Politecnico di Milano

SCUOLA DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE
Corso di Laurea Magistrale in Ingegneria Chimica

Tesi di laurea magistrale

Metabolic network-based modeling of micro-organisms: Evaluation of black- and grey-box flux model structures

Relatore:
Prof. Flavio Manenti

Correlatori:
Prof. dr. ir. Jan Van Impe
Prof. dr. ir. Filip Logist
Ir. Dominique Vercammen

Candidato:
Giorgio Ferrigno
Matricola 786983

Anno Accademico 2013–2014

In collaboration with:

KU Leuven
Faculty of Engineering Sciences
Department of Chemical Engineering
BioTeC - Bioprocess Technology and Control
W. de Croylaan 46
B-3001 Leuven (Belgium)
Tel: +32 16/321466 - Fax: +32 16/322991

Prof. dr. ir. **Jan Van Impe**
Prof. dr. ir. **Filip Logist**
Ir. **Dominique Vercammen**

Ringraziamenti

Sembra proprio che sia infine arrivato alla conclusione di questo percorso. Come poi ci sia arrivato, è un'altra storia. E' stato un percorso lungo (molto lungo, a ripensarci adesso), fatto dall'alternarsi di momenti belli, momenti difficili, momenti di grandi soddisfazioni e di altrettanto grandi delusioni, momenti di stress, momenti di crescita, momenti di voler tornare bambino, momenti di sconforto e chi più ne ha più ne metta. Tanti, tantissimi momenti, nel complesso.

Ingegneria è stata senza dubbio un'esperienza che mi ha cambiato la vita. Sono e mi sento un'altra persona rispetto a quel ragazzino che cinque anni fa (quasi sei, in realtà) ha varcato le soglie del Politecnico di Milano, senza avere ancora ben chiaro in mente cosa ci facesse in quel posto. Quest'esperienza mi ha dato tanto, e nonostante mi abbia anche tolto tanto (sonno, luce, salute, spensieratezza, ma per fortuna i capelli ancora no), ha contribuito a fare di me la persona che sono oggi. E alla fine, cosa importa se per qualcuno sono diventato il ritratto vivente di quella barzelletta vivente che è l'ingegnere? Mentre compilo il questionario del Politecnico di valutazione finale del percorso formativo, alla domanda "sceglieresti ancora questo percorso di studi?", dopo qualche istante di riflessione, rispondo di si.

Segue adesso la lunga lista delle persone cui un grazie è d'obbligo, ma anche un immenso piacere.

Ringrazio prima di tutto e tutti la mia famiglia, mio padre Angelo e mia madre Addina, che sono stati con me in ogni istante di questo faticoso percorso. Parlo non solo di ingegneria, ma anche di tutto quel che la ha preceduta. La mia vita, insomma. Un grazie particolare perchè hanno saputo incoraggiarmi sempre senza mai farmi sentire sotto pressione, combinazione che credo sia difficile da bilanciare per dei genitori.

Ringrazio Erika, la bella abruzzese che ha stoicamente sopportato buona parte dei miei scleri e dei miei momenti di sconforto, spronandomi a continuare (a volte con dolcezza, altre con modi leggermente più bruschi, ma comunque molto efficaci) ed infondendomi ogni giorno nuova motivazione.

Ringrazio le persone che mi hanno aiutato in questo lavoro di tesi. Il mio relatore, Flavio Manenti, persona sempre entusiasta ed energica, oltre che molto impegnata. I dottorandi Francesco Rossi e Davide Papasidero, che mi hanno aiutato e dato consiglio. Tutto il gruppo di ricerca BioTeC della KU Leuven, che mi ha accolto in terra belga e mi ha trattato per sei mesi come uno di "famiglia". In particolare ringrazio i professori Jan Van Impe e Filip Logist, che mi hanno supervisionato, sempre gentili ed estremamente disponibili; Mattia Vallerio, per avermi fatto da cicerone e da appoggio in quel fantastico universo che la cittadina di Leuven; ultimo ma non ultimo, Dominique Vercammen, che ha partorito la stragrande maggioranza

delle idee che stanno dietro questo lavoro e che mi ha dedicato tanto tempo e tanta pazienza. Ringrazio poi i miei compagni di corso, con cui ho condiviso tanti dei sopracitati momenti, e che sono uno dei principali motivi per cui di certo mi mancherà la vita universitaria. Rimpiangerò le pause caffè con gli ormai dottorandi Luca Dietz e Roberto Abbiati, che tante volte hanno alleviato il mio stress, facendomi recuperare un po' di sorriso.
Infine ringrazio tutti i miei amici, perchè ciascuno di loro ha avuto la sua parte, ed a ciascuno di loro devo una piccola fetta di questo momento.
La scuola non finisce mai, dicono. E così, dopo essere stato un veterano dell'università, dopo essermi finalmente sentito alla fine, mi ritroverò in men che non si dica ad essere nuovamente un fanciullo da qualche altra parte. Ma qualsiasi cosa mi riserverà il futuro, non smetterò mai di sorridere ripensando a questi anni di università, e lo devo, ancora una volta, a tutti voi. Grazie.

Introduzione

I micro-organismi sono oggigiorno largamente utilizzati in molti campi industriali, come industria alimentare, farmaceutica, biocombustibili ecc. Con la crescente diffusione delle biotecnologie, è nata una nuova scienza che aspira ad una comprensione quantitativa degli organismi, chiamata *microbiologia predittiva*. Considerando la grande complessità, variabilità e non linearità dei sistemi biologici, una comprensione quantitativa richiede necessariamente l'uso di modelli computazionali. Per descrivere i microorganismi, devono essere combinate informazioni provenienti da scale dimensionali diverse, dalla macroscala, che considera l'intera popolazione batterica, alla microscala, che descrive ciò che avviene dentro e nell'immediato intorno di una singola cellula. Mentre le variabili alla macroscala sono facilmente misurabili, caratterizzare sperimentalmente la microscala è ben più complicato. Nonostante questo, una recente tecnica sperimentale chiamata *analisi isotopica* permette di misurare le variabili intracellulari, ovvero le velocità di reazione, dette *flussi*, e le concentrazioni dei metaboliti. Tale metodo consiste nel nutrire la cellula con atomi di carbonio identificabili, registrando come questi si muovono all'interno della cellula attraverso il suo metabolismo.
Un approccio di modellazione di sistemi biologici particolarmente promettente è basato sui *network metabolici* (reti metaboliche). I network metabolici sono mappe del metabolismo cellulare, che riportano informazioni stechiometriche riguardo alla maggior parte dei metaboliti che partecipano alla crescita della cellula ed alle reazioni che li legano. Sfruttando i network metabolici, è possibile formulare un *modello primario*, ovvero un modello che descrive le naturali dinamiche dei metaboliti cellulari, nonostante la scarsità di informazioni sperimentali precise a livello intracellulare. Questo richiede molti meno dati rispetto alla determinazione delle equazioni cinetiche di tutte le numerose reazioni intracellulari. Ciò è reso possibile grazie ad alcune assunzioni semplificative, che sono alla base dell'*analisi dei flussi metabolici* (MFA): la popolazione batterica è considerata essere omogenea, considerando la concentrazione dei metaboliti nell'ambiente microscopico che circonda la cellula pari alla rispettiva concentrazione macroscopica; le dinamiche intracellulari sono trascurate, assumendo che il sistema sia sempre in uno stato pseudo stazionario, in equilibrio con l'ambiente extracellulare. L'estensione della MFA a sistemi in stato metabolico non stazionario è chiamata *analisi dinamica dei flussi metabolici* (dMFA).
Il principale obiettivo cui la microbiologia predittiva aspira è la formulazione di un modello capace di catturare ed esprimere la complessità del microorganismo, prevedendo la reazione del sistema biologico a stimoli ambientali esterni. Tale modello permetterebbe di implementare strategie di controllo in tempo reale per ottimizzare ogni variabile ed aspetto del processo, spingendo la coltura cellulare a soddisfare gli interessi industriali. Questo studio è da inten-

dersi alla luce di tale prospettiva, dato che aspira a risolvere il modello dinamico primario del sistema biologico sfruttando solo i dati disponibili online durante il processo. Il problema sta nel caratterizzare completamente il modello in funzione delle variabili extracellulari. Secondo la dMFA, infatti, sia le variabili extracellulari che quelli intracellulari variano nel tempo, ma mentre le prime possono essere determinate in tempo reale durante la fermentazione, l'analisi isotopica per la misurazione delle seconde è un metodo discontinuo. Bisogna quindi trovare una relazione che correli la variazione dei flussi all'evoluzione dinamica delle concentrazioni extracellulari. Due diversi approcci per esprimere questa relazione sono stati sperimentati. Il cosiddetto *approccio a scatola grigia* è basato sull'*analisi di bilanciamento dei flussi* (FBA), che segue una logica evoluzionistica. Secondo le teorie darwiniane, infatti, il tempo ha trasformato i microorganismi in perfetti ottimizzatori del loro stesso metabolismo, e i flussi devono variare ottimizzando in ogni istante una funzione obbiettivo biologica. Al contrario, il secondo approccio è semplicemente basato su un *modello a scatola nera*, molto più generale e flessibile, la cui validità per sistemi in mezzo omogeneo e sotto moderate condizioni ambientali è stata largamente confermata. Non di meno, il limite di questi modelli è la robustezza, ovvero l'estensione a condizioni ambientali diverse ed in situazioni più realistiche.
I risultati di questi approcci sono testati su due casi: un piccolo network simulato, i cui dati sono stati generati artificialmente imponendo semplici relazioni tra i flussi e le concentrazioni; un network più grande, che descrive una popolazione di *E. coli* ingegnerizzata per la produzione di 1,3-propandiolo (PDO). Questa sostanza è usata industrialmente nella produzione di plastiche, come additivo per compositi, rivestimenti, additivi ecc. e come solvente. Questo esempio ricalca quindi una reale fermentazione industriale.
Lo studio si muove in un campo di ricerca molto recente ed innovativo, esplorando le potenzialità di un approccio modellistico mai testato prima. Molte problematiche relative all'uso della FBA sono evidenziate ed esaminate, mentre l'approccio sperimentale ha prodotto alcuni risulti utili. Il limite di questo approccio rimane sempre l'estrapolazione fuori dalle condizioni sperimentali, ma rappresenta comunque un modello di partenza da migliorare in futuro ed arricchire di informazioni biologiche e meccanicistiche. La prospettiva di implementare un modello in tempo reale efficiente e preciso, capace di rivoluzionare i bioprocessi, è ancora lontana, ma molto si può ancora sperare da un campo di ricerca fresco ed in continua crescita come appare la microbiologia predittiva.

Summary

Micro-organisms nowadays are largely used in many fields of industry, such as food industry, pharmaceutical industry, biofuels etc. With the spread of biotechnology, a new science which aims to quantitative understanding of organisms was born, called *predictive microbiology*. Considering the enormous complexity, variability and nonlinearity of biological systems, a quantitative understanding necessarily requires computational models. To describe micro-organisms, information at different scales must be collected, from the macroscopic scale of the reactor, which considers the entire microbial population, to the microscopic scale, which describes what happens inside and in the immediate around of a single cell. While measurements of the macroscopic variables are easily available, to experimentally characterize the microscopic scale is much more difficult. Nevertheless, a recent technique called *isotopomer analysis* allows to measure the intracellular states, i.e., reaction rates, called *fluxes*, and metabolite concentrations. This method consists of feeding labeled carbon atoms to the cell, registering their movements through the cellular metabolism.

A kind of modeling for biological system which is very promising is based on *metabolic networks*. Metabolic networks are a blueprint of cellular metabolism, reporting stoichiometric information about most of the metabolites which participate to the cellular growth and the reactions between them. Using metabolic networks, a *primary model*, i.e., a model which describes the natural dynamics of the cellular metabolite, can be formulated which deals with the lack of precise experimental information at the intracellular scale. In fact, it requires a much inferior number of data than the determination of the kinetic equations of all the numerous intracellular reactions. This is possible thanks to some assumptions, which are at the base of *metabolic flux analysis* (MFA): the bacterial population is considered to be homogeneous, assuming the concentration of metabolites in the microscopic medium around the cell to be equal to the macroscopic, and then measurable, concentration; the intracellular dynamics are disregarded, assuming the system to be always at the metabolic pseudo-steady state, in equilibrium with the extracellular variables. These are quite strong hypotheses, but they are sufficiently satisfied considering simple, liquid media under moderate environmental conditions. The extension of MFA to systems not at the metabolic steady state is called *dynamic metabolic flux analysis* (dMFA).

The main goal which predictive microbiology aims to is to formulate a model able to unravel the complexity of the micro-organism, predicting the response of the biological system to the environmental stimuli. Such a model would allow to implement online control strategies to optimize every variable and aspect of the process, pushing the bacterial culture to accomplish the industrial interests. This study moves toward this prospect, aiming to solve

only with online inputs the primary dynamic model of the biological system. The issue is to completely characterize the model as function of the extracellular variables. According with dMFA, both the extracellular and the intracellular variables continuously vary in time, but while the extracellular states can be determined in real time during the fermentation process, the isotopomer analysis for measuring the intracellular ones is a discontinuous method. A relation which links the variation of the fluxes to the dynamic evolution of the extracellular concentrations must be found. Two different approaches to find this relation will be tested. The so-called *grey-box approach* is based on *flux balance analysis* (FBA), which follows an evolutionary logic. According with the Darwinian theory, the time has transformed the micro-organism to a perfect optimizer of its own metabolism, and the fluxes should vary to always optimizing a biological objective function. On the contrary, the second approach is simply based on a *black-box model*, much more general and flexible, and whose validity on homogeneous media under moderate environmental conditions has already been widely confirmed. Nevertheless, the limit of this kind of model is the robustness, its validity under a wider range of experimental conditions and in more realistic situations. The results of these approaches will be tested on two case studies: a small-scale simulated network, called *toy network*, whose data were artificially generated with simple relations between fluxes and concentrations; a bigger network which describes an *E.coli* population engineered to produce 1,3-propanediol (PDO). This chemical is used in industry to produce plastic, as an additive to a variety of products such as coatings, composites, adhesives etc and as a solvent. Hence, the case study resembles a real industrial fermentation.

