✓		✓				✓	✓	✓	
Open source	✓	✓							✓	✓	✓	
Curated pathway/network content		✓		✓		✓	✓	✓				
Standard file format support	✓		✓		✓				✓	✓		✓
User-defined networks/pathways	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Functionality to infer new pathways	✓		✓			✓		✓	✓			
GO/pathway enrichment analysis	✓	✓	✓				✓	✓				
Automated graph layout	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓
Complex criteria for visual properties		✓				✓	✓	✓		✓	✓	✓
Multiple visual styles	✓		✓	✓		✓	✓			✓		
Advanced node selection	✓		✓	✓		✓	✓	✓	✓	✓	✓	✓
Customizable gene/protein database		✓	✓			✓		✓	✓			
Rich graphical annotation		✓	✓				✓	✓				✓
Statistical network analysis	✓		✓				✓	✓	✓		✓	
Extensible functionality: plugins or API	✓		✓		✓	✓	✓	✓	✓			
Quantitative pathway simulation					✓	✓						

FIGURE 4.2: Comparison of network analysis softwares available in literature. CY, Cytoscape (www.cytoscape.org); GM, GenMAPP (www.genmapp.org/introduction.html); VA, VisANT (visant.bu.edu); OS, Osprey (biodata.mshri.on.ca/osprey); CD, CellDesigner (www.celldesigner.org); AR, Ariadne Genomics Pathway Studio (www.elsevier.com/online-tools/pathway-studio/about); IN, Ingenuity Pathways Analysis (www.ingenuity.com/products/ipa); GG, GeneGo (sbp.qfab.org/software-architecture/genego); PI, PIANA (sbi.imim.es/piana); PR, ProViz (cbi.labri.fr/eng/proviz.htm); BL, BioLayout (www.biolayout.org); PA, PATIKA (www.patika.org/software). Figure from [107].

Core functions and plugins Cytoscape is able to perform many different core tasks involving several aspects of the network analysis such as the visual representation of the graph (along with labels of data), the selection of nodes and links, the integration of external attributes, and the manual or automatic definition of the network layout exploiting several algorithms that can be tuned *ad hoc* [107] (see Figure 4.3 for some examples). The software offers the possibility to create a network from scratch, as well as to load networks from the most widely used formats for Systems Biology such as SBML, SIF (Simple Interaction File), GML, XGMML or using built-in functionalities such as the Pathway Commons Web Service Client or the import from a text file or an Excel spreadsheet.

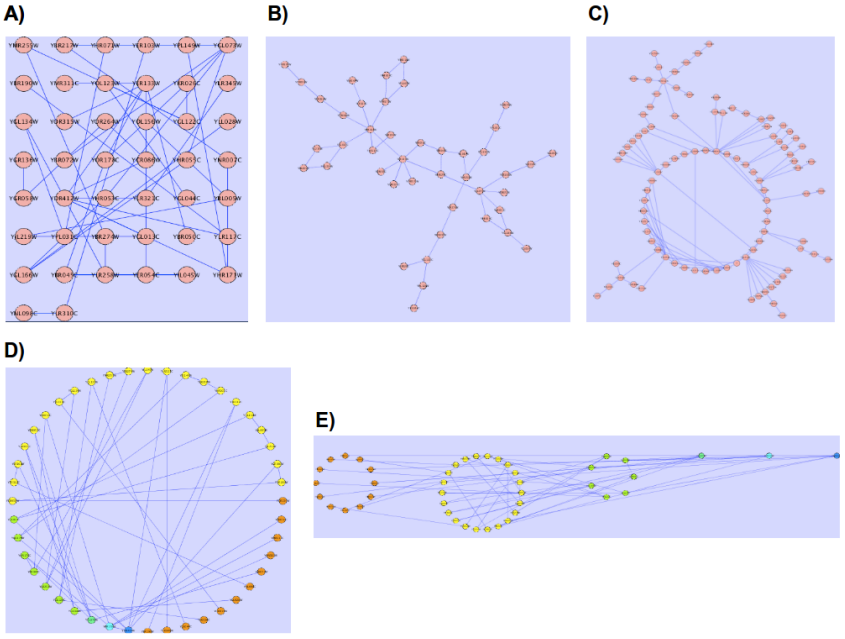


FIGURE 4.3: Some examples of Cytoscape layouts: A) Grid Layout; B) Force Directed Layout; C) Spring-Embedded Layout; D) Attribute Circle Layout; E) Group Attributes Layout. Modified from http://cytoscape.org/manual

Moreover, Cytoscape core functions can be extended through additional modules (called plugins) available mainly through the online *Cytoscape App Store* where they are organized in categories such as data import from DBs, data visualization, topological analysis, network enrichment, clustering, layout, ontologies, identification of modules and motifs.

As stated at the beginning of this section, a key function of Cytoscape is the ability to integrate in the network some attributes (i.e. additional information on links and nodes) to further describe properties of the elements of the network and their relation. These attributes are pairs $(name, value)$ where names of both nodes and links are associated to several kind of data such as: text, numerical values or external references.

In Cytoscape, through the *data-to-visual attribute mapping*, attributes can be mapped to the network exploiting visual properties of nodes and links. Indeed, the graphical representation of these elements can be modified using VizMapper, the visual editor of Cytoscape; a tool that allows personalize node and link properties such as: color, size, shape, label, etc. In particular, the visual mapping of attributes can be performed via three different methods [200]: (I) the passthrough mapper is used to map an item to the corresponding unique value (e.g. the name of a node); (II) the continuous mapper is used to obtain a visual item (e.g. size of the line indicating a link) proportional to a value in a continuous range (e.g. the flux value); (III) the discrete mapper, is the

choice when mapping discrete values (e.g. different types of nodes) on "discrete" visual attributes (e.g. different shapes).

Lastly, the software is able to filter both nodes and links on the basis of a given attribute value or on the combination of several attributes. It is also possible to filter nodes and links selecting their first neighbors.

These core functionalities are automatically exploited by plugins. In the next section I will introduce a plugin used to map fluxes distributions obtained by means of constraint-based analysis on a metabolic network.

4.3.2.1 The CyFluxViz plugin

FluxViz or (CyFluxViz in recent releases) [106] is an open-source Cytoscape plug-in for the visualization of flux distributions in metabolic networks. The tool has been firstly developed as a frontend for FASIMU [151], a software for FBA, but since it relies exclusively on network structure and flux data, it can be used with any simulation tool (in this thesis, notably with data from the COBRA toolbox).

Figure 4.4 illustrates the global workflow of the tool. After the loading of a net-

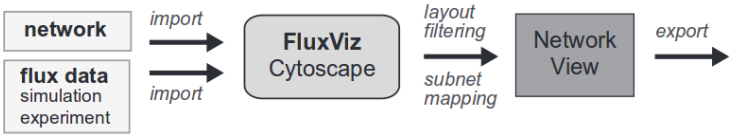


FIGURE 4.4: The FluxViz workflow. Figure from [106].

work, the first step of the procedure is the import of a table containing a list of pairs $(reactionidentifier, fluxvalue)$. Subsequently, a view of the network is generated by means of a layout (both automatic or manual). In this step the plugin automatically set the mapping properties of edges and returns a visual representation where the size of lines corresponding to a link are proportional to the flux value. The output network can be further analyzed by Cytoscape core functionalities and plugins (e.g. to perform a topological analysis) and a snapshot of the network (or of a filtered subnetwork) can be exported as image. Lastly, it is possible to modify the mapping function in order to represent additional information such as gene expression data.

FluxViz is a powerful tool to obtain a bird-eye view of flux distributions but it is also able to setup deeper analyses. However, there are some bugs in its implementation and the plugin has not been updated to meet the requirements of the last version of

Cytoscape (versions 3.X). For these reasons, and due to the fact that all its capabilities can be implemented exploiting the core functions of Cytoscape, in this thesis I decided not to use FluxViz, choosing instead to exploit the built-in tools of Cytoscape to perform network visualization. Nevertheless, analyzing FluxViz has been useful to investigate the feasibility of a Cytoscape plugin with features useful to the pipeline object of this thesis.

In Figure 4.5 is illustrated a network visualization for the yeast CM developed in Section 3.5 and analyzed by means of the eeFBA approach. It is possible to notice that the network is represented by a substrate-enzyme bipartite graph where two classes of nodes (reactions, green diamonds and metabolites, blue circles) have been defined.

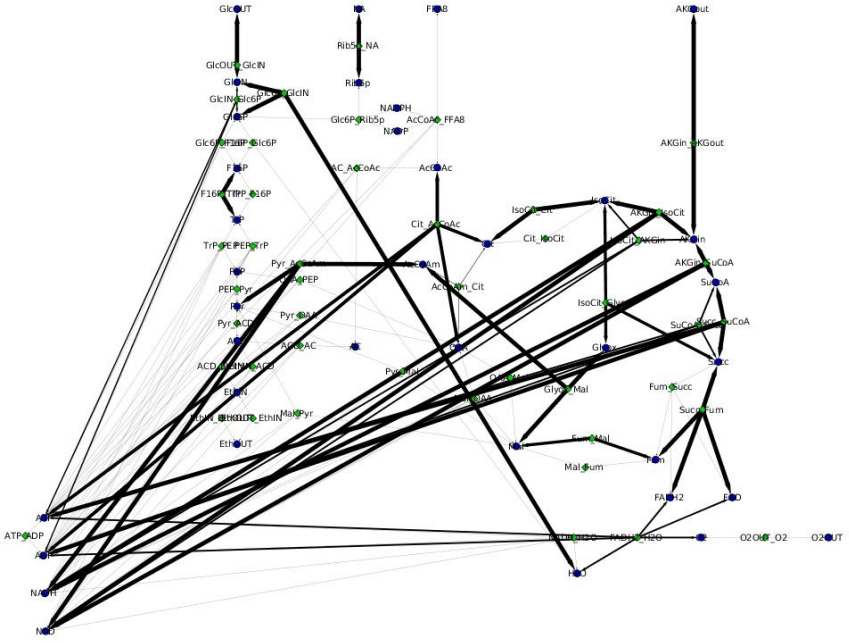


FIGURE 4.5: Visualization of flux distribution for the yeast core model.

The metabolic network has been arranged exploiting a manual layout representing the main metabolic pathways in a form that can be easily recognized by a biologist/modeler (e.g. the TCA cycle is a circle in the middle of the network, while glycolysis is a line at his left side), cofactors have been positioned on the extreme left to help readability. Moreover, exploiting the core functions of Cytoscape has been possible to map, on links connecting reaction and metabolites nodes (and thereby representing the reaction itself), the flux value derived from the average flux distribution for a "sub-phenotypes" emerged with the clustering of the C$\ominus$ ensemble (see Section 3.5.3, and Figure 3.9).

A further visualization of the yeast CM has been developed starting from the previous one but including, this time, the mapping of the degree node. In Figure 4.6 the size of the nodes (both reactions and metabolites, i.e. diamonds and circles) is proportional to the node degree calculated using the NetworkAnalyzer plugin (see Section 4.4.1).

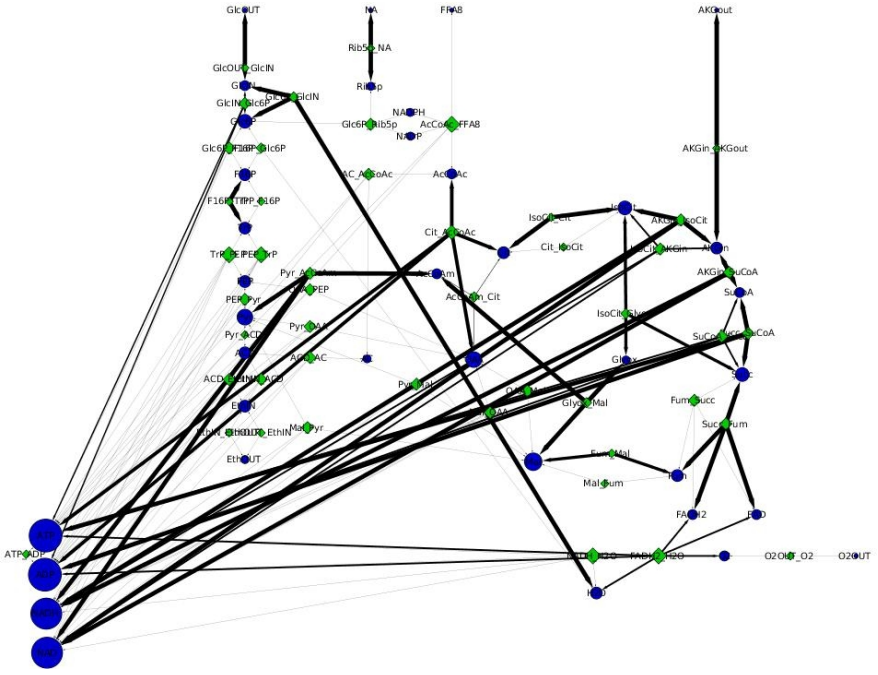


FIGURE 4.6: Visualization of flux distribution and node degree for the yeast core model.

Following the same approach, a visual representation has been developed also for the HMR "reference" CM developed in Section 3.4.2. In particular in Figure 4.7 (and in Figure 4.8 after the filtering of cofactors) is illustrated a portion of the whole network where link thickness is proportional to the value of the z-score for that reaction when comparing two different tested conditions. Moreover, the green/red color indicates the sign of the score.

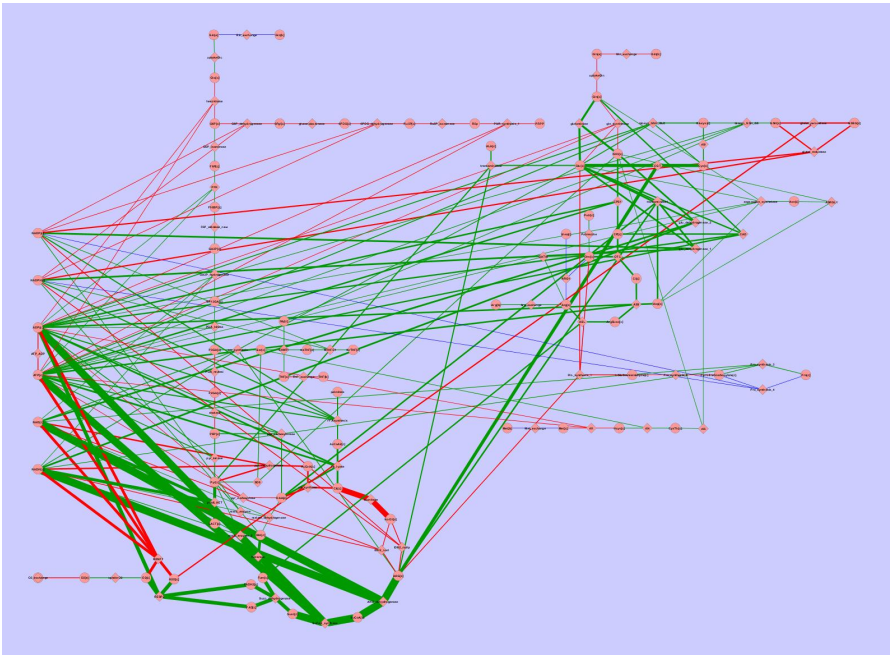


FIGURE 4.7: Visualization of flux distribution for the HMR core model. The color of the links indicate the sign of the z-score, while the thickness identify the value.

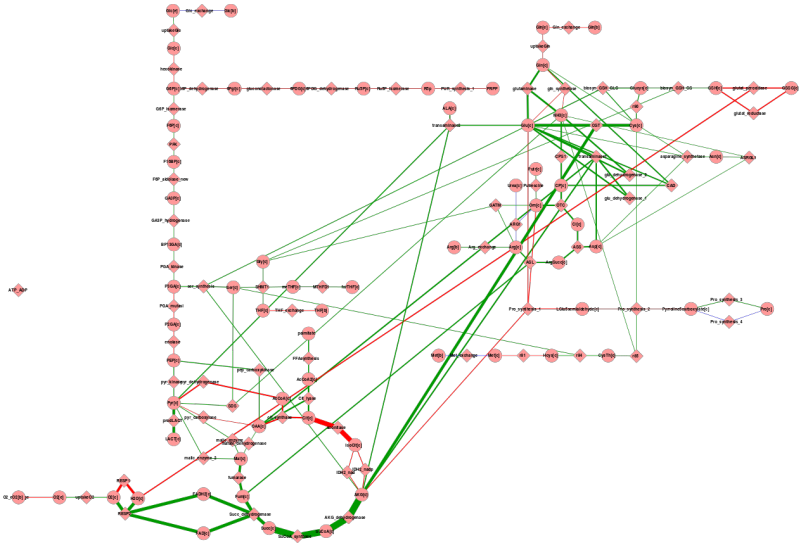


FIGURE 4.8: Visualization of flux distribution for the HMR core model without cofactors. The color of the links indicate the sign of the z-score, while the thickness identify the value.

4.4 Network metrics and correlations in fluxes distributions

4.4.1 Network metrics of genome-wide and core models

Graph theory measures illustrated in Section 4.1 have been exploited to perform a topological analysis networks, with the aim of investigating their structural properties[3]. The analysis has been conducted on both the genome-wide and core metabolic networks defined using a substrate graph (see Section 2.3 for a definition) and exploiting the NetworkAnalyzer plugin of Cytoscape.
Overall, it emerged that both in the normal and in the cancer condition, the models have the same structural characteristics. More in particular, this type of approach highlighted four important properties of these networks: hierarchical and scale-free topology, modularity, ultra small-world property, disassortative nature.

The first outcome is the hierarchical topology that, by definition, integrates at the same time two elements. The first element is the presence of modules, which can be inferred by the topological parameter average coefficient clustering, the second element is the presence of modules, suggested by the fact that in Figure 4.9 $C(k) \propto 1/k$, meaning that, given a node, other nodes connected to it will be likely to form a clique. In the context of the metabolic networks, we suppose that each module corresponds to each of the different pathways characterizing the metabolism of a cell.

The second result is the presence of a scale-free topology, inferred by the topological parameter node degree, in which the most part of the metabolites takes part in few reactions, establishing therefore a low number of interactions, while a small number of metabolites (hubs), are characterized by a high number of connections as it is possible to notice in Figures 4.10, 4.11, 4.12, 4.13. From a biological point of view, the presence of hubs is of particular relevance since these species could correspond to some possible target for the development of anti-cancer therapies. In this work, the detected hubs correspond to the cofactors, which are the most involved species in the reactions, because of their implication in the enzymatic mechanisms. However, in none of the studied cases, we found strictly specific hubs for the cancer cells.

A third feature of these metabolic networks is the ultra small-world property that can be inferred by the parameter path length and allows to give indications on the speed of the system response with respect to an external perturbation. In Figures 4.14, 4.15, 4.17, 4.16, are illustrated the shortest path lengths for the four analyzed GW networks.

[3]Disclaimer: analyses in Subsection 4.4.1 have been performed mainly M. Di Filippo (SYSBIO - Centre of Systems Biology, Milan - Italy) under the supervision of the author.

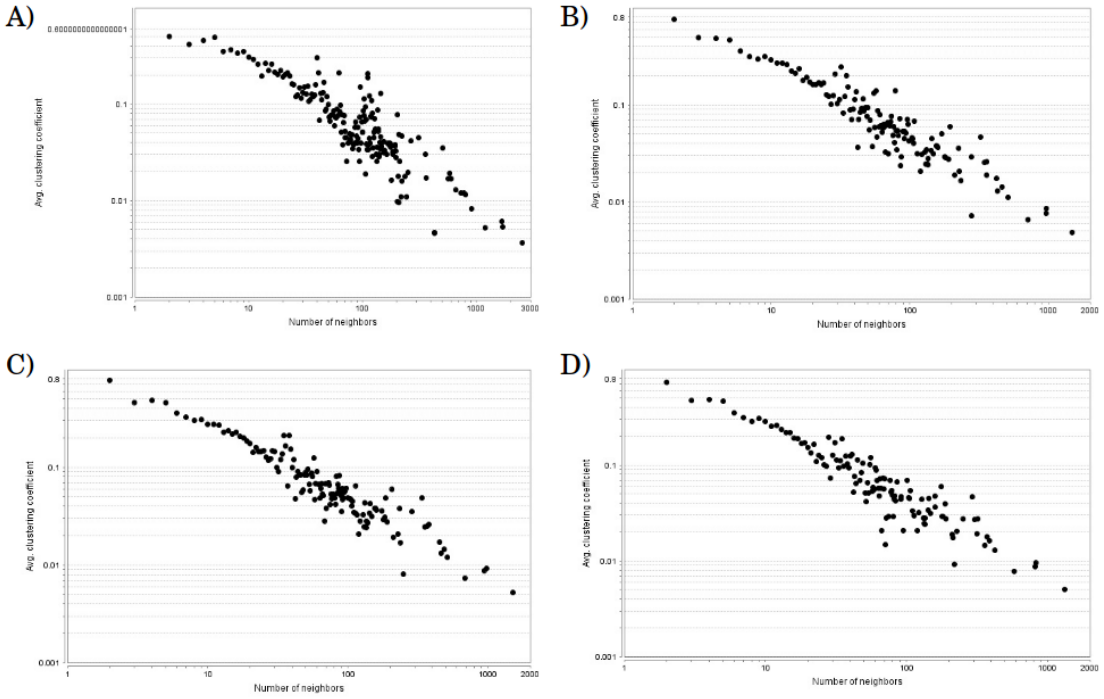


FIGURE 4.9: Clustering coefficient plot for HMA GW networks in logarithmic scale: on the x-coordinate is the number of nodes establishing a link with a given node, while on the y-coordinateis represented the average clustering coefficient. A) HMR network, B) Breast network, C) Liver network, D) lung network.

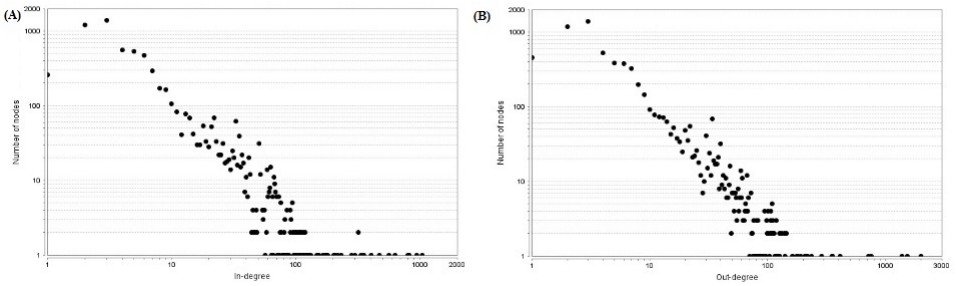


FIGURE 4.10: In and out degree distributions for the GW HMR model.

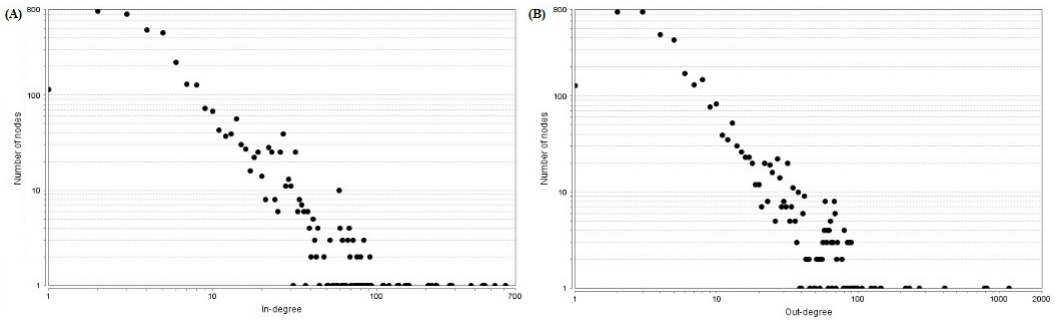


FIGURE 4.11: In and out degree distributions for the GW breast cancer cell model.

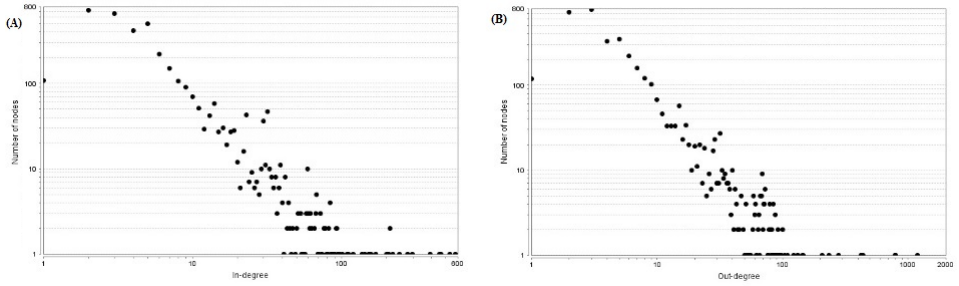


FIGURE 4.12: In and out degree distributions for the GW liver cancer cell model.

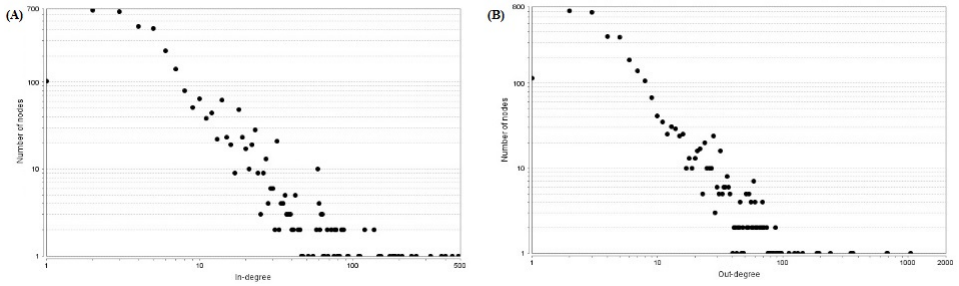


FIGURE 4.13: In and out degree distributions for the GW lung cancer cell model.

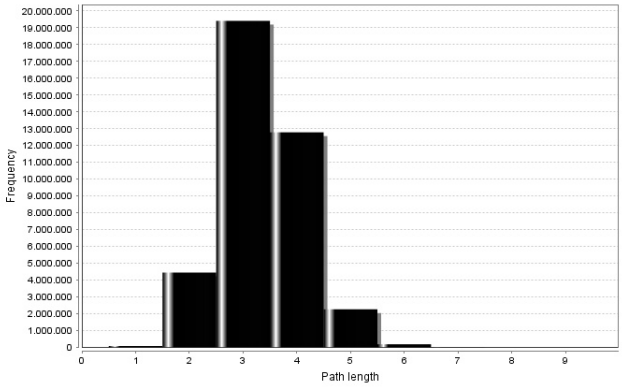


FIGURE 4.14: Histogram representing the shortest path length distribution for the GW HMR network. On the x-coordinate is reported the path length, while on the y-coordinate the frequency.

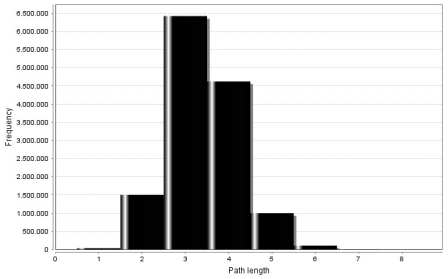


FIGURE 4.15: Histogram representing the shortest path length distribution for the GW breast cancer cell network. On the x-coordinate is reported the path length, while on the y-coordinate the frequency.

In particular, the ultra small-world property implies that, inside a network, each couple of elements are linked by a low number of connections. This ensures a fast transmission of the informations in the network and, at biological level, it means that the local perturbations on the metabolites concentrations could reach the entire network very quickly.