The study moves in a very recent and innovative research field, exploring the potentiality of a previously untested modeling approach. Many problematic issue about the application of FBA will be highlighted and examined, while the black-box approach will return some interesting results. The limit of this approach remains the extrapolation outside the range of experimental conditions, but still it represents a starting model to be improved in future and enriched with biological and mechanistic information. The prospect of the implementation of a real-time, efficient and effective model which revolutionizes the bioprocesses is still far nowadays, but much can still be expected from a fresh and ongoing research area such as predictive microbiology.

Contents

List of Figures

List of Tables

Chapter 1

Literature study

1.1 Introduction

This study is focused on predictive microbiology. What is predictive microbiology? McMeekin (McMeekin et al., 2002) defined it as a quantitative science that enables users to objectively evaluate the effect of processing, distribution and storage operations on the microbiological safety and quality of foods. Since the previous quotation was included in the first book on the subject, predictive microbiology can be considered a quite recent science, still growing and improving nowadays. The adjective "quantitative" still expresses more an approach than a real state of knowledge. Quoting Lord Kelvin:

When you can measure what you are speaking about and express it in numbers you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind."

Starting from the first quantitative practice of many food microbiologists of enumerating all the micro-organisms in all the stages of food storage, which was a slow and expensive procedure, predictive microbiology is currently adopting the mathematical model approach, evolving toward a more exact science.

Although the previous definition referred to food, predictive microbiology can be meant as a general approach to the study of micro-organisms, and consequently extended beyond food industry to all the biotechnological applications, i.e., all the processes which involve the use of micro-organisms. The intent of predictive microbiology is to schematize and simplify the high complexity of the organism, applying a reductionist approach. This approach consists of identifying a limited number of environmental variables able to considerably influence the cellular behavior, i.e., such that the response of the organism can be expressed as function of these extracellular stimuli. Developing such an approach would make it possible to reach a

deep understanding of the intracellular mechanisms.
The necessary condition to apply this scientific approach to the study of living organisms is the hypothesis that they can be described as mathematical systems. This is expressed by two general premises, the basis of predictive microbiology science:

- the factors that influence the organism's responses are limited, and the responses are reproducible under the same conditions;
- based on past observations, it is possible to predict the micro-organism's behavior.

The results that were obtained in predictive microbiology since the first studies confirmed the effectiveness of this kind of scientific approach on living organisms, and they contributed during the years to make predictive microbiology a recognized and promising science.

"[...] the growth of bacterial cultures, despite the immense complexity of the phenomena to which it testifies, generally obeys relatively simple laws [...] The accuracy, the ease, the reproducibility of bacterial growth constant determinations is remarkable and probably unparalleled, so far as biological quantitative characteristics are concerned." (Monod, 1949)

1.2 Food safety

1.2.1 Origins

Predictive microbiology was born in relatively recent years, as an answer to food safety issues. Its history and development appears to be strongly bound to those of this field, and that is why this historical review will be particularly focused on the food industry case study. The use of predictive microbiology in food safety mainly aims to prevent or at least minimize the growth of micro-organisms in foods, a variable and sometimes heterogeneous medium. The characteristics of the so-called *"food predictive microbiology"* and the differences with the other main field of application of this science, biotechnology, will be explained later (Subsection 1.4).
Methods to limit and prevent the growth of micro-organisms in food were applied since thousands of years: refrigeration, use of salt, thermal treatment, etc. (McMeekin et al., 2002). Of course until relatively recent time the approach remained merely empirical, without having any knowledge or quantitative understanding of what was happening at cellular scale.
The first quantitative approach in the study of micro-organisms, consisting of mathematical models, arrived with the attempt of characterizing the growth and death of micro-organisms in time at constant environmental conditions. These models belong to the primary model class (Subsection 1.5.2). The first appearance of this approach probably dates back to 1922,

with Esty and Meyer's log-linear model for thermal death of *Clostridium botulinum* type A (McMeekin et al., 2002; Baranyi and Roberts, 2004).
A great advance came in 1936 with Scott (Scott, 1937), who investigated the dependency of the specific death rate of micro-organisms on environmental conditions such as water activity (an index between 0 and 1 that counts the relative lack or abundance of water in the medium). This additional step was important because it introduced the secondary models, i.e., models that include environmental changes in primary ones (Subsection 1.5.2). Scott also gave a great contribution to predictive microbiology as we know it, since he was the first to understand the potentiality of using collected data of microbial responses to predict food safety issues:

"A knowledge of the rates of growth of certain micro-organisms at different temperatures is essential to studies of the spoilage of chilled beef. Having these data it should be possible to predict the relative influence on spoilage exerted by the various organisms at each storage temperature. Further, it would be possible to predict the possible extent of the changes in populations that various organisms may undergo during the initial cooling off of the sides of beef in the meatworks when the meat surfaces are frequently at temperatures very favourable to microbial proliferation." (Scott, 1937)

Parallel to food predictive microbiology, also biotechnology studies started and had a fast development in the middle of the twentieth century: famous are the models for fermentation industry proposed by Monod (Monod, 1949). Although food predictive microbiology and biotechnology have many common characteristics and they aim to a progressive unification, there are also some differences in their fundamental hypothesis that made these two sciences to develop distinguished models.

1.2.2 Modern predictive microbiology

Publications in food predictive microbiology remained limited until the 1960-1970's. Then, many research studies in different areas started, e.g., fish spoilage (Olley and Ratkowsky, 1973a,b) and prevention of microbial intoxications (Genigeorgis, 1981; Roberts et al., 1981), giving a great impulse to the modeling approach. The biggest exploit of the predictive microbiology approach for food industry arrived in the 1980s, with prioritisation of food safety research by the governments in the USA, UK, other EU countries, Australia and New Zealand (McMeekin et al., 2002) and the creation of the first validated, commercialized database for kinetic data collection, FoodMicroModel™, in 1988 (Baranyi and Roberts, 2004).
Which were the main reasons that caused such a fast increase of trust toward this science, almost ignored before? Some possible causes, pointed out by different authors in the past, were:

(i) the marked increase during the 1980's in the incidence of major food poisoning outbreaks, particularly in ready-to-eat food, which led to an acutely increased public awareness of the requirement for a safe and wholesome food supply (Ross et al., 1999); *(ii)* the realization by many food microbiologists that the traditional methods for food quality and safety were inefficient, and, due to the long time to obtain results, their predictive potentiality was limited (McMeekin, 1993); *(iii)* the advent, development and widespread diffusion of computers, which provided a superior and widely accessible computational power (Buchanan, 1991) and permitted the creation of huge databases to collect kinetic data and model information for a wide range of micro-organisms (FoodMicroModel, *ComBase* [UK-US], SymPrevius [FR], the Pathogen Modeling Program (PMP) [US]); *(iv)* the change of mentality of people about food during the 1980's, with the diffusion of the green culture paradox, that asked for more healthy, more safe but less processed foods: *"Consumers are demanding miracle foods that are totally natural, have zero calories, zero fats and cholesterol, delicious taste, total nutrition, low price, environmentally friendly production, "green" packaging [...] and that guarantee perfect bodies, romance and immortality."*(Carol Brookins, Global Food and Agriculture Summit, 1999). Deepening the first statement, a great factor of change that made more immediate the need of a revolution in food industry, as in many other fields, was the globalization. In a world characterized by condensation, stratification and mobility of human population, that was experiencing an unprecedented rate of change as result of scientific and technological advances, the emergence and reemergence of foodborne microbial pathogens is not surprising. Typical bacterial features of adaptation and exploitation, due to their small size, speed of reproduction, phenotypic plasticity and genetic promiscuity, permit the micro-organisms to colonize almost every conceivable habitat on earth and to adapt to every environmental change (Lederberg, 1997). To reduce the contamination risk, fundamental importance had to go to food quality control and surveillance strategies. Predictive microbiology offered a feasible alternative to the traditional microbiological end-product testing to estimate shelf-life and safety, used until that moment in food industry. This detection method was expensive and ineffective, since it was based on a retrospective approach (McMeekin et al., 2002). It was particularly inefficient against very low infectious dose pathogens, a problem emerging in that period. Predictive microbiology instead promised the possibility of preventing the bacterial growth on food products with a theoretic approach, acting in a proactive way. Important elements of a proactive approach are the accumulation of quantitative information and kinetic data and an increased understanding of microbial physiology (McMeekin et al., 1997). According to Ross (Ross et al., 1999), the development of predictive models would greatly reduce the need for microbial examinations, and would enable quality and safety predictions to be quick and inexpensive.

Henceforward, predictive microbiology has known a constant growth of interest and development over years.

1.2.3 Applications

Ross (Ross et al., 1999) identified some useful applications for food industry that the future development of predictive microbiology could make possible: *(i)* prediction of the influence of product formulation and environmental conditions on food shelf life and safety; *(ii)* formulation of new processes and products to reach the required level of safety; *(iii)* evaluation of microbial responses to normal processing, that is subject to risk assessment due to its intrinsic variability, and to eventual lapses in processing or storage control, determining the appropriate remedial action;*(iv)* understanding the effect of processing at microbial scale through mathematical models, which enables the identification of the steps that contribute more to the overall risk and permits to minimize redundant processing while food safety is still guaranteed. Although these possible use patterns of predictive microbiology are promising, the real application in food industry is still far. Problems are currently present at different levels, starting already from practical issues such as experimental design and measuring of some parameters, e.g., water activity, and moving toward even more philosophical criticisms about the legitimacy of the predictive microbiology approach.

1.2.4 Criticisms

Many problems are connected with the high uncertainty of biological systems. A difficult issue is the characterization of the lag-phase, what McMeekin called "growth/no growth interface" (McMeekin et al., 2002). Since the best option to keep a limited concentration of micro-organisms is to prevent their growth, the lag-phase is very important in food safety and shelf-life determination of foods. The duration of this phase depends on many factors and it is very hard to be modeled. It can depend on singular environmental conditions, e.g., temperature, pH or water activity, but also on a combination of these. Considering the combined effect of many environmental factors, the Hurdle concept, formulated by Professor Leistner (Leistner, 1978, 1992), aims to determine the minimum value for each factor to prevent microbial growth. This phase is also characterized by an inherent variability, that has been proven to increase fast with the increase of response time. This variability is due to the precedent history of the population, which is difficult or impossible to be characterized for every single case. At a certain time and under certain conditions, cells contaminating food can be found to be damaged and require repair before starting growth, or they can have entered a suspended animation state or can be dead. In other words, the cells need an adaptation gap

to get used to the new environment, before they start to reproduce themselves. In situations characterized by variability and uncertainty, the development of good mechanistic models is impossible, and the formulation of good empirical kinetic models improbable (McMeekin et al., 2002). These considerations introduce the need of a probabilistic approach: under which sets of conditions could the food product be considered safe with a sufficient certainty? Parallel to the development of kinetic models, which are concerned with the determination of rates of response of the organisms, probability models were developed, that consider the probability or the likelihood of some event within a fixed period of time.
Although large databases have been created including primary and secondary models for many micro-organism species, the understanding of the intrinsic mechanisms that determine cellular responses to environmental changes is still quite poor. Most of the secondary models proposed are black-box models, i.e., merely based on experimental data, since their formulation requires less time and data. Even if these models were proven to be useful in practical applications, they should still represent a temporary, intermediate step, moving toward more biologically significant expressions. According to McMeekin (McMeekin et al., 2002), the recent trends toward increased use of black-box models like *artificial neural networks* may inhibit the search for mechanistic and biologically relevant models.

1.3 Biotechnology

1.3.1 History and applications

Although the term predictive microbiology is historically connected with food safety applications, the general approach of this science is far more general, and it could be applied to study every system that involves cellular populations. The biggest field other than food industry that involves the use of micro-organisms is biotechnology, i.e., the science that includes every technological application of biology and that uses living systems and organisms to develop or make useful products. From this last point of view, biotechnology exists since thousands of years, and it contributed largely to the growth and development of human kind, since even the cultivation of plants and agriculture could be considered the earliest biotechnology applications. Even focusing on the modern concept of biotechnology, which concerns the use of micro-organisms for useful productions, biotechnology dates back to Babylonians and Sumerians, since these populations started to use yeast to prepare alcohol. This was the beginning of a long history of fermentation processes, e.g., production of wine, vinegar, cheese, yogurt, bread, etc. Many of these applications are still fundamental in our society.
A fermentation process is defined as the overall set of biochemical reaction mechanisms to extract energy and form products under anaerobic conditions, even if aerobic processes are

also sometimes called fermentations. The word fermentation comes from the latin verb *fervere*, describing the action of yeast on malt or sugar.
In the mid-nineteenth century, Louis Pasteur pointed out the role of micro-organisms not only in food processes, but also in chemical industry for the production of fuels and fine chemicals. The branch of biotechnology that concerns industrial applications is called *white biotechnology*. The quantitative knowledge of biological phenomena and mechanisms enables biotechnology to optimize every aspect of the fermentation processes, using organisms to produce chemicals on industrial scale. Other examples of white biotechnology are for example the use of enzymes as catalyst to produce valuable chemicals or destroy hazardous or polluting ones. These considerations sketch the profile of an alternative chemical industry, more in line with the present environmental issues.
Industrial fermentation knew an incredible enhancement more recently, from the 1970's on, thanks to the development of genetic engineering and gene mounting. The first successes in this field came from the works of Paul Berg, Herbert W. Boyer and Stanley N. Cohen. This new science made it possible to force bacteria toward the production of chemicals of interest, e.g., amino acids.
Nowadays biotechnology is widely used in industrial fermentation, such as the production of alcohols and acetone, and in pharmaceutical industry (Najafpour, 2007). Moreover the rising demand for biofuels, an important alternative to petroleum-derived fuels, is giving a great boost both to the industrial production of ethanol and to genetic applications in agriculture.

1.3.2 Bioreactors

The heart of a biochemical fermentation process is the bioreactor or *fermenter* (Figure 1.1). A properly designed bioreactor should provide a controlled environment to optimize cellular growth and formation of products according with the particular biological system employed. The performance of the bioreactor depends on many variables, given the complexity of cellular systems, and many factors have to be controlled and regulated. Some of the fundamental features to consider, which can influence the efficiency of the fermentation, are: *(i)* biomass concentration, which should remain high enough to provide satisfactory yield; *(ii)* sterile conditions, to obtain a pure culture; *(iii)* agitation, to distribute and uniform the substrate and other concentrations in the available volume; *(iv)* heat transfer, to maintain the temperature as constant as possible and optimal for cellular growth; *(v)* creation of the correct shear condition, not too high share rate to avoid damages at the culture but high enough to prevent phenomena as flocculation, aggregation of the cells and cellular growth on the reactor walls; *(vi)* nutrient supply; *(vii)* product removal; *(viii)* product inhibition; *(ix)* aeration. To guarantee that the reactor operates respecting all of these requirements, sensors are needed to

measure at least pH, temperature and dissolved oxygen. Compared with a normal chemical process, a fermentation usually works at milder conditions, but it also presents much more strict operative ranges. Living organisms, in fact, are really sensible to the environmental conditions, and even a small change of these can cause great damages to the microbial culture, inhibiting the micro-organisms' growth or killing them. From these considerations the fundamental role that process control assumes in these kinds of processes emerges.

To design a bioreactor is a complex task, relying on scientific and engineering principles and many rules of thumb. Some fundamental aspects that require critical decisions for the bioreactor design are *(i)* the configuration, e.g., air-lift or stirred tank, and *(ii)* the mode of operation, continuous or discontinuous. As for normal chemical reactors, for every mode of operation there are many different configurations of bioreactors, each one with its own advantages and drawbacks, that must be weighted on the particular system to treat. The different configurations will not be treated in this study, and for every mode of operation only a stirred tank reactor will be considered.

Most of the bio-processes are discontinuous (Figure 1.2, a), since this mode of operation adapts well to the intrinsic dynamic behavior of cellular growth: lag phase, growth, stabilization and death (Section 1.5). The reactors for this kind of process can be fully discontinuous, i.e., batch reactors, or semi-discontinuous, i.e., fed-batch reactors. Batch reactors simply consist of a vessel, usually jacketed and mixed, and they have the dual advantage of low capital and operating costs. The working principle of this equipment is quite basic: the reactants are charged inside the tank and this is closed; when the desired environmental conditions are reached the reaction starts; the reaction proceeds for a determined time, then the process is stopped and the reactor is opened, yielding the final products. This equipment works cyclically, repeating this sequence of operations every time. To remove the products and eventually clean the vessel, after every cycle a dead time must be considered, which reduces the productivity of the reactor. Since it behaves as a closed system, it is mathematically described by a dynamic model. A critical aspect of this bioreactor is the agitation: turbulence is required to enhance material and heat exchange, but it can also cause foaming, which may lead to unknown contamination in the vessel. This can be prevented both chemically, adding antifoam agents to the medium, and mechanically, with foam breakers near the liquid surface. Baffles are often used to reduce vortexing and to improve the heat exchange in large volume vessels (Najafpour, 2007; Doran, 1995).

The fed-batch reactor (Figure 1.2, b) is very similar to the previous one, and most of the

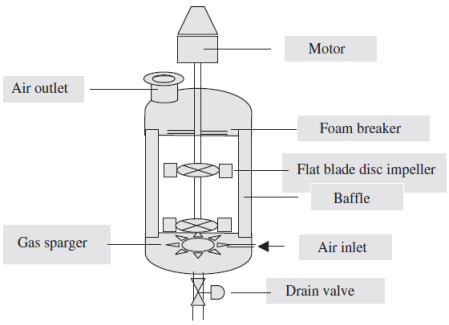


Figure 1.1: Stirred tank bioreactor.

considerations are still valid. The difference is in the working principle. The reactor in fact is closed only on one side: the reactants are fed during the whole process cycle, but there are no streams coming out. The level in the reactor increases over time, and the model is still necessarily dynamic. This working principle is useful because it maintains the basic characteristics of a batch reactor, but it permits to gradually feed the substrate. This way it avoids too high concentrations of substrate at the beginning of the cycle, which may lead to quickly reaching the stationary phase. A too fast growth in fact can cause the rise of limiting effects, e.g., an oxygen demand required for the growth which is too high for the mass-transfer capability of the reactor or the accumulation of substances which have an inhibitory effect on cell growth (Doran, 1995).
Even continuous reactors (Figure 1.2, c) are possible for biotechnology applications. The continuous bioprocesses are few, e.g., brewing, baker's yeast production and waste treatment. Considering a stirred tank configuration, a continuous operation is possible if the bacteria are suspended in the liquid medium and not immobilized on a support, otherwise it would be a problem to remove the dead cells. In this equipment the level is kept constant by controlling the input and output flow rates, and the mathematical model that describes the system is stationary (Doran, 1995).

1.4 Food predictive microbiology vs biotechnology

Since the approach of predictive microbiology is general and in theory extendable to both food safety and biotechnology, why are the models developed for these fields of application different? Some differences that explain the distinction between the two sciences (Baranyi and Roberts, 1994) are highlighted here.

- **Objective:** The aim of food predictive microbiology is to prevent or at least minimize bacterial growth. On the contrary, biotechnology usually attempts to maximize micro-

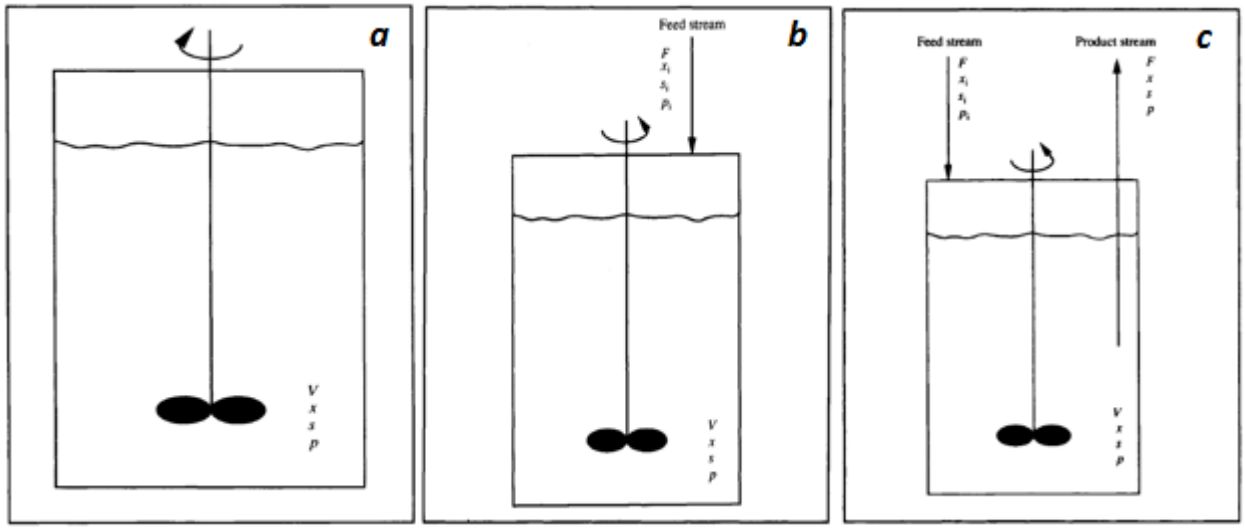


Figure 1.2: Batch, fed-batch and continuous stirred tank reactor.

bial growth to a certain extent or product formation.

- **Bacterial concentration:** Due to the industrial scale of the bioreactors and to the optimal conditions for cellular growth, bacterial concentration in bioprocesses is typically greater than $10^6 - 10^7\, cell/ml$. On the contrary, microbial concentration in foods is generally very low. As a consequence, some methods validated and commonly applied in biotechnology applications cannot be used for food safety issues.

- **Model phase:** Food predictive microbiology models focus on the lag-phase, investigating the inhibitory effects of environmental factors. Biotechnology models instead usually describe the exponential growth phase. Monod's model (Monod, 1942) for example, commonly used in biotechnology to represent the transition from the exponential to the stationary phase, loses its significance in food safety because substrate limitation is rarely important unless food spoilage is reached.

- **Medium:** The bioreactor is a controlled environment, in which many variables are controlled and manipulated. It is generally a mixed liquid medium, and it is almost everywhere homogeneous, unless for inherent non-ideal behaviors. Foods can be considered homogeneous only if they are liquid, and to collect physicochemical information about this "environment" is far more difficult. Hence food safety requires more simplifying hypotheses, empirical elements and mathematical-statistical methods.

Although the efforts of food microbiologists and biochemical engineers are directed toward different phases of microbial growth, unifying their works would make it possible to characterize the entire life of the cellular population. Furthermore, global approaches that treat the growth curve without focusing on particular phases could be useful for both fields of application.

1.5 Modeling

1.5.1 Empirical vs mechanistic models

A possible definition for a model is *"the description of a system, theory, or phenomenon that accounts for its known or inferred properties and may be used for further study of its characteristics"* (McMeekin et al., 2008). In engineering applications, a model is a simplified representation of the relationship between the effects, and the factors that are considered to be the causes of those effects. Translating this relationship as a mathematical function, the effects are the dependent variables, observations or responses, while the causes are the independent variables or inputs.

There are two main approaches to formulate a model starting from a set of measurements: *(i)* the first approach is simply based on the experimental data, trying to fit them; *(ii)* according with the second approach, a model should be based on theoretical considerations and should express some intrinsic relationship between dependent and independent variables. Models that belong to the first class are called black-box or empirical models, since they are general and they can be applied to any set of data coming from any system, without requiring any specific knowledge about system features. For the models of the latter class instead, called mechanistic models, measurements should be used just to validate the model after its formulation.

Few models are purely mechanistic, while fully empirical models are useful only in a small number of applications, since they provide information limited to the very specific conditions at which they were obtained. Most of the models could be classified halfway: empirical models that include some knowledge about the specific system or models based on biological considerations with empirical parameters. This way, both types of models can be used not only for prediction, which is generally the first aim of model formulation, but also for validation of some base hypotheses. Modeling in general does not only have an immediate purpose, limited to certain conditions and circumstances, but it can also be used to systematically improve the deep understanding of a system. Mechanistic models are preferred for this aim, since

they contain more *a priori* knowledge about the system, but to obtain a mechanistic model is much more time expensive and requires much more data for validation than to perform a regression for a black-box model. These observations justify the large use of black-box models in predictive microbiology.

1.5.2 Biological models

The different categories of models used in predictive microbiology will now be defined and analysed singularly. According with a general classification, biological models can be divided into: *(i) primary models*, which describe the dynamics of growth and death of populations of microorganisms in a constant environment; *(ii) secondary models*, which express the parameters of primary ones as function of relevant environmental factors; *(iii) tertiary models*, which collect the information from the primary and secondary models, incorporate the algorithm into a computer software package and enable practical applications and predictions.
Primary models are the most important models in predictive microbiology, and chronologically also the first being formulated. Generally the life of a bacterial population is characterized by different phases. Models were formulated both focused on singular phases and trying to give a global description of the cellular life-cycle. The initial stationary phase is called *lag* (Figure 1.3, Lag phase). During the lag phase growth has not started yet, since cells are not ready to reproduce. It can be thought of as a period of adaptation of the cell to the new environment that surrounds it. The extent of this temporary lapse depends on many factors which are not easy to isolate, and on the previous history of the organism, introducing a great variability which is usually modeled with a probabilistic approach. Since generally limited information is available on the *pre-inoculation* period, many models simply disregard this phase. After this phase the cell begins to reproduce, and the exponential growth starts (Figure 1.2, Growth phase). At invariant environmental conditions and while growth is not limited, the rate at which bacterial doubling occurs is constant. This second phase is also called *log phase*, since on a semi-logarithmic kinetic curve, i.e., a graph with natural logarithm of the cell number versus time, it is approximately linear. The third phase (Figure 1.2, Stationary phase) occurs when the bacterial population reaches the maximum carrying capacity of the environment and some growth-limiting factor gains importance, e.g., lack of essential nutrients, accumulation of inhibitory substances or a concentration of cells which is too high. The growth rate drops, until it becomes equal to the death rate: a new stationary phase results. Finally, when the substrate concentration is no longer high enough to permit cellular survival, the death phase begins (Figure 1.2, Death phase), with the number of cells decreasing over time. The corresponding behavior on the kinetic curve is called *tailing-off*.

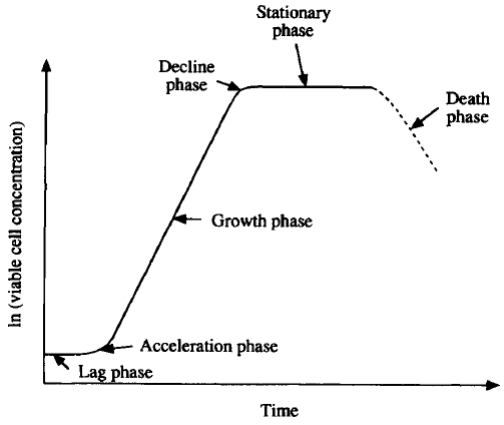


Figure 1.3: Typical batch growth curve.