Lastly, the analyzed networks showed a disassortative nature. This property, which can be inferred by the topological parameter neighborhood connectivity, implies that in a metabolic network the most part of the interactions are established between hub and metabolites having few edges. The disassortativity also implicates that two hubs avoid to connect to each other, probably because the removal of one of them from the network has, as a consequence, a considerable negative effect on the entire network structure. Therefore, if two hubs with a common interaction were removed, the obtained effect would be much more catastrophic in terms of connectivity among all the elements of the network.

Overall, this analysis, performed on both the normal and tumor genome-scale and core networks, has produced the same results (as an example node degree distribution for CM model of HMR is provided in figure 4.18), confirming that the "model reduction" performed in Section 3.4.2 is able to lower model complexity (see definition in Chapter 1), while maintaining the same topological properties.

However, the fact that comparable results have been obtained in all the evaluated models suggest that the interaction-based approach has not predictive ability, in this context, because it does not allow to highlight the redistribution of metabolic fluxes at the basis of cellular transformation. For this reason, constraint-based approaches, or mechanism-based approaches (when applicable) are the best choice to investigate structures and paths behind the onset of cancer.

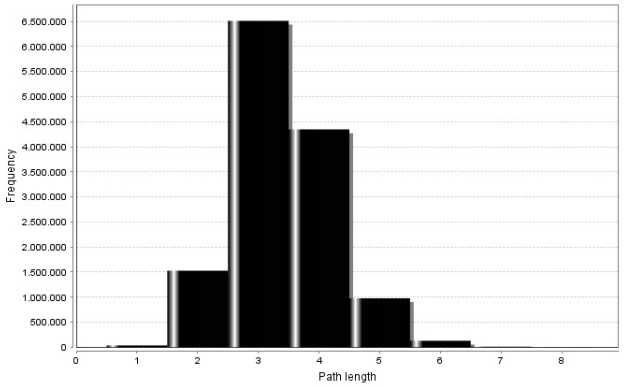


FIGURE 4.16: Histogram representing the shortest path length distribution for the GW liver cancer cell network. On the x-coordinate is reported the path length, while on the y-coordinate the frequency.

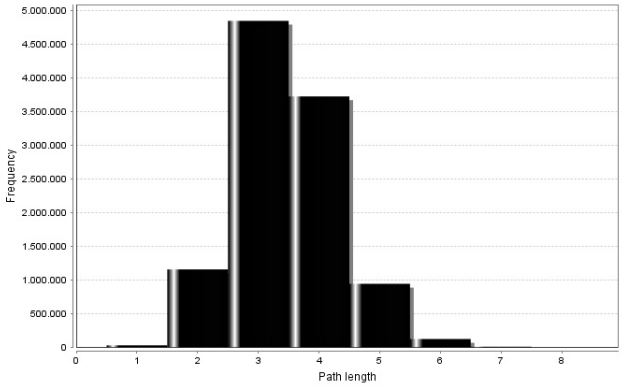


FIGURE 4.17: Histogram representing the shortest path length distribution for the GW lung cancer cell network. On the x-coordinate is reported the path length, while on the y-coordinate the frequency.

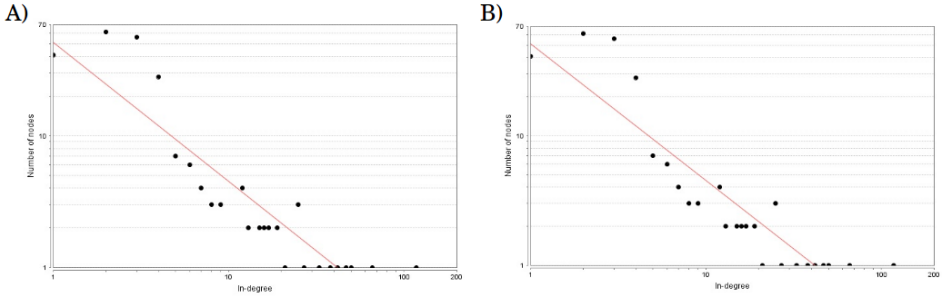


FIGURE 4.18: In and out degree distributions for the GW HMR model.

4.4.2 Node-flux correlation

Network visualizations realized in Subsection 4.3.2.1 suggested the possibility of a correlation between the degree of a node (metabolite) and the value of the sum of reaction fluxes pointing at the node. To investigate this hypothesis I initially identified the value of fluxes for the yeast CM described in Chapter 3, by averaging the flux values obtained in the Crabtree-negative ensemble; then I calculated the Pearson correlation coefficient ϱ between the two variables "node degree" and "cumulative flux": the computed value has been $\varrho = 0.98$, indicating a strong correlation (Figure 4.19), (here significant correlation is for $\varrho > 0.3$).

To verify that this remarkable value is not due to the fact that a "metabolic hub" (e.g. ATP) is correlated to high flux value due to the high number of pointing reactions singularly carrying a low flux, I calculated the Pearson correlation coefficient also in the case when the sum of fluxes pointing at a node is normalized dividing it by the node degree. As expected, in this case the correlation coefficient lowered to a "moderated" value $\varrho = 0.62$ (Figure 4.19).

Moreover I performed the same analysis evaluating the HMR CM described in Section 3.4.1, a model having approximately 5 times more reactions and metabolites with respect to the yeast CM. In this last case the correlation coefficient has further lowered (moderate, $\varrho = 0.47$, Figure 4.21) probably due to the fact that some particular structures called "bottlenecks" are present in the network (i.e. nodes establishing few connections but connecting hubs, and hence carrying a high flux value).

Attempts to investigate relations between flux distributions and topology has been proposed in literature (e.g. in [201–203]). However, to the best of my knowledge, the findings emerged from analyses performed in this section have not been confirmed by other literature studies. For this reason investigations on the correlation between node degree and flux value need to be extended to GW models in order to establish their correctness and relevance.

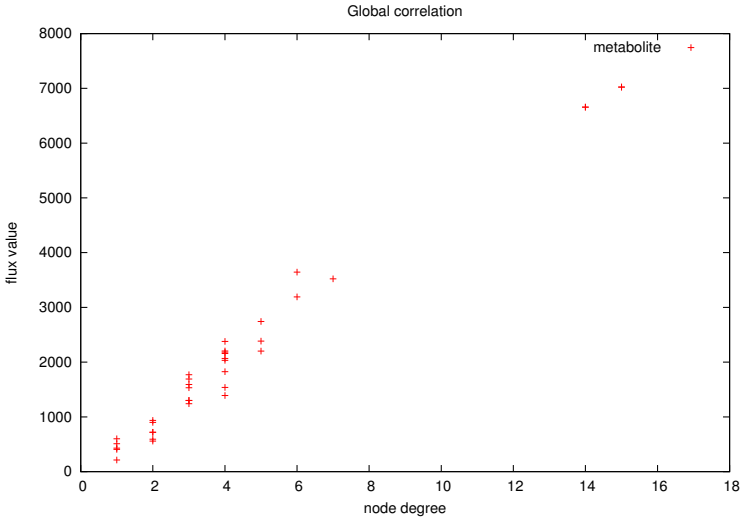


FIGURE 4.19: Scatter plot illustrating the correlation between node degree and flux value in the yeast CM.

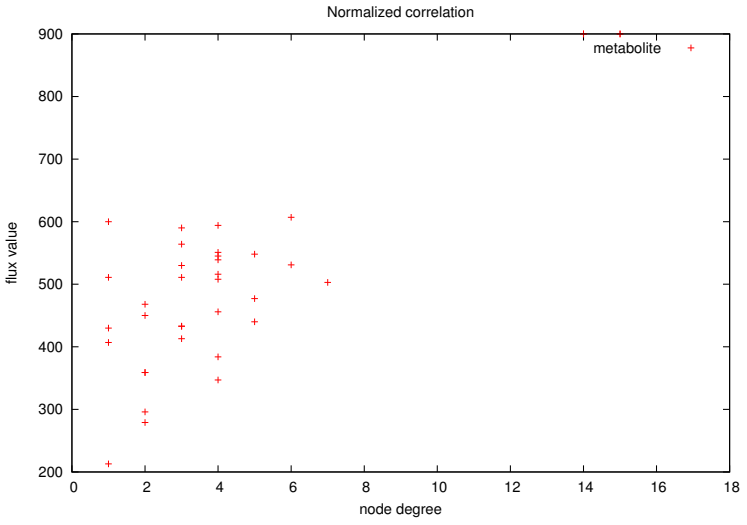


FIGURE 4.20: Scatter plot illustrating the correlation between node degree and flux value in the yeast CM, normalizing the flux value on the degree.

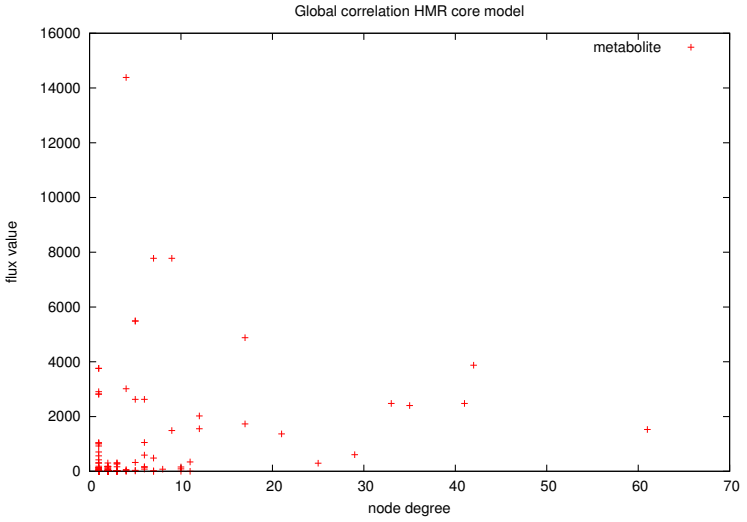


FIGURE 4.21: Scatter plot illustrating the correlation between node degree and flux value in the HMR CM.

Chapter 5

Mechanism-based analysis

As stated in Chapter 2, mechanism-based approaches in Systems Biology are the most informative methods since they allow to fully reconstruct the spatio-temporal behavior of biological complex systems. In this chapter I will illustrate some of the main mechanistic approaches exploited in Systems Biology lingering, in particular, on those used in this work of thesis (i.e. stochastic approaches for the Post-Replication Repair model in Subsection 2.2.1.1, and differential equations to estimate parameters from the yeast CM, see Section 3.5). Moreover I will discuss parameter estimation methods and a devised strategy to identify parameters for mechanistic modeling starting from eeFBA outputs.

5.1 Mechanistic approaches in Systems Biology

Thanks to mechanistic approaches it is possible to simulate the temporal evolution of the different molecular species of a biological system. Strikingly, with these models it is also possible to verify effects of system's perturbations when initial conditions or kinetic parameters are modified (see for example the PSA performed in Chapter 2 applied to the PRR pathway). However, in order to perform these simulations it is necessary to establish the correct interactions among elements and to define the proper value for model parameters (i.e. kinetic constants and and initial quantities for species). The retrieval of these parameters is, in Systems Biology, a challenging task that often limits the applicability of mechanism-based methods only to "small" systems (see Figure 1.3).

To determine which mechanistic method is the most suitable for a given biological system, some aspects of the model and of the system itself must be taken into account; among these: size (in terms of number of reactions and species), abundance of every

molecular species, relevance of biological noise due to stochastic fluctuations of species [204] influencing the dynamics of the system (e.g. bistabilities).
In the case of a model/system of remarkable dimensions, molecular quantities on average $\gtrsim 10^3$ for the different species, and a scarce relevance for the biological noise; the system can be described by means of deterministic methods (Section 5.1.1). If otherwise the model/system is rather "small", quantities are on average $\lesssim 10^3$ for the different species, and biological noise has relevant effects; it is preferable to investigate the system through stochastic methods (Section 5.1.2). Lastly, if aspects are halfway between the two previous cases, it should be evaluated if the system is analyzable through hybrid approaches (Section 5.1.3) combining features both from deterministic and stochastic methods.

5.1.1 Deterministic approaches

The use of a set of ordinary differential equations (ODEs) is the most common deterministic approach for the mechanistic modeling of biochemical systems. Differential equations (one for each of the N chemical species involved in the system) are defined to evaluate the changes in times of species concentration: the dynamic of the system is simulated solving the set of N ODEs, given an initial state $\mathbf{X}_0$ and a set of concentration values.

When dealing with simple ODEs it is possible to obtain an exact solution, instead complex ODEs need to be solved numerically by using linear approximations of smooth curves over infinitesimal time intervals in order to calculate, for each time step, the value of chemical species concentrations.

Dating back to XVII and XVIII centuries, many different and increasingly accurate methods have been developed for the approximation of ODEs. Nowadays are publicly available software libraries to solve them efficiently, among these an effective library is LSODA (Livermore solver for ODEs with automatic method) [38].

ODEs have been widely used for modeling biochemical networks (see Table 2.2 for applications to metabolism); another relevant method for mechanistic modeling of biochemical systems are the so called S-systems. This approach is based on the fact that equations describing the dynamics of biochemical systems can be written as sum of of logarithmic or power law functions terms. Starting from this, efficient methods to calculate power law approximations of ODEs systems have been developed.

Taylor series approximations of ODEs [205] have been used to estimate the reaction rate at steady state. In this way it is possible to obtain a system of non-linear equations that

can be transformed into linear equations when using a logarithmic coordinate system in which the slope of the line is the kinetic order of the reaction. These linear equations can be numerically solved in a computationally efficient way. Globally, S-systems allow a great simplification of the system under evaluation exploiting a fast method that enables a rapid evaluation of unknown parameter values.

A last method in the context of differential equations is represented by partial differential equations (PDEs) [206] that, using partial derivatives, defines spatial and temporal dependencies. This is a well understood and solid mathematical formalism that exploits numerical methods to solve the PDEs in efficient way. Another strong point of PDEs is the determination of both the temporal and spatial dynamic of the system. However this method has some drawbacks due to the fact that it is somehow complicated to implement/generalize for a standard biochemical system and it is not able to model state of discontinuous transitions.

5.1.2 Stochastic approaches

In the case of biochemical systems exhibiting a low number of molecular quantities (i.e. few units of each species) it is preferable to adopt a discrete and stochastic description of the system. Here the term "discrete" refers to the domain of the chemical species description where each species varies by an integer number of molecules. Instead, the term "stochastic" underlines the probabilistic behavior of molecular dynamics; an aspect that is not evaluated in the context of a deterministic treatise.

In particular, in the context of the stochastic approaches, it is possible to define the Chemical Master Equation (CME) [207, 208], that is the equation describing the temporal evolution associated to the state of the system. For complex systems, the CME is analytically unsolvable and numerically intractable. It is however possible to identify trajectories for the dynamic of the system trough Monte Carlo [209] simulation techniques such as the *stochastic simulation algorithm* (SSA) [32, 210] that is able to provide the dynamics of the system by applying a single reaction for each simulation step.

SSA is able to reproduce exact realization of the CME, however the computational cost can be particularly high because it simulates one reaction after the other (intended as collision of molecules), and even simple biochemical system involve a great number of molecules and the execution of many reactions.

This problem lead to the development of approximate algorithms having an outstanding advantage in terms of reduced simulation time. The algorithm used in the case of the PRR pathway described in Section 2.2.1.1 is the so called tau-leaping [31, 211], that

instead of keeping track of every single reaction, exploits a time interval τ in which a certain number of reactions is performed simultaneously.

5.1.3 Hybrid approaches

As stated in the previous section, due to high computational costs, stochastic methods such as tau-leaping and SSA are not adequate to simulate biochemical systems involving a wide number of species or reactions, and in this case the deterministic approach can provide a more efficient description of the system.

However in some cases there is the possibility that a system encompass some species having a large amount of molecules and others having only few molecules. In these cases both the deterministic and the stochastic approaches are not efficient in depicting the system dynamics, either being too slow or too inaccurate. To solve this issue, and to perform more efficient simulations, a class of hybrid modeling techniques [212] has been developed.

These techniques partition (*a priori* or dynamically) the set of model reactions into two groups: the first group is modeled using a stochastic approach while the second one is modeled deterministically [28]. The *a priori* determination of the two sets is done exploiting some biological knowledge about the system. The dynamic determination is performed evaluating the amounts of the chemical species involved in the reactions and their propensities to be executed (see [210]): if the molecules of the species are less than a given threshold, all the reactions in which they are involved as reactants are stochastically modeled; otherwise, they are treated as deterministic.

As a rule of thumb, in a comprehensive model (i.e. a model including both a metabolic and a gene regulatory components), Alfonsi *et al.* [213] suggested to use a stochastic approach for the modeling of the gene regulatory component, and a deterministic modeling for the metabolic part. This is due to the fact that metabolic pathways involves high-numbered species that slow down the simulation algorithms like SSA and tau-leaping. In other works (e.g. in [212]) it was suggested that metabolism can be simulated efficiently by means of hybrid algorithms [28].

5.2 Bridging the gap from constraint-based to mechanism-based models

In Chapter 3 it has been discussed how constraint-based methods, and in particular FBA, have proven to be useful and accurate to calculate the flux of metabolites through

reactions of a metabolic network. Despite of this, constraint-based methods alone have not been able to explain mechanics of events and their temporal evolution, suggesting that other *in silico* methods should be applied. Due to the complex nature of biologic processes, *in silico* methods should consider multiple approaches to investigate systems. Multi-level analysis is today a hot research topic in different areas, that still require an appropriate theoretical formalization of the method and the development of computational tools for the integration of the different modeling perspectives.

As an example, in [214] the author presents a statistical approach that integrate high throughput data and analyze dynamical mechanisms of metabolic networks under mild perturbations. In particular in [214] it has been used a statistic framework able to determine how fast metabolites can reach the steady state. In this context a "feasible kinetic library" has been defined starting from high throughput metabolome technology and it has been preferred to the determination of accurate kinetic information difficult to retrieve. The devised approach has been tested on a core metabolic model and emerging results will be useful to explore the relationship between dynamic and physiology in metabolic reconstruction (possibly GW) and to overcome the chronic lack of kinetic information.

A further example can be found in [100, 215]. Here authors delineated a method to define kinetic models for metabolic networks exploiting only the information deriving from reaction stoichiometries. As illustrated in Chapter 3, FBA is able to determine the value of fluxes through every reaction in the model (i.e. the flux distribution). In the devised method these fluxes are allowed to vary dynamically according to linlog kinetics, where the linlog approximation [216] is used to simplify reaction rate laws in metabolic networks. This method stems from metabolic control analysis describing the effect of metabolite levels on flux as a linear sum of logarithmic terms and providing a good approximation near a chosen reference state.

Another method in bridging constraint-based and mechanism-based approaches has been developed in [91] where authors exploit an ensemble modeling (EM) to deal with large-scale kinetic modeling through the reduction of the size of the parameter space by means of experimental data (flux values, intracellular metabolite concentrations, thermodynamic constraints for the directionality of the reactions). The first step in the EM procedure is the definition of a kinetic model predicting the experimentally observed phenotypic characteristics. At this stage, the additional biological data are used to screen the models until a minimal set of kinetic models are obtained. The EM has been successfully used in modeling many different metabolic network; in particular an outstanding work is presented in [74] where authors modeled a metabolic network for cancer cells.

Although these methods try to establish a bridge between constraint-based and mechanism-based methods, the most accurate way to connect the two modeling approaches is the estimation of the kinetic parameters (and, if necessary, the molecular concentrations of metabolites) starting from the metabolic flux distributions. In Section 5.3 I will review the most commonly used methods for parameter estimation, in particular in Section 5.3.1 I will describe in detail the Particle Swarm Optimization (PSO) technique and its applications to the determination of kinetic parameters starting from metabolic fluxes, and in Section 5.6 I will introduce a novel and efficient version of the PSO algorithm that has been called "Proactive Particles in Swarm Optimization".

5.3 Parameter estimation

In the context of Systems Biology, parameter estimation can be defined as "the ability to calibrate a model in order to reproduce, through simulation, experimental results in the most accurate way" [217]. In biochemical systems, the parameter estimation problem can be defined in terms of a non linear programming task subject to constraints.
Moreover, due to the fact that optimization problems applied to biochemical systems are usually multimodal, and in order to avoid local solutions, it is worth to underline that exploited methods are here usually global optimizations. Indeed, local optima could lead to misinterpretations in the calibration of models resulting in a bad fit between simulations and experimental data [218].

Methods developed for global optimization can be categorized in deterministic and stochastic. Whereas stochastic methods do not have a convergence theorem assuring the retrieval of the best global solution, the deterministic methods are able to guarantee the identification of the global optimum in a finite time. However, the computational requirements of these latter strategies are extremely high and increase drastically with the dimensionality of the problem.

For this reason, even at the cost of not having the best solution guaranteed, stochastic methods are widely used to efficiently identify candidate best solutions (and in many cases the retrieved optimal solution is close enough to the real one).
Moreover, these methods can be used as a "black box" because the original problem does not need to be transformed. A fact that helps to integrate the optimizer with an external software performing a further task.

Hereafter I will briefly illustrate some the main groups of stochastic methods for global optimization in the context of Systems Biology:

- *Simulated annealing* is a physics inspired method that mimic the cooling process of metals in the fact that atoms assume the most stable configuration during a slow decrease of metal temperature [219].

- *Evolutionary computation* methods are biology inspired strategies relying on mechanism governing the evolution of biological entities [220] (i.e. reproduction, mutation, and individual fitness). Evolutionary computation methods mimic natural evolutionary process through the generation of individuals that are more and more "fit", i.e. solutions that are more and more close to the global best of the optimization problem. The most used methods in this context are: Genetic Algorithms [221, 222] (used in Section 3.5.1 to populate the ensembles of solutions in the eeFBA algorithm), Evolutionary Programming [220], and Evolution Strategies [223].

- *Swarm intelligence* methods are algorithms based on the concept of collective behavior of agents exploring a solution space. Over years many bio-inspired methods have been developed; among these in ant colony optimization [224] simulated "ants" keep track of position and quality of solution to attract more ants in following iterations (a mechanism mimicking pheromones), while Particle Swarm Optimization (PSO) [225] mimic the movement of a bird flock attracted by food or repelled by a predator.

In the next section (5.3.1) I will describe in detail the PSO strategy used to estimate parameters for the kinetic modeling in this work of thesis.

5.3.1 Particle Swarm Optimization

In the context of swarm intelligence methods for global optimization, Particle Swarm Optimization (PSO) is a population-based meta-heuristics where N particles (i.e. the solutions) belong to a swarm "flying" in a M-dimensional search space to identify the optimal solution in a cooperative manner. A particle i is described by two vectors: $\mathbf{x}_i \in \mathbb{R}^M$ that is the position in the search space, and $\mathbf{v}_i \in \mathbb{R}^M$ that represent the velocity. Each particle is initially positioned accordingly to a uniform random distribution over the search space.

During the optimization process, particle velocity is influenced by the best position individuated by the particle itself ($\mathbf{b}_i \in \mathbb{R}^M$), and the best position identified collectively by the swarm ($\mathbf{g} \in \mathbb{R}^M$) [110].

The PSO evaluates two distinct components for the update of the velocity:

- the social factor $c_{soc} \in \mathbb{R}^+$ determines the attraction of the particle towards the best position identified by the swarm $\mathbf{g}$ and quantifies the extension of the velocity update due to the best position;
- the cognitive factor $c_{cog} \in \mathbb{R}^+$ determines the tendency of the particle to remain in the vicinity of its best position $\mathbf{b}_i$.

Both the social and cognitive factor are multiplied by two vectors $\mathbf{R}_1$ and $\mathbf{R}_2$ of random numbers sampled from the uniform distribution in the unit interval $(0, 1)$, due to the fact that a deterministic movement of particles could entrap particles in a local optima.

Instead, velocity is calibrated by means of an inertia weight $w \in \mathbb{R}^+$ to counteract the chaotic movement of particles. The velocity update for the i-th particle during the t iteration can be defined as:

$$\mathbf{v}_i(t) = w \cdot \mathbf{v}_i(t-1) + c_{soc} \cdot \mathbf{R}_1 \left(\mathbf{x}_i(t-1) - \mathbf{g}(t-1)\right) + c_{cog} \cdot \mathbf{R}_2 \left(\mathbf{x}_i(t-1) - \mathbf{b}_i(t-1)\right). \quad (5.1)$$

Following the definition of velocity, and exploiting Equation 5.1 it is possible to calculate the position of particles as:

$$\mathbf{x}_i(t) = \mathbf{x}_i(t-1) + \mathbf{v}_i(t), \text{ for all } i = 1, \ldots, N. \quad (5.2)$$

In PSO a fitness function f is exploited to evaluate the proximity of every particle to the global best. Fitness values define a hyper-surface defined "fitness landscape".

The fitness function drives the evolution of the whole swarm since it is used, iteration by iteration, to evaluate the fitness of each particle and then to update the values of $\mathbf{b}_i$ and $\mathbf{g}$. To prevent that particles could exit the feasible solution space, boundaries are defined for example by using a random bounce when the particle reaches the limit of the search space [226]. Also the velocity of particles is regulated exploiting a maximum value $v_{max_m} \in \mathbb{R}^+$ along each m-th dimension of the search space, with $m = 1, \ldots, M$ [227].

The main drawback of this method is linked to the dependence of performances by the adequate selection of the parameters of the algorithm such as: N, c_{soc}, c_{cog}, w and the vector of maximum velocity values $\mathbf{v}_{max}$ [227]. In many cases a wrong parametrization leads to poor performances in terms of quality of the solution and convergence speed.

5.4 MetaFluxAnalysis

As a first step towards the estimation of parameters for mechanistic analysis, I performed a "feasibility study" implementing "MetaFluxAnalysis", a LabVIEW tool to determine metabolic fluxes starting from mechanistic simulations.

LabVIEW is a development environment for a visual programming language defined by National Instruments [108]. A LabVIEW tool is composed mainly of two virtual instruments (VI). The first is a block scheme (Figure 5.1) where the different blocks (control flow, constants and functions) are graphically represented and connected to represent the source code. The second VI is a front panel (Figure 5.2) used to manage inputs (called Controls) and outputs (Indicators).

The block scheme of MetaFluxAnalysis defines the instructions to calculate the following outputs:

- the flux intensity graph, a heatmap illustrating the value of all the fluxes in the metabolic network at every simulated time point;
- the dynamics of a single flux, i.e. the "amplitude" of the flux over time in terms of units of molecules produced by the reaction in each time step;
- the minimum and maximum value of calculated fluxes;
- the histogram distribution, a diagram illustrating the count of "amplitudes" split in a given number of bins representing the range of values calculated at the previous step.

These outputs are calculated taking as input a file that contains the molecular quantities of every metabolite at each simulated time step and some additional values such as the indexes identifying each species and reaction in the stoichiometric matrix and the vectors defining the values of the reaction constants, the molecular quantities of each species at the beginning of the simulation, and, if needed, the species whose quantity is maintained through the simulation.