The dynamic behavior resulting from the succession of all these phases and characterizing the growth of every bacterial culture justifies the wide use of batch and fed-batch bioreactors in biotechnology applications. The particular shape of the entire kinetic curve for bacterial growth in constant environmental conditions on a non-logarithmic graph can be well approximated with a sigmoid function. This was the form chosen by the Gompertz empirical model, which has been the most popular primary model until the mid-90s (Zwietering et al., 1990). An alternative strategy for the formulation of a primary model, which is still nowadays largely studied, is based on stoichiometric considerations. It will be presented later on in section 1.6. To formulate a secondary model, a reductionist approach has to be applied to determine a limited number of factors that can be considered to notably influence the cellular response. These factors were individuated, in decreasing importance order, as temperature, pH, water activity, concentration of preservatives and antimicrobials, and composition of the atmosphere. These factors do not act individually on the cellular growth, but they interact between each other. This feature is the base of the Hurdles concept (Subsection 1.2.4), that studies the synergic effects of environmental factors on growth inhibition (Leistner, 1978, 1992). A number of mechanistic models were proposed to describe temperature effects, e.g., the square-root model by Ratkowsky (Ratkowsky et al., 1983), based on thermodynamics, Arrhenius-based models (Broughall et al., 1983) and the Cardinal Temperature model of Rosso (Rosso et al., 1994). Nevertheless, most of the secondary models nowadays are still black-box models, particularly when the simultaneous variation of more than one variable is considered. Many studies suggested a quadratic polynomial regression to take into account the interaction terms without using too complex models. Geeraerd (Geeraerd et al., 2004) proposed an alternative quadratic polynomial regression able to also include physical constraints, moving toward grey-box models.

1.5.3 Black-box models

The use of black-box models is very common in predictive microbiology, particularly for secondary models. The formulation of an empirical model requires to select an appropriate functional form, a regression method and an error function. To perform a good regression, good experimental data are needed. Before the regression step, an optimal experimental design is also important. To perform a regression means to find the value of the parameters that minimizes the error function on the whole set of experimental data, i.e., such that the response of the empirical model is as close as possible to the experimental values. Mathematically, this is translated as solving an optimization problem. In this sense, the regression provides the best model between the infinite possible ones for the selected model shape. There are many regression methods, each one with its own strengths and drawbacks: the easiest and the most common one is the least-squares estimation.

There are some general rules to correctly formulate a black-box model: *(i)* parsimony; *(ii)* experimental data analysis; *(iii)* stochastic specification; *(iv)* validation. To explain the concept of parsimony, consider first a least-squares estimation method. In the linear case, if some hypotheses on the experimental data are satisfied, this method exhibits many desirable properties, e.g., the parameters selected are unbiased estimators of the real ones, they are the most efficient estimators, i.e., they possess the minimum variance between all the possible estimators, and they are normally distributed. Since these properties are not valid for the nonlinear case, it would be auspicable to always work with the easiest models available, since they are closer to linear behavior and to the cited properties. For nonlinear models it is also possible to perform a *parameterization*, i.e., a reorganization of the shape of the parameters that does not affect the prediction power of the model but could improve some of its properties. More generally, a trade-off problem must be considered for any regression. Depending on the error function chosen to be minimized, it is often true that the more complex a model is, i.e., the higher the number of terms and parameters, the better it fits the data. Theoretically, for a number of parameters equal to the number of measurements, perfect match is reached. But data fitting is not the aim of a regression model, since otherwise it would be limited to describe that particular set of measurements used to train it. A model is useful for practical applications if it can predict the response of the system under conditions that are at least slightly different from the original ones, and the higher the complexity of the model is, the more difficult it will be to generalize it. This is a problem also treated by philosophy in the past. The principle of parsimony embodied in Ockhams Razor, enunciated in the Middle Ages by William of Ockham, can be expressed as “Entities are not to be multiplied beyond necessity” (Ratkowsky, 1993).

Before concluding a regression step on a set of experimental data, it is often necessary to modify these data to better satisfy some hypotheses and for experimental design reasons. Some of the most common issues of dealing with the set of experimental data are: *(i)* particularly if the data are raw, i.e., they directly come from the field, they can include gross errors. A big issue consists of the individuation, study and possible elimination of the so-called outliers, data that are not well fitted by the model. Together with the experimental data, the error on the data should also be provided to the regression. *(ii)* The error on the data should be stochastic, i.e., with null mean and constant variance all over the experimental range. This last property is called homoscedasticity. *(iii)* The problem of collinearity or multi-collinearity between the data consists of using terms that are linear combinations of others: this should be avoided because it makes the global matrix of the regression problem, obtained by multiplying the matrix which contains all the experimental values for the inputs with the transposed matrix of the regression coefficients, ill conditioned. Consequently, to calculate the jacobian of this matrix, which is necessary to solve the optimization problem, becomes difficult. *(iv)* Unless the main interest is focused only around a portion of the experimental range, because the model can provide good fit on the entire experimental range and catch different behaviors, the experimental data should uniformly cover the overall range. If the data distribution is not uniform on the whole range, data can be modified to improve their distribution.
Since the estimators of the parameters obtained from the regression are stochastic variables, they posses an intrinsic variability. This variability should always be reported. If the variance on a parameter is too large, the estimators cannot be considered meaningful (Ratkowsky, 1993).
The validation step is useful before the regression and essential after it. The validation performed before regression is part of a procedure called *cross-validation*. The first step of cross-validation is called *training*, in which the regression is done by using only a fraction of the original set of data. This step is reiterated many times, selecting every iteration a different set of training data. Also the criterion to select these data can change. Then the proper *validation* step arrives, in which the sum of the objective values on all the iterations is compared between the different models. Cross-validation can help to select the shape of the model if no previous information is available about it, and it takes into account not only data fitting but also the capability of the model to be generalized. Outside the cross-validation context instead, the term validation is generally used to point out the phase, after regression, of verification of the estimated model, by trying to use it to fit sets of data different from the training set.

1.6 Metabolic network-based models

1.6.1 Metabolic networks

As was told, large use of empirical or semi-empirical models is made in predictive microbiology, both to describe growth and inactivation of species and to characterize the parameters of these models as function of the environmental conditions. Grey-box models, i.e., empirical models which also include biological information, are flexible, since for secondary models they can be extended to more variables with a multiplicative approach, and they guarantee good prediction properties while even being parsimonious. Nevertheless, they also have some drawbacks: *(i)* they don't perform well outside the range used for training, i.e., the extrapolation results are no longer accurate; *(ii)* although these models were proven to provide a good description of microbial dynamics under certain non-stressing conditions, avoiding for example rapid change of environmental factors, they fail when they are applied to more realistic and complex systems. In fact they consider simple liquid systems controlled by few variables, disregarding many complex phenomena able to modify the dynamic cellular behavior, e.g., background flora, microbial competition, stress and stress adaptation and physico-chemical properties of the medium. The influence of these and many other factors is considered by the so-called *completeness error*, which is one of the largest sources of error in predictive microbiology (McMeekin and Ross, 2002). From these considerations the necessity emerges to make the existing models more applicable and reliable, moving toward deterministic models. It is necessary to look inside the black-box and unravel the underlying biological mechanisms (Brul and Westerhoff, 2007). But to catch the intrinsic complexity of living organisms, the study at macroscopic scale, i.e., the level of the overall population, is not enough. Information must be considered also at the microscopic scale, i.e., at what happens inside the cell.
A viable way to do it is to exploit metabolic network-based models, making use of the results of previous studies that precisely characterize the metabolic networks of many species: *Metabolic networks are a blueprint of the reactions that occur inside the micro-organisms during the biochemical process* (Van Impe et al., 2012). These reactions can be divided in intracellular reactions, i.e., between intracellular metabolites, and transport reactions, i.e., reactions that bond the extracellular species to the intracellular metabolism. While for most cellular networks, e.g., signaling or protein-protein interaction networks, not all the metabolites or reactions are known and the relative studies are still focused on the identification of these elements, metabolic networks are a notable exception. In many metabolic networks the interaction topology, i.e., *most of the reactions that take a role in metabolism, the enzymes that catalyze them, the genes that encode the enzymes and how they interact stoichiometrically within a biochemical network* (Schuetz et al., 2007), is well established and allows to

describe almost the entire microbial genomes with mechanistic models (Schuetz et al., 2007). Although the structure of a network is quite intuitive, real networks are incredibly complicated to study, since the number of molecular components that interact is enormous, and the way they interact is highly nonlinear (Burgard and Maranas, 2003).

The primary models obtained by metabolic networks can be considered mechanistic models, since they don't include experimental parameters but they are just dynamic mass balances for the extracellular metabolites based on stoichiometric information of the metabolic reactions. The metabolic network approach contributes to make predictive microbiology a more exact science, moving from empirical to mechanistic models.

1.6.2 Flux Balance Analysis

The solution of the mechanistic model derived from a metabolic network requires to determine the rate of all the reactions that participate to cellular metabolism, both intracellular and transport reactions. Looking at the network structure, the reactions appear to link the metabolites. The reaction rates could be interpreted as material fluxes that move from the reactants to the products, and that is why in metabolic networks analysis these quantities are called *fluxes*. The fluxes are related to extrinsic and intrinsic factors via kinetic expressions, each one including a set of empirical parameters. Since the number of reactions in a network is usually notable, and for every reaction many kinetic parameters must be determined, even for small networks, e.g., in the order of 50 reactions, the operation of determining the parameter values would be experimentally and computationally very demanding. Techniques for model reduction are essential to enable the use of these models for simulation, prediction and control purposes. Furthermore, to obtain experimental data for these systems is very difficult, and their quality is often poor (a large experimental error must be considered). While the extracellular metabolite concentrations, which are involved in uptake and secretion fluxes, can be quite easily measured, it is more difficult to collect information about what happens inside the cell. The analysis method to obtain data about the intracellular components is relatively recent, and it is called *isotopomer analysis*. It consists of feeding 13C-labeled substrate and then analysing the labeling state by nuclear magnetic resonance (NMR) and/or gas chromatography/mass spectroscopy (GC/MS) measurements (Burgard and Maranas, 2003). An alternative to the kinetic approach to obtain the distribution of fluxes is possible, but a strong hypothesis is necessary, whose basic concept is well expressed by two sentences:

"Living organisms have evolved to maximize their chance of survival" (Darwin, 1899)

"Experimental evidence suggests that organisms have developed control structures to ensure

optimal growth in response to environmental constraints" (Edwards et al., 2001)

The first sentence motivates the redundancy of pathways that link the enzymes in a metabolic network, so that the removal of a single one due to for example environmental changes or limitations will not prevent an organism's ability to produce key components (Burgard and Maranas, 2003). In mathematical terms, this is translated by the fact that, even when adding physical and chemical constraints to the model, the distribution of fluxes will always identify a space of solutions. The second sentence instead allows to determine a single point in this feasible space. Always in accordance with Darwin's evolution theory, cellular organisms have maximized their growth performance as a response to selective pressure (Gianchandani et al., 2008). The flux distribution is obtained by solving a mathematical optimization problem, whose objective function expresses the organism's trial to survive or grow, depending on the environmental conditions and limitations.
The evolution hypothesis and the derived optimization problem, which are fundamental steps in the reductionist approach to enable the solution of metabolic networks-based models, constitute the basis of a method called *Flux Balance Analysis* (FBA).

1.6.3 The objective function

As we have seen in the previous section, FBA allows to compute the flow of cellular metabolites through the metabolic network, making it possible to predict the growth rate of the organism or the rate of production of a particular metabolite of interest (Feist and Palsson, 2010). As the constraint-based reconstruction of the genome-scale network is often still a large mathematical space of fluxes, this method requires an objective function to determine a flux distribution that corresponds to an optimal network state.
To determine the objective function means to understand what a micro-organism tries to do in a given environment. According with the evolutionary logic, i.e., everything in biology should be viewed through the eyes of evolution (Feist and Palsson, 2010), the answer implies some optimal performance based on the organism's past history. Common objective functions are for example maximization of biomass, maximization of ATP yield, minimization of glucose consumption etc. These functions were proven to work well on different micro-organisms and in different situations (Schuetz et al., 2007).
Many studies have been carried out in the past decades to investigate the objective function optimization with different metabolic networks in different conditions. These studies can be subdivided in two main categories: *(i)* studies examining which hypothesized cellular objective function best predicts cellular behavior through metabolic network optimization and comparison with experimental data and *(ii)* studies utilizing computational algorithms

to determine best-fit cellular objective functions. The first studies resulted in the knowledge that objective functions for an organism are likely condition-dependent and training data specific. To obtain an objective function able to describe the aim of the organism, numerous input, output and intracellular training-data fluxes must be analysed, with a case-by-case logic, in order to find the best overall predictive function. Notable is the study of Schuetz (Schuetz et al., 2007), that use a combinatorial engineering approach on *E.coli* to compare the performance of 11 objective functions, with different constraints, under 6 different growth conditions. The second kind of studies, in addition to the immediate predictive use, can be used to confirm the importance of some previously hypothesized objective function, improving the metabolic network understanding (Burgard and Maranas, 2003).
Even though there is a wide number of studies which explore the objective function in every possible situation, many others are still to appear: we are only beginning to decipher what cellular objectives actually are, and the search of these objectives is an ongoing area of research (Feist and Palsson, 2010).