The core of the procedure to calculate the value of the flux at every simulated time step is based on the equations deriving the flux according to the mass action hypothesis:

$$v_i = k_i \prod_{w=1}^{M} [\chi_w]^{\alpha_{wi}} \tag{5.3}$$

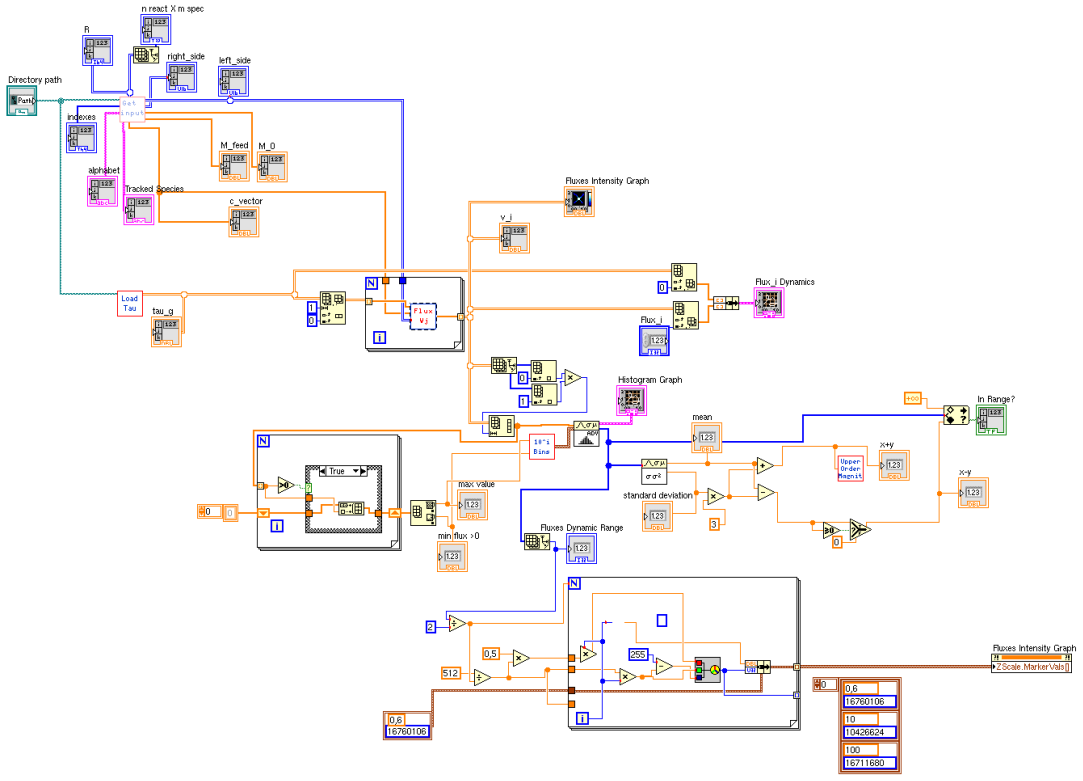


FIGURE 5.1: MetaFluxAnalysis block scheme in LabVIEW.

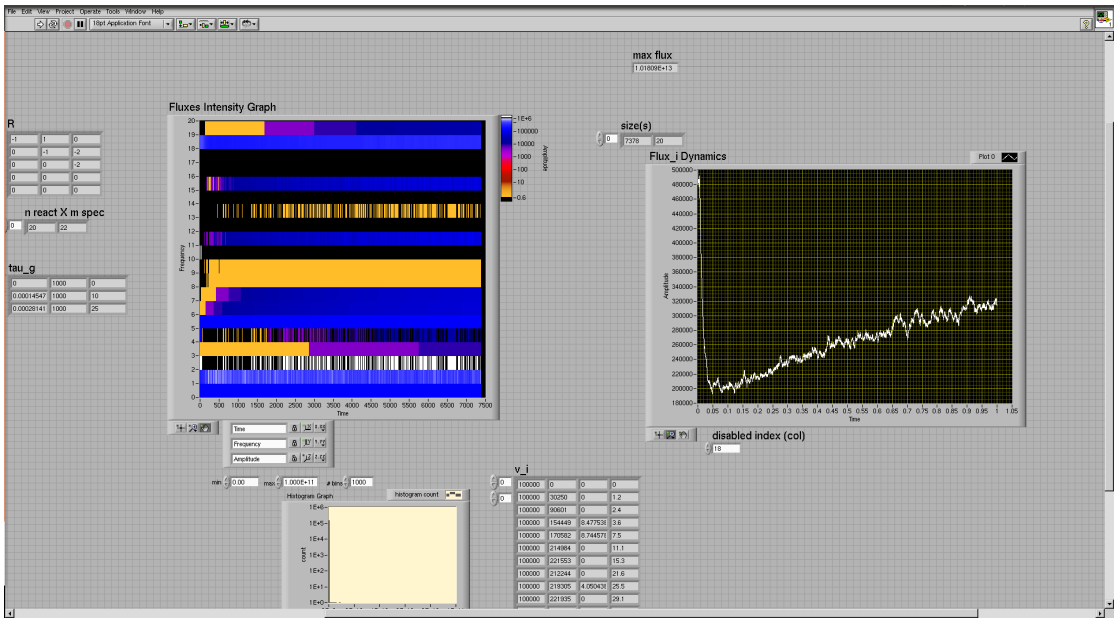


FIGURE 5.2: MetaFluxAnalysis front panel in LabVIEW.

where v_i is the flux through the reaction i, k is the rate constant of reaction i and $\prod_{w=1}^{M}[\chi_w]^{\alpha_{wi}}$, is the product of concentrations of species i raised to the stoichiometric coefficient α_{wi}.

The software tool implemented in this section has been tested on a metabolic toy model (with random parametrization for molecular quantities and reaction constants), mechanistically simulated exploiting a mass action approach and the SSA algorithm introduced in Section 5.1.2. A snapshot of the outputs of MetaFluxAnalysis is represented in Figure 5.2. Overall this approach has been useful to investigate the relation between fluxes and kinetic constants, moreover it could be exploited to determine the value of metabolic fluxes in a model having a mechanistic description, and to provide a an instantaneous graphical overview of the fluxes dynamics.

5.5 Estimating kinetic constants for the eeFBA model of yeast

In order to estimate kinetic constants for the eeFBA model of yeast is necessary to enrich the model adding all the information on metabolic concentrations and kinetic constants retrievable in literature. For this reason I exploited two tools: the Yeast Metabolome Database and the KiPar information retrieval software (Section 5.5.1).

5.5.1 Integrating data of metabolic concentrations

The Yeast Metabolome Database (YMDB) [228] contains a wide range of information on the metabolome of *S. cerevisiae* such as compound description, names and synonyms, structural information, physical-chemical data, reference Nuclear Magnetic Resonance (NMR) and Mass Spectrometry (MS) spectra, growth conditions and substrates, pathway information, enzyme data, gene/protein sequence data, as well as numerous hyperlinks. Globally, from the database it is possible to retrieve information for more than 2000 metabolites and 1000 proteins connected with metabolism.

In the context of the present work, the most remarkable feature of YMDB is the availability of an extended documentation of experimental intracellular and extracellular metabolite concentration data obtained by means of detailed MS and NMR metabolomic analyses deriving both from literature and *ad hoc* performed experiments. In Appendix B it is provided the list of identified concentrations for metabolites of the yeast CM.

In view of an extension of the work to human/mammalian models, it is worth to underline that the YMDB has been developed on the same line of the Human Metabolome Database [162].

To retrieve the widest number of kinetic constants from literature, I exploited "KiPar" [229], an information retrieval tool. This platform has been devised to retrieve textual

documents containing information (such as kinetic parameters) for the kinetic modeling of metabolic pathways. A further step performed by KiPar is the compilation of a list of annotations regarding kinetic parameters and reactions/pathways where they are involved.
The input of KiPar is defined through the use of biological ontologies and databases (KEGG and Gene Ontology) related to pathways/reactions of interest. This strategy has been chosen to overtake the terminological variability of biological sublanguages.
These ontologies or concepts are used to query the literature in order to search the kinetic information for the model parametrization. To perform an efficient query, KiPar makes use of the Entrez search and retrieval system implemented for the NCBI databases PubMed and PubMed Central [230].
At this stage, a local database is used to store the collected information and it is queried to retrieve information from the selected documents. A scoring system is eventually used to weight matching concepts of each type considered and then the list of documents with relative scores is returned to the user.

5.5.2 Estimation of kinetic constants with a particle swarm optimizer (PSO)

From each cluster identified in Section 3.5.3 through a constraint-based technique, it is possible to derive an average flux distribution at steady state that is used as the target for the estimation of kinetic constants in the mechanism-based mass action CM of yeast metabolism. This is performed exploiting the relation between fluxes and constants expressed by the Equation 5.3.

The mass action kinetic constant is estimated by means of a particle PSO coupled with deterministic simulations based on ODEs. As first approach, the goal is to identify a set of plausible constants, keeping into account that their value could be far from the experimentally measured one.
Equation 5.3 has infinite $(k_i,[\chi_w])$ pairs of solutions for a single v_i, implying that as the result of the method will be parametric, a subsequent phase to screen the ensemble of obtained solutions will be necessary.

In order to estimate the rate constants with the PSO procedure, the main hypothesis is here that the dynamics of the metabolic system will reach, when simulated, a steady state at which the concentration of all the involved metabolites will be stable over time. Under this hypothesis it possible to define a fitness function based on the output of a mechanism-based simulation of the metabolic model where the time course for a generic metabolite χ_w will be similar to the one represented in Figure 5.3 where, after a transient

phase, the concentration of χ_w will reach a steady state.
For the definition of the fitness function it is sufficient to evaluate the concentration $[\chi_w(t)]$ at two different time points: t_1, t_2 chosen after that when the steady state is reached, such that $t_2 > t_1$. It is also possible to define $D_w = |[\chi_w(t_1)] - [\chi_w(t_2)]|$, the distance between the hypothetic steady state reached at t_1 and the actual concentration reached at t_2; if $D = 0$ it means that the steady state has actually been reached, otherwise if $D_w > 0$, $[\chi_w]$ is still far from the steady state. Given these premises the fitness function is defined as:

$$f = min \sum_{w=1}^{\varrho} D_w \tag{5.4}$$

that is the minimization of the distance D_w for each one of the ϱ reactions in the metabolic model.

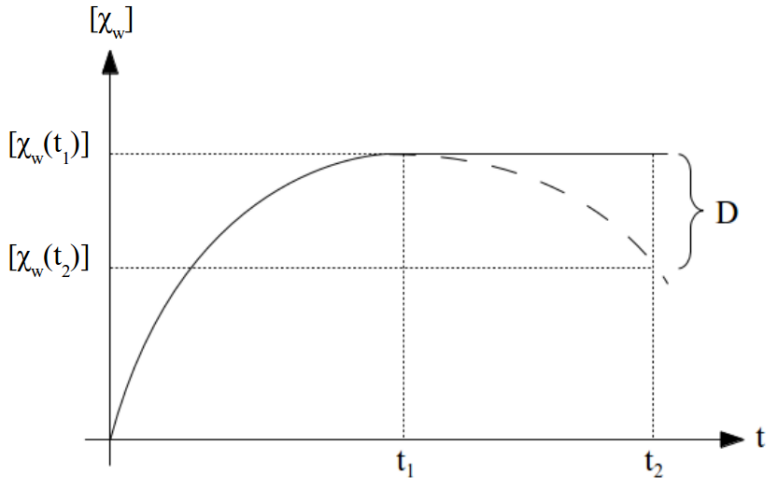


FIGURE 5.3: Temporal evolution for the concentration of a metabolite χ_w during a mechanism-based simulation.

To evaluate the fitness function it is needed to initialize the unknown k_w and $[\chi_w]$ assigning $[\chi_w]$ them a random value for each particle/reaction and then performing a mechanism-based simulation (for the ODE system derived for the CM of yeast) till t_2. After this, the fitness function is evaluated applying Equation 5.4 and updating values of $\mathbf{b}_w$, $\mathbf{g}$ and $\mathbf{v}_w$ accordingly to the PSO algorithm that will terminate when the value of f will not change for a given number of iterations.

To perform the parameter estimation, I decided to exploit the PSO implementations and toolboxes already available under the MATLAB environment. This choice has been done to maintain uniformity of development in the pipeline (eeFBA has been developed in the same environment), moreover MATLAB provides an accessible, stable and reliable environment to develop/modify the code and it supports by many libraries. However, it is a well known fact that MATLAB implementations have poor performances when

compared with other programming languages (see Figure 1 and Table 1 in [231] for a performance comparison between MATLAB and other programming languages on benchmarks).

It has been possible to identify at least four different implementations for the PSO algorithm in MATLAB (http://www.mathworks.com/matlabcentral/fileexchange/) [1]. Each of these has been tested on selected benchmark functions among the canonical ones (Ackeley, Alpine and Rastrigin) where the evaluation of performances is in terms of iterations to confluence. From the results of the comparison, the "Particle Swarm Optimization (PSO) by Pramit Biswas" emrged as the best choice.

I tested the parameter estimation procedure using the selected implementation of the PSO in MATLAB and many different values for parameters both of the PSO algorithm $N, c_{soc}, c_{cog}, w, \mathbf{v}_{max}$ and for t_1 and t_2. Unfortunately, for the yeast CM, after 4000 iterations of the PSO, the convergence has never been reached.

This fact is probably due to some concomitants factors: as already stated, MATLAB, in spite of a high usability, is a computationally inefficient environment, moreover the fitness landscape generated by the problem could be potentially of difficult exploration and the tested parameter settings could be inappropriate for the problem under evaluation.

To overcome this inability to reconstruct parameters, I contributed to define a more efficient version of the PSO developed exploiting Fuzzy Logic for the determination of PSO parameters. In Section 5.6 I will introduce the theoretical definition and the testing of this novel version of the PSO (for greater details, the interested reader should refer to [110]).

5.6 Proactive Particles in Swarm Optimization: a fuzzification of PSO

As illustrated in Section 5.3.1, PSO performance is highly dependent on the proper tuning of its parameters $N, c_{soc}, c_{cog}, w, v_{max}$. The setting of these parameters is dependent on the problem under evaluation and their numerical value could be established only with a precise knowledge of the fitness landscape shape; as a result, the search for a good parameter set is a complex and time consuming procedure. For this reason the definition and implementation of self-tuning and adaptive modifications of PSO is today a hot research topic [232–234].

[1] Particle Swarm Optimization Toolbox by Brian Birge 22 Apr 2005; Particle Swarm Optimization (PSO) by Pramit Biswas 17 Sep 2013 (Updated 08 May 2014); Particle Swarm Optimization (PSO) algorithm by Milan Rapai 24 Nov 2008; Another Particle Swarm Toolbox by Sam 01 Dec 2009 (Updated 01 Apr 2014)

In particular, Fuzzy Logic (FL) [235] has been used to determine the behavior of the swarm and to dynamically select the settings for the PSO (see [236] for an extensive review of methods). In this context a first approach has been introduced in [237] where a Fuzzy Rule-Based System (FRBS) has been used to determine the inertia weight for the whole swarm, based on the performance of the best candidate solution and its inertia in every iteration.
Another fuzzy approach has been introduced in [238] where the Fuzzy Adaptive Turbulence in Particle Swarm Optimization algorithm has been devised to deal with the premature convergence problem: here FL has been used to slow the convergence of the swarm to the global best by means of an adaptive tuning of the minimum velocity of particles.

A Fuzzy Particle Swarm Optimization (FPSO) algorithm has also been described in [239] as a method to tune the inertia weight and a new parameter modulating the velocity of particles and named "learning coefficient".

It is worth noticing that all the works existing in literature exploit FL assigning to each of the PSO parameters (w, c_{soc}, c_{cog}) an identical value for every particle of the swarm. Instead in the present thesis and in [110], exploiting a FRBS it has been possible to determine a specific setting for the parameters of the PSO tuning their value for each particle. In this way individuals of the swarm become proactive optimizing agents and hence the devised algorithm has been named *Proactive Particles in Swarm Optimization* (PPSO). In this novel approach the setting of each particle, at each iteration, has been determined using the FRBS to compute two indexes: the distance from the global best, and a normalized fitness incremental factor.

Fuzzy Logic Fuzzy Logic (FL) was introduced by L. Zadeh in [235] as a "mathematical tool to cope with uncertainty and to provide a conceptual framework for the use of ambiguous linguistic variables dealing with imprecise and approximate reasoning" (see [240]). By means of FL, vague linguistic constructs can be represented with the goal to build and automatic reasoning and inference system.

In this context, fuzzy set theory (FST) [235] is a generalization of the classical set theory (considering only "crisp" values i.e. assuming only the Boolean values true or false).

In classical (crisp) set theory, an element can be part of a certain set only with Boolean values meaning that the element either do or do not belong to a certain set. Following this reasoning it is possible to define a membership function for the x element of the crisp set C. However, in the "real world" it is not likely to assign an element to a set with value TRUE or FALSE (e.g. a priori it is not possible to determine if the temperature

35°C is hot with value TRUE). To deal with this kind of sets FL uses a mathematical construct called FST by means of which it is possible to define a membership function for the fuzzy set A as in crisp set theory:

Definiton 1. Let X be a set. A fuzzy subset $A \subseteq X$ is defined by a membership function $f(a) : X \to [0, 1]$.

Meaning that an element a of X belongs to the fuzzy subset A proportionally to a number in the interval [0, 1] representing its "degree of membership". A pivotal point of the FST is the definition of the shape of the membership function, using for this task the advice of an expert.

In order to apply FL to a given system, the system has to be defined in natural language coupling the membership function with linguistic variables defined as follows:

Definiton 2. Let V be a variable, X the range of values of the variable and T_V a finite set of fuzzy (sub)sets. A linguistic variable can be defined as a triplet (V, X, T_V).

In fuzzy approaches the central part of the process is the fuzzy inference system (FIS) that is the procedure to retrieve a crisp "defuzzified" output using as input linguistic variables, fuzzy rules, and fuzzy reasoning.

To calculate the output of a FIS, given the inputs, a standard procedure has been defined as follows:

1. *Determination of a list of fuzzy rules*; a FRBS of linguistic statements to encode the behaviour of the FIS in the classification of a certain feature or to control a certain output. The canonical form of a fuzzy rule is the following:

   if (input *1* is membership function *1*) **and/or** (input *2* is membership function *2*) **then** (output n is output membership function n).

   The ensemble of the rules of a fuzzy system is called the decision matrix.

2. *Fuzzification of the inputs* to link crisp inputs to the corresponding values on membership functions in the range [0, 1].

3. *Definition of an inference engine* that, starting from fuzzified inputs use fuzzy rules to establish the classification or the fuzzy output. The outcome of the application of the fuzzy rules in the decision matrix depends on the type of fuzzy implication chosen (e.g. Mamdani [241] or Sugeno [242]). Moreover all the different rules must be applied to the decision matrix in order to obtain an output distribution. For this reason all the fuzzy output of the FIS must be aggregated using an operator for the union.

4. *A unique crisp value must be returned as output* of the FIS (e.g. to determine how to tune a controller) through the defuzzification process that can be achieved choosing among several different definitions of defuzzification such as the "method of the mean of maxima" (MeOM) and the "center of gravity method" (COG) [243].

Proactive PSO In the Proactive Particles in Swarm Optimization (PPSO)[2], contrary to previous attempts in literature (see Section 5.3.1), parameters (w, c_{soc}, c_{cog}) involving particles of the swarm are individually tuned at each iteration to determine the velocity. To this end, the Equation 5.1 is modified in the following way:

$$\mathbf{v}_i(t) = w_i(t-1) \cdot \mathbf{v}_i(t-1) + c_{soc_i}(t-1) \cdot \mathbf{R}_1 \left(\mathbf{x}_i(t-1) - \mathbf{g}(t-1)\right) + c_{cog_i}(t-1) \cdot \mathbf{R}_2 \left(\mathbf{x}_i(t-1) - \mathbf{b}_i(t-1)\right), \tag{5.5}$$

where $w_i(t), c_{soc_i}(t)$ and $c_{cog_i}(t)$ are the parameters of i-th particle at iteration t.

Moreover in this implementation, two values representing (I) the distance of the particle from the global best $\mathbf{g}$, and (II) a function measuring its fitness improvement with respect to the previous iteration, have been defined to perform a dynamic fuzzy estimation of w, c_{soc} and c_{cog}.

In formal terms, (I) can be written as a function having domain in $\mathbb{R}^m$:

$$\delta(\mathbf{x}_i(t), \mathbf{x}_j(t)) = \sqrt{\sum_{m=1}^{M} (x_{i,m}(t) - x_{j,m}(t))^2}, \tag{5.6}$$

here i and j are the two particles, $x_{i,m}, x_{j,m}$ are the m-th components of the position vectors $\mathbf{x}_i, \mathbf{x}_j$, respectively, for some $i, j = 1, \ldots, N$.

While (II), named "normalized fitness incremental factor" is a function having codomain in $(-1, 1)$ and calculated according to the current and the previous positions of particle i and the corresponding fitness values:

$$\phi(\mathbf{x}_i(t), \mathbf{x}_i(t-1)) = \frac{\min\{f(\mathbf{x}_i(t)), f_\triangle\} - \min\{f(\mathbf{x}_i(t-1), f_\triangle)\}}{|f_\triangle|} \cdot \frac{\delta(\mathbf{x}_i(t), \mathbf{x}_i(t-1))}{\delta_{max}}, \tag{5.7}$$

[2] Disclaimer: the mathematical formalization of the PPSO has been realized in collaboration with D. Besozzi (Department of Informatics, University of Milan), P. Cazzaniga (Department of Human and Social sciences, University of Bergamo) and G. Pasi (Department of Informatics, Systems and Communication, University of Milan – Bicocca), while implementation and testing of PPSO has been realized by M. S. Nobile (Department of Informatics, Systems and Communication, University of Milan – Bicocca) using the Python language. The standard PSO algorithm, exploited to compare performances, has been implemented in plain vanilla Python code. The fuzzy engine used for the implementation is "pyfuzzy" (http://pyfuzzy.sourceforge.net) and NumPy (http://www.numpy.org) (see [244]).

where δ_{max} is the length of the diagonal of the hyperrectangle defined by the search space, and $f_\triangle$ represents the estimated worst fitness value for the optimization problem under investigation, and whose evaluation shortly follows.

The correct estimation of the worst fitness value is comparable to the solution of the optimization problem due to the fact the shape of the fitness landscape for the problem is unknown in most of the cases. For this reason, fitness values for all particles is calculated in the first iteration of PPSO along with their position. From this, $f_\triangle$ is assumed to be the worst value at that iteration. Thereafter, by means of the min functions in Equation 5.7, it is possible to clamp fitness values worse than $f_\triangle$ in the optimization phase. Moreover in the same equation, since a minimization problem is here evaluated, the first factor considers the improvement of the fitness value of the i-th particle, normalizing to $[-1, 1]$ by dividing by $|f_\triangle|$.
In this case, a low value of $\phi(\mathbf{x}_i(t), \mathbf{x}_i(t-1))$ in $[-1, 1]$ corresponds to a lower fitness value of particle i when compared to its value in the previous iteration, indicating a better solution for the optimization problem. In Equation 5.7 the second term has been introduced to weigh ϕ evaluating the distance between the current and the previous position of the particle. This factor can assume values $[0, 1]$ due to the normalization obtained by dividing by δ_{max}.

A FRBS of 9 fuzzy rules (Table 5.1) has been defined to establish the values $w_i(t), c_{soc_i}(t)$ and $c_{cog_i}(t)$, for each particle $i = 1, \ldots, N$ at each iteration t.
Two linguistic variables named "distance from $\mathbf{g}$" (δ_i) and "normalized fitness incremental factor" (ϕ_i) have been used in the antecedent, while the output variables in the consequent of rules correspond to the PSO parameters described in Section 5.3.1 and have been named *Inertia*$_i$, *Social*$_i$ and *Cognitive*$_i$.

Rule n.	*Rule definition*
1	IF (ϕ_i IS Worse OR δ_i IS Medium OR δ_i IS High) THEN Inertia$_i$ IS Low
2	IF (ϕ_i IS Unvaried OR δ_i IS Low) THEN Inertia$_i$ IS Medium
3	IF ϕ_i IS Better THEN Inertia$_i$ IS High
4	IF (ϕ_i IS Better OR δ_i IS Medium) THEN Social$_i$ IS Low
5	IF ϕ_i IS Unvaried THEN Social$_i$ IS Medium
6	IF (ϕ_i IS Worse OR δ_i IS Low OR δ_i IS High) THEN Social$_i$ IS High
7	IF δ_i IS High THEN Cognitive$_i$ IS Low
8	IF (ϕ_i IS Unvaried OR ϕ_i IS Worse OR δ_i IS Low OR δ_i IS Medium) THEN Cognitive$_i$ IS Medium
9	IF ϕ_i IS Better THEN Cognitive$_i$ IS High

TABLE 5.1: Fuzzy rules used by PPSO. Table from [110].

Following Equation 5.6, the numeric values of the distance between $\mathbf{x}_i$ and $\mathbf{g}$ define the universe of discourse of δ_i and its base variable is defined the interval $[0, \delta_{max}]$.

Output variable	*Term*	*Value*
$Inertia_i$	Low	0.3
	Medium	0.5
	High	1.0
$Social_i$	Low	0.1
	Medium	1.5
	High	3.0
$Cognitive_i$	Low	0.1
	Medium	1.5
	High	3.0

TABLE 5.2: Defuzzification of output variables. Table from [110].

Linguistic values *Low*, *Medium* and *High*, form the term set of δ_i allowing in this way to characterize the proximity of particles to the global best. In Figure 5.4 it is shown the membership functions linked to the linguistic values of δ_i. Values defining the shape of the three membership functions ($\delta_1, \delta_2, \delta_3 \in [0, \delta_{max}]$) have been set in accordance to the dimensions of the search space. Lastly, the domain expertise on PSO has been exploited to set the following parameters: $\delta_1 = 0.05 \cdot \delta_{max}$, $\delta_2 = 0.1 \cdot \delta_{max}$, $\delta_3 = 0.15 \cdot \delta_{max}$.

Following Equation 5.7, values of function ϕ of particle i with respect to the previous iteration, determine the universe of discourse of ϕ_i and its base variable is defined in the interval $[-1, 1]$. Linguistic values *Better*, *Unvaried* and *Worse*, form the term set of ϕ_i defining the improvement of a particle with respect to its value in the previous iteration. In Figure 5.5 it is shown the membership functions associated to the linguistic values of ϕ_i. According to the domain expertise, the values $\phi_1, \phi_2 \in [-1, 1]$ are set to $\phi_1 = -0.0025$ and $\phi_2 = 0.0025$ and define the shape of the fuzzy set associated with the *Unvaried* linguistic value.

Based on the Sugeno inference method [242], a FRBS has been defined to set fuzzy rules with fuzzy inputs and crisp outputs. The defuzzification process is synthesized in Table 5.2: here the output variables ($Inertia_i$, $Social_i$, $Cognitive_i$) can assume three different linguistic values: *Low*, *Medium* and *High*, each one modeled as a fuzzy singleton.

Sugeno is used to calculate the final numerical value of an output variable (given a set $\mathcal{R}$ of R rules, all having the same output variable in their consequent) as the weighted average of the output of all rules in $\mathcal{R}$, accordingly to:

$$output = \frac{\sum_{r=1}^{R} \rho_r z_r}{\sum_{r=1}^{R} \rho_r}, \tag{5.8}$$

where ρ_r is the membership degree of the input variable of the r-th rule, and z_r represents the output crisp value for the r-th rule, as reported in Table 5.2.

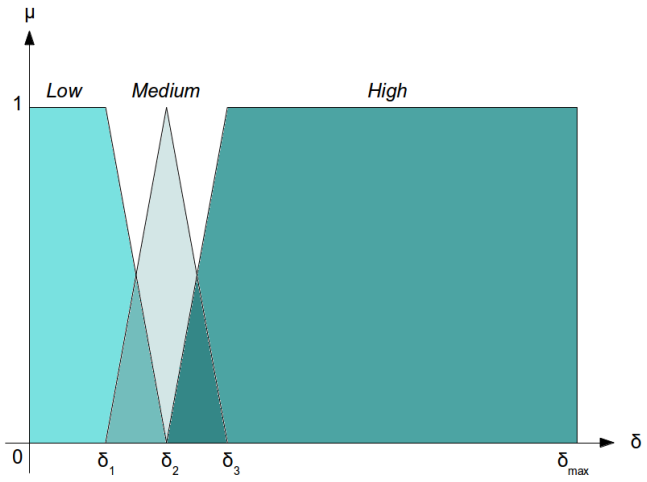


FIGURE 5.4: Distance from **g**: membership functions. Here the shape of the membership function is trapezoid for *Low* and *High*, while *Medium* has a triangular shape. Figure modified from [110].