Chapter 2

Materials and methods

2.1 Introduction

In this chapter the materials and methods used to produce the results of this study are presented. The knowledge of these procedures is fundamental for the understanding of the next chapter (3).
In the first section (Section 2.2) the model used for the simulations is introduced. Since it is a stoichiometry-based model, the starting point will be the metabolic network and the stoichiometric matrix derived from it. All the reductive hypotheses are highlighted and properly motivated. Then the focus moves to the optimization step, formulating the biologic optimization problem and defining the objective function and its basic constraints. Later on, a general overview of the optimization problems is provided (Section 2.3). After presenting the general characteristics of a simple optimization problem, particular attention goes to describing specific kind of problems encountered during the study, e.g., bilevel optimization problems and problems with complementarity constraints. Finally, the last section is focused on the black-box model formulation (Section 2.4). Different kinds of regression methods are presented, with their advantages and drawbacks. The entire procedure to correctly formulate an empirical model is briefly analyzed, starting from the experimental data and finishing with the model validation. This topic alone is incredibly wide. Since the aim of this study is not an analysis of the black-box model formulation, but black-box models are instruments inserted in a wider procedure, this part provides a brief description of the problems. To have more information about the empirical model formulation and how to manage experimental data, *data mining* studies can be consulted (Buzzi-Ferraris and Manenti, 2010).

2.2 Metabolic reaction network-based modeling

2.2.1 Stoichiometric model

A cellular organism can be schematized as a mathematically open system, that exchanges both matter and energy with the environment that surrounds it. Molecular species are taken up by the cell, they are consumed by the metabolism reactions to produce other components and finally the products leave the cell. Metabolic networks provide a schematic representation of both metabolites and the reactions between them (Figure 2.1). This kind of system, although easily explained above, is actually incredibly complicated to study, since the number of molecular components that interact is enormous, and the way they interact is highly nonlinear (Burgard and Maranas, 2003).

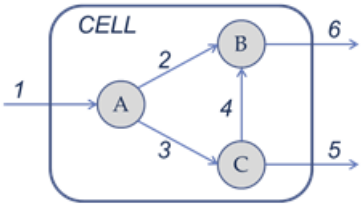


Figure 2.1: A schematized organism: (1) is the uptake flux; (2),(3),(4) are the intracellular reactions that link the intracellular metabolites A, B and C; (5),(6) are the desorption fluxes.

A biological system can be described at three different levels, which allow to catch different aspects of the phenomena involved: *(i)* **the macroscopic level**, on which the characteristics and the behavior of the overall cellular population are described; *(ii)* **the mesoscopic level**, on which small populations or part of a bigger population are studied; *(iii)* **the microscopic level**, on which information at cellular or intracellular level is considered. The link between macroscopic and mesoscopic scale is very important, since mesoscopic scale allows to catch the heterogeneity of the system, which can no longer be considered to be made up of identical cells. To properly describe a biological system, information from both microscopic and macroscopic level, moving through the mesoscopic one, have to be considered and linked together.
At the macroscopic scale, the system is described with the following dynamic equations:

$$\frac{dN_{macro}}{dt} = \mu \cdot N_{macro} \tag{2.1}$$

$$\frac{d\mathbf{C}_{S,macro}}{dt} = -\sigma \cdot N_{macro} \tag{2.2}$$

$$\frac{d\mathbf{C}_{P,macro}}{dt} = \pi \cdot N_{macro} \tag{2.3}$$

$$\sigma = \frac{\mu}{Y_{X/S}} + \frac{\pi}{Y_{P/S}} \tag{2.4}$$

N_{macro} is the macroscopic number of cells, expressed as a concentration $[CFU/ml]$ (CFU = Colony Forming Unit); $\mathbf{C}_{S,macro}\,[mol/ml]$ and $\mathbf{C}_{P,macro}\,[mol/ml]$ are the macroscopic concentrations of substrates and products, respectively; $\mu\,[1/h]$ is the specific rate of cell growth; σ and $\pi\,[\frac{mol}{CFUh}]$ are the specific rates of substrate consumption and product formation; finally $Y_{X/S}$ and $Y_{P/S}$ are the yield coefficients of biomass on substrate and product on substrate. This system can be also rewritten in a more compact form as:

$$\frac{d\mathbf{C}_{macro}}{dt} = \mathbf{R}_{macro} \cdot N_{macro} \tag{2.5}$$

where $\mathbf{C}_{macro}$ is a vector containing N_{macro}, $\mathbf{C}_{S,macro}$ and $\mathbf{C}_{P,macro}$; $\mathbf{R}_{macro}$ is a vector containing the respective specific rates.

Moving to the microscopic level, the system of equations becomes:

$$\frac{d\mathbf{C}_{int}}{dt} = \mathbf{S}_{int} \cdot \mathbf{v} - \mu \cdot \mathbf{C}_{int} \tag{2.6}$$

$$\frac{d\mathbf{C}_{ext}}{dt} = \mathbf{S}_{ext} \cdot \mathbf{v} \cdot N \tag{2.7}$$

where $\mathbf{C}_{ext}$ and $\mathbf{C}_{int}$ are the microscopic concentrations of extracellular and intracellular metabolites, respectively, expressed in $[mol/CFU]$; N is the microscopic number of cells, considered as an extracellular metabolite; $\mathbf{v}$ is a vector of all the metabolic fluxes; finally $\mathbf{S}$ is a matrix of dimensions $n \times m$ called the *stoichiometric matrix* of the network, with n the number of metabolites and m the number of reactions. It can also be split in two parts, corresponding to extracellular and intracellular components. The term $\mu\,\mathbf{C}_{int}$ is called "dilution term": physically speaking, the more the number of cells increases, the less the intracellular concentration of metabolites will be, when their total amount is kept constant. The contribution of this term is much smaller than the stoichiometric term, and thus it will be disregarded in the following treatise.

The two ODE systems are linked together through kinetic equations. These equations relate the fluxes to extrinsic and intrinsic factors through a set of parametrized expressions f with parameters $\mathbf{\Phi}$:

$$\mathbf{v} = f(\mathbf{C}_{ext}, \mathbf{C}_{int}, T, pH, \ldots, \mathbf{\Phi}) \tag{2.8}$$

The determination of the kinetic models appears to be quite an insurmountable step at the present moment (1.6.2).

The problems faced now are: how to find an alternative way to kinetic expressions to compute the extracellular concentrations at the microscopic scale? How to correlate the microscale to the macroscale, obtaining the macroscopic concentrations from the microscopic ones? Two hypotheses are needed. *(i)* **Average cell**: all the cells are equal to one average cell, i.e., they all have the same flux distribution. This means $N_{macro} = N$. *(ii)* **Homogeneity of the population**: the microscopic concentrations are homogeneous all over the medium, $\mathbf{C}_{ext} = \mathbf{C}_{macro}$. This is sufficiently satisfied assuming a mixed liquid medium.

The new system we obtain when applying these hypotheses is:

$$\frac{d\mathbf{C}_{macro}}{dt} = \mathbf{S}_{ext} \cdot \mathbf{v} \cdot N_{macro} \tag{2.9}$$

$$\frac{d\mathbf{C}_{int}}{dt} = \mathbf{S}_{int} \cdot \mathbf{v} \tag{2.10}$$

$$\mathbf{v} = f(\mathbf{C}_{macro}, \mathbf{C}_{int}, T, pH, \ldots, \Phi) \tag{2.11}$$

A further hypothesis can be done to simplify the system. *(iii)* **Pseudo-steady state:** observing that the intracellular dynamics are much faster than the extracellular/macroscopic dynamics, they can be disregarded and the intracellular concentrations can be assumed at pseudo-steady state: $\mathbf{S}_{int}\mathbf{v} = 0$. The dynamic equations for the intracellular metabolites are converted into algebraic equations. Furthermore, $\mathbf{S}_{int}$ is a redundant matrix, i.e., not all the reactions it describes are independent. Most of them actually are linear combination of a few independent fluxes. The original set of fluxes can be re-written as $\mathbf{v} = \mathbf{K} \cdot \mathbf{u}$, where $\mathbf{u}$ are called *free fluxes*, and $\mathbf{K}(n\text{x}(n-m))$ is a suitable basis for the null space of $\mathbf{S}_{int}$. The null space of a matrix can be easily computed using the MATLAB® function `null`. The null space, and then the set of free fluxes also, is not unique. After these assumptions, the final system appears as:

$$\frac{d\mathbf{C}_{macro}}{dt} = \mathbf{S}_{ext} \cdot \mathbf{K} \cdot \mathbf{u} \cdot N_{macro} \tag{2.12}$$

$$\mathbf{u} = f(\mathbf{C}_{macro}, T, pH, \ldots, \mathbf{\Phi}) \tag{2.13}$$

This procedure is fully explained and motivated by Van Impe (Van Impe et al., 2012).

2.2.2 Optimization problem

The concentration of the extracellular metabolites depends on the intracellular dynamics (Subsection 2.2.1). To compute this concentration, it would be necessary to link the macro and the microscale through kinetic equations. Since the amount of reactions in a network is relevant, and for each reaction more kinetic parameters have to be determined, a large amount of experimental data would be needed to determine all kinetic expressions. Our present possibilities make it difficult to obtain such a set of experimental data.
How to overcome this problem and determine the dynamic profile of the fluxes? Once the pseudo-steady state hypothesis is set (Subsection 2.2.1), the intracellular flux distribution can be determined by solving a stationary problem for every time point. A continuous time profile can then be obtained by interpolating all the individual time points, but the problem to determine the flux distribution from the stationary problem still remains. If $n - m$ is the number of free fluxes obtained by simplifying the original set of fluxes, i.e., the dimension of the null-space of the original stoichiometric matrix $\mathbf{S}$, the problem consists of determining a unique flux distribution. Mathematically this means finding a unique point in an unbounded space $\mathbb{R}^{n-m}$. It is now shown how it is possible to bound and reduce the feasible space using general physical considerations.

- **Irreversibility matrix:** Not all the fluxes $\mathbf{v}$ can assume every value. Some of them are bounded to only positive values, since the corresponding reaction in the network is irreversible. This can be expressed mathematically as:

$$\mathbf{IR} \cdot \mathbf{K} \cdot \mathbf{u} \geq 0 \tag{2.14}$$

  The $\mathbf{IR}$ matrix is called *irreversibility matrix*. It is a matrix of dimension $(n_{irr} \,\mathrm{x}\, n)$, with n_{irr} the number of irreversible fluxes and n the total number of fluxes. Each row identifies an irreversible reaction, and it has one on the column corresponding to the irreversible flux, zero for the non-irreversible ones. The following irreversibility matrix, for example, would be applied on a network with $n = 4$ fluxes, among which 1, 2 and 4 are irreversible.

$$\mathbf{IR} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2.15}$$

Practically, this matrix multiplied for the vector of fluxes selects some rows from the entire vector **v**, so that these fluxes are bounded to positive values. This constraint modifies the feasible space, forcing it to assume a particular shape called *convex polyhedral cone* (Figure 2.2).

- **Upper bounds:** The polyhedral cone selected by the irreversibility constraint is still an unbounded space. The value of the fluxes can be bounded *(i)* with physical considerations, e.g., there is a maximal value on the substrate uptake flux due to transport limitations; *(ii)* with experimental statistical considerations, e.g., a value above a certain limit was never observed experimentally for a certain flux. Despite the different nature of these considerations, they can be mathematically expressed in the same way, as:

$$\mathbf{K} \cdot \mathbf{u} \leq \mathbf{UB} \tag{2.16}$$

where **UB** is a vector of dimension n. Imposing the previously presented constraints, the feasible area is reduced to a completely bounded space. It is called *Bounded convex polyhedral cone* (Figure 2.2).

The size of the feasible space can be reduced by applying the constraints just introduced, but still infinite feasible flux distributions remains. A unique solution can be found by searching the optimal value of a specific objective function in the feasible area. This step requires a further and important hypothesis: according with Darwin's evolution theories (Darwin, 1899), the cellular organism has become a perfect optimizer of its own metabolism, pushing it to achieve at best some objectives to maximize its chance of survival (Subsection 1.6.2). Finally, the stationary problem to determine the flux distribution for every time point can be written as a constrained optimization problem:

$$\min_{\mathbf{u}} \quad f(\mathbf{u}) \tag{2.17}$$

$$s.t. \quad \mathbf{IR} \cdot \mathbf{K} \cdot \mathbf{u} \geq 0 \tag{2.18}$$

$$\mathbf{UB} - \mathbf{K} \cdot \mathbf{u} \geq 0 \tag{2.19}$$

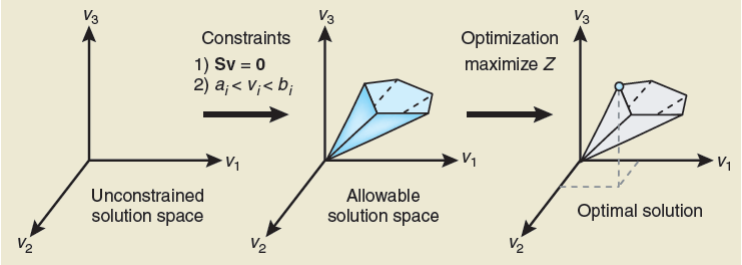


Figure 2.2: Conceptual basis of the FBA problem: (A) the entire space $\mathbb{R}^{n-m}$ of free flux distribution; (B) the bounded convex polyhedral cone, introducing the irreversible and upper bound constraints; (C) the solution of the optimization problem, finding the optimum of the objective function in the feasible area.

2.2.3 Dynamic Metabolic Flux Analysis (dMFA)

Leighty (Leighty and Antoniewicz, 2011) highlighted the pseudo state assumption used in the model as one of the reasons why flux balance analysis is not commonly used in industry yet. Since most of the industrial bioprocesses are fed-batch fermentations, the PSS assumption is in contrast with the inherent dynamic nature of a system where cells continually adapt to a changing of environmental conditions. In Equation (2.12) both the fluxes and concentrations are continuous functions of time. The *in vivo* experiments provide isolate points in time of what are called *states*, i.e., external metabolite concentrations or fluxes. From these isolated points, a unique smooth profile must be obtained on time. This step can be considered as a problem called *Dynamic Metabolic Flux Analysis* (dMFA).