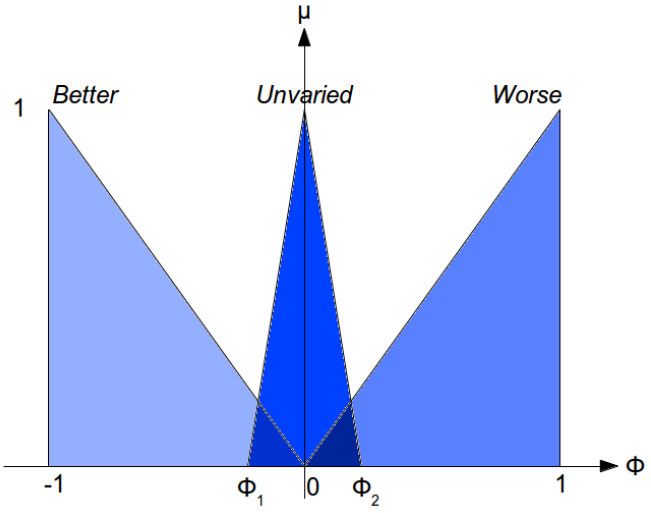


FIGURE 5.5: Normalized fitness incremental factor: membership functions. Here all the membership functions (*Better*, *Unvaried* and *Worse*) have a triangular shape. Figure modified from [110].

The FRBS is defined by rules listed in Table 5.1. The FRBS has been divided into 3 groups according to the output variable. $Inertia_i$ (rules 1–3), $Social_i$ (rules 4–6) and $Cognitive_i$ (rules 7–9). Accordingly to the value of the input linguistic variables (ϕ_i and δ_i), each rule defines the changes in the settings of the PPSO and in Figure 5.6 the three resulting surfaces for $Inertia_i$, $Social_i$ and $Cognitive_i$ calculated using the Sugeno method are shown.

The first group of rules is aimed at tuning the variable $Inertia_i$ (that determines the contribution of the previous velocity of a particle to its current velocity) in order to find better solutions for the optimization problem. This is done in order to increase the

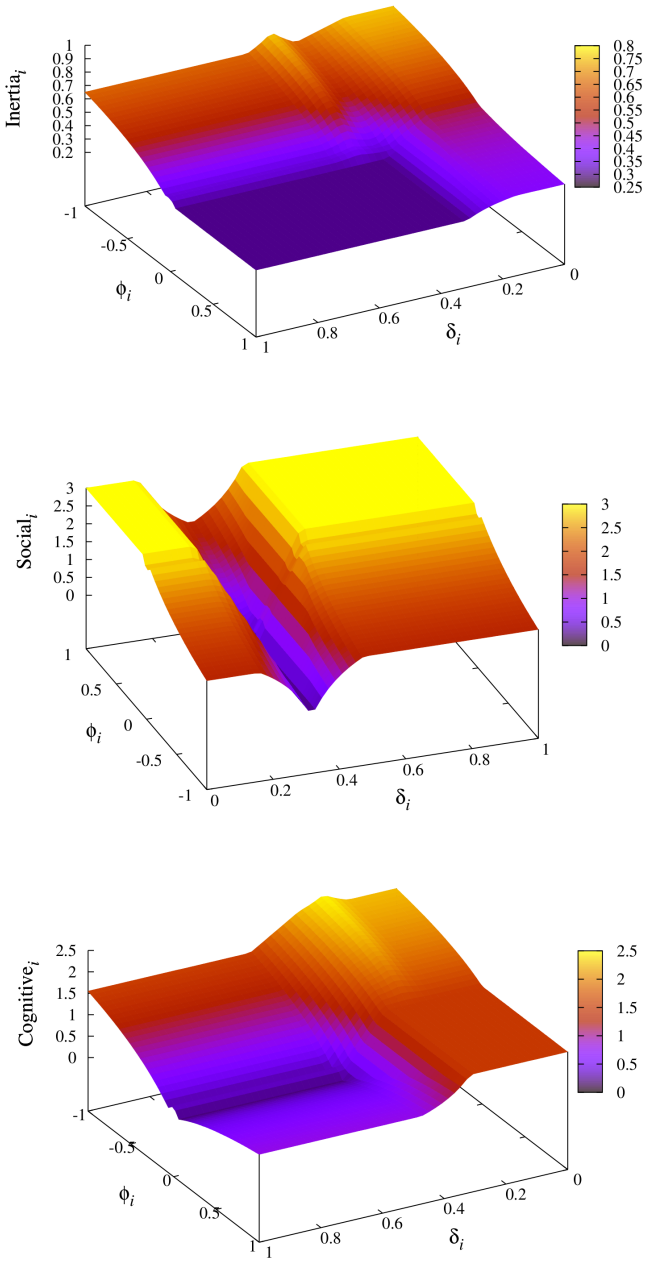


FIGURE 5.6: 3D plots describing the surfaces obtained for the inertia value (*top*), the social factor (*center*) and the cognitive factor (*bottom*) at different values of δ_i (here $\delta_{max} = 1$) and ϕ_i. Surfaces have been obtained thanks to the FRBS illustrated in Table 5.1. Figure modified from [110].

value of the variable when the particle shows a good performance (in terms of fitness), and to lower it in a different case. This is obtained by means of the imposition of a *Low* value when ϕ_i is *Worse*, or the distance δ_i from the global best is *High* or *Medium*. A *High* value is instead set when the particle is following the right direction (i.e., ϕ_i is *Better*). Lastly, $Inertia_i$ assumes a neutral (*Medium*) value, in the case there are no relevant variations in fitness or the distance from the global best is *Low*.

The second group of rules has been defined to control the strength of the social information sharing among particles. This is obtained through the setting of the variable $Social_i$, that tunes the attraction of particle i to $\mathbf{g}$ (the global best of the swarm). Here the particle should ignore the information retrieved from the swarm (setting $Social_i$ to *Low*) when it is finding better solutions (i.e., ϕ_i is *Better*), or is positioned near $\mathbf{g}$ (i.e., δ_i is *Medium*). In the inverse case, the particle should set $Social_i$ to *High* following "the advice" of the other components of the swarm when it is not able to find better solutions or it is positioned far from $\mathbf{g}$. Moreover, if ϕ_i is *Unvaried* (i.e. if no relevant changes in the fitness value are seen) an intermediate value is assigned by the FRBS to the $Social_i$.

The third group of rules is defined to determine the value for the variable $Cognitive_i$, weighting the attraction of the particle to $\mathbf{b}_i$ (personal best). In the case of a *High* distance of the particle from $\mathbf{g}$, its movement towards $\mathbf{b}_i$ should be limited by setting $Cognitive_i$ to *Low*. Instead, the tendency of the particle to move towards $\mathbf{b}_i$ should be balanced (with respect to the effect of the social component) by setting an intermediate value of $Cognitive_i$, when particle is not improving its fitness value or it is not far from $\mathbf{g}$. Lastly, if ϕ_i is *Better* (i.e. better optimizations are seen), the particle should be encouraged to perform a local exploration around its current position within the search space by setting $Cognitive_i$ to *High*.

Summing up the above described rules, Figure 5.6 shows the three resulting surfaces for $Inertia_i$, $Social_i$ and $Cognitive_i$ computed by means of the Sugeno method.

Analyzing the surface relative to $Inertia_i$ (Figure 5.6, top), it emerges how the maximum value is reached at the minimum value of ϕ_i and δ_i . The value of $Inertia_i$ then decreases both increasing the distance δ_i, and, strongly, ϕ_i. A minimum value of $Inertia_i$ is found in a large hollow approximately delimited by $0 < \phi_i < 1$ and $0.35 < \delta_i < 1$.

The surface associated to $Social_i$ (Figure 5.6, center), shows how the maximum value for this output is reached for a large plateau in $0 < \phi_i < 1$ and almost for any δ_i (except for a deep dip centered at δ_i=$\sim$0.3). The large plateau undergoes a smooth decrease in the range $-1 < \phi_i < 0$, for all the values of δ_i except for the already mentioned dip. This slope determines a lowering of the $Social_i$ value gradually from the maximum value 3 to 1.55.

The bottom plot in Figure 5.6, depicts the surface obtained for $Cognitive_i$. In this last case, the highest value for the output is found in a peak located around minimum values for both ϕ_i and δ_i. A small plateau can be found in the region delimited by $0 < \delta_i < 0.3$ and $0 < \phi_i < 1$. The remaining part of the plot is characterized by a gentle and wide dip centered in ϕ_i=0 where $Cognitive_i$ has a minimum value.

Moreover, in the tuning of the PPSO, the size N of the swarm has been determined in an automatic fashion by means of an heuristic described in [245] where $N = \lfloor 10 + 2\sqrt{M} \rfloor$. Lastly it has been defined, along each component of the search space, the maximum velocity of particle as $v_{max_m} = 0.2 \cdot |b_{max_m} - b_{min_m}|$, for $m = 1, \ldots, M$.

Comparative evaluation of PSO and PPSO PPSO has been tested comparing its performance with a standard implementation of PSO evaluating 12 reference benchmark functions parametric in the number of dimensions $M \in \mathbb{N}$ and illustrated in Table 5.3.

Function	*Equation*	*Search space*	*Value in global minimum*
Ackley	$f(\mathbf{x}) = 20 + e - 20\exp(-.2\sqrt{\frac{1}{M}\sum_{m=1}^{M} x_m^2}) - \exp(\frac{1}{n}\sum_{m=1}^{M}\cos(2\pi x_m))$	$[-30, 30]^M$	$f(\mathbf{0}) = 0$
Alpine 1	$f(\mathbf{x}) = \sum_{m=1}^{M} \lvert x_m \sin(x_m) + .1x_m \rvert$	$[-10, 10]^M$	$f(\mathbf{0}) = 0$
Bohachevsky	$f(\mathbf{x}) = \sum_{m=1}^{M-1}(x_m^2 + 2x_{m+1}^2 - .3\cos(3\pi x_m) - .4\cos(4\pi x_{m+1}) + .7)$	$[-15, 15]^M$	$f(\mathbf{0}) = 0$
Griewank	$f(\mathbf{x}) = \frac{1}{4000}\sum_{m=1}^{M} x_m^2 - \prod_{m=1}^{M}\cos(\frac{x_m}{\sqrt{m}}) + 1$	$[-600, 600]^M$	$f(\mathbf{0}) = 0$
Michalewicz	$f(\mathbf{x}) = -\sum_{m=1}^{M}\sin(x_m)\sin^{2k}(\frac{mx_m^2}{\pi})$, with $k = 10$ in this work	$[0, \pi]^M$	$f(\mathbf{0}) = -1.8013$
Mishra 1	$f(\mathbf{x}) = (1 + \alpha_M)^{\alpha_M}, \alpha_M = M - \sum_{m=1}^{M-1} x_m$	$[0, 1]^M$	$f(\mathbf{1}) = 2$
Ferretti 1	$f(\mathbf{x}) = 30 + \sum_{m=1}^{M} \lfloor x_m \rfloor$	$[-5.12, 5.12]^M$	$f(\mathbf{-5.12}) = -6M + 30$
Quintic	$f(\mathbf{x}) = \sum_{m=1}^{M} \lvert x_m^5 - 3x_m^4 + 4x_m^3 + 2x_m^2 - 10x_m - 4 \rvert$	$[-10, 10]^M$	$f(\mathbf{-1}) = 0$
Rastrigin	$f(\mathbf{x}) = 10M + \sum_{m=1}^{M}(x_m^2 - 10\cos(2\pi x_m))$	$[-5.12, 5.12]^M$	$f(\mathbf{0}) = 0$
Rosenbrock	$f(\mathbf{x}) = \sum_{m=1}^{M-1}[100(x_m^2 - x_{m+1})^2 + (x_m - 1)^2]$	$[-2048, 2048]^M$	$f(\mathbf{1}) = 0$
Schwefel 26	$f(\mathbf{x}) = 418.98M - \sum_{m=1}^{M} x_m \sin(\sqrt{\lvert x_m \rvert})$	$[-512, 512]^M$	$f(\mathbf{420.9}) = 0$
Xin-She Yang 2	$f(\mathbf{x}) = \sum_{m=1}^{M} \lvert x_m \rvert [\exp(\sum_{m=1}^{M}\sin(x_m^2)]^{-1}$	$[-2\pi, 2\pi]^M$	$f(\mathbf{0}) = 0$

TABLE 5.3: Benchmark functions. Table modified from [110].

To perform the comparison it has been exploited the *Average Best Fitness* (ABF), a value calculated as the mean of the global best particle fitness value found at each iteration t. The value has been evaluated over a number Θ of runs using both PSO or PPSO:

$$\text{ABF} = \frac{1}{\Theta}\sum_{\theta=1}^{\Theta} f(\mathbf{g}_\theta(t)), \tag{5.9}$$

where $\mathbf{g}_\theta(t)$ is the global best found at iteration t during the θ-th run using either PSO or PPSO. In the present case Θ has been set to 30.

For what concerns the PSO, the following values have been set for the parameters of the algorithm:

- inertia w linearly decrementing from 0.9 to 0.4;
- cognitive factor $c_{cog} = 1.9$;
- social factor $c_{soc} = 1.9$.

While, both for PSO and PPSO the values of N and v_{max_m} have been determined on the basis of the heuristic illustrated in Section 5.6. For both algorithms, the damping boundary condition was used and the number of iterations was fixed to 400.

All the benchmark functions in Table 5.3 have been tested with PSO and PPSO by setting $M = 100$. In Figure 5.7 values for the ABF are illustrated both for the canonical PSO algoritm (red dashed lines) and for the PPSO (green solid lines) In all tested cased the ABF value is lower in the PPSO case respect to the canonical PSO, thereby indicating that PPSO outperforms PSO in terms of convergence to better solutions.

In Figure 5.7 are illustrated comparisons for $M = 100$. Here, the canonical PSO is hardly reaching the optimal solution, whereas with PPSO, the ABF is constantly improving

Globally, in this section it has been shown how the developed PPSO algorithm, exploiting FL, has better performances with respect to standard PSO when tested on 12 standard multi-dimensional benchmark functions: for all the tested benchmarks, PPSO exhibit both faster convergence and better average fitness values than standard PSO.

The next step, not included in this work of thesis, but currently under implementation, is the exploitation of the efficient performances of the PPSO for the estimation of the kinetic constants described in Section 5.5.

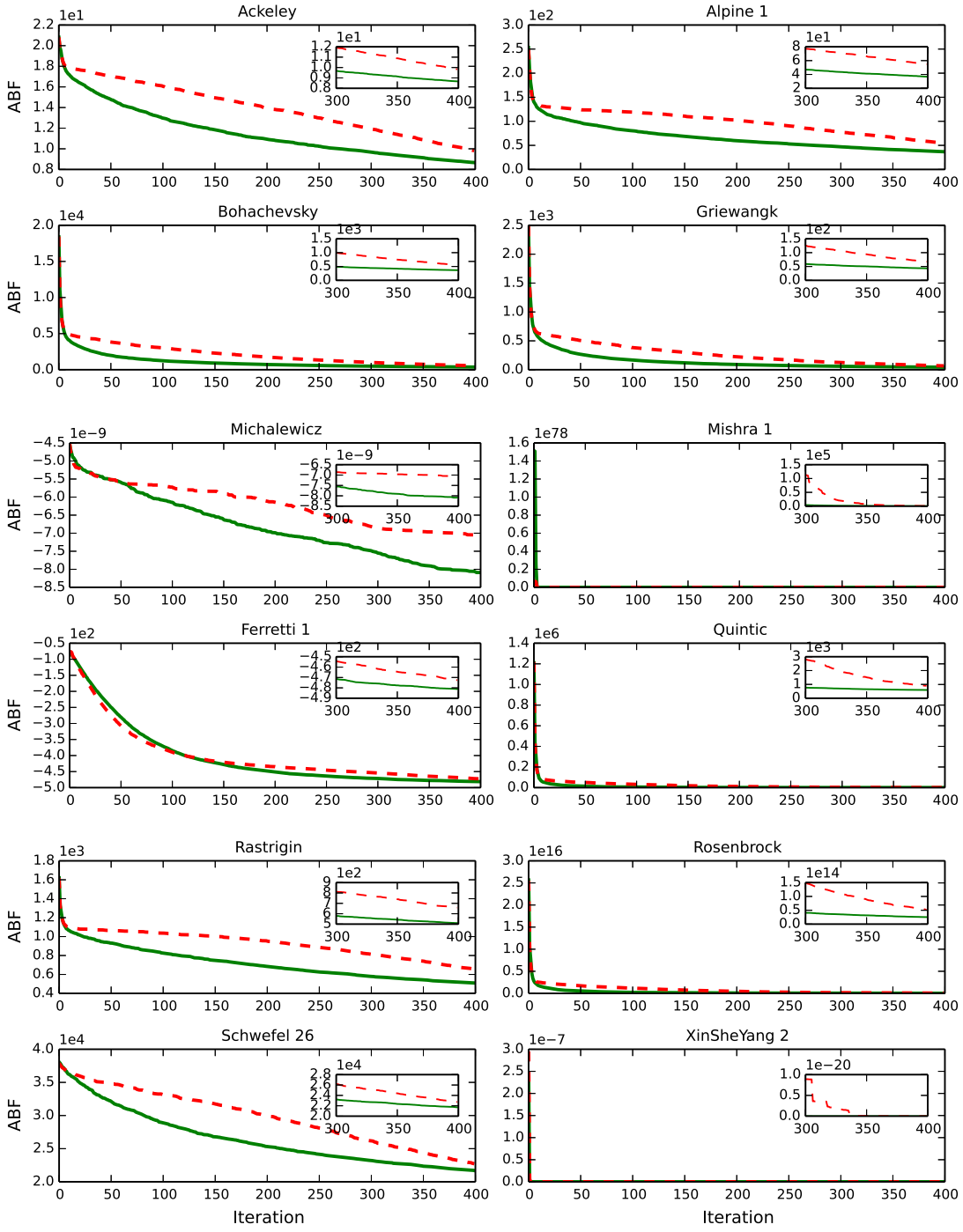


FIGURE 5.7: In this figure is shown the evaluation of PPSO performance (green) compared to the standard PSO (red). ABF values for the benchmark functions in Table 5.3 are illustrated on x axis for $\Theta = 30$ runs, where the size of M was 100. Small plots zoom on the final 100 iterations. Figure from [110].

Chapter 6

Conclusions and perspectives

6.1 Conclusions

The complex nature of biological systems highlight the need for methods that are able to investigate the system under different aspects. In this context, in the scientific community is emerging an increasing interest for the definition and application of methods able to integrate information deriving from multiple "levels" of analysis in order to obtain a thorough comprehension of the system under evaluation. The need for a multi-level analysis has been underlined by the fact that, at present, there is no computational "silver bullet" for the understanding biological complex systems such as metabolism.

To overtake limitations of the single modeling methods, in this thesis I defined and implemented a computational approach dealing with complexity in biological systems through the definition of a computational pipeline exploiting all the three different "levels" of Systems Biology analyses (i.e. interaction-based, constraint-based and mechanism-based).

The main novelty of the approach is the attempt to gain, from every level, a different type of information, i.e. identification of flux distributions and metabolic sub-phenotypes from the ensemble evolutionary FBA (eeFBA); information on network structural properties, correlations between flux values and topological metrics from graph theory approaches; estimation of kinetic constants and simulation of dynamics by means of mechanistic approaches. Moreover, to provide a better communication of the experimental results, I also redefined a network visualization strategy able to overlay flux values and topological metrics to network structure.

More in detail, in Chapter 3, the eeFBA approach has been exploited to explore the space of the randomly generated objective functions, by means of a filtering procedure

selecting solutions matching a metabolic phenotype, or by means of a genetic algorithm to identify both solutions that best matches the reference phenotype and a pool of solutions (individuals) characterizing the ensemble properties of the phenotype.
The procedure has been tested on a yeast core model from which it was possible to retrieve two ensembles of solutions matching a definition for Crabtree-positive and negative phenotypes based on the evaluation of the oxygen uptake flux (proxy of the oxidative phosphorylation) and the ethanol secretion flux (proxy for the fermentative metabolism). With a hierarchical clustering it has been possible to identify, inside each phenotype, some sub-clusters revealing potential sub-phenotypes.
Moreover, through a Kolmogorov-Smirnov test it has been possible to highlight metabolic fluxes that are enhanced in the Crabtree-positive and negative phenotypes. In particular, accordingly to literature data, the glycolytic pathway is enhanced in the Crabtree-positive phenotype, while reactions linking glycolysis with the TCA cycle and reactions producing building blocks are up-regulated in the Crabtree-negative case.
Lastly, the classical FBA approach has been used to assess the correctness of predictions emerging from the reduction of genome-wide metabolic models of three different types of cancer cells (iLiverCancer1715, iBreastCancer1771, iLungCancer1472) and a "reference" cell (HMR). In the reduced models, flux of metabolites through the reactions has been investigated to understand up-and down-regulations in metabolic pathways involved in the redistribution of fluxes in the cancer condition. In particular, results underlined that flux distributions significantly differ both between the reference and cancer models and among the three cancer models.

In Chapter 4, an interaction-based approach based on graph theory has been used to analyze topological measures of the genome-wide models exploited in Chapter 3 and on the derived "core" versions. This modeling framework confirmed that, in the analyzed networks, it is possible to retrieve some key features typical of biological networks such a scale-free and hierarchical topology (indicating the presence of modules). At the same time these networks exhibit the ultra small-world property and the disassortative nature. Remarkably, topological measures assumed comparable values in all the evaluated models confirming that the interaction-based approach has not predictive ability due to the fact that it does not allow to highlight the redistribution of metabolic fluxes at the basis of cellular transformation.
A further investigation has involved the evaluation of the correlation between node degree and flux value emerged from FBA both from yeast core model (evaluating the average of fluxes from the Crabtree-negative ensemble) and the HMR model. In both cases analyses pointed out a significant value of correlation suggesting a key role of hubs in sustaining the flux of metabolites through metabolic reactions.

In Chapter 5, in the context of mechanism-based analyses I developed MetaFluxAnalysis, a LabVIEW tool to determine metabolic fluxes starting from mechanistic simulations. This has been a first step towards the estimation of kinetic constants from a flux distribution obtained with FBA. The flux distribution is the target for the estimation of kinetic constants by means of a Particle Swarm Optimizer (PSO). Unfortunately, the exploited MATLAB PSO implementation has not been able to provide a set of kinetic constants due to convergence problems. For this reason a novel version of the PSO algorithm named Proactive Particles in Swarm Optimization (PPSO) has been developed exploiting Fuzzy Logic to automatically tune particle parameters (inertia, social and cognitive components).
From the comparative evaluation of PSO and PPSO it emerged that PPSO has better performances when evaluating 12 standard multi-dimensional benchmark functions, confirming the effectiveness of the method.

6.2 Perspectives

The next step in the application of the computational pipeline will be the conclusion of the kinetic constants estimation by means of the PPSO algorithm. Once putative kinetic constants have been estimated, the following task will be the mechanistic simulation of a metabolic system (e.g. the yeast CM exploited in this thesis).
It is worth to underline that besides being able to simulate the kinetic evolution of the metabolic system (i.e. the transient state [246, 247] and the variations of the steady state due the modification of both the molecular quantities and the kinetic constants), mechanism-based approach can be used as a framework to investigate the sensitivity of the kinetic constants and to determine which reactions are the most relevant governing points of the system behavior.

A further future development of the eeFBA approach will be its application to curated genome-wide models. In this case, because of the higher computational requirements due to the expected extension of the random set of objective functions (performed to mitigate the effects of a possible under sampling of the solution space), or due to the increase of the number of individuals evaluated by the genetic algorithm in the sampling and searching methods of the eeFBA, high-performance computing (HPC) capabilities will be probably required (see [248] for a discussion on perspectives for the integration of Systems Biology and HPC).

Theoretical approaches defined in the present dissertation are currently being applied to develop a model to investigate cancer cell proliferation [171], a system level property that can be fully understood only exploiting strategies able to tackle the complexity of

the underlying metabolic network. In this context, recent studies suggested that the enhanced proliferation is sustained by an extensive metabolic rewiring involving an up-regulated glycolytic flux and an increased uptake of glutamine [167].
The investigation of cancer metabolic rewiring by means of a computational model is a fundamental step towards the understanding of how this metabolic phenotype rises and is a starting point to devise strategies to counteract the phenomenon. To this end I am contributing to develop a CM of metabolism simulated by means of a constraint-based approach leading to the prediction of fluxes distribution underlying the metabolic rewiring in pseudo-hypoxic conditions and in the case of glutamine limitation.

As widely discussed in Chapter 2, dynamic mechanism-based models are considered the best way to achieve a detailed comprehension of metabolic processes. However they require knowledge not only of the metabolic network and stoichiometry of reactions, but also of the kinetic rate of each reaction and of the initial concentration of all the metabolites. By exploiting the computational pipeline here devised I aim at identifying values for these parameters to analyze the system exploiting the three different frameworks proposed by Stelling in [16] and with the final goal to develop a computational model able to describe the interplay between enhanced glycolysis and its related pathways and glutamine utilization pathways, a condition that may be relevant in sustaining tumor forming ability.

The modeling of cancer metabolic rewiring, and in particular the desirable identification of sub-phenotypes with the eeFBA approach, is part of a new vision postulating the transition from a Systems Biology to a Systems Medicine perspective [156].
The need for a Systems Medicine approach is highlighted by the many limitations of current clinical approaches such as the evidence that many drugs are effective only in a reduced portion of patients or the fact that targeting some relevant functions is ineffective for the remission of the disease. Moreover, from the systemic point of view it is more and more evident that biological basis of many diseases can be fully understood only in a network perspective where altered mechanisms can affect topologically or functionally remote factors possibly influenced by individual variability [249].

The final goal of the paradigm shift is therefore the realization of the personalized medicine [250] through the implementation of a "virtual twin" [251], i.e. an accurate, multi-scale and personal, computational description of the human physiology.

In this context the evaluation, quantification and the management of the complexity (as illustrated in this thesis) is a pivotal aspect to fruitfully integrate -omics data (with potential incompatibilities due to the structure of the data, and to the conflicting evidences emerging from studies performed under different experimental settings) and to perform accurate and reliable simulations of the whole body physiology.

Appendix A

Flux distributions in reference and cancer CMs

In this appendix flux distributions emerging from FBA are illustrated for the four core models (CMs) derived in Chapter 3 from the corresponding genome-wide models. The first column of the hereafter shown table, indicates the structure of the metabolic reaction, while in the second column the value of the flux in the "reference" model is reported (HMR model). Columns from 3 to 5 list the value of fluxes for the tissue specific cancer models (i.e. liver, breast and lung).