In literature one approach to solve the problem (Niklas et al., 2011) is to perform a regression on the measurements, obtaining additional time points apart from the original ones. For each of these points a standard static FBA problem is solved (2.17), computing the flux values. By representing the dynamic problem as a series of disconnected points, each one statically solved, important information on the dynamic nature of the system is lost. Another approach (Leighty and Antoniewicz, 2011) consists of combining the fluxes with the biomass concentration to non-specific fluxes, which are then parametrized as piecewise linear functions. The dynamic problem is this way turned in a non-dynamic, non-linear parameter estimation problem, that can be solved analytically. Also this approach presents some drawbacks, e.g., the fact that the conversion from non-specific fluxes to biologically descriptive fluxes results in a loss of information and the non-smoothness of the profile obtained from the piecewise linear description. To overcome most of the presented disadvantages, a new method has been investigated by Vercammen, solving a dynamic input optimization problem with a least-squares objective function, that represents the true non-linear dynamic problem:

$$\min_{\mathbf{u(t)},\mathbf{x}_0} \quad \sum_{i=1}^{n_{time}} \sum_{j=1}^{n_{out}} (\frac{y_j(t_i) - m_{ij}}{\sigma_{ij}})^2 \tag{2.20}$$

$$s.t. \quad \frac{d\mathbf{x}(t)}{dt} = \mathbf{S}_e \cdot \mathbf{K} \cdot \mathbf{u}(t) \cdot \mathbf{X} \cdot \mathbf{x}(t) \tag{2.21}$$

$$\mathbf{x}(0) = 0 \tag{2.22}$$

$$\mathbf{y}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)) \tag{2.23}$$

$$\mathbf{IR} \cdot \mathbf{u}(t) \geq 0 \tag{2.24}$$

This optimization problem requires state-of-the-art tools for this kind of problems. The method consists of B-spline parametrization with incremental spline knot insertion.
Although the results obtained from the dMFA solution will be used in this work, the procedure will not be described in detail.

2.2.4 Objective function

The underdetermined problem of identifying the distribution of fluxes is solved by defining a feasible area based on biological constraints, and optimizing an objective function inside this space. Still there is the problem of identifying the objective function, since no information is a priori available about what the cell is trying to accomplish in every instant. In other words, the problem of determining the objective function answers the question: which particular combination of fluxes is the cell trying to optimize in that particular environmental frame? Can this function be justified with biological considerations?
In literature, a list of objective functions with their respective biological meaning is provided by Schuetz (Schuetz et al., 2007) (Figure 2.3). In this article, the flux distributions obtained when optimizing different objective functions with a direct biological interpretation are compared with the experimental flux distributions obtained in different environmental situations. This way it is verified which objective function appears to better describe the behaviour of the micro-organism for a specific situation or which describes the behaviour of the micro-organism in a wider range of conditions.

An alternative approach is also presented in literature, called *ObjFind* (Burgard and Maranas, 2003). Because the fluxes have to approximate the corresponding experimental values, they can be correctly determined by adding a step of minimization of the least squares error between the estimated and the experimental values. Additional coefficients are then added to the objective function. A possible objective function could be simply made up of a summation of fluxes in which each term is multiplied with a coefficient. The global problem is a bilevel

Table III Objective functions implemented in constraint-based FBA

Objective function[a]	Mathematical definition	Explanation	Rationale	Reference
Max biomass[b]	$\max \frac{v_{biomass}}{v_{glucose}}$	Maximization of biomass yield	Evolution drives selection for maximal biomass yield ($Y_{X/S}$)	(van Gulik and Heijnen, 1995; Edwards and Palsson, 2000b; Price *et al*, 2004)
Max ATP	$\max \frac{v_{ATP}}{v_{glucose}}$	Maximization of ATP yield	Evolution drives maximal energetic efficiency ($Y_{ATP/S}$)	(van Gulik and Heijnen, 1995; Ramakrishna *et al*, 2001)
$\text{Min} \sum v^2$	$\min \sum_{i=1}^{n} v_i^2$	Minimization of the overall intracellular flux	Postulates maximal enzymatic efficiency for cellular growth (analogous to minimization of the Euclidean norm)	(Bonarius *et al*, 1996; Blank *et al*, 2005a)
Max ATP per flux unit[c]	$\max \frac{v_{ATP}}{\sum_{i=1}^{n} v_i^2}$	Maximization of ATP yield per flux unit	Cells operate to maximize ATP yield while minimizing enzyme usage	(Dauner and Sauer, 2001)
Max biomass per flux unit[c]	$\max \frac{v_{biomass}}{\sum_{i=1}^{n} v_i^2}$	Maximization of biomass yield per flux unit	Cells operate to maximize biomass yield while minimizing enzyme usage	
Min glucose	$\min \frac{v_{glucose}}{v_{biomass}}$	Minimization of glucose consumption	Evolution drives selection for most efficient usage of substrate	(Oliveira *et al*, 2005)
Min reaction steps[c]	$\min \sum_{i=1}^{n} y_i, y_i \in \{0,1\}$	Minimization of reaction steps	Cells minimizes number of reaction steps to produce biomass	(Melendez-Hevia and Isidoro, 1985)
Max ATP per reaction step[c]	$\min \frac{v_{ATP}}{\sum_{i=1}^{n} y_i^2}, y_i \in \{0,1\}$	Maximization of ATP yield per reaction step	Cells operate to maximize ATP yield per reaction step	
Min redox potential[d,e]	$\min \frac{\sum v_{NADH}}{v_{glucose}}$	Minimization of redox potential[f]	Cells decrease number of oxidizing reactions thus conserving their energy or using their energy in the most efficient way possible	(Knorr *et al*, 2007)
Min ATP production[d,e]	$\min \frac{\sum v_{ATP}}{v_{glucose}}$	Minimization of ATP producing fluxes[g]	Cells grow while using the minimal amount of energy, thus conserving energy	(Knorr *et al*, 2007)
Max ATP production[d,e]	$\max \frac{\sum v_{ATP}}{v_{glucose}}$	Maximization of ATP producing fluxes[h]	Cells produce as much ATP as possible	(Heinrich *et al*, 1997; Ebenhoh and Heinrich, 2001; Knorr *et al*, 2007)

[a]Both maximization of biomass objectives (absolute and per flux unit) require no *a priori* assumptions. For all other objectives the specific growth rate was set to the experimentally determined value under each condition.
[b]Often also referred to as optimization of growth rate (Price *et al*, 2004).
[c]*n* refers to the overall number of reactions in the network, that is 98 in the present case.
[d]Reaction name is that specified in Supplementary Table 1; '_R' refers to the *reverse* reaction.
[e]All reversible reactions in Supplementary Table 1 were converted to two irreversible reactions resulting in a final stoichiometric model of 60 metabolites and 151 reactions.
[f]Reactions: *gapA, aceEF, maeA, sucAB, mdh, udhA, fdhF, fdoGHI, fdnGHI, ldhA, adhE,R, mhpF_R, adhP_R, adhC_R, maeB, zwf, gnd, icd, pntAB, frdABCD, sdhAB, ddh, sdhABCD_R.*
[g]Reactions: *pgk, pykA, pykF, sucCD, atpAH, ackA, adeB, ldcD, purT.*
[h]Reactions: *pgk, pykA, pykE, sucCD, atpAH, ackA, ackB, ldcD, purT.*

Figure 2.3: A list of objective functions with their biological explanation (Schuetz et al., 2007).

optimization problem, where the fluxes are the optimization variables of the inner problem, which optimizes the biological objective function, and the CoIs are the optimization variables of the outer problem, which minimizes the distance between the estimated fluxes and the experimental fluxes. In this sense, the outer problem can be interpreted as a regression, while the mechanistic information is contained in the inner problem. Since in the original **ObjFind** method the objective function is a linear combination of the fluxes, both the objective function and the constraints are linear functions of the optimization variables. The inner problem, when formulated in this way, is a linear programming (LP) optimization problem:

$$\min_{\mathbf{c}\in\mathbb{R}^q} \quad \sum_{i=1}^{n-m} (u_i - u_i^{exp})^2 \tag{2.25}$$

$$\text{s.t.} \quad \sum_{i=1}^{q} c_i = 1 \tag{2.26}$$

$$\max_{\mathbf{u}\in\mathbb{R}^{n-m}} \quad f(\mathbf{u},\mathbf{c}) \tag{2.27}$$

$$\text{s.t.} \quad \mathbf{IR}\cdot\mathbf{K}\cdot\mathbf{u} \geq 0 \tag{2.28}$$

$$\mathbf{UB} - \mathbf{K}\cdot\mathbf{u} \geq 0 \tag{2.29}$$

where n is the total number of fluxes $\mathbf{v}$ in the network and m is the number of intracellular metabolites, so that $n-m$ is the number of free fluxes u; q is the number of CoIs $\mathbf{c}$, which are normalized to one; f is the objective function, which depends on the free fluxes through the CoIs; $\mathbf{IR}$ is the irreversibility matrix, while $\mathbf{K}\cdot\mathbf{u} = \mathbf{v}$; finally, $\mathbf{UB}$ is a vector of dimension n which contains the upper bounds for every flux of $\mathbf{v}$. The optimization solver generally requires to set upper bounds for each optimization variable. To set the upper bound constraint is different because it bounds not only the free fluxes $\mathbf{u}$, which are optimization variables, but every flux $\mathbf{v}$ which can be obtained as a combination of the free fluxes. Being an LP, this bilevel problem is solved though reformulation with the strong duality theorem (2.3.3). The aspect of the optimization problem after reformulation is:

$$\min_{\mathbf{c},\mathbf{u}\in\mathbb{R}^{n-m},\,\boldsymbol{\lambda}_1\in\mathbb{R}^p,\,\boldsymbol{\lambda}_2\in\mathbb{R}^n} \sum_{i=1}^{n-m}(u_i - u_i^{exp})^2 \tag{2.30}$$

$$\text{s.t.} \sum_{i=1}^{n-m} c_i = 1 \tag{2.31}$$

$$\mathbf{c}^T \cdot \mathbf{u} = -\boldsymbol{\lambda}_2^T \cdot \mathbf{UB} \tag{2.32}$$

$$\mathbf{IR} \cdot \mathbf{K} \cdot \mathbf{u} \geq 0 \tag{2.33}$$

$$\mathbf{UB} - \mathbf{K} \cdot \mathbf{u} \geq 0 \tag{2.34}$$

$$\mathbf{c} - (\mathbf{IR} \cdot \mathbf{K})^T \cdot \boldsymbol{\lambda}_1 + \mathbf{K}^T \cdot \boldsymbol{\lambda}_2 = 0 \tag{2.35}$$

$$(c_j \geq 0 \; \forall j \in 1, \cdots n-m)$$

$$(\lambda_{1,i} \geq 0 \; \forall j \in 1, \cdots p)$$

$$(\lambda_{2,j} \geq 0 \; \forall j \in 1, \cdots n)$$

where p is the number of irreversible fluxes, i.e., the number of rows of the **IR** matrix, and λ_1 and λ_2 are dual variables or *lagrangian multipliers*. In this case the optimization problem is an LP.

As highlighted during the ObjFind study, this formulation mainly adapts to find the optimization coefficients **c** for a given flux distribution. Since the fluxes are not free to vary anymore, the biological meaning of the objective function falls entirely on the coefficients. Considering a linear objective function, these coefficients can be interpreted as the weights of the fluxes in the summation, and they are called *Coefficients of Importance* (CoIs). The set of coefficients establishes the relative importance of each flux in the network. To clarify this concept, the example of the biomass function follows. It is intuitive, in an evolutionary logic, that a possible objective function for a micro-organism is the maximization of its own growth. Since, based on metabolic network biological analysis, it is possible to also formulate a combination of fluxes which expresses the biomass growth, the ObjFind approach could be used both to verifying if the distribution of CoIs resembles the hypothesized one, confirming and enhancing the biological understanding of the network, or, inserting the biomass growth in the network as an additional flux, to verify if it assumes an important weight in the summation.

On the contrary, the inverse problem, i.e., using a set of CoIs to obtain the fluxes, does not perform well. In fact, the optimization of the CoI-based fluxes subject to stoichiometric balances results in many different flux distributions, due to the *degeneracy* of the LP optimization problem. Most of these solutions are not biologically meaningful, and they don't represent physical solutions. In practice, in accordance with the ObjFind method, to obtaining the

fluxes from the coefficients is not possible.

2.3 Optimization

2.3.1 Material and software

In this section the formulation of an optimization problem is introduced, particularly focusing on constrained bilevel optimization. Although this kind of problem is the heart of the FBA method previously presented, and its solution is a fundamental step for the use of metabolic network-based models, a deep mathematical treatise is not the main aim of this study. For more complete information about optimization, this study refers to the "Script for Numerical Optimization" by Moritz Diehl (Diehl, 2009), while the solution of mathematical problems with complementarity constraints takes the entire 11^{th} chapter of "NONLINEAR OPTIMIZATION, Concepts, Algorithms and Applications to Chemical Processes" by Lorenz T. Biegler (Biegler, 2010).
Two different numerical computing environments were tested to solve the optimization problem: MATLAB® and CasADi. Exploiting the characteristics of each environment, the optimization problem was solved with different strategies.
The MATLAB® optimization toolbox offers many optimization solvers, each one for a specific kind of problem. In particular, the solvers available for a constrained optimization problem with a generic objective function are: `linprog` for LP, `quadprog` for QP and `fmincon` for NLP. The computational effort required to solve them increases in the same order.
For the aims of this study, the main advantage of using MATLAB® instead of CasADi is the possibility of solving the bilevel optimization problem without reformulating it, using embedded functions. At each time and variance iteration, the least squares can be minimized with the free fluxes obtained by maximizing the objective function with its constraints. The kind of internal problem, and consequently the type of solver between the cited ones, depends uniquely on the selected objective function. The advantage of not writing the KKT conditions consists of avoiding the solution of the MPCC problem.
Using CasADi instead, the formulation of the problem is more complex, since the original bilevel problem must necessarily be reformulated using the duality theorem for LP or the KKT conditions for NLP (2.3.3). The reformulated problem was easily implemented in CasADi using symbolic optimization variables, which allow to compute the derivatives with automatic differentiation instead of numerical differentiation, as MATLAB® does. Using symbolic variables, the different operations are performed analytically, and only at the end the numerical values are substituted. Theoretically, automatic differentiation requires more computational time than numerical differentiation, but it is more precise, since it is exact

up to the limit of machine precision (2.3.4). Furthermore, the optimization solver used in CasADi, named IPOPT (**I**nterior **P**oint **OPT**imizer), offers a wider range of options than MATLAB®'s `optimset`. By modifying the default options the user is allowed to play with many parameters of the optimization. The reformulation of the bilevel problem introduces the difficulty of solving an MPCC problem. If the solver proceeds too slowly or does not find a feasible solution, it is convenient to try to relax the complementarity constraints with one of the methods proposed by literature (Subsection 2.3.5).

2.3.2 Constrained bilevel optimization

The form of a standard constrained optimization problem is:

$$\min_{x\in\mathbb{R}^n} f(x) \tag{2.36}$$

$$\text{s.t. } g(x) = 0 \tag{2.37}$$

$$h(x) \geq 0 \tag{2.38}$$

with $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$, and $h : \mathbb{R}^n \to \mathbb{R}^q$ smooth functions.
A constrained bilevel problem is a combination of two different constrained problems embedded. The solution of the internal problem is a necessary condition for the solution of the other one. To mathematically express this correlation, the solution of the inner problem is used as a constraint for the outer problem. The general form of a constrained bilevel optimization problem is:

$$\min_{x\in\mathbb{R}^n} \quad f_{outer}(x, y) \tag{2.39}$$

$$\text{s.t.} \quad g_{outer}(x, y) = 0 \tag{2.40}$$

$$h_{outer}(x, y) \geq 0 \tag{2.41}$$

$$\min_{y\in\mathbb{R}^p} \quad f_{inner}(y) \tag{2.42}$$

$$\text{s.t.} \quad g_{inner}(y) = 0 \tag{2.43}$$

$$h_{inner}(y) \geq 0 \tag{2.44}$$

This kind of problem can be equivalently solved by reformulating the inner problem with some additional conditions. These conditions have to substitute the inner problem in the constraints of the outer one. The equivalent conditions depend on the type of inner problem under study (2.3.3).

2.3.3 Inner problem reformulation: duality and KKT conditions

To solve a bilevel optimization problem, it is necessary to reformulate the inner problem with equivalent conditions. These conditions depend on which class of optimization problems the inner one belongs to: LP or NLP. The quadratic programming (QP) problems are a particular case of NLP, and the reformulation is the same. For LP, given a constrained optimization problem in the form (2.36), called *Primal problem*, which solution is indicated as p^*, the *Lagrange function* is defined as:

$$\mathcal{L}(x, \lambda, \mu) = f(x) - \lambda^T g(x) - \mu^T h(x) \tag{2.45}$$

Where $\lambda \in \mathbb{R}^m$ and $\mu \geq 0 \in \mathbb{R}^q$ are the *Lagrange multipliers* or *dual variables*. Since $\mu \geq 0$, from the definition of the Lagrange function it results that if x^* is a solution of the problem, then $\mathcal{L}(x^*, \lambda, \mu) \leq f(x^*)$.
For fixed λ and μ, the unconstrained infimum of the Lagrange function over x is called "Lagrange dual function":

$$q(\lambda, \mu) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda, \mu) \tag{2.46}$$

It is easy to prove that $q(\lambda, \mu) \leq p^*$.
The dual problem is the following:

$$d^* = \max_{\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^q} \quad q(\lambda, \mu) \tag{2.47}$$

$$s.t. \qquad \mu \geq 0 \tag{2.48}$$

For every possible primal problem, *weak duality* is valid:

$$d^* \leq p^* \tag{2.49}$$

If the Primal problem is also convex, then the more powerful *strong duality* theorem is valid:

$$d^* = p^* \tag{2.50}$$

Using the duality theorem, the LP inner problem can be rewritten as its dual problem, which can be added as a constraint to the outer problem.
Considering a constrained nonlinear optimization problem (NLP), the first order necessary optimality condition (FONC) can be expressed using a series of equations called *Karush-Kuhn-Tucker* (KKT) conditions. Some basic definitions are needed. *(i)* **Active Set:** an inequality constraint $h_i \geq 0$ is called *active* at x^* iff $h_i(x^*) = 0$; otherwise it is called *inactive*.

The set $\mathcal{A}(x^*)$ of i active inequality constraints is called the *active set. (ii)* **LICQ:** the linear independence constraint qualification (LICQ) for the constrained problem holds at x^* iff all vectors $\nabla g_i(x^*)$ for $i \in \{1, \ldots, m\}$ & $\nabla h_i(x^*)$ for $i \in \mathcal{A}(x^*)$ are linearly independent (usually this condition is always satisfied and it is not necessary to verify it).
If x^* is a local minimum of the optimization problem (2.36) and LICQ holds at x^*, then $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^q$ exist, where:

$$\begin{aligned} \nabla f(x^*) - \nabla g(x^*)\lambda^* - \nabla h(x^*)\mu^* &= 0 && (2.51)\\ g(x^*) &= 0 && (2.52)\\ h(x^*) &\geq 0 && (2.53)\\ \mu &\geq 0 && (2.54)\\ \mu_i^* h_i(x^*) &= 0,\ i = 1, \ldots, q && (2.55) \end{aligned}$$

The KKT conditions are useful to indirectly solve a bilevel optimization problem. Since these conditions are the FONC for the constrained problem (they are equivalent to the condition $\nabla f(x^*) = 0$ for unconstrained optimization problem), the solution of the inner optimization problem can be guaranteed by imposing these conditions as constraints for the outer problem.

2.3.4 Automatic differentiation (AD)

Solving an optimization problem always requires to calculate the jacobian (∇) or the hessian (∇^2) of a function f. The derivatives can be evaluated in different ways. *(i)* **Finite differences**

$$\nabla f(x)^T p \approx \frac{f(x + tp) - f(x)}{t} \tag{2.56}$$

Critical in this method is the choice of t. A trade-off must be reached between small values of t, which make the derivatives more precise but subject to numerical noise, and big values, which make the linearization influence prevalent. A rule of thumb to obtain a good compromise consists of choosing $t = \sqrt{\varepsilon_{macheps}}$, where $\varepsilon_{macheps}$ is the precision of the function evaluation. As a limit, $\varepsilon_{macheps}$ is equal to the machine precision. This means that the derivative uses only half of the digits compared to the function evaluation, losing precision. This problem becomes even more critical with the second order derivative. *(ii)* **Automatic differentiation (AD)**: automatic differentiation has the advantage to compute the derivatives up to machine precision. It makes use of symbolic expressions that concatenate different basic operations. Each operation can be computed subsequently by the calculator. The algorithm for function evaluation via elementary operations is reported. This algorithm is chosen because it is particularly useful to show the principle of automatic differentiation.

Algorithm: Function Evaluation via Elementary Operations

Input : $x_1, \ldots, x_n$

Output : x_{n+m}

for $i = (n+1)$ **to** $(n+m)$ **do**

$x_i \longleftarrow \varphi_i(x_1, \ldots, x_{n-1})$

end for

Where φ_i is an elementary operation and x_i is a combination of the function variables obtained by the precedent iterations. Automatic differentiation uses the chain rule for derivatives, separately differentiating each elementary operation. Two possible strategies exist for AD, with two different algorithms, respectively:

- **Forward Mode**
  It is based on the following chain rule formula:

$$\frac{dx_{n+i}}{dt} = \sum_{j<n+i} \frac{\partial \varphi_{n+i}}{\partial x_j} \frac{dx_j}{dt}, \; i = 1, \ldots, m \tag{2.57}$$

  The original function f is decomposed to elementary operations φ_i. Both the variables and the operations are differentiated with respect to a virtual time, and they are called *dot quantities*. The algorithm is the following:

Algorithm: Forward automatic differentiation

Input : $\dot{x}_1, \ldots, \dot{x}_n$

Output : $\dot{x}_{n+m}$

for $i = 1$ **to** m **do**

$\dot{x}_{n+i} \longleftarrow \sum_{j<n+1} \frac{\partial \varphi_{n+1}}{\partial x_j} \dot{x}_j$

end for

  The virtual time expedient has the advantage that each intermediate variable is used in the following iteration. Storing all variables at the beginning is not necessary. Computing derivatives with the forward AD algorithm is slightly more computationally expensive than with finite differences, but it is exact up to machine precision.

- **Backward Mode**

  It is based on the alternative chain rule formula:

$$\frac{df}{dx_i} = \sum_{j>max(i,n)} \frac{df}{dx_j} \frac{\partial \varphi_j}{\partial x_j} \tag{2.58}$$

  Instead of *dot quantities* $\dot{x}$, *bar quantities* $\bar{x}$ are used: partial derivatives of the final output, i.e., the derivative of the entire function f with respect to the intermediate quantity variable ($\bar{x}_j = \frac{df}{dx_i}$).

Algorithm: Reverse automatic differentiation

Input : all $\dfrac{\partial \varphi_j}{\partial x_i}$

Output : $\bar{x}_1, \ldots, \bar{x}_n$

$\bar{x}_1, \ldots, \bar{x}_{n+m-1} \longleftarrow 0$

$\bar{x}_{n+m} \longleftarrow 1$

for $j = n + m$ down to $n + 1$ **do**

for all $i < j$ **do**

$\bar{x}_i \longleftarrow \bar{x}_i + \bar{x}_j \dfrac{\partial \varphi_j}{\partial x_i}$

end for

end for

  The same considerations as in the forward algorithm are valid, but two additional observation can be done: *(i)* the reverse AD algorithm requires more space to store all the intermediate variables x_i at the beginning of the procedure; *(ii)* it is also far more efficient than forward AD for large n.

2.3.5 Inequality constrained optimization: complementarity

Equation (2.55), is called *complementarity* condition. It must be considered together with equations (2.53) and (2.54), which are inequality constraints. If these inequality constraints are present, the optimization problem becomes a mathematical problem with complementarity constraints (MPCC). The analytic formulation of the complementarity condition is:

$$h_{i(x*)} \perp \mu \tag{2.59}$$

This is equivalent to writing one of the following conditions:

$$h_i(x^*) = 0 \textbf{ OR } \mu_i = 0 \tag{2.60}$$
$$h_i(x^*)\mu_i = 0 \ (\mathbf{h_i}^T \cdot \boldsymbol{\mu} = 0) \tag{2.61}$$
$$h_i(x^*)\mu_i \leq 0 \ (\mathbf{h_i}^T \cdot \boldsymbol{\mu} \leq 0) \tag{2.62}$$
$$(h_i(x^*) \geq 0), \ (\mu \geq 0), \ i = 1, \ldots, m$$

The complementarity constraints violate the LICQ (Subsection 2.3.3), i.e., linearly independent gradients for equality and inequality constraints. For every feasible solution x^*, $h_i(x^*) = 0$ and $h_i(x^*) \cdot \mu = 0$. Since the constraint qualification does not hold anymore, the multipliers of the MPCC are unbounded and non-unique. Consequently to these considerations, directly solving a MPCC can result to be hard. Different algorithms have been proposed to avoid these problems, reformulating the MPCC with alternative *relaxed* forms. All these algorithms are iterative. They start with solving the problem with very low precision, and they gradually increase the precision by using the previous solution as starting point. This way they gradually lead the solution to the real value. The final precision must be set by the user.
Different reformulations are reported by Biegler (Biegler, 2010):

$$\begin{aligned} \textbf{Reg}(\epsilon): \quad \min \quad & f(x) && (2.63)\\ \text{s.t.} \quad & g(x) = 0 && (2.64)\\ & h(x) \geq 0 && (2.65)\\ & \mu = 0 && (2.66)\\ & h_i(x) \cdot \mu_i \leq \epsilon, \text{ for } i = 1, \ldots, m && (2.67) \end{aligned}$$

$$\begin{aligned} \textbf{RegComp}(\epsilon): \quad \min \quad & f(x) && (2.68)\\ \text{s.t.} \quad & g(x) = 0 && (2.69)\\ & h(x) \geq 0 && (2.70)\\ & \mu = 0 && (2.71)\\ & \mathbf{h}(x)^T \cdot \boldsymbol{\mu} \leq \epsilon && (2.72) \end{aligned}$$

$$\mathbf{RegEq}(\epsilon): \quad \min \quad f(x) \tag{2.73}$$
$$\text{s.t.} \quad g(x) = 0 \tag{2.74}$$
$$h(x) \geq 0 \tag{2.75}$$
$$\mu = 0 \tag{2.76}$$
$$h_i(x) \cdot \mu_i = \epsilon, \text{ for } i = 1, \ldots, m \tag{2.77}$$

where ϵ is a positive scalar which gradually approaches to 0, e.g., concatenating the values for the different iterations in a vector, $\boldsymbol{\epsilon} = [10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, \ldots]$. The value of the last iteration must reach the desired precision.
Slightly different is the subsequent reformulation, called *Penalty* formulation:

$$\mathbf{PF}(\rho): \quad \min \quad f(x) + \rho \cdot \mathbf{h}(x)^T \cdot \boldsymbol{\mu} \tag{2.78}$$
$$\text{s.t.} \quad g(x) = 0 \tag{2.79}$$
$$h(x) \geq 0 \tag{2.80}$$
$$\mu = 0 \tag{2.81}$$

In this case the additional vector ρ assumes positive increasing values, $\rho = [10, 10^2, 10^3, 10^4, \ldots]$, until the solution reaches the desired precision.
The number of elements of the ϵ or ρ vectors and their value must be decided case-by-case to reach a compromise between increase of the computational time and improvements of the solution. Studies on the efficiency of the different reformulations have been led, and some results are reported in Figure 2.4.