Reactions	HMR model	Liver Cancer model	Breast Cancer model	Lung Cancer model
glucose[s] → glucose[c]	88,50025942	367,3906121	346,9188093	599,797916
ATP[c] + glucose[c] → ADP[c] + glucose-6-phosphate[c]	88,50025942	367,3906121	346,9188093	599,797916
glucose-6-phosphate[c] → fructose-6-phosphate[c]	88,50025942	367,3906121	346,9188093	599,797916
ATP[c] + fructose-6-phosphate[c] → ADP[c] + fructose-1,6-bisphosphate[c]	83,45954903	352,9138514	332,4325152	580,5023231
fructose-1,6-bisphosphate[c] → DHAP[c] + GAP[c]	83,45954903	352,9138514	332,4325152	580,5023231
DHAP[c] ↔ GAP[c]	78,41883865	352,9138514	332,4325152	561,2067303
GAP[c] + NAD[c] + Pi[c] → 1,3-bisphospho-D-glycerate[c] + H[c] + NADH[c]	151,7969669	698,5893225	657,6218834	1103,117868
1,3-bisphospho-D-glycerate[c] + ADP[c] → 3-phospho-D-glycerate[c] + ATP[c]	0	698,5893225	657,6218834	1103,117868
1,3-bisphospho-D-glycerate[c] → 2,3-bisphospho-D-glycerate[c]	151,7969669	-	-	-
2,3-bisphospho-D-glycerate[c] + H2O[c] → 3-phospho-D-glycerate[c] + Pi[c]	151,7969669	-	-	-
3-phospho-D-glycerate[c] → 2-phospho-D-glycerate[c]	58,71125853	564,9197185	523,8642541	946,2022284
2-phospho-D-glycerate[c] → H2O[c] + PEP[c]	58,71125853	564,9197185	523,8642541	946,2022284
ADP[c] + PEP[c] → ATP[c] + pyruvate[c]	58,71125853	564,9197185	523,8642541	946,2022284
H[c] + NADH[c] + pyruvate[c] → L-lactate[c] + NAD[c]	0	223,7461825	152,7124929	871,6275454
H[c] + NADH[c] + pyruvate[c] → D-lactate[c] + NAD[c]	0	0	0	0
L-lactate[c] → L-lactate[s]	0	223,7461825	152,7124929	871,6275454
D-lactate[c] → D-lactate[s]	0	0	0	0
DHAP[c] + erythrose-4-phosphate[c] ↔ sedoheptulose-1,7-bisphosphate[c]	5,040710385	0	0	19,29559284
ATP[c] + ribose-5-phosphate[c] → AMP[c] + PRPP[c]	15,12213115	21,71514101	21,72944105	57,88677852
glucose-6-phosphate[c] + NADP[c] ↔ glucono-1,5-lactone-6-phosphate[c] + H[c] + NADPH[c]	0	0	0	0
glucono-1,5-lactone-6-phosphate[c] + H2O[c] → 6-phospho-D-gluconate[c]	0	0	0	0
6-phospho-D-gluconate[c] + NADP[c] → CO2[c] + H[c] + NADPH[c] + ribulose-5-phosphate[c]	0	0	0	0
ribulose-5-phosphate[c] ↔ ribose-5-phosphate[c]	10,08142077	14,47676067	14,48629404	38,59118568
fructose-6-phosphate[c] + GAP[c] ↔ D-xylulose-5-phosphate[c] + erythrose-4-phosphate[c]	5,040710385	7,238380336	7,243147018	19,29559284
D-xylulose-5-phosphate[c] ↔ ribulose-5-phosphate[c]	10,08142077	14,47676067	14,48629404	38,59118568
GAP[c] + sedoheptulose-7-phosphate[c] ↔ D-xylulose-5-phosphate[c] + ribose-5-phosphate[c]	5,040710385	7,238380336	7,243147018	19,29559284
GAP[c] + sedoheptulose-7-phosphate[c] ↔ erythrose-4-phosphate[c] + fructose-6-phosphate[c]	0	-7,238380336	-7,243147018	0
ADP[c] + sedoheptulose-1,7-bisphosphate[c] ↔ ATP[c] + sedoheptulose-7-phosphate[c]	5,040710385	0	0	19,29559284
CoA[m] + NAD[m] + pyruvate[m] → acetyl-CoA[m] + CO2[m] + H[m] + NADH[m]	150,480784	0	0	0
ATP[m] + H[m] + HCO3-[m] + pyruvate[m] → ADP[m] + OAA[m] + Pi[m]	0	256,8650979	286,7878036	2129,154672
acetyl-CoA[m] + H2O[m] + OAA[m] → citrate[m] + CoA[m]	150,480784	0	0	0
citrate[m] ↔ isocitrate[m]	0	0	-299,6337474	-270,1459903
isocitrate[m] + NAD[m] → AKG[m] + CO2[m] + H[m] + NADH[m]	0	-	-	-
isocitrate[m] + NADP[m] → AKG[m] + CO2[m] + H[m] + NADPH[m]	0	-	-	-
isocitrate[m] + NAD[m] ↔ AKG[m] + CO2[m] + H[m] + NADH[m]	-	0	0	-270,1459903
isocitrate[m] + NADP[m] ↔ AKG[m] + CO2[m] + H[m] + NADPH[m]	-	-199,8162148	-299,6337474	-
AKG[m] + CoA[m] + NAD[m] → CO2[m] + H[m] + NADH[m] + succinyl-CoA[m]	519,6058158	48,54131611	18,8197293	270,1459903
GDP[m] + Pi[m] + succinyl-CoA[m] ↔ CoA[m] + GTP[m] + succinate[m]	519,6058158	48,54131611	18,8197293	270,1459903
FAD[m] + succinate[m] ↔ FADH2[m] + fumarate[m]	519,6058158	48,54131611	18,8197293	270,1459903
fumarate[m] + H2O[m] ↔ malate[m]	300,961568	371,2156973	280,8140181	345,9665426
malate[m] + NAD[m] ↔ H[m] + NADH[m] + OAA[m]	150,480784	371,2156973	280,8140181	-1783,18813

Continued on next page

Reactions	HMR model	Liver Cancer model	Breast Cancer model	Lung Cancer model
malate[m] + NAD[m] → CO2[m] + H[m] + NADH[m] + pyruvate[m]	0	0	0	2129,154672
malate[m] + NADP[m] → CO2[m] + H[m] + NADPH[m] + pyruvate[m]	150,480784	0	0	0
ATP[c] + citrate[c] + CoA[c] → acetyl-CoA[c] + ADP[c] + OAA[c] + Pi[c]	139,1493155	199,8162148	199,9477996	176,7466129
fumarate[c] + H2O[c] ↔ malate[c]	225,0351401	-313,4971609	-253,4695539	9,072484036
malate[c] + NAD[c] ↔ H[c] + NADH[c] + OAA[c]	225,0351401	-313,4971609	-253,4695539	9,072484036
GTP[c] + OAA[c] → PEP[c] + GDP[c] + CO2[c]	-	0	0	-
citrate[c] ↔ isocitrate[c]	0	-199,8162148	0	-
isocitrate[c] + NADP[c] → AKG[c] + CO2[c] + H[c] + NADPH[c]	0	0	-	-
isocitrate[c] + NADP[c] → H[c] + NADPH[c] + oxalosuccinate[c]	-	-	0	-
oxalosuccinate[c] → AKG[c] + CO2[c]	-	-	0	-
H2O[c] + Ppi[c] → 2 Pi[c]	1507,428888	67,5771657	66,96313845	97,87172871
5 H[m] + NADH[m] + ubiquinone[m] → 4 H[c] + NAD[m] + ubiquinol[m]	953,250357	0	0	345,9665426
FADH2[m] + ubiquinone[m] → FAD[m] + ubiquinol[m]	916,6305718	238,8836195	299,2749852	528,6194987
2 ferricytochrome-C[m] + 2 H[m] + ubiquinol[m] → 2 ferrocytochrome-C[m] + 4 H[c] + ubiquinone[m]	1879,070456	252,0796356	312,4796912	886,2585232
4 ferrocytochrome-C[m] + 8 H[m] + O2[m] → 4 ferricytochrome-C[m] + 4 H[c] + 2 H2O[m]	939,5352282	126,0398178	156,2398456	443,1292616
ADP[m] + 4 H[c] + Pi[m] → ATP[m] + 4 H[m] + H2O[m]	2851,683233	156,7184564	255,1221305	1588,862692
H[c] + pyruvate[c] → H[m] + pyruvate[m]	0	256,8650979	286,7878036	0
citrate[m] → citrate[c]	150,480784	-	299,6337474	270,1459903
isocitrate[m] → isocitrate[c]	-	199,8162148	-	-
H[c] + Pi[c] ↔ H[m] + Pi[m]	3680,232945	-	-	-
ADP[c] + ATP[m] ↔ ADP[m] + ATP[c]	0	-	-	-270,1459903
H[s] ↔ H[c]	22,8222203	30,13595287	30,15579827	26,65663346
CO2[c] ↔ CO2[m]	-820,1090951	408,798092	567,6018218	-
CoA[c] ↔ CoA[m]	5,588887057	-	-	-
O2[s] → O2[c]	1000	208,9136882	239,1682909	516,4350036
O2[c] → O2[m]	939,5352282	126,0398178	156,2398456	443,1292616
GDP[c] + GTP[m] ↔ GDP[m] + GTP[c]	3370,372471	-52,92151619	-12,84594381	-
ADP[c] + GTP[c] ↔ ATP[c] + GDP[c]	3370,372471	-52,92151619	-12,84594381	-38,67873965
ADP[m] + GTP[m] ↔ ATP[m] + GDP[m]	-2850,766656	101,4628323	31,66567311	270,1459903
H2O[s] ↔ H2O[c]	-310,0952067	195,2657261	295,7387908	6,472486166
H2O[c] ↔ H2O[m]	-4296,650859	-	-	-
riboflavin[s] ↔ riboflavin[c]	0	0	-5,55383232728890e-29	2,83222534537321e-29
ATP[c] + riboflavin[c] → ADP[c] + FMN[c]	0	0	-5,55383232728890e-29	2,83222534537321e-29
ATP[c] + FMN[c] → FAD[c] + Ppi[c]	0	0	-5,55383232728890e-29	2,83222534537321e-29
FAD[c] + H[c] + NADPH[c] → FADH2[c] + NADP[c]	397,024756	190,3423034	280,4552559	258,4735084
FAD[m] + FADH2[c] ↔ FAD[c] + FADH2[m]	397,024756	190,3423034	280,4552559	258,4735084
AKG[c] + Pi[m] ↔ AKG[m] + Pi[c]	528,5047212	-269,752865	-249,148345	194,3254381
NH3[c] → NH3[s]	562,979771	707,2877507	817,7963379	687,0063997
NH3[c] ↔ NH3[m]	9,357194155	110,6284946	1,90581567040500e-13	0
fumarate[m] + Pi[c] ↔ fumarate[c] + Pi[m]	218,6442478	-322,6743811	-261,9942888	-75,82055223
CO2[c] + H2O[c] → H[c] + HCO3-[c]	48,31173694	69,37488965	69,42057501	61,36527397
CO2[m] + H2O[m] → H[m] + HCO3-[m]	0,458288739	257,5231933	286,7878036	2129,154672
glutamine[s] → glutamine[c]	706,0189519	1000	1000	1000

Continued on next page

Reactions	HMR model	Liver Cancer model	Breast Cancer model	Lung Cancer model
glutamine[c] + H2O[c] → glutamate[c] + NH3[c]	524,880328	749,7692653	749,6044812	701,3019512
glutamate[c] + H[c] → glutamate[m] + H[m]	0	0	0	0
glutamine[c] + H[c] → glutamine[m] + H[m]	0	0	-	-
glutamine[m] + H2O[m] → glutamate[m] + NH3[m]	0	0	-	-
ATP[c] + glutamate[c] + H2O[c] → ADP[c] + glutamate[s] + Pi[c]	0	-	-	-
glutamine[c] + H2O[c] + PRPP[c] → 5-phosphoribosylamine[c] + glutamate[c] + Ppi[c]	5,932603611	8,519124894	8,524734982	46,21429662
aspartate[c] + GTP[c] + IMP[c] → adenylosuccinate[c] + GDP[c] + Pi[c]	0	0	0	38,67873965
adenylosuccinate[c] ↔ AMP[c] + fumarate[c]	0	0	0	38,67873965
5-phosphoribosylamine[c] + ATP[c] + glycine[c] → ADP[c] + GAR[c] + Pi[c]	5,932603611	8,519124894	8,524734982	46,21429662
10-formyl-THF[c] + GAR[c] → N-formyl-GAR[c] + THF[c]	5,932603611	8,519124894	8,524734982	46,21429662
ATP[c] + glutamine[c] + H2O[c] + N-formyl-GAR[c] → 5-phosphoribosylformylglycinamidine[c] + ADP[c] + glutamate[c] + Pi[c]	5,932603611	8,519124894	8,524734982	46,21429662
5-phosphoribosylformylglycinamidine[c] + ATP[c] → ADP[c] + AIR[c] + Pi[c]	5,932603611	8,519124894	8,524734982	46,21429662
AIR[c] + CO2[c] ↔ 5-phosphoribosyl-4-carboxy-5-aminoimidazole[c]	5,932603611	8,519124894	8,524734982	46,21429662
5-phosphoribosyl-4-carboxy-5-aminoimidazole[c] + aspartate[c] + ATP[c] → ADP[c] + Pi[c] + SAICAR[c]	5,932603611	8,519124894	8,524734982	46,21429662
SAICAR[c] ↔ AICAR[c] + fumarate[c]	5,932603611	8,519124894	8,524734982	46,21429662
10-formyl-THF[c] + AICAR[c] ↔ FAICAR[c] + THF[c]	5,932603611	8,519124894	8,524734982	46,21429662
FAICAR[c] ↔ H2O[c] + IMP[c]	5,932603611	8,519124894	8,524734982	46,21429662
IMP[c] + NADP[c] + NH3[c] ↔ GMP[c] + H[c] + NADPH[c]	-1448,473116	8,519124894	8,524734982	7,535556967
ATP[c] + GMP[c] ↔ ADP[c] + GDP[c]	0,750308087	1,077430539	1,078140058	0,953036762
GDP[c] → dGDP[c] + H2O[c]	0,750308087	1,077430539	1,078140058	0,953036762
ADP[c] + dGDP[c] ↔ ATP[c] + dGMP[c]	0,750308087	1,077430539	1,078140058	0,953036762
2 ADP[c] ↔ AMP[c] + ATP[c]	-1475,491193	-21,71514101	-21,0709123	-95,98340261
ADP[c] → dADP[c] + H2O[c]	1,12476352	1,615142614	1,616206232	1,428667772
ADP[c] + dADP[c] ↔ ATP[c] + dAMP[c]	1,12476352	1,615142614	1,616206232	1,428667772
H2O[c] + IMP[c] + NAD[c] → H[c] + NADH[c] + xanthosine-5-phosphate[c]	1454,40572	0	0	0
ATP[c] + NH3[c] + xanthosine-5-phosphate[c] → AMP[c] + GMP[c] + Ppi[c]	1454,40572	0	0	0
2 ATP[c] + glutamine[c] + H[c] + H2O[c] + HCO3-[c] → 2 ADP[c] + carbamoyl-phosphate[c] + glutamate[c] + Pi[c]	9,189527543	13,19601612	13,20470607	11,67248191
aspartate[c] + carbamoyl-phosphate[c] ↔ N-carbamoyl-L-aspartate[c] + Pi[c]	9,189527543	13,19601612	13,20470607	11,67248191
N-carbamoyl-L-aspartate[c] ↔ S-dihydroorotate[c] + H2O[c]	9,189527543	13,19601612	13,20470607	11,67248191
S-dihydroorotate[c] + ubiquinone[m] ↔ orotate[c] + ubiquinol[m]	9,189527543	13,19601612	13,20470607	11,67248191
orotate[c] + PRPP[c] ↔ orotidine-5-phosphate[c] + Ppi[c]	9,189527543	13,19601612	13,20470607	11,67248191
orotidine-5-phosphate[c] → CO2[c] + UMP[c]	9,189527543	13,19601612	13,20470607	11,67248191
ATP[c] + UMP[c] ↔ ADP[c] + UDP[c]	6,44678122	9,257475803	9,263572117	8,188662235
UDP[c] → dUDP[c] + H2O[c]	1,12476352	1,615142614	1,616206232	1,428667772
dUMP[c] + 5,10-methylene-THF[c] ↔ dihydrofolate[c] + dTMP[c]	1,12476352	1,615142614	1,616206232	1,428667772
dUDP[c] + ADP[c] ↔ dUMP[c] + ATP[c]	1,12476352	1,615142614	1,616206232	1,428667772
ADP[c] + UTP[c] ↔ ATP[c] + UDP[c]	-5,3220177	-7,642333189	-7,647365885	-6,759994463
ADP[c] + CTP[c] ↔ ATP[c] + CDP[c]	5,3220177	7,642333189	7,647365885	6,759994463
ADP[c] + CDP[c] ↔ ATP[c] + CMP[c]	4,571709613	6,56490265	6,569225827	5,806957701
ATP[c] + NH3[c] + UTP[c] → ADP[c] + CTP[c] + Pi[c]	5,3220177	7,642333189	7,647365885	6,759994463

Continued on next page

Reactions	HMR model	Liver Cancer model	Breast Cancer model	Lung Cancer model
CDP[c] → dCDP[c] + H2O[c]	0,750308087	1,077430539	1,078140058	0,953036762
dCDP[c] + ADP[c] ↔ dCMP[c] + ATP[c]	0,750308087	1,077430539	1,078140058	0,953036762
glutamate[m] + H2O[m] + NAD[m] ↔ AKG[m] + H[m] + NADH[m] + NH3[m]	0	-309,7866141	-299,6337474	0
glutamate[m] + H2O[m] + NADP[m] ↔ AKG[m] + H[m] + NADPH[m] + NH3[m]	-8,898905416	199,8162148	299,6337474	0
AKG[c] + aspartate[c] ↔ glutamate[c] + OAA[c]	-364,1844556	113,680946	53,52175431	-185,8190969
glutamate[m] + OAA[m] ↔ AKG[m] + aspartate[m]	0	628,0807952	567,6018218	345,9665426
aspartate[m] + glutamate[c] + H[c] → aspartate[c] + glutamate[m] + H[m]	0	628,0807952	567,6018218	345,9665426
aspartate[c] → fumarate[c] + NH3[c]	0	0	0	0
2 ATP[m] + H[m] + HCO3-[m] + NH3[m] → 2 ADP[m] + carbamoyl-phosphate[m] + Pi[m]	0,458288739	0,658095376	-	-
carbamoyl-phosphate[m] + ornithine[m] ↔ citrulline[m] + Pi[m]	0,458288739	0,658095376	-	-
citrulline[m] + H[c] + ornithine[c] → citrulline[c] + H[m] + ornithine[m]	0,458288739	0,658095376	-	-
citrulline[c] + H[c] + ornithine[m] → citrulline[m] + H[m] + ornithine[c]	0	0	-	-
aspartate[c] + ATP[c] + citrulline[c] → AMP[c] + argininosuccinate[c] + Ppi[c]	0,458288739	0,658095376	-	-
argininosuccinate[c] ↔ arginine[c] + fumarate[c]	0,458288739	0,658095376	-	-
arginine[c] + H2O[c] → urea[c] + ornithine[c]	0	0	-	-
ornithine[s] → ornithine[c]	-	-	427,5801385	0
ornithine[c] → CO2[c] + putrescine[c]	0	268,8685564	427,5801385	0
putrescine[c] → putrescine[s]	0	268,8685564	427,5801385	0
urea[c] → urea[s]	0	0	-	-
arginine[c] + glycine[c] ↔ ornithine[c] + guanidinoacetate[c]	0	0	-	-
H2O[c] + arginine[c] → NH3[c] + citrulline[c]	0	-	-	-
ornithine[m] + AKG[m] ↔ glutamate[m] + Lglu5semialdehyde[m]	7,81908002052499e-19	1,56381600410500e-18	-	-
H[m] + NADH[m] + glutamate[m] ↔ H2O[m] + Lglu5semialdehyde[m] + NAD[m]	8,898905416	109,9703993	-	-
Lglu5semialdehyde[m] ↔ Lglu5semialdehyde[c]	8,898905416	-	-	-
Lglu5semialdehyde[c] ↔ 1-Pyrroline5carboxylate[c] + H2O[c]	8,898905416	12,77868734	12,78710247	11,30333545
1-Pyrroline5carboxylate[c] + H[c] + NADH[c] → proline[c] + NAD[c]	0	12,77868734	0	0
1-Pyrroline5carboxylate[c] + H[c] + NADPH[c] → proline[c] + NADP[c]	8,898905416	-	12,78710247	11,30333545
glutamate[c] + 4-methyl-2-oxopentanoate[c] ↔ AKG[c] + leucine[c]	5,505053751	7,905170061	7,910375843	6,99248574
glutamate[c] + 2-oxo-3-methylvalerate[c] ↔ AKG[c] + isoleucine[c]	1,834552176	2,634387891	2,636122711	2,330236999
aspartate[c] + ATP[c] + glutamine[c] + H2O[c] → AMP[c] + asparagine[c] + glutamate[c] + Ppi[c]	5,96334249	-	-	-
asparagine[c] + H2O[c] → aspartate[c] + NH3[c]	0	-	-	-
glutamate[c] + phenylpyruvate[c] ↔ AKG[c] + phenylalanine[c]	4,587079052	-	-	-
O2[c] + phenylalanine[c] + tetrahydrobiopterin[c] → H2O[c] + tyrosine[c] + dihydrobiopterin[c]	2,752526876	-	-	-
glutamate[c] + mercaptopyruvate[c] ↔ AKG[c] + cysteine[c]	0,596613693	0,856727821	0,857292001	0,757815079
4-methylthio-2-oxobutanoic-acid[c] + 2 H[c] + glutamine[c] → glutamate[c] + methionine[c]	0,917974699	-	-	-
glutamate[c] + 4-methylthio-2-oxobutanoic-acid[c] → AKG[c] + methionine[c]	0	1,31819714	1,31906521	1,166005871
3-phospho-D-glycerate[c] + NAD[c] ↔ 3-phosphonooxypyruvate[c] + H[c] + NADH[c]	93,08570838	133,669604	133,7576293	156,9156394
3-phosphonooxypyruvate[c] + glutamate[c] ↔ AKG[c] + 3-phosphoserine[c]	93,08570838	133,669604	133,7576293	156,9156394
H2O[c] + 3-phosphoserine[c] → Pi[c] + serine[c]	93,08570838	133,669604	133,7576293	156,9156394
serine[c] + THF[c] ↔ 5,10-methylene-THF[c] + glycine[c] + H2O[c]	74,73739217	107,3217123	107,3923868	133,6097199
serine[c] → pyruvate[c] + NH3[c]	0	0	0	0
glutamine[c] + pyruvate[c] → 2-oxoglutaramate[c] + alanine[c]	58,71125853	84,30843807	84,3639576	74,57468294
2-oxoglutaramate[c] + H2O[c] → AKG[c] + NH3[c]	58,71125853	84,30843807	84,3639576	-

Continued on next page

Reactions	HMR model	Liver Cancer model	Breast Cancer model	Lung Cancer model
serine[c] → serine[m]	0	-	-	0
glycine[m] ↔ glycine[c]	0	-	-	0
THF[m] + serine[m] ↔ 5,10-methylene-THF[m] + H2O[m] + glycine[m]	0	0	4,74044086878493e-29	0
folate[s] → folate[c]	-2,00198482106296e-14	-	-	-
folate[c] → folate[m]	0	7,44536339118936e-29	-2,33573474355777e-29	0
formate[m] → formate[c]	0	-	0	-
H[m] + NADPH[m] + folate[m] → NADP[m] + dihydrofolate[m]	0	7,52324185198316e-29	-2,33573474355777e-29	0
H[m] + NADPH[m] + dihydrofolate[m] ↔ NADP[m] + THF[m]	0	7,25199159220308e-29	-2,33573474355777e-29	0
ATP[m] + THF[m] + formate[m] ↔ 10-formyl-THF[m] + ADP[m] + Pi[m]	0	-	-6,32058782504658e-29	-
10-formyl-THF[m] + H[m] ↔ 5,10-methenyl-THF[m] + H2O[m]	0	-7,20428107894788e-29	-6,03749898051165e-29	0
5,10-methenyl-THF[m] + NADH[m] ↔ 5,10-methylene-THF[m] + NAD[m]	-141,5818786	-7,44536339118936e-29	-4,74044086878493e-29	0
5,10-methenyl-THF[m] + NADPH[m] ↔ 5,10-methylene-THF[m] + NADP[m]	141,5818786	-	-	-
H[c] + NADH[c] + folate[c] ↔ NAD[c] + dihydrofolate[c]	-2,00198482106296e-14	-2,77847254161773e-14	-2,76720364574264e-14	7,87217655715256e-15
H[c] + NADPH[c] + folate[c] ↔ NADP[c] + dihydrofolate[c]	0	0	0	0
H[c] + NADH[c] + dihydrofolate[c] ↔ NAD[c] + THF[c]	0	0	0	0
H[c] + NADPH[c] + dihydrofolate[c] ↔ NADP[c] + THF[c]	1,12476352	1,615142614	1,616206232	1,428667772
ATP[c] + THF[c] + formate[c] ↔ 10-formyl-THF[c] + ADP[c] + Pi[c]	5,246567725	7,533988218	7,538949566	-39,75245891
10-formyl-THF[c] + H2O[c] + NADP[c] → CO2[c] + H[c] + NADPH[c] + THF[c]	66,99398915	96,20230815	96,26566013	-
10-formyl-THF[c] + H[c] ↔ 5,10-methenyl-THF[c] + H2O[c]	-73,61262865	-105,7065697	-105,7761805	-132,1810521
5,10-methenyl-THF[c] + NADPH[c] ↔ 5,10-methylene-THF[c] + NADP[c]	-1997,936164	-387,9434655	-490,9736465	-529,6594979
5,10-methenyl-THF[c] + NADH[c] ↔ 5,10-methylene-THF[c] + NAD[c]	1924,323535	282,2368958	385,197466	397,4784458
acetoacetyl-CoA[c] + CoA[c] ↔ 2 acetyl-CoA[c]	-31,47940635	-45,20392931	-45,2336974	-39,98495018
acetoacetyl-CoA[c] + acetyl-CoA[c] + H2O[c] → CoA[c] + HMG-CoA[c]	31,47940635	45,20392931	45,2336974	39,98495018
2 H[c] + HMG-CoA[c] + 2 NADPH[c] → 2 NADP[c] + CoA[c] + R-mevalonate[c]	31,47940635	45,20392931	45,2336974	39,98495018
R-mevalonate[c] + ATP[c] → ADP[c] + R-5-phosphomevalonate[c]	31,47940635	45,20392931	45,2336974	39,98495018
R-5-diphosphomevalonate[c] + ADP[c] ↔ ATP[c] + R-5-phosphomevalonate[c]	-31,47940635	-45,20392931	-45,2336974	-39,98495018
isopentenyl-pPP → dimethylallyl-PP[c]	10,49313545	15,06797644	15,07789913	13,32831673
R-5-diphosphomevalonate[c] + ATP[c] → ADP[c] + CO2[c] + isopentenyl-pPP[c] + Pi[c]	31,47940635	45,20392931	45,2336974	39,98495018
dimethylallyl-PP[c] + isopentenyl-pPP[c] → geranyl-PP[c] + Ppi[c]	10,49313545	15,06797644	15,07789913	13,32831673
geranyl-PP[c] + isopentenyl-pPP[c] → farnesyl-PP[c] + Ppi[c]	10,49313545	15,06797644	15,07789913	13,32831673
2 farnesyl-PP[c] ↔ Ppi[c] + presqualene-PP[c]	5,246567725	7,533988218	7,538949566	6,664158364
H[c] + NADPH[c] + presqualene-PP[c] → NADP[c] + Ppi[c] + squalene[c]	5,246567725	7,533988218	7,538949566	6,664158364
H[c] + NADPH[c] + O2[c] + squalene[c] → H2O[c] + NADP[c] + squalene-2,3-oxide[c]	5,246567725	7,533988218	7,538949566	6,664158364
squalene-2,3-oxide[c] → lanosterol[c]	5,246567725	7,533988218	7,538949566	6,664158364
H[c] + NADPH[c] + lanosterol[c] + O2[c] → 4,4-dimethyl-14alpha-hydroxymethyl-5alpha-cholesta-8,24-dien-3beta-ol[c] + H2O[c] + NADP[c]	5,246567725	7,533988218	7,538949566	6,664158364
H[c] + NADPH[c] + 4,4-dimethyl-14alpha-hydroxymethyl-5alpha-cholesta-8,24-dien-3beta-ol[c] + O2[c] → 4,4-dimethyl-14alpha-formyl-5alpha-cholesta-8,24-dien-3beta-ol[c] + 2 H2O[c] + NADP[c]	5,246567725	7,533988218	7,538949566	6,664158364
H[c] + NADPH[c] + 4,4-dimethyl-14alpha-formyl-5alpha-cholesta-8,24-dien-3beta-ol[c] + O2[c] → 4,4-dimethyl-5alpha-cholesta-8,14,24-trien-3beta-ol[c] + H2O[c] + NADP[c] + formate[c]	5,246567725	7,533988218	7,538949566	6,664158364
4,4-dimethyl-5alpha-cholesta-8,14,24-trien-3beta-ol[c] + H[c] + NADPH[c] → 14-demethyllanosterol[c] + NADP[c]	5,246567725	7,533988218	7,538949566	6,664158364
14-demethyllanosterol[c] + H[c] + NADPH[c] + O2[c] → 4-alpha-hydroxymethyl-4beta-methyl-5alpha-cholesta-8,24-dien-3beta-ol[c] + H2O[c] + NADP[c]	5,246567725	7,533988218	7,538949566	6,664158364

Continued on next page

Reactions	HMR model	Liver Cancer model	Breast Cancer model	Lung Cancer model
4-alpha-hydroxymethyl-4beta-methyl-5alpha-cholesta-8,24-dien-3beta-ol[c] + H[c] + NADPH[c] + O2[c] → 4-alpha-formyl-4beta-methyl-5alpha-cholesta-8,24-dien-3beta-ol[c] + 2 H2O[c] + NADP[c]	5,246567725	7,533988218	7,538949566	6,664158364
4-alpha-formyl-4beta-methyl-5alpha-cholesta-8,24-dien-3beta-ol[c] + H[c] + NADPH[c] + O2[c] → 4-alpha-carboxy-4beta-methyl-5alpha-cholesta-8,24-dien-3beta-ol[c] + H2O[c] + NADP[c]	5,246567725	7,533988218	7,538949566	6,664158364
4-alpha-carboxy-4beta-methyl-5alpha-cholesta-8,24-dien-3beta-ol[c] + NAD[c] → 3-keto-4-methylzymosterol[c] + CO2[c] + H[c] + NADH[c]	0	0	0	0
4-alpha-carboxy-4beta-methyl-5alpha-cholesta-8,24-dien-3beta-ol[c] + NADP[c] → 3-keto-4-methylzymosterol[c] + CO2[c] + H[c] + NADPH[c]	5,246567725	7,533988218	7,538949566	6,664158364
3-keto-4-methylzymosterol[c] + 3 H[c] + NADP[c] → 4-alpha-methylzymosterol[c] + NADPH[c]	5,246567725	7,533988218	7,538949566	6,664158364
4-alpha-methylzymosterol[c] + H[c] + NADPH[c] + O2[c] → 4-alpha-hydroxymethyl-5alpha-cholesta-8,24-dien-3beta-ol[c] + H2O[c] + NADP[c]	5,246567725	7,533988218	7,538949566	6,664158364
4-alpha-hydroxymethyl-5alpha-cholesta-8,24-dien-3beta-ol[c] + H[c] + NADPH[c] + O2[c] → 4-alpha-formyl-5alpha-cholesta-8,24-dien-3beta-ol[c] + 2 H2O[c] + NADP[c]	5,246567725	7,533988218	7,538949566	6,664158364
4-alpha-formyl-5alpha-cholesta-8,24-dien-3beta-ol[c] + H[c] + NADPH[c] + O2[c] → 4-alpha-carboxy-5alpha-cholesta-8,24-dien-3beta-ol[c] + H2O[c] + NADP[c]	5,246567725	7,533988218	7,538949566	6,664158364
4-alpha-carboxy-5alpha-cholesta-8,24-dien-3beta-ol[c] + NADP[c] → 5-alpha-cholesta-8,24-dien-3-one[c] + CO2[c] + H[c] + NADPH[c]	5,246567725	7,533988218	7,538949566	6,664158364
5-alpha-cholesta-8,24-dien-3-one[c] + H[c] + NADPH[c] → NADP[c] + zymosterol[c]	5,246567725	7,533988218	7,538949566	6,664158364
zymosterol[c] → 5-alpha-cholesta-7,24-dien-3beta-ol[c]	5,246567725	7,533988218	7,538949566	6,664158364
5-alpha-cholesta-7,24-dien-3beta-ol[c] + H[c] + NADPH[c] + O2[c] → 7-dehydrodesmosterol[c] + 2 H2O[c] + NADP[c]	5,246567725	7,533988218	7,538949566	6,664158364
7-dehydrodesmosterol[c] + H[c] + NADPH[c] → desmosterol[c] + NADP[c]	5,246567725	7,533988218	7,538949566	6,664158364
desmosterol[c] + H[c] + NADPH[c] → cholesterol[c] + NADP[c]	5,246567725	7,533988218	7,538949566	6,664158364
ATP[c] + acetyl-CoA[c] + HCO3-[c] + H[c] → ADP[c] + Pi[c] + malonyl-CoA[c]	39,1222094	56,17887353	56,21586894	49,69279206
acetyl-CoA[c] + ACP[c] → acetyl-ACP[c] + CoA[c]	5,588887057	8,025553362	8,03083842	7,098970294
malonyl-CoA[c] + ACP[c] → malonyl-ACP[c] + CoA[c]	39,1222094	56,17887353	56,21586894	49,69279206
acetyl-ACP[c] + malonyl-ACP[c] → ACP[c] + acetoacetyl-ACP[c] + CoA[c]	5,588887057	8,025553362	8,03083842	7,098970294
acetoacetyl-ACP[c] + NADPH[c] + H[c] → NADP[c] + R-3-hydroxybutanoyl-ACP[c]	5,588887057	8,025553362	8,03083842	7,098970294
R-3-hydroxybutanoyl-ACP[c] → but-2-enoyl-ACP[c] + H2O[c]	5,588887057	8,025553362	8,03083842	7,098970294
but-2-enoyl-ACP[c] + H[c] + NADPH[c] → NADP[c] + butyryl-ACP[c]	5,588887057	8,025553362	8,03083842	7,098970294
butyryl-ACP[c] + malonyl-ACP[c] → ACP[c] + 3-oxohexanoyl-ACP[c] + CO2[c]	5,588887057	8,025553362	8,03083842	7,098970294
3-oxohexanoyl-ACP[c] + H[c] + NADPH[c] → NADP[c] + D-3-hydroxyhexanoyl-ACP[c]	5,588887057	8,025553362	8,03083842	7,098970294
D-3-hydroxyhexanoyl-ACP[c] → 2E-hexenoyl-ACP[c] + H2O[c]	5,588887057	8,025553362	8,03083842	7,098970294
2E-hexenoyl-ACP[c] + H[c] + NADPH[c] → hexanoyl-ACP[c] + NADP[c]	5,588887057	8,025553362	8,03083842	7,098970294
hexanoyl-ACP[c] + malonyl-ACP[c] → ACP[c] + 3-oxooctanoyl-ACP[c] + CO2[c]	5,588887057	8,025553362	8,03083842	7,098970294
3-oxooctanoyl-ACP[c] + H[c] + NADPH[c] → R-3-hydroxyoctanoyl-ACP[c] + NADP[c]	5,588887057	8,025553362	8,03083842	7,098970294
R-3-hydroxyoctanoyl-ACP[c] → 2E-octenoyl-ACP[c] + H2O[c]	5,588887057	8,025553362	8,03083842	7,098970294
2E-octenoyl-ACP[c] + H[c] + NADPH[c] → NADP[c] + octanoyl-ACP[c]	5,588887057	8,025553362	8,03083842	7,098970294
malonyl-ACP[c] + octanoyl-ACP[c] → ACP[c] + 3-oxodecanoyl-ACP[c] + CO2[c]	5,588887057	8,025553362	8,03083842	7,098970294
3-oxodecanoyl-ACP[c] + H[c] + NADPH[c] → NADP[c] + R-3-hydroxydecanoyl-ACP[c]	5,588887057	8,025553362	8,03083842	7,098970294
R-3-hydroxydecanoyl-ACP[c] → 2E-decenoyl-ACP[c] + H2O[c]	5,588887057	8,025553362	8,03083842	7,098970294
2E-decenoyl-ACP[c] + H[c] + NADPH[c] → NADP[c] + decanoyl-ACP[c]	5,588887057	8,025553362	8,03083842	7,098970294
decanoyl-ACP[c] + malonyl-ACP[c] → ACP[c] + 3-oxododecanoyl-ACP[c] + CO2[c]	5,588887057	8,025553362	8,03083842	7,098970294

Continued on next page

Reactions	HMR model	Liver Cancer model	Breast Cancer model	Lung Cancer model
3-oxododecanoyl-ACP[c] + H[c] + NADPH[c] → NADP[c] + D-3-hydroxydodecanoyl-ACP[c]	5,588887057	8,025553362	8,03083842	7,098970294
D-3-hydroxydodecanoyl-ACP[c] → 2E-dodecenoyl-ACP[c] + H2O[c]	5,588887057	8,025553362	8,03083842	7,098970294
2E-dodecenoyl-ACP[c] + H[c] + NADPH[c] → NADP[c] + dodecanoyl-ACP[c]	5,588887057	8,025553362	8,03083842	7,098970294
dodecanoyl-ACP[c] + malonyl-ACP[c] → ACP[c] + 3-oxotetradecanoyl-ACP[c] + CO2[c]	5,588887057	8,025553362	8,03083842	7,098970294
3-oxotetradecanoyl-ACP[c] + H[c] + NADPH[c] → NADP[c] + HMA[c]	5,588887057	8,025553362	8,03083842	7,098970294
HMA[c] → 2E-tetradecenoyl-ACP[c] + H2O[c]	5,588887057	8,025553362	8,03083842	7,098970294
2E-tetradecenoyl-ACP[c] + NADPH[c] + H[c] → NADP[c] + tetradecanoyl-ACP[c]	5,588887057	8,025553362	8,03083842	7,098970294
tetradecanoyl-ACP[c] + malonyl-ACP[c] → ACP[c] + 3-oxohexadecanoyl-ACP[c] + CO2[c]	5,588887057	8,025553362	8,03083842	7,098970294
3-oxohexadecanoyl-ACP[c] + H[c] + NADPH[c] → NADP[c] + R-3-hydroxypalmitoyl-ACP[c]	5,588887057	8,025553362	8,03083842	7,098970294
R-3-hydroxypalmitoyl-ACP[c] → 2E-hexadecenoyl-ACP[c] + H2O[c]	5,588887057	8,025553362	8,03083842	7,098970294
2E-hexadecenoyl-ACP[c] + H[c] + NADPH[c] → hexadecanoyl-ACP[c] + NADP[c]	5,588887057	8,025553362	8,03083842	7,098970294
H2O[c] + hexadecanoyl-ACP[c] → ACP[c] + palmitate[c]	5,588887057	8,025553362	8,03083842	7,098970294
biomass_synthesis'	139,7221764	200,638834	200,7709605	177,4742574
Ex_glucose[s]'	-88,50025942	-367,3906121	-346,9188093	-599,797916
Ex_O2[s]'	-1000	-208,9136882	-239,1682909	-516,4350036
Ex_folate[s]'	2,00198482106296e-14		-	-
Ex_D-lactate[s]'	0	0	0	0
Ex_L-lactate[s]'	0	223,7461825	152,7124929	871,6275454
Ex_urea[s]'	0	0	-	-
Ex_biomass[s]'	139,7221764	-	-	-
Ex_cancer-biomass[s]'	-	200,638834	200,7709605	177,4742574
Ex_glutamate[s]'	0	-	-	-
Ex_riboflavin[s]'	0	0	5,55383232728890e-29	-2,83222534537321e-29
Ex_H2O[s]'	310,0952067	-195,2657261	-295,7387908	-6,472486166
Ex_H[s]'	-22,8222203	-30,13595287	-30,15579827	-26,65663346
Ex_glutamine[s]'	-706,0189519	-1000	-1000	-1000
Ex_NH3[s]'	562,979771	707,2877507	817,7963379	687,0063997
Ex_putrescine[s]'	0	268,8685564	427,5801385	0
Ex_ornithine[s]'	-	-	-427,5801385	0
Ex_ornithine[c]'	-0,458288739	-269,5266517	-	-
Ex_guanidinoacetate[c]'	0	0	-	-
Ex_mercaptopyruvate[c]'	-0,596613693	-0,856727821	-0,857292001	-0,757815079
Ex_phenylpyruvate[c]'	-4,587079052	-	-	-
Ex_dihydrobiopterin[c]'	2,752526876	-	-	-
Ex_tetrahydrobiopterin[c]'	-2,752526876	-	-	-
Ex_4-methylthio-2-oxobutanoic-acid[c]'	-0,917974699	-1,31819714	-1,31906521	-1,166005871
Ex_4-methyl-2-oxopentanoate[c]'	-5,505053751	-7,905170061	-7,910375843	-6,99248574
Ex_2-oxo-3-methylvalerate[c]'	-1,834552176	-2,634387891	-2,636122711	-2,330236999
Ex_citrate[c]'	11,33146851	0	99,68594789	93,39937742
Ex_CoA[m]'	5,588887057	-	-	-
Ex_CO2[c]'	917,5541354	-	-	-
Ex_ADP[c]'	11,95602664	17,16866503	17,17997109	131,2226911
Ex_ATP[c]'	-13,5390789	-19,44190302	-19,45470607	-94,55473484

Continued on next page

Reactions	HMR model	Liver Cancer model	Breast Cancer model	Lung Cancer model
Ex_folate[c]'	-	2,77847254161772e-14	2,76720364574265e-14	-7,87217655715256e-15
Ex_formate[c]'	-	0	-	46,41661727
Ex_glutamate[c]'	-	0	0	0
Ex_AKG[c]'	-	386,7644439	426,4710334	159,6558419
Ex_Lglu5semialdehyde[c]'	-	-12,77868734	-12,78710247	-11,30333545
Ex_Lglu5semialdehyde[m]'	-	109,9703993	-	-
Ex_CoA[c]'	-	8,025553362	8,03083842	7,098970294
Ex_serine[m]'	-	0	-4,74044086878493e-29	
Ex_glycine[m]'	-	0	4,74044086878493e-29	
Ex_10-formyl-THF[m]'	-	7,20428107894788e-29	-	-
Ex_2-oxoglutaramate[c]'	-	-	-	74,57468294

TABLE A.1: Flux distributions in reference and tumoral CMs.

Appendix B

Metabolites concentrations from the YMDB database

The table hereafter shown, illustrates the value of metabolic concentrations retrieved from the Yeast Metabolome Database (YMDB) [228]. In detail, the first column reports the name of the metabolite, while the second column indicates the interval of concentrations in a given medium (column 3) and with an associated oxigen condition (column 4). Lastly, column 5 indicates the literature reference of the retrieved data using the PubMed unique identifier.

Metabolite	Interval	Medium	Oxigen condition	Reference
Acetate	25775 ± 1289 µM	YEB media with 0.5 mM glucose	aerobic	Experimentally Determined
	18500 ± 16500 µM	Synthetic medium with 1% glucose and 0.1% yeast extract	aerobic	PMID: 16623706
Acetyl-CoA[c]	Not Available			
Acetyl-CoA[m]	Not Available			
Acetaldehyde	50 ± 50 µM	Synthetic medium with 1% glucose and 0.1% yeast extract	aerobic	PMID: 16623706 Link_out
ADP	420 ± 110 µM	20 ml 2% (wt/vol) glucose, 0.5% (wt/vol) ammonium sulfate, 0.17% (wt/vol) yeast nitrogen base without amino acids (Difco, Detroit, MI) and 100 mM potassium phthalate at pH 5.0, supplemented with required nutrients (40 mg/L uracil, 40 mg/L Ltryptophan, 60	aerobic	PMID: 11135551
	1400 ± 800 µM	Minimal medium supplemented with ammonia salts and glucose	aerobic and anaerobic;resting cells	PMID: 4578278
	950 ± 350 µM	Minimal medium supplemented with ammonia salts and (glucose or galactose)	aerobic;growing cells	PMID: 4578278
	320 ± 20 µM	Synthetic medium with 2% glucose	aerobic;growing cells	PMID: 6229402
	530 ± 100 µM	Synthetic medium with 2% glucose	aerobic;resting cells	PMID: 6229402
	530 ± 140 µM	Synthetic medium with 2% glucose	anaerobic;resting cells	PMID: 6229402
	1100 ± 300 µM	Synthetic medium with 2% galactose	aerobic;resting cells	PMID: 6229402
Alphaketoglutarate[in]	410 ± 195 µM	Synthetic medium with 20 g/L glucose	aerobic	PMID: 12584756
Alphaketoglutarate[out]	5500 ± 500 µM	Minimal medium supplemented with ammonia salts and glucose	aerobic;resting cells	PMID: 4578278
	1300 ± 0 µM	Minimal medium supplemented with ammonia salts and glucose	anaerobic;resting cells	PMID: 4578278
	10 ± 0 µM	Minimal medium supplemented with ammonia salts and galactose	aerobic;growing cells	PMID: 4578278
	2500 ± 2300 µM	Minimal medium supplemented with ammonia salts and glucose	aerobic;growing cells	PMID: 4578278
	5000 ± 200 µM	Synthetic medium with 2% glucose	aerobic;growing cells	PMID: 6229402
	210 ± 10 µM	Synthetic medium with 2% glucose	aerobic;resting cells	PMID: 6229402
	1500 ± 400 µM	Synthetic medium with 2% glucose	anaerobic;resting cells	PMID: 6229402
	3700 ± 700 µM	Synthetic medium with 2% galactose	aerobic;resting cells	PMID: 6229402
ATP	2800 ± 320 µM	20 ml 2% (wt/vol) glucose, 0.5% (wt/vol) ammonium sulfate, 0.17% (wt/vol) yeast nitrogen base without amino acids (Difco, Detroit, MI) and 100 mM potassium phthalate at pH 5.0, supplemented with required nutrients (40 mg/L uracil, 40 mg/L Ltryptophan, 60	aerobic	PMID: 11135551
	2650 ± 1750 µM	Minimal medium supplemented with ammonia salts and glucose	aerobic and anaerobic;resting cells	PMID: 4578278
	1250 ± 150 µM	Minimal medium supplemented with ammonia salts and (glucose or galactose)	aerobic;growing cells	PMID: 4578278
	1900 ± 100 µM	Synthetic medium with 2% glucose	aerobic;growing cells	PMID: 6229402
	1800 ± 100 µM	Synthetic medium with 2% glucose	aerobic;resting cells	PMID: 6229402
	1600 ± 300 µM	Synthetic medium with 2% glucose	anaerobic;resting cells	PMID: 6229402
	1500 ± 200 µM	Synthetic medium with 2% galactose	aerobic;resting cells	PMID: 6229402
Citrate	2400 ± 100 µM	Minimal medium supplemented with ammonia salts	aerobic;resting cells	PMID: 4578278
	13500 ± 9500 µM	Minimal medium supplemented with ammonia salts and glucose	aerobic and anaerobic;resting cells	PMID: 4578278

Continued on next page

Metabolite	Interval	Medium	Oxigen condition	Reference
	700 ± 0 µM	Minimal medium supplemented with ammonia salts and (glucose or galactose)	aerobic;growing cells	PMID: 4578278
	5200 ± 500 µM	Synthetic medium with 2% glucose	aerobic;growing cells	PMID: 6229402
	210 ± 10 µM	Synthetic medium with 2% glucose	aerobic;resting cells	PMID: 6229402
	1500 ± 400 µM	Synthetic medium with 2% glucose	anaerobic;resting cells	PMID: 6229402
	3700 ± 700 µM	Synthetic medium with 2% galactose	aerobic;resting cells	PMID: 6229402
Ethanol[in]	325823 ± 68594 µM	YEPD media 200g/L glucose	anaerobic;30C;12h	PMID: 3513699
	583052 ± 120040 µM	YEPD media 200g/L glucose	anaerobic;30C;24h	PMID: 3513699
	1406185 ± 291526 µM	YEPD media 200g/L glucose	anaerobic;30C;48h	PMID: 3513699
	1825 ± 91 µM	YEB media with 0.5 mM glucose	aerobic	Experimentally Determined
Ethanol[out]	868282 ± 8000 µM	hops, malted barley	anaerobic	PMID: 16448171
	291526 ± 17149 µM	YEPD media 200g/L glucose	anaerobic;30C;12h	PMID: 3513699
	943173 ± 17149 µM	YEPD media 200g/L glucose	anaerobic;30C;24h	PMID: 3513699
	1920643 ± 171486 µM	YEPD media 200g/L glucose	anaerobic;30C;48h	PMID: 3513699
	1500 ± 1000 µM	Synthetic medium with 1% glucose and 0.1% yeast extract	aerobic	PMID: 16623706
Fructose 1,6-bp	90 ± 70 µM	Minimal medium supplemented with ammonia salts	aerobic;resting cells	PMID: 4578278
	3800 ± 3200 µM	Minimal medium supplemented with ammonia salts and glucose	aerobic and anaerobic;resting cells	PMID: 4578278
	50 ± 0 µM	Minimal medium supplemented with ammonia salts and ethanol	aerobic;growing cells	PMID: 4578278
	3500 ± 1000 µM	Minimal medium supplemented with ammonia salts and (glucose or galactose)	aerobic;growing cells	PMID: 4578278
	1700 ± 10 µM	Synthetic medium with 2% glucose	aerobic;growing cells	PMID: 6229402
	410 ± 70 µM	Synthetic medium with 2% glucose	aerobic;resting cells	PMID: 6229402
	530 ± 130 µM	Synthetic medium with 2% glucose	anaerobic;resting cells	PMID: 6229402
	2700 ± 600 µM	Synthetic medium with 2% galactose	aerobic;resting cells	PMID: 6229402
FAD	Not Available			
FADH2	Not Available			
Fatty Acid 16:1	Not Available			
Fatty Acid 16:0	Not Available			
Fatty Acid 18:1	Not Available			
Fatty Acid 18:0	Not Available			
Fatty Acid 14:0	Not Available			
Fatty Acids total	Not Available			
Fumarate	Not Available			
Glucose-6P	2050 ± 110 µM	20 ml 2% (wt/vol) glucose, 0.5% (wt/vol) ammonium sulfate, 0.17% (wt/vol) yeast nitrogen base without amino acids (Difco, Detroit, MI) and 100 mM potassium phthalate at pH 5.0, supplemented with required nutrients (40 mg/L uracil, 40 mg/L Ltryptophan, 60	aerobic	PMID: 11135551
	2300 ± 200 µM	Synthetic medium with 2% glucose	aerobic;growing cells	PMID: 6229402
	560 ± 80 µM	Synthetic medium with 2% glucose	aerobic;resting cells	PMID: 6229402
	340 ± 130 µM	Synthetic medium with 2% glucose	anaerobic;resting cells	PMID: 6229402
	1360 ± 600 µM	Synthetic medium with 2% galactose	aerobic;resting cells	PMID: 6229402
Glucose[in]	1500 ± 100 µM	1% wt/vol glucose 2%wt vol yeast nitrogen base	aerobic	PMID: 9457857

Continued on next page

Metabolite	Interval	Medium	Oxigen condition	Reference
	2122 ± 106 µM	YEB media with 0.5 mM glucose	aerobic	Experimentally Determined
	1500 ± 500 µM	Synthetic medium with 1% glucose and 0.1% yeast extract	aerobic	PMID: 16623706
Glucose[out]	Not Available			
Glyoxylate	Not Available			
H2O	60.4 (±0.2) %			Bionumbers [1]
Isocitrate	Not Available			
Malate	2515 ± 126 µM	YEB media with 0.5 mM glucose	aerobic	Experimentally Determined
dCTP	18 ± 0 µM	YEPD medium	aerobic	PMID: 14573610
dATP	44 ± 0 µM	YEPD medium	aerobic	PMID: 14573610
dGTP	15 ± 0 µM	YEPD medium	aerobic	PMID: 14573610
dTTP	70 ± 0 µM	YEPD medium	aerobic	PMID: 14573610
NAD	950 ± 150 µM	Minimal medium supplemented with ammonia salts and glucose	aerobic and anaerobic;resting cells	PMID: 4578278
	1300 ± 300 µM	Minimal medium supplemented with ammonia salts and (glucose or galactose)	aerobic;growing cells	PMID: 4578278
NADH	70 ± 0 µM	Minimal medium supplemented with ammonia salts	aerobic;resting cells	PMID: 4578278
	250 ± 0 µM	Minimal medium supplemented with ammonia salts and glucose	aerobic;resting cells	PMID: 4578278
	500 ± 0 µM	Minimal medium supplemented with ammonia salts and glucose	anaerobic;resting cells	PMID: 4578278
	1025 ± 775 µM	Minimal medium supplemented with ammonia salts and (glucose or galactose)	aerobic;growing cells	PMID: 4578278
NADP	325 ± 305 µM	Minimal medium supplemented with ammonia salts and glucose	aerobic;resting cells	PMID: 4578278
	85 ± 65 µM	Minimal medium supplemented with ammonia salts and (glucose or galactose)	aerobic;growing cells	PMID: 4578278
NADPH	100 ± 0 µM	Minimal medium supplemented with ammonia salts and glucose	aerobic;resting cells	PMID: 4578278
	100 ± 50 µM	Minimal medium supplemented with ammonia salts and (glucose or galactose)	aerobic;growing cells	PMID: 4578278
O2[in]	Not Available			
O2[out]	Not Available			
Oxaloacetate	25 ± 25 µM	Minimal medium supplemented with ammonia salts and (glucose or galactose)	aerobic;growing cells	PMID: 4578278
Phosphoenolpyruvate	15 ± 15 µM	Synthetic medium with 2% glucose	aerobic;growing cells	PMID: 6229402
Pyruvate	3390 ± 540 µM	20 ml 2% (wt/vol) glucose, 0.5% (wt/vol) ammonium sulfate, 0.17% (wt/vol) yeast nitrogen base without amino acids (Difco, Detroit, MI) and 100 mM potassium phthalate at pH 5.0, supplemented with required nutrients (40 mg/L uracil, 40 mg/L Ltryptophan, 60	aerobic	PMID: 11135551
	3380 ± 169 µM	YEB media with 0.5 mM glucose	aerobic	Experimentally Determined
	130 ± 0 µM	Minimal medium supplemented with ammonia salts	aerobic;resting cells	PMID: 4578278
	140 ± 80 µM	Minimal medium supplemented with ammonia salts and glucose	aerobic and anaerobic;resting cells	PMID: 4578278
	5250 ± 4750 µM	Minimal medium supplemented with ammonia salts and (glucose or galactose)	aerobic;growing cells	PMID: 4578278
	1600 ± 10 µM	Synthetic medium with 2% glucose	aerobic;growing cells	PMID: 6229402
	440 ± 40 µM	Synthetic medium with 2% glucose	aerobic;resting cells	PMID: 6229402
	340 ± 110 µM	Synthetic medium with 2% glucose	anaerobic;resting cells	PMID: 6229402
	1300 ± 300 µM	Synthetic medium with 2% galactose	aerobic;resting cells	PMID: 6229402

Continued on next page

[1] http://bionumbers.hms.harvard.edu//bionumber.aspx?id=103689&ver=11

Metabolite	Interval	Medium	Oxigen condition	Reference
Ribose-5P	Not Available			
Succinate	600 ± 30 µM	YEB media with 0.5 mM glucose	aerobic	Experimentally Determined
Succynil-CoA	Not Available			
TrioseP DHAP	330 ± 10 µM	Synthetic medium with 2% glucose	aerobic;growing cells	PMID: 6229402
	120 ± 10 µM	Synthetic medium with 2% glucose	aerobic;resting cells	PMID: 6229402
	460 ± 70 µM	Synthetic medium with 2% galactose	aerobic;resting cells	PMID: 6229402
TrioseP Gly3P	46 ± 4 µM	Synthetic medium with 2% glucose	aerobic;resting cells	PMID: 6229402
	100 ± 13 µM	Synthetic medium with 2% galactose	aerobic;resting cells	PMID: 6229402

TABLE B.1: Metabolites concentrations from the YMDB database.

Bibliography

[1] M. Newman, "Complex systems: A survey," *Am J Phys*, vol. 79, no. arXiv: 1112.1440, pp. 800–810, 2011.

[2] Y. Bar-Yam, *Dynamics of complex systems*, vol. 213. Addison-Wesley Reading, MA, 1997.

[3] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Phys Rep*, vol. 424, no. 4, pp. 175–308, 2006.

[4] C. Gershenson, *Complexity: 5 Questions*. Copenhagen, Denmark, Denmark: Automatic Press /VIP, 2008.

[5] Y. Bar-Yam, "Introducing complex systems," in *International Conference on Complex Systems, Nashua, NH*, 2001.

[6] Y. Bar-Yam, "Multiscale variety in complex systems," *Complexity*, vol. 9, no. 4, pp. 37–45, 2004.

[7] Y. Bar-Yam, "Multiscale complexity/entropy," *Adv Complex Syst*, vol. 7, no. 01, pp. 47–63, 2004.

[8] G. Nicolis and C. Nicolis, *Foundations of Complex Systems: Emergence, Information and Predicition*. World Scientific, 2012.

[9] K. Saetzler, C. Sonnenschein, and A. M. Soto, "Systems biology beyond networks: generating order from disorder through self-organization," in *Seminars in cancer biology*, vol. 21, pp. 165–174, Elsevier, 2011.

[10] L. Chong and L. Ray, "Whole-istic Biology.," *Science*, vol. 295, no. 5560, p. 1661, 2002.

[11] L. Alberghina and H. V. Westerhoff, *Systems Biology: Definitions and Perspectives*. Topics in Current Genetics, Springer, 2005.

[12] H. Kitano, "Systems Biology: a brief overview," *Science*, vol. 295, no. 5560, pp. 1662–1664, 2002.

[13] N. J. Eungdamrong and R. Iyengar, "Computational approaches for modeling regulatory cellular networks," *Trends Cell Biol*, vol. 14, no. 12, pp. 661–669, 2004.

[14] T. Meng, S. Somani, and P. Dhar, "Modelling and simulation of biological systems with stochasticity," *In Silico Biol*, vol. 4, no. 3, pp. 293–309, 2004.

[15] I. Chou and E. Voit, "Recent developments in parameter estimation and structure identification of biochemical and genomic systems," *Math Biosci*, vol. 219, pp. 57–83, Mar. 2009.

[16] J. Stelling, "Mathematical models in microbial systems biology," *Curr Opin Microbiol*, vol. 7, no. 5, pp. 513–8, 2004.

[17] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, 2000.

[18] A. L. Hopkins, "Network pharmacology: the next paradigm in drug discovery," *Nat Chem Biol*, vol. 4, no. 11, pp. 682–690, 2008.

[19] M. Ashyraliyev, Y. Fomekong-Nanfack, J. A. Kaandorp, and J. G. Blom, "Systems biology: parameter estimation for biochemical models," *Febs Journal*, vol. 276, no. 4, pp. 886–902, 2009.

[20] B. Bollobás, *Modern Graph Theory*, vol. 184. New York, NY, USA: Springer, 1998.

[21] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *Nature*, vol. 406, no. 6794, pp. 378–382, 2000.

[22] R. Montañez, M. A. Medina, R. V. Solé, and C. Rodríguez-Caso, "When metabolism meets topology: Reconciling metabolite and reaction networks," *Bioessays*, vol. 32, no. 3, pp. 246–256, 2010.

[23] J. R. Karr, J. C. Sanghvi, D. N. Macklin, M. V. Gutschow, J. M. Jacobs, B. B. Jr., N. Assad-Garcia, J. I. Glass, and M. W. Covert, "A whole-cell computational model predicts phenotype from genotype," *Cell*, vol. 150, no. 2, pp. 389–401, 2012.

[24] J. D. Orth, I. Thiele, and B. Ø. Palsson, "What is flux balance analysis?," *Nat Biotechnol*, vol. 28, no. 3, pp. 245–8, 2010.

[25] N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. Ø. Palsson, "Global reconstruction of the human metabolic network based on genomic and bibliomic data," *PNAS*, vol. 104, no. 6, pp. 1777–82, 2007.

[26] I. Thiele, N. Swainston, R. M. T. Fleming, A. Hoppe, S. Sahoo, M. K. Aurich, H. Haraldsdottir, M. L. Mo, O. Rolfsson, M. D. Stobbe, S. G. Thorleifsson, R. Agren, C. Bölling, S. Bordel, A. K. Chavali, P. Dobson, W. B. Dunn, L. Endler, D. Hala, M. Hucka, D. Hull, D. Jameson, N. Jamshidi, J. J. Jonsson, N. Juty, S. Keating, I. Nookaew, N. Le Novère, N. Malys, A. Mazein, J. a. Papin, N. D. Price, E. Selkov, M. I. Sigurdsson, E. Simeonidis, N. Sonnenschein, K. Smallbone, A. Sorokin, J. H. G. M. van Beek, D. Weichart, I. Goryanin, J. Nielsen, H. V. Westerhoff, D. B. Kell, P. Mendes, and B. Ø. Palsson, "A community-driven global reconstruction of human metabolism," *Nat Biotechnol*, Mar. 2013.

[27] F. Amara, R. Colombo, P. Cazzaniga, D. Pescini, A. Csikász-Nagy, M. Muzi-Falconi, D. Besozzi, and P. Plevani, "In vivo and in silico analysis of PCNA ubiquitylation in the activation of the post replication repair pathway in *S. cerevisiae*," *BMC Syst Biol*, vol. 7, no. 1, p. 24, 2013.

[28] P. Cazzaniga, C. Damiani, D. Besozzi, R. Colombo, M. S. Nobile, D. Gaglio, D. Pescini, S. Molinari, G. Mauri, L. Alberghina, and M. Vanoni, "Computational strategies for a system-level understanding of metabolism," *Metabolites*, vol. 4, no. 4, pp. 1034–87, 2014.

[29] M. Giannattasio, C. Follonier, H. Tourrière, F. Puddu, F. Lazzaro, P. Pasero, M. Lopes, P. Plevani, and M. Muzi-Falconi, "Exo1 competes with repair synthesis, converts NER intermediates to long ssDNA gaps, and promotes checkpoint activation," *Mol Cell*, vol. 40, no. 1, pp. 50–62, 2010.

[30] A. Aboussekhra and I. Al-Sharif, "Homologous recombination is involved in transcription-coupled repair of UV damage in *Saccharomyces cerevisiae*," *EMBO J*, vol. 24, no. 11, pp. 1999–2010, 2005.

[31] Y. Cao, D. T. Gillespie, and L. Petzold, "Efficient step size selection for the tau-leaping simulation method," *J Chem Phys*, vol. 124, no. 4, p. 044109, 2006.

[32] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *J Phys Chem*, vol. 81, no. 25, pp. 2340–2361, 1977.

[33] D. Besozzi, P. Cazzaniga, G. Mauri, and D. Pescini, "BioSimWare: a software for the modeling, simulation and analysis of biological systems," in *Membrane Computing, 11th International Conference, CMC 2010, Jena, Germany, August 24-27, Revised selected papers* (M. Gheorghe, T. Hinze, G. Păun, G. Rozenberg, and A. Salomaa, eds.), pp. 119–143, LNCS 6501, 2010.

[34] M. Morris, "Factorial sampling plans for preliminary computational experiments," *Technometrics*, vol. 33, no. 2, pp. 161–174, 1991.

[35] F. Campolongo, J. Cariboni, and A. Saltelli, "An effective screening design for sensitivity analysis of large models," *Environ Modell Softw*, vol. 22, no. 10, pp. 1509–1518, 2007.

[36] F. Campolongo, A. Saltelli, and J. Cariboni, "From screening to quantitative sensitivity analysis. A unified approach," *Comput Phys Commun*, vol. 182, no. 4, pp. 978–988, 2011.

[37] I. Sobol', "On the distribution of points in a cube and the approximate evaluation of integrals," *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, vol. 7, no. 4, pp. 784–802, 1967.

[38] L. Petzold, "Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations," *SIAM J Sci Stat Comp*, vol. 4, no. 1, pp. 136–148, 1983.

[39] O. Wolkenhauer, M. Ullah, W. Kolch, and C. Kwang-Hyun, "Modeling and simulation of intracellular dynamics: choosing an appropriate framework," *IEEE T Nanobiosci*, vol. 3, no. 3, pp. 200–207, 2004.

[40] A. Degasperi and S. Gilmore, "Sensitivity analysis of stochastic models of bistable biochemical reactions," in *Formal Methods for Computational Systems Biology (SFM-08:Bio)* (M. Bernardo, P. Degano, and G. Zavattaro, eds.), LNCS 5016, (Berlin, Heidelberg), pp. 1–20, 2008.

[41] J. Zhao, H. Yu, J. Luo, Z. Cao, and Y. Li, "Complex networks theory for analyzing metabolic networks," *Chinese Sci Bull*, vol. 51, no. 13, pp. 1529–1537, 2006.

[42] A. M. Feist, M. J. Herrgård, I. Thiele, J. L. Reed, and B. Ø. Palsson, "Reconstruction of biochemical networks in microorganisms," *Nat Rev Microbiol*, vol. 7, no. 2, pp. 129–143, 2008.

[43] D. M. Hendrickx, M. M. W. B. Hendriks, P. H. C. Eilers, A. K. Smilde, and H. C. J. Hoefsloot, "Reverse engineering of metabolic networks, a critical assessment," *Mol Biosyst*, vol. 7, no. 2, pp. 511–520, 2011.

[44] P. D. Karp, M. Riley, S. M. Paley, and A. Pelligrini-Toole, "Ecocyc: an encyclopedia of *Escherichia coli* genes and metabolism," *Nucleic Acids Res*, vol. 24, no. 1, pp. 32–39, 1996.

[45] P. D. Karp, C. A. Ouzounis, and S. M. Paley, "Hincyc: a knowledge base of the complete genome and metabolic pathways of *H. influenzae*," in *Proc of International Conference on Intelligent Systems for Molecular Biology (ISMB)*, vol. 4, pp. 116–124, 1996.

[46] M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Res*, vol. 28, no. 1, pp. 27–30, 2000.

[47] D. De Martino, F. Capuani, M. Mori, A. De Martino, and E. Marinari, "Counting and correcting thermodynamically infeasible flux cycles in genome-scale metabolic networks," *Metabolites*, vol. 3, no. 4, pp. 946–966, 2013.

[48] A. Sims and B. Folkes, "A kinetic study of the assimilation of 15N-ammonia and the synthesis of amino acids in an exponentially growing culture of *Candida utilis*," *P Roy Soc B-Biol Sci*, vol. 159, no. 976, pp. 479–502, 1964.

[49] S. Danø, P. G. Sørensen, and F. Hynne, "Sustained oscillations in living cells," *Nature*, vol. 402, no. 6759, pp. 320–322, 1999.

[50] F. Achcar, E. J. Kerkhoven, The SilicoTryp Consortium, B. M. Bakker, M. P. Barrett, and R. Breitling, "Dynamic modelling under uncertainty: The case of *Trypanosoma brucei* energy metabolism," *PLoS Comput Biol*, vol. 8, p. e1002352, January 2012.

[51] R. Agren, J. M. Otero, and J. Nielsen, "Genome-scale modeling enables metabolic engineering of *Saccharomyces cerevisiae* for succinic acid production," *J Ind Microbiol Biotechnol*, vol. 40, no. 7, pp. 735–747, 2013.

[52] M. R. Andersen, M. L. Nielsen, and J. Nielsen, "Metabolic model integration of the bibliome, genome, metabolome and reactome of *Aspergillus niger*," *Mol Syst Biol*, vol. 4, no. 1, 2008.

[53] J. N. Bazil, G. T. Buzzard, and A. E. Rundell, "Modeling mitochondrial bioenergetics with integrated volume dynamics," *PLoS Comput Biol*, vol. 6, no. 1, p. e1000632, 2010.

[54] D. A. Beard, "A biophysical model of the mitochondrial respiratory system and oxidative phosphorylation," *PLoS Comput Biol*, vol. 1, no. 4, p. e36, 2005.

[55] I. Chang, M. Heiske, T. Letellier, D. Wallace, and P. Baldi, "Modeling of mitochondria bioenergetics using a composable chemiosmotic energy transduction rate law: theory and experimental validation," *PLoS ONE*, vol. 6, p. e14820, September 2011.

[56] M. Cloutier and P. Wellstead, "The control systems structures of energy metabolism," *J R Soc Interface*, vol. 7, pp. 651–665, 2010.

[57] M. E. Csete and J. C. Doyle, "Bow ties, metabolism and disease," *Trends Biotechnol*, vol. 22, no. 9, pp. 446–450, 2004.

[58] C. Damiani, D. Pescini, R. Colombo, S. Molinari, L. Alberghina, M. Vanoni, and G. Mauri, "An ensemble evolutionary constraint-based approach to understand the emergence of metabolic phenotypes," *Nat Comput*, vol. 13, no. 3, pp. 321–331, 2014.

[59] S. K. Dasika, S. T. Kinsey, and B. R. Locke, "Facilitated diffusion of myoglobin and creatine kinase and reaction–diffusion constraints of aerobic metabolism under steady-state conditions in skeletal muscle," *Biotechnol Bioeng*, vol. 109, no. 2, pp. 545–558, 2012.

[60] D. Degenring, C. Froemel, G. Dikta, and R. Takors, "Sensitivity analysis for the reduction of complex metabolism models," *J Process Contr*, vol. 14, no. 7, pp. 729 – 745, 2004.

[61] A. Dräger, M. Kronfeld, M. J. Ziller, J. Supper, H. Planatscher, and J. B. Magnus, "Modeling metabolic networks in *C. glutamicum*: a comparison of rate laws in combination with various parameter optimization strategies," *BMC Syst Biol*, vol. 3, no. 5, 2009.

[62] J. Edwards and B. Ø. Palsson, "The *Escherichia coli* mg1655 in silico metabolic genotype: its definition, characteristics, and capabilities.," *PNAS*, vol. 97, no. 10, pp. 5528–5533, 2000.

[63] J. S. Edwards and B. Ø. Palsson, "Systems properties of the *Haemophilus influenzae* Rd metabolic genotype," *J Biol Chem*, vol. 274, no. 25, pp. 17410–17416, 1999.

[64] S. Federowicz, D. Kim, A. Ebrahim, J. Lerman, H. Nagarajan, B.-k. Cho, K. Zengler, and B. Ø. Palsson, "Determining the control circuitry of redox metabolism at the genome-scale," *PLoS Genet*, vol. 10, no. 4, p. e1004264, 2014.

[65] D. A. Fell and A. Wagner, "The small world of metabolism," *Nat Biotechnol*, vol. 18, no. 11, pp. 1121–1122, 2000.

[66] O. Folger, L. Jerby, C. Frezza, E. Gottlieb, E. Ruppin, and T. Shlomi, "Predicting selective drug targets in cancer through metabolic networks," *Mol Syst Biol*, vol. 7, p. 501, 2011.

[67] C. Gille, C. Bölling, A. Hoppe, S. Bulik, S. Hoffmann, K. Hübner, A. Karlstädt, R. Ganeshan, M. König, K. Rother, and *et al.*, "Hepatonet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology," *Mol Syst Biol*, vol. 6, no. 1, 2010.

[68] F. Guillaud, S. Dröse, A. Kowald, U. Brandt, and E. Klipp, "Superoxide production by cytochrome bc_1 complex: A mathematical model," *BBA-Bioenergetics*, vol. 1837, pp. 1643–1652, 2014.

[69] T. Hao, H.-W. Ma, X.-M. Zhao, and I. Goryanin, "Compartmentalization of the Edinburgh Human Metabolic Network," *BMC Bioinformatics*, vol. 11, p. 393, 2010.

[70] B. D. Heavner, K. Smallbone, N. D. Price, and L. P. Walker, "Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance," *Database-Oxford*, vol. 2013, p. 1, 2013.

[71] M. J. Herrgård, N. Swainston, P. Dobson, W. B. Dunn, K. Y. Arga, M. Arvas, N. Blüthgen, S. Borger, R. Costenoble, M. Heinemann, and *et al.*, "A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology," *Nat Biotechnol*, vol. 26, no. 10, pp. 1155–60, 2008.

[72] I. Kareva, "Prisoner's dilemma in cancer metabolism," *PLoS ONE*, vol. 6, no. 12, p. e28576, 2011.

[73] E. J. Kerkhoven, F. Achcar, V. P. Alibu, R. J. Burchmore, I. H. Gilbert, M. Trybiło, N. N. Driessen, D. Gilbert, R. Breitling, B. M. Bakker, and M. P. Barrett, "Handling uncertainty in dynamic models: the pentose phosphate pathway in *Trypanosoma brucei*," *PLoS Comput Biol*, vol. 9, no. 12, p. e1003371, 2013.

[74] T. Khazaei, A. McGuigan, and R. Mahadevan, "Ensemble modeling of cancer metabolism," *Front Physiol*, vol. 3, p. 135, 2012.

[75] H. Knoop, M. Gründel, Y. Zilliges, R. Lehmann, S. Hoffmann, W. Lockau, and R. Steuer, "Flux balance analysis of cyanobacterial metabolism: the metabolic network of *Synechocystis* sp. pcc 6803," *PLoS Comput Biol*, vol. 9, no. 6, p. e1003081, 2013.

[76] M. König, S. Bulik, and H. G. Holzhütter, "Quantifying the contribution of the liver to glucose homeostasis: a detailed kinetic model of human hepatic glucose metabolism.," *PLoS Comput Biol*, vol. 8, no. 6, p. e1002577, 2012.

[77] E. Metelkin, I. Goryanin, and O. Demin, "Mathematical modeling of mitochondrial adenine nucleotide translocase," *Biophys J*, vol. 90, no. 2, pp. 423–432, 2006.

[78] M. K. Monaco, T. Z. Sen, P. D. Dharmawardhana, L. Ren, M. Schaeffer, S. Naithani, V. Amarasinghe, J. Thomason, L. Harper, J. Gardiner, and *et al.*, "Maize metabolic network construction and transcriptome analysis," *Plant Genome*, vol. 6, no. 1, 2013.

[79] M. Oshiro, H. Shinto, Y. Tashiro, N. Miwa, T. Sekiguchi, M. Okamoto, A. Ishizaki, and K. Sonomoto, "Kinetic modeling and sensitivity analysis of xylose metabolism in *Lactococcus lactis* io-1," *J Biosci Bioeng*, vol. 108, pp. 376–384, Nov 2009.

[80] T. Österlund, I. Nookaew, S. Bordel, and J. Nielsen, "Mapping condition-dependent regulation of metabolism in yeast through genome-scale modeling," *BMC Syst Biol*, vol. 7, no. 1, p. 36, 2013.

[81] J. M. Otero, D. Cimini, K. R. Patil, S. G. Poulsen, L. Olsson, and J. Nielsen, "Industrial systems biology of *Saccharomyces cerevisiae* enables novel succinic acid cell factory," *PLoS ONE*, vol. 8, no. 1, p. e54144, 2013.

[82] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[83] M. C. Reed, R. L. Thomas, J. Pavisic, S. J. James, C. M. Ulrich, and H. F. Nijhout, "A mathematical model of glutathione metabolism," *Theor Biol Med Model*, vol. 5, p. 8, Jan 2008.

[84] O. Resendis-Antonio, A. Checa, and S. Encarnación, "Modeling core metabolism in cancer cells: surveying the topology underlying the Warburg effect," *PLoS ONE*, vol. 5, no. 8, p. e12383, 2010.

[85] O. Resendis-Antonio, M. Hernández, Y. Mora, and S. Encarnación, "Functional modules, structural topology, and optimal activity in metabolic networks," *PLoS Comput Biol*, vol. 8, no. 10, p. e1002720, 2012.

[86] S. Sahoo, L. Franzson, J. J. Jonsson, and I. Thiele, "A compendium of inborn errors of metabolism mapped onto the human metabolic network," *Mol Biosyst*, vol. 8, no. 10, pp. 2545–2558, 2012.

[87] H. Shinto, Y. Tashiro, M. Yamashita, G. Kobayashi, T. Sekiguchi, T. Hanai, Y. Kuriya, M. Okamoto, and K. Sonomoto, "Kinetic modeling and sensitivity analysis of acetone-butanol-ethanol production," *J Biotechnol*, vol. 131, pp. 45–56, Aug 2007.

[88] T. Shlomi, T. Benyamini, E. Gottlieb, R. Sharan, and E. Ruppin, "Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warburg effect," *PLoS Comput Biol*, vol. 7, no. 3, p. e1002018, 2011.

[89] E. Simeonidis, E. Murabito, K. Smallbone, and H. V. Westerhoff, "Why does yeast ferment? A flux balance analysis study," *Biochem Soc T*, vol. 38, no. 5, pp. 1225–1229, 2010.

[90] B. Teusink, J. Passarge, C. A. Reijenga, E. Esgalhado, C. C. van der Weijden, M. Schepper, M. C. Walsh, B. M. Bakker, K. van Dam, H. V. Westerhoff, and *et al.*, "Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry," *Eur J Biochem*, vol. 267, no. 17, pp. 5313–5329, 2000.

[91] L. M. Tran, M. L. Rizk, and J. C. Liao, "Ensemble modeling of metabolic networks," *Biophys J*, vol. 95, no. 12, pp. 5606–5617, 2008.

[92] A. Varma and B. Ø. Palsson, "Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110," *Appl Environ Microbiol*, vol. 60, no. 10, pp. 3724–3731, 1994.

[93] A. Vazquez, J. Liu, Y. Zhou, and Z. N. Oltvai, "Catabolic efficiency of aerobic glycolysis: the Warburg effect revisited," *BMC Syst Biol*, vol. 4, p. 58, 2010.

[94] A. Wagner and D. A. Fell, "The small world inside large metabolic networks," *P Roy Soc B-Biol Sci*, vol. 268, no. 1478, pp. 1803–1810, 2001.

[95] N. Xu, L. Liu, W. Zou, J. Liu, Q. Hua, and J. Chen, "Reconstruction and analysis of the genome-scale metabolic network of *Candida glabrata*," *Mol Biosyst*, vol. 9, no. 2, pp. 205–216, 2013.

[96] K. Yugi, Y. Nakayama, A. Kinoshita, and M. Tomita, "Hybrid dynamic/static method for large-scale simulation of metabolism," *Theor Biol Med Model*, vol. 2, p. 42, 2005.

[97] K. Yugi, "Dynamic kinetic modeling of mitochondrial energy metabolism," in *E-Cell System*, pp. 105–142, Springer, 2013.

[98] J. Zhao, H. Yu, J.-H. Luo, Z.-W. Cao, and Y.-X. Li, "Hierarchical modularity of nested bow-ties in metabolic networks," *BMC Bioinformatics*, vol. 7, no. 1, p. 386, 2006.

[99] L. Zhou, M. Aon, T. Almas, S. Cortassa, R. Winslow, and B. O'Rourke, "A reaction-diffusion model of ROS-induced ROS release in a mitochondrial network," *PLoS Comput Biol*, vol. 6, no. 1, p. e1000657, 2010.

[100] K. Smallbone, E. Simeonidis, D. S. Broomhead, and D. B. Kell, "Something from nothing: bridging the gap between constraint-based and kinetic modelling," *FEBS J*, vol. 274, pp. 5576–85, Nov. 2007.

[101] S. Kauffman, "A proposal for using the ensemble approach to understand genetic regulatory networks," *J Theor Biol*, vol. 230, no. 4, pp. 581–590, 2004.

[102] J. Schellenberger and B. Ø. Palsson, "Use of randomized sampling for analysis of metabolic networks," *J Biol Chem*, vol. 284, no. 9, pp. 5457–5461, 2009.

[103] K. R. Patil, I. Rocha, J. Förster, and J. Nielsen, "Evolutionary programming as a platform for in silico metabolic engineering," *BMC Bioinformatics*, vol. 6, p. 308, Jan. 2005.

[104] D. Whitley, "A genetic algorithm tutorial," *Stat Comput*, vol. 4, no. 2, pp. 65–85, 1994.

[105] S. A. Becker, A. M. Feist, M. L. Mo, G. Hannum, B. Ø. Palsson, and M. J. Herrgard, "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox," *Nat Protoc*, vol. 2, no. 3, pp. 727–38, 2007.

[106] M. Konig and H.-G. Holzhütter, "Fluxviz—cytoscape plug-in for visualization of flux distributions in networks," in *Genome Informatics 2010: The 10th Annual International Workshop on Bioinformatics and Systems Biology (IBSB 2010): Kyoto University, Japan, 26-28 July 2010*, no. 24, p. 96, World Scientific, 2010.

[107] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Res*, vol. 13, no. 11, pp. 2498–504, 2003.

[108] C. Elliott, V. Vijayakumar, W. Zink, and R. Hansen, "National instruments LabVIEW: a programming environment for laboratory automation and measurement," *JALA*, vol. 12, no. 1, pp. 17–24, 2007.

[109] R. Eberhart and J. Kennedy, "A new optimiser using particle swarm theory," in *Proc. of the Sixth International Symposium on Micro Machine and Human Science*, pp. 39–43, IEEE service center, 1995.

[110] M. S. Nobile, G. Pasi, P. Cazzaniga, D. Besozzi, R. Colombo, and G. Mauri, "Proactive particles in swarm optimization:a self-tuning algorithm based on fuzzy logic," in *Accepted to FUZZ-IEEE 2015. The 2015 IEEE International Conference on Fuzzy Systems*, (Istanbul, Turkey), 2015.

[111] N. E. Lewis, H. Nagarajan, and B. Ø. Palsson, "Constraining the metabolic genotype–phenotype relationship using a phylogeny of *in silico* methods," *Nat Rev Microbiol*, vol. 10, no. 4, pp. 291–305, 2012.

[112] "COBRA Methods." http://cobramethods.wikidot.com/methods.

[113] K. Raman and N. Chandra, "Flux balance analysis of biological systems: applications and challenges," *Brief Bioinform*, vol. 10, no. 4, pp. 435–49, 2009.

[114] J. M. Lee, E. P. Gianchandani, and J. A. Papin, "Flux balance analysis in the era of metabolomics," *Brief Bioinform*, vol. 7, no. 2, pp. 140–150, 2006.

[115] C. H. Schilling, D. Letscher, and B. Ø. Palsson, "Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective," *J Theor Biol*, vol. 203, no. 3, pp. 229–248, 2000.

[116] S. Schuster and C. Hilgetag, "On elementary flux modes in biochemical reaction systems at steady state," *J Biol Sys*, vol. 2, no. 02, pp. 165–182, 1994.

[117] K. A. Hunt, J. P. Folsom, R. L. Taffs, and R. P. Carlson, "Complete enumeration of elementary flux modes through scalable, demand-based subnetwork definition," *Bioinformatics*, vol. 30, no. 11, pp. 1569–1578, 2014.

[118] L. F. De Figueiredo, A. Podhorski, A. Rubio, C. Kaleta, J. E. Beasley, S. Schuster, and F. J. Planes, "Computing the shortest elementary flux modes in genome-scale metabolic networks," *Bioinformatics*, vol. 25, no. 23, pp. 3158–3165, 2009.

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

[120] S. Gudmundsson and I. Thiele, "Computationally efficient flux variability analysis," *BMC Bioinformatics*, vol. 11, no. 1, p. 489, 2010.

[121] R. Mahadevan, J. S. Edwards, and F. J. Doyle III, "Dynamic flux balance analysis of diauxic growth in *Escherichia coli*," *Biophys J*, vol. 83, no. 3, pp. 1331–1340, 2002.

[122] C. Bro, B. Regenberg, J. Förster, and J. Nielsen, "In silico aided metabolic engineering of *Saccharomyces cerevisiae* for improved bioethanol production," *Metab Eng*, vol. 8, no. 2, pp. 102–111, 2006.

[123] A. P. Burgard and C. D. Maranas, "Optimization-based framework for inferring and testing hypothesized metabolic objective functions," *Biotechnol Bioeng*, vol. 82, no. 6, pp. 670–677, 2003.

[124] R. Ramakrishna, J. S. Edwards, A. McCulloch, and B. Ø. Palsson, "Flux-balance analysis of mitochondrial energy metabolism: consequences of systemic stoichiometric constraints," *Am J Physiol-Reg I*, vol. 280, no. 3, pp. R695–R704, 2001.

[125] M. W. Covert, C. H. Schilling, and B. Ø. Palsson, "Regulation of gene expression in flux balance models of metabolism," *J Theor Biol*, vol. 213, no. 1, pp. 73–88, 2001.

[126] M. J. Herrgård, B.-S. Lee, V. Portnoy, and B. Ø. Palsson, "Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*," *Genome Res*, vol. 16, no. 5, pp. 627–635, 2006.

[127] A. Feist and B. Ø. Palsson, "The biomass objective function," *Curr Opin Microbiol*, vol. 13, no. 3, pp. 344–349, 2010.

[128] A. Mardinoglu, F. Gatto, and J. Nielsen, "Genome-scale modeling of human metabolism - a systems biology approach," *Biotechnol J*, pp. 1–12, Apr. 2013.

[129] R. Schuetz, N. Zamboni, M. Zampieri, M. Heinemann, and U. Sauer, "Multidimensional optimality of microbial metabolism," *Science*, vol. 336, no. 6081, pp. 601–604, 2012.

[130] S. Bordel, R. Agren, and J. Nielsen, "Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes," *PLoS Comput Biol*, vol. 6, p. e1000859, Jan. 2010.

[131] W. W. Soon, M. Hariharan, and M. P. Snyder, "High-throughput sequencing for biology and medicine," *Mol Syst Biol*, vol. 9, no. 1, 2013.

[132] A. Levchenko, "Dynamical and integrative cell signaling: challenges for the new biology," *Biotechnol Bioeng*, vol. 84, no. 7, pp. 773–782, 2003.

[133] S. Devoid, R. Overbeek, M. DeJongh, V. Vonstein, A. A. Best, and C. Henry, "Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED," in *Systems Metabolic Engineering*, pp. 17–45, Springer, 2013.

[134] V. S. Kumar, M. S. Dasika, and C. D. Maranas, "Optimization based automated curation of metabolic reconstructions," *BMC Bioinformatics*, vol. 8, no. 1, p. 212, 2007.

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

[136] M. Latendresse, M. Krummenacker, M. Trupp, and P. D. Karp, "Construction and completion of flux balance models from pathway databases," *Bioinformatics*, vol. 28, no. 3, pp. 388–396, 2012.

[137] M. Latendresse, "Efficiently gap-filling reaction networks," *BMC Bioinformatics*, vol. 15, no. 1, p. 225, 2014.

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

[139] R. Agren, S. Bordel, A. Mardinoglu, N. Pornputtapong, I. Nookaew, and J. Nielsen, "Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT," *PLoS Comput Biol*, vol. 8, no. 5, p. e1002518, 2012.

[140] M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober, H. Wernerus, L. Björling, and F. Ponten, "Towards a knowledge-based human protein atlas," *Nat Biotechnol*, vol. 28, no. 12, pp. 1248–1250, 2010.

[141] D. McCloskey, B. Ø. Palsson, and A. M. Feist, "Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*," *Mol Syst Biol*, vol. 9, no. 1, 2013.

[142] M. A. Oberhardt, B. Ø. Palsson, and J. A. Papin, "Applications of genome-scale metabolic reconstructions," *Mol Syst Biol*, vol. 5, no. 1, 2009.

[143] J. Monk, J. Nogales, and B. Ø. Palsson, "Optimizing genome-scale network reconstructions," *Nat Biotechnol*, vol. 32, no. 5, pp. 447–452, 2014.

[144] J. Schellenberger, N. E. Lewis, and B. Ø. Palsson, "Elimination of thermodynamically infeasible loops in steady-state metabolic models," *Biophys J*, vol. 100, no. 3, pp. 544–553, 2011.

[145] K. Smallbone, H. L. Messiha, K. M. Carroll, C. L. Winder, N. Malys, W. B. Dunn, E. Murabito, N. Swainston, J. O. Dada, F. Khan, and *et al.*, "A model of yeast glycolysis based on a consistent kinetic characterisation of all its enzymes," *FEBS Lett*, vol. 587, no. 17, pp. 2832–2841, 2013.

[146] R. Diaz-Ruiz, M. Rigoulet, and A. Devin, "The Warburg and Crabtree effects: on the origin of cancer cell energy metabolism and of yeast glucose repression," *BBA-Bioenergetics*, vol. 1807, no. 6, pp. 568–576, 2011.

[147] A. Mogilner, R. Wollman, and W. F. Marshall, "Quantitative modeling in cell biology: What is it good for?," *Dev Cell*, vol. 11, no. 3, pp. 279–287, 2006.

[148] "Biomet toolbox." http://biomet-toolbox.org/.

[149] J. Schellenberger, R. Que, R. M. Fleming, I. Thiele, J. D. Orth, A. M. Feist, D. C. Zielinski, A. Bordbar, N. E. Lewis, S. Rahmanian, and *et al.*, "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0," *Nat Protoc*, vol. 6, no. 9, pp. 1290–1307, 2011.

[150] J. Boele, B. G. Olivier, and B. Teusink, "FAME, the flux analysis and modeling environment," *BMC Syst Biol*, vol. 6, no. 1, p. 8, 2012.

[151] A. Hoppe, S. Hoffmann, A. Gerasch, C. Gille, and H.-G. Holzhütter, "FASIMU: flexible software for flux-balance computation series in large metabolic networks," *BMC Bioinformatics*, vol. 12, no. 1, p. 28, 2011.

[152] I. Rocha, P. Maia, P. Evangelista, P. Vilaça, S. Soares, J. P. Pinto, J. Nielsen, K. R. Patil, E. C. Ferreira, and M. Rocha, "OptFlux: an open-source software platform for *in silico* metabolic engineering," *BMC Syst Biol*, vol. 4, no. 1, p. 45, 2010.

[153] P. D. Karp, S. M. Paley, M. Krummenacker, M. Latendresse, J. M. Dale, T. J. Lee, P. Kaipa, F. Gilham, A. Spaulding, L. Popescu, and *et al.*, "Pathway tools version 13.0: integrated software for pathway/genome informatics and systems biology," *Brief Bioinform*, p. bbp043, 2009.

[154] R. Agren, L. Liu, S. Shoaie, W. Vongsangnak, I. Nookaew, and J. Nielsen, "The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*," *PLoS Comput Biol*, vol. 9, no. 3, p. e1002980, 2013.

[155] A. Gevorgyan, M. E. Bushell, C. Avignone-Rossa, and A. M. Kierzek, "SurreyFBA: a command line tool and graphics user interface for constraint-based modeling of genome-scale metabolic reaction networks," *Bioinformatics*, vol. 27, no. 3, pp. 433–434, 2011.

[156] G. Clermont, C. Auffray, Y. Moreau, D. M. Rocke, D. Dalevi, D. Dubhashi, D. R. Marshall, P. Raasch, F. Dehne, P. Provero, *et al.*, "Bridging the gap between systems biology and medicine," *Genome Med*, vol. 1, no. 9, p. 88, 2009.

[157] A. Mardinoglu, R. Agren, C. Kampf, A. Asplund, I. Nookaew, P. Jacobson, A. J. Walley, P. Froguel, L. M. Carlsson, M. Uhlen, *et al.*, "Integration of clinical data with a genome-scale metabolic model of the human adipocyte," *Mol Syst Biol*, vol. 9, no. 1, 2013.

[158] H. Ma, A. Sorokin, A. Mazein, A. Selkov, E. Selkov, O. Demin, and I. Goryanin, "The Edinburgh human metabolic network reconstruction and its functional analysis," *Mol Syst Biol*, vol. 3, p. 135, 2007.

[159] P. Romero, J. Wagg, M. L. Green, D. Kaiser, M. Krummenacker, and P. D. Karp, "Computational prediction of human metabolic pathways from the complete human genome," *Genome Biol*, vol. 6, no. 1, p. R2, 2004.

[160] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: back to metabolism in KEGG," *Nucleic Acids Res*, vol. 42, no. D1, pp. D199–D205, 2014.

[161] M. Uhlén, E. Björling, C. Agaton, C. A.-K. Szigyarto, B. Amini, E. Andersen, A.-C. Andersson, P. Angelidou, A. Asplund, C. Asplund, *et al.*, "A human protein atlas for normal and cancer tissues based on antibody proteomics," *Mol Cell Proteomics*, vol. 4, no. 12, pp. 1920–1932, 2005.

[162] D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, *et al.*, "HMDB: the human metabolome database," *Nucleic Acids Res*, vol. 35, no. suppl 1, pp. D521–D526, 2007.

[163] D. S. Wishart, C. Knox, A. C. Guo, R. Eisner, N. Young, B. Gautam, D. D. Hau, N. Psychogios, E. Dong, S. Bouatra, *et al.*, "HMDB: a knowledgebase for the human metabolome," *Nucleic Acids Res*, vol. 37, no. suppl 1, pp. D603–D610, 2009.

[164] D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, *et al.*, "HMDB 3.0—the human metabolome database in 2013," *Nucleic Acids Res*, p. gks1065, 2012.

[165] J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. Forman, and F. Bray, "GLOBOCAN 2012 v1. 0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, Fr Int Agency Res Cancer," 2013.

[166] F. Bray, J.-S. Ren, E. Masuyer, and J. Ferlay, "Estimates of global cancer prevalence for 27 sites in the adult population in 2008," *Int J Cancer*, vol. 132, no. 5, pp. 1133–1145, 2013.

[167] L. Alberghina and D. Gaglio, "Redox control of glutamine utilization in cancer," *Cell Death Dis*, vol. 5, no. 12, p. e1561, 2014.

[168] C. Li, M. Donizelli, N. Rodriguez, H. Dharuri, L. Endler, V. Chelliah, L. Li, E. He, A. Henry, M. I. Stefan, and *et al.*, "BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models," *BMC Syst Biol*, vol. 4, p. 92, Jun 2010.

[169] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, *et al.*, "Tissue-based map of the human proteome," *Science*, vol. 347, no. 6220, p. 1260419, 2015.

[170] R. J. DeBerardinis, A. Mancuso, E. Daikhin, I. Nissim, M. Yudkoff, S. Wehrli, and C. B. Thompson, "Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis," *PNAS*, vol. 104, no. 49, pp. 19345–50, 2007.

[171] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *Cell*, vol. 144, no. 5, pp. 646–74, 2011.

[172] J. W. Locasale and L. C. Cantley, "Altered metabolism in cancer," *BMC Biol*, vol. 8, p. 88, 2010.

[173] M. G. Vander Heiden, L. C. Cantley, and C. B. Thompson, "Understanding the Warburg effect: the metabolic requirements of cell proliferation," *Science*, vol. 324, no. 5930, pp. 1029–33, 2009.

[174] R. J. DeBerardinis, J. J. Lum, G. Hatzivassiliou, and C. B. Thompson, "The biology of cancer: metabolic reprogramming fuels cell growth and proliferation," *Cell Metab*, vol. 7, no. 1, pp. 11–20, 2008.

[175] J. Hooda, D. Cadinu, M. M. Alam, A. Shah, T. M. Cao, L. A. Sullivan, R. Brekken, and L. Zhang, "Enhanced heme function and mitochondrial respiration promote the progression of lung cancer cells," *PLoS One*, vol. 8, no. 5, p. e63402, 2013.

[176] R. H. De Deken, "The Crabtree effect: a regulatory system in yeast," *J Gen Microb*, vol. 44, no. 2, pp. 149–156, 1966.

[177] J. P. Barford and R. J. Hall, "An examination of the Crabtree Effect in *Saccharomyces cerevisiae*: the role of respiratory adaptation," *J Gen Microb*, vol. 114, 1979.

[178] H. Van Urk, W. Voll, W. Scheffers, and J. Van Dijken, "Transient-state analysis of metabolic fluxes in crabtree-positive and crabtree-negative yeasts," *Appl Environ Microb*, vol. 56, no. 1, pp. 281–287, 1990.

[179] D. Porro, L. Brambilla, and L. Alberghina, "Glucose metabolism and cell size in continuous cultures of *Saccharomyces cerevisiae*," *FEMS microbiol lett*, vol. 229, no. 2, pp. 165–171, 2003.

[180] M. Papini, I. Nookaew, M. Uhlén, and J. Nielsen, "*Scheffersomyces stipitis*: a comparative systems biology study with the Crabtree positive yeast *Saccharomyces cerevisiae*," *Microb Cell Fact*, vol. 11, p. 136, Jan. 2012.

[181] M. Mitchell, *An Introduction to Genetic Algorithms*. The MIT Press, 1996.

[182] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 2, pp. 241–254, 1967.

[183] A. Hagman, T. Säll, C. Compagno, and J. Piskur, "Yeast "make-accumulate-consume" life strategy evolved as a multi-step process that predates the whole genome duplication," *PloS One*, vol. 8, no. 7, p. e68734, 2013.

[184] S. Supudomchok, N. Chaiyaratana, and C. Phalakomkule, "Co-operative co-evolutionary approach for flux balance in *Bacillus subtilis*," in *Evolutionary Computation, 2008. CEC 2008. (IEEE World Congress on Computational Intelligence). IEEE Congress on*, pp. 1226–1231, 2008.

[185] D. Segre, A. DeLuna, G. M. Church, and R. Kishony, "Modular epistasis in yeast metabolism," *Nat Genet*, vol. 37, no. 1, pp. 77–83, 2005.

[186] L. Alberghina, D. Gaglio, R. M. Moresco, M. C. Gilardi, C. Messa, and M. Vanoni, "A systems biology road map for the discovery of drugs targeting cancer cell metabolism," *Curr Pharm Design*, 2013.

[187] M. Chiu, L. Ottaviani, M. G. Bianchi, R. Franchi-Gazzola, and O. Bussolati, "Towards a metabolic therapy of cancer?," *Acta bio-medica: Atenei Parmensis*, vol. 83, no. 3, pp. 168–176, 2012.

[188] T. Soga, "Cancer metabolism: key players in metabolic reprogramming," *Cancer Sci*, vol. 104, no. 3, pp. 275–281, 2013.

[189] H. Kitano, K. Oda, T. Kimura, Y. Matsuoka, M. Csete, J. Doyle, and M. Muramatsu, "Metabolic syndrome and robustness tradeoffs," *Diabetes*, vol. 53, no. suppl 3, pp. S6–S15, 2004.

[190] I. A. Razinkov, B. L. Baumgartner, M. R. Bennett, L. S. Tsimring, and J. Hasty, "Measuring competitive fitness in dynamic environments," *J Phys Chem B*, vol. 117, no. 42, pp. 13175–13181, 2013.

[191] A.-L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nat Rev Genet*, vol. 5, no. 2, pp. 101–13, 2004.

[192] "Network science book project, http://barabasilab.neu.edu/networksciencebook/."

[193] M. Suderman and M. Hallett, "Tools for visually exploring biological networks," *Bioinformatics*, vol. 23, no. 20, pp. 2651–2659, 2007.

[194] G. A. Pavlopoulos, A.-L. Wegener, and R. Schneider, "A survey of visualization tools for biological network analysis," *BioData Min*, vol. 1, no. 1, p. 12, 2008.

[195] G. Michal, "Biochemical pathways (poster)," *Boehringer Mannheim, Penzberg*, 1993.

[196] N. Gehlenborg, S. I. O'Donoghue, N. S. Baliga, A. Goesmann, M. A. Hibbs, H. Kitano, O. Kohlbacher, H. Neuweger, R. Schneider, D. Tenenbaum, *et al.*, "Visualization of omics data for systems biology," *Nat methods*, vol. 7, pp. S56–S68, 2010.

[197] A. Kostromins and E. Stalidzans, "Paint4Net: COBRA Toolbox extension for visualization of stoichiometric models of metabolism," *Biosystems*, vol. 109, no. 2, pp. 233–239, 2012.

[198] M. Hucka *et al*, "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models," *Bioinformatics*, vol. 19, no. 4, pp. 524–531, 2003.

[199] M. S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P.-L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker, and G. D. Bader, "Integration of biological networks and gene expression data using Cytoscape," *Nat Protoc*, vol. 2, no. 10, pp. 2366–82, 2007.

[200] "Tutorial Cytoscape, http://wiki.cytoscape.org/cytoscape_user_manual/visual_styles."

[201] A. P. Burgard, E. V. Nikolaev, C. H. Schilling, and C. D. Maranas, "Flux coupling analysis of genome-scale metabolic network reconstructions," *Genome Res*, vol. 14, no. 2, pp. 301–312, 2004.

[202] A. Samal, S. Singh, V. Giri, S. Krishna, N. Raghuram, and S. Jain, "Low degree metabolites explain essential reactions and enhance modularity in biological networks," *BMC bioinformatics*, vol. 7, no. 1, p. 118, 2006.

[203] S. Singh, A. Samal, V. Giri, S. Krishna, N. Raghuram, and S. Jain, "Flux-based classification of reactions reveals a functional bow-tie organization of complex metabolic networks," *Phys Rev E*, vol. 87, no. 5, p. 052708, 2013.

[204] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, "Stochastic gene expression in a single cell," *Science*, vol. 297, pp. 1183–1186, Aug 2002.

[205] M. A. Savageau, "Biochemical systems theory: operational differences among variant representations and their significance," *J Theor Biol*, vol. 151, no. 4, pp. 509–530, 1991.

[206] A.-M. Wazwaz, *Partial Differential Equations*. CRC Press, 2002.

[207] D. McQuairre, "Stochastic approach to chemical kinetics," *J Applied Probability*, vol. 4, pp. 413–478, 1967.

[208] D. Gillespie, "A rigorous derivation of the chemical master equation," *Physica A*, vol. 188, pp. 404–425, 1992.

[209] D. Gillespie, *Markov Processes: An Introduction for Physical Scientists*. Academic Press, 1991.

[210] D. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *J Comp Phys*, vol. 22, pp. 403–434, 1976.

[211] D. Gillespie, "Approximate accelerated stochastic simulation of chemically reacting systems," *J Chem Phys*, vol. 115, pp. 1716–1733, 2001.

[212] J. Pahle, "Biochemical simulations: stochastic, approximate stochastic and hybrid approaches," *Brief Bioinform*, vol. 10, no. 1, pp. 53–64, 2009.

[213] A. Alfonsi, E. Cancès, G. Turinici, B. Di Ventura, and W. Huisinga, "Adaptive simulation of hybrid stochastic and deterministic models for biochemical systems," in *ESAIM: Proc*, vol. 4, pp. 1–13, 2005.

[214] O. Resendis-Antonio, "Filling kinetic gaps: dynamic modeling of metabolism where detailed kinetic information is lacking," *PLoS One*, vol. 4, no. 3, p. e4967, 2009.

[215] K. Smallbone, E. Simeonidis, N. Swainston, and P. Mendes, "Towards a genome-scale kinetic model of cellular metabolism," *BMC Syst Biol*, vol. 4, p. 6, Jan. 2010.

[216] V. Hatzimanikatis and J. E. Bailey, "Effects of spatiotemporal variations on metabolic control: approximate analysis using (log) linear kinetic models," *Biotechnol Bioeng*, vol. 54, no. 2, pp. 91–104, 1997.

[217] C. G. Moles, P. Mendes, and J. R. Banga, "Parameter estimation in biochemical pathways: a comparison of global optimization methods," *Genome Res*, vol. 13, pp. 2467–2474, 2003.

[218] J. R. Banga, "Optimization in computational systems biology," *BMC Syst Biol*, vol. 2, no. 1, p. 47, 2008.

[219] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by Simulated Annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[220] T. Bäck, D. B. Fogel, and Z. Michalewicz, *Evolutionary computation 1: Basic algorithms and operators*, vol. 1. CRC Press, 2000.

[221] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, 1989.

[222] Z. Michalewicz, *Genetic algorithms + data structures = evolution programs.* Springer-Verlag London, UK, 1996.

[223] H.-G. Beyer and H.-P. Schwefel, "Evolution strategies–a comprehensive introduction," *Nat Comput*, vol. 1, no. 1, pp. 3–52, 2002.

[224] M. Dorigo and M. Birattari, "Ant colony optimization," in *Encyclopedia of Machine Learning*, pp. 36–39, Springer, 2010.

[225] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, vol. 4, (Piscataway, NJ), pp. 1942–1948, 1995.

[226] S. Xu and Y. Rahmat-Samii, "Boundary conditions in particle swarm optimization revisited," *IEEE Antennas Propag.*, vol. 55, pp. 760–765, 2007.

[227] R. Poli, "Analysis of the publications on the applications of particle swarm optimisation," *J Artif Evol Appl*, vol. 2008, pp. 2–10, 2008.

[228] T. Jewison, C. Knox, V. Neveu, Y. Djoumbou, A. C. Guo, J. Lee, P. Liu, R. Mandal, R. Krishnamurthy, I. Sinelnikov, *et al.*, "YMDB: the yeast metabolome database," *Nucleic Acids Res*, p. gkr916, 2011.

[229] I. Spasić, E. Simeonidis, H. L. Messiha, N. W. Paton, and D. B. Kell, "KiPar, a tool for systematic information retrieval regarding parameters for kinetic modelling of yeast metabolic pathways," *Bioinformatics*, vol. 25, no. 11, pp. 1404–1411, 2009.

[230] R. J. Roberts, "PubMed Central: The GenBank of the published literature," *PNAS*, vol. 98, no. 2, pp. 381–382, 2001.

[231] J. Bezanson, S. Karpinski, V. B. Shah, and A. Edelman, "Julia: A fast dynamic language for technical computing," *arXiv preprint arXiv:1209.5145*, 2012.

[232] M. Clerc, "Particle swarm optimization," in *International scientific and technical encyclopaedia*, Hoboken: Wiley, 2006.

[233] G. Papa, "Parameter-less algorithm for evolutionary-based optimization," *Comput. Optim. Appl.*, vol. 56, no. 1, pp. 209–229, 2013.

[234] M. Tomassini, L. Vanneschi, J. Cuendet, and F. Fernández, "A new technique for dynamic size populations in genetic programming," in *Proc. CEC2004. IEEE Congress on Evolutionary Computation*, vol. 1, (Portland, OR), pp. 486–493, 2004.

[235] L. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338 – 353, 1965.

[236] A. Rezaee Jordehi and J. Jasni, "Parameter selection in particle swarm optimisation: a survey," *J. Exp. Theor. Artif. Intell.*, vol. 25, no. 4, pp. 527–542, 2013.

[237] Y. Shi and R. C. Eberhart, "Fuzzy adaptive particle swarm optimization," in *Evolutionary Computation, 2001. Proceedings of the 2001 Congress on*, vol. 1, pp. 101–106, IEEE, 2001.

[238] A. Abraham and H. Liu, "Turbulent particle swarm optimization using fuzzy parameter tuning," in *Foundations of Computational Intelligence Volume 3*, pp. 291–312, Springer, 2009.

[239] D. Tian and N. Li, "Fuzzy particle swarm optimization algorithm," in *Artificial Intelligence, 2009. JCAI'09. International Joint Conference on*, pp. 263–267, IEEE, 2009.

[240] T. J. Ross, *Fuzzy logic with engineering applications*. New York: McGraw-Hill, 1995.

[241] E. H. Mamdani, "Application of fuzzy logic to approximate reasoning using linguistic synthesis," *IEEE T Comput*, vol. 100, no. 12, pp. 1182–1191, 1977.

[242] M. Sugeno, *Industrial Applications of Fuzzy Control*. New York, NY: Elsevier Science Inc., 1985.

[243] W. Van Leekwijck and E. E. Kerre, "Defuzzification: criteria and classification," *Fuzzy Set Syst*, vol. 108, no. 2, pp. 159–178, 1999.

[244] T. E. Oliphant, "Python for scientific computing," *Comput Sci Eng*, vol. 9, no. 3, pp. 10–20, 2007.

[245] N. Hansen, R. Ros, N. Mauny, M. Schoenauer, and A. Auger, "Impacts of invariance in search: When CMA-ES and PSO face ill-conditioned and non-separable problems," *Appl Soft Comput*, vol. 11, no. 8, pp. 5755–5769, 2011.

[246] M. R. Antoniewicz, "Dynamic metabolic flux analysis—tools for probing transient states of metabolic networks," *Curr Op Biotech*, vol. 24, no. 6, pp. 973–978, 2013.

[247] J. H. van Heerden, M. T. Wortel, F. J. Bruggeman, J. J. Heijnen, Y. J. Bollen, R. Planqué, J. Hulshof, T. G. O'Toole, S. A. Wahl, and B. Teusink, "Lost in transition: start-up of glycolysis yields subpopulations of nongrowing cells," *Science*, vol. 343, no. 6174, p. 1245114, 2014.

[248] E. Murabito, R. Colombo, C. Wu, M. Verma, S. Rehman, J. Snoep, S.-L. Peng, N. Guan, X. Liao, and H. V. Westerhoff, "SupraBiology 2014: Promoting UK-China collaboration on Systems Biology and High Performance Computing," *Quant Biol*, pp. 1–8, 2015.

[249] H. V. Westerhoff and B. Ø. Palsson, "The evolution of molecular biology into systems biology," *Nat Biotechnol*, vol. 22, no. 10, pp. 1249–1252, 2004.

[250] W.-L. Liao and F.-J. Tsai, "Personalized medicine: a paradigm shift in healthcare," *BioMedicine*, vol. 3, no. 2, pp. 66–72, 2013.

[251] A. Kolodkin, F. C. Boogerd, N. Plant, F. J. Bruggeman, V. Goncharuk, J. Lunshof, R. Moreno-Sanchez, N. Yilmaz, B. M. Bakker, J. L. Snoep, *et al.*, "Emergence of the silicon human and network targeting drugs," *Eur J Pharm Sci*, vol. 46, no. 4, pp. 190–197, 2012.