2.4 Linear regression

2.4.1 Pre-processing of data

To regress means literally to come back: in mathematics and statistics, to make a regression consists of trying to obtain a mathematical expression that describes how some variables, called *dependent variables*, vary as function of others, called *independent variables*.
The first step to perform a regression is to find an appropriate functional form suitable for the kind of relationship to represent. The regression step does not modify the structure of the model. The functional expression can be derived from physical considerations, as it usually happens for kinetic expressions, or it can be simply guessed and compared with different possible shapes to select the best one. The comparison is performed during the cross-validation

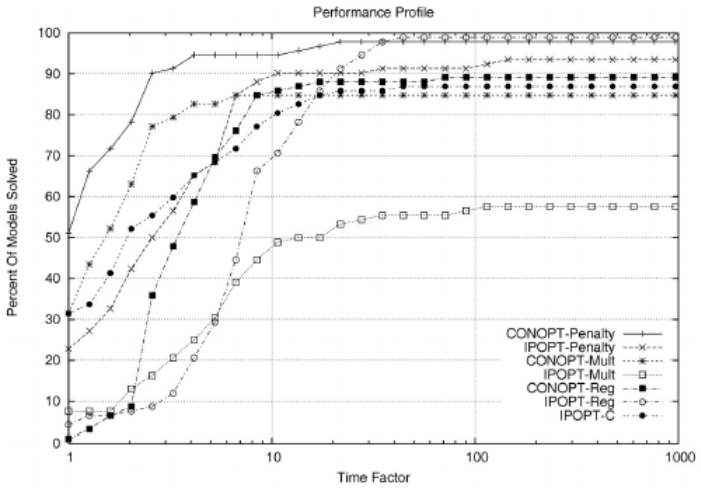


Figure 2.4: A comparison of the performances of the different MPCC formulations tested on a collection of 92 relevant problems with two reliable solvers, IPOPT and CONOPT (Biegler, 2010).

step (Subsection 2.4.5).

Before facing the real regression problem, it can be useful to have a look at the set of data to be used. Pre-processing consists of optional additional operations on the data to make the following steps easier. Pre-processing includes: *(i)* mean-centering; *(ii)* scaling; *(iii)* improvement of the distribution of the data on the experimental range.

In most regression problems, the variables are used in the mean-centered form (Figure 2.5, B). This form is better conditioned than the original one. For each column of the $\mathbf{X}$ matrix of the independent variables and/or of the $\mathbf{Y}$ matrix of the dependent variables, i.e., the experimental set of each independent/dependent variable, each element is diminished with the average of that column.

Scaling can be useful or even necessary depending on the set of data (Figure 2.5, C). It is necessary if the variables are expressed in different units. It is useful if there are differences of orders of magnitude between the values of the variables, since in this case it makes the problem better conditioned. Scaling is usually done by dividing every element of a column by the standard deviation computed on the population of that column. A different kind of scaling also exists, consisting of a sort of weighting. The element is not divided by the standard deviation but by a selected parameter which takes into account the importance of that term. If a variable is considered to influence the model in a lower degree, it will be weighted less in the parameters identification step.

To properly perform a regression on a set of data, these data have to be more or less equally

distributed on the experimental range. The distribution of the data can be easily visualized by subdividing the range of each variable in smaller intervals and plotting the number of experiments which fall in every interval. If the distribution on the whole range is visibly far from being homogeneous, it can be improved slightly by modifying each column.
Other pre-processing operations exist, but they are not presented here, since they were not used in this study. In the following treatise every x or y variable is assumed to be mean-centered and variance-scaled already.

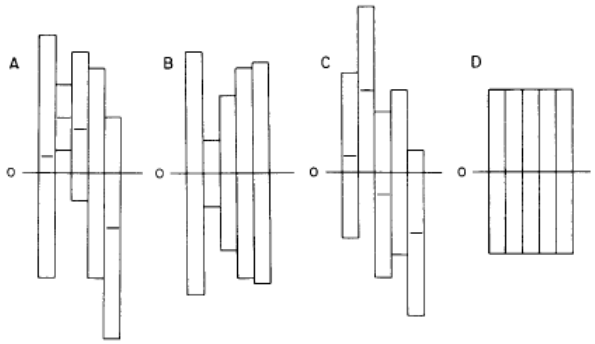


Figure 2.5: Data preprocessing. The data for each variable are represented by a variance bar and its center: a) Most raw data look like this; b) The results after mean-centering only; c) The result after variance-scaling only; d) The result after mean-centering and variance-scaling.

2.4.2 Multiple Linear Regression (MLR)

In this study only linear regression was used. A linear regression is a regression problem with a linear model. It is called multiple linear regression (MLR) if it is used to estimate more than one dependent variable, i.e., if $\mathbf{Y}$ is no longer a column vector but it is a matrix. To present multiple linear regression problems, the general expression of a linear model is first introduced:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{2.82}$$

or

$$\mathbf{Y} = \mathbf{F(X)B} + \mathbf{E} \tag{2.83}$$

Looking at the first equation: $\mathbf{Y}$ $(n \times p)$ has p independent variables on the columns and n repeated experiments for each variable on the rows; $\mathbf{X}$ $(n \times m)$ has a similar structure to $\mathbf{Y}$, but with as many rows as the number of independent variables m; $\mathbf{B}$ $(m \times p)$ contains

the set of estimated parameters for each variable x_i on the rows and y_i on the columns; finally, $\mathbf{E}$ $(n \times m)$ is the matrix of the residuals between the predicted values $\mathbf{Z} = \mathbf{XB}$ and the experimental ones $\mathbf{Y}$ (Figure 2.6). The second equation is equivalent to the previous one, but it highlights that the word *linear* refers to the parameters B of the model, not necessary to the independent variables. The $\mathbf{F(X)}$ term is a general functional expression of the independent variables, linear or not. In this case, m is no longer the number of the independent variables but it is the number of terms of the function. The shape of the model $\mathbf{F(X)}$ is a priori selected by using knowledge of the specific system or with an additional cross-validation step (Subsection 2.4.5). The following treatise always refers to the $\mathbf{X}$ matrix of the independent variables presuming it can also be a combination of the columns of the original $\mathbf{X}$.

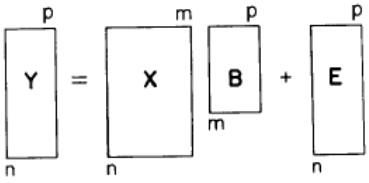


Figure 2.6: A representation of the dimensions of the MLR problem.

The regression problem corresponds to the particular case $n > m$. In this case the available information is not enough to reduce the feasible area to a unique solution, i.e., a point, and an infinite number of feasible solutions is possible, i.e., a space. The problem is called *underdetermined*, and it is solved as an optimization problem, selecting the set of parameters $\mathbf{B}$ which provides the best fit of the experimental data, i.e., which minimizes a particular selected error function. An error function consists of a summation of the distances between the experimental data and the respective points of the estimated model. Many error functions exist, each one with its own properties. The most common one is the *least squares* function, which inserted in the optimization problem provides:

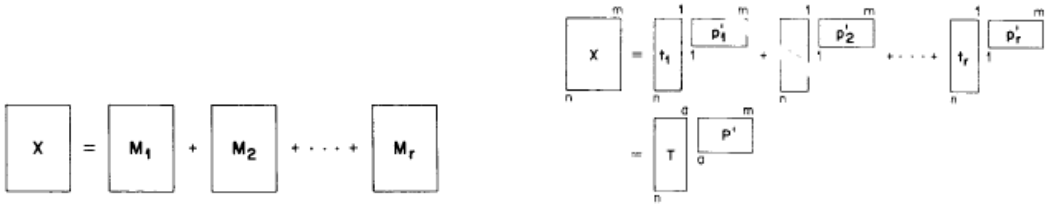
$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \parallel z_{i,j} - y_{i,j} \parallel_2^2 \tag{2.84}$$

This optimization problem can be analytically solved. The analytic solution is:

$$\mathbf{B} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{Y} = \mathbf{X}^+ \cdot \mathbf{Y} \tag{2.85}$$

The $\mathbf{X}^+$ matrix is called *pseudo-inverse*. Obtaining the pseudo-inverse is possible only if the $\mathbf{X}^T \cdot \mathbf{X}$, called *covariance matrix*, is invertible, i.e., the matrix is positive definite ($\mathbf{X}^T \cdot \mathbf{X} \succ 0$). Problems with the inversion can derive from linear dependency between different variables. This situation is called *collinearity* or *multicollinearity*.

The LP optimization problem of the linear regression can be modified to take into account



(a) (b)

Figure 2.7: representation of the dimensions of the matrix $\mathbf{X}$: a) $\mathbf{X}$ is expressed as summation of r matrices of rank 1; b) $\mathbf{X}$ is expressed as summation of PCA components.

physical considerations. These considerations can be mathematically expressed by adding additional constraints. The new solution $\mathbf{B}$ respects the imposed constraints, but it is not analytic anymore. The optimization problem then needs a specific solver for constrained LP or NLP, depending on whether the constraints are linear or not.

2.4.3 Principal Components Analysis (PCA)

The rank of a matrix is a number expressing the true underlying dimensionality of a matrix. Assuming the rank of $\mathbf{X}$ to be r, $\mathbf{X}$ can be written as a sum of r matrices of rank 1. Each of these matrices is obtained by the product of two vectors $\mathbf{t}_i(n \times 1)$ and $\mathbf{p}_i(m \times 1)$ (Figure 2.7):

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_2' + \cdots + \mathbf{t}_r\mathbf{p}_r' \tag{2.86}$$

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}' \tag{2.87}$$

The vectors $\mathbf{t}_i$ and $\mathbf{p}_i$, respectively called *scores* and *loading*, possess a precise physical meaning. This meaning is evident when visualizing the bidimensional case, with only two independent variables, $\mathbf{x}_1$ and $\mathbf{x}_2$. When representing the elements of the $\mathbf{X}$ matrix on the $\mathbf{x}_1$ and $\mathbf{x}_2$ plane, PCA finds the line which best fits the data. The original bidimensional $\mathbb{R}^2$ space is reduced to only one variable, and the distance of the data from this new axis is considered as noise. PCA tries to explain most of the variance on the data with a reduced number of components. The maximum number of components is equal to the number of original independent variables. In this specific case the model explains 100% of the variance. Most likely, the largest part of the variance can be explained by using just a few components, while the components which contribute just a little can be disregarded according with a fixed criterion.

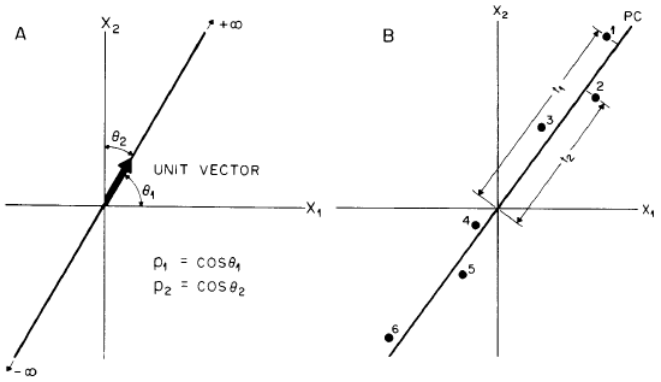


Figure 2.8: A principal component analysis in the case of two variables: (A) loadings are the angle cosines of the direction vector; (B) scores are the projection of the sample points on the PCA direction. The data are mean-centered.

Looking at the figure (Figure 2.8), the loadings ($m \times 1$) are the cosines of the new components with respect to the axes of the original space, and the scores ($n \times 1$) are the projection of each data on the corresponding component.
There are many algorithms to perform PCA on a set of data, each one with advantages and drawbacks for particular applications. The NIPALS algorithm 2.4.6 is important because it is one of the most complete and elegant algorithms for prediction. Furthermore Partial least squares (PLS) regression is based on this algorithm (Subsection 2.4.4). In the NIPALS algorithm scores and loadings are iteratively computed pair-by-pair for one component at a time, starting from the most important, i.e., the one that explains most of the variance.

PCA can also be used to obtain predictions of $\mathbf{Y}$, and then it is called principal components regression (PCR). In this context, PCA can be interpreted as a step to re-organize $\mathbf{X}$ before the regression, in order to reduce its dimension and improve its properties. In fact, since the scores $\mathbf{T}$ are the new axes for the regression and they are orthogonal by definition, the problem of multicollinearity does not exist anymore after PCA.

$$\mathbf{Y} = \mathbf{T} \cdot \mathbf{B} + \mathbf{E} \tag{2.88}$$

$$\mathbf{B} = (\mathbf{T}^T \cdot \mathbf{T})^{-1} \cdot \mathbf{T}^T \cdot \mathbf{Y} \tag{2.89}$$

However, PCA mostly remains a method to re-organize the independent variable set. It is not optimized for predictions, since it does not take into account the Y.

2.4.4 Partial Least Squares (PLS)

The Partial least squares regression is based on the NIPALS algorithm 2.4.6. The method consists of two relations, one called *external*, that individually reorganizes the independent and the dependent variables, decomposing them in their respective scores and loadings, and the latter called *internal*, that links both $\mathbf{X}$ and $\mathbf{Y}$ blocks. While principal components analysis treats each block separately, the partial least squares method gives the blocks information about each other, making the model to slightly rotate. The fact that both the independent and the dependent variables are considered at the same time and they influence each other makes PLS particularly useful for prediction.
PLS is now analyzed step by step. An oversimplified algorithm for PLS would perform two separate PCAs, using the NIPALS algorithm on both $\mathbf{X}$ and $\mathbf{Y}$:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}' + \mathbf{E} \tag{2.90}$$

$$\mathbf{Y} = \mathbf{U} \cdot \mathbf{Q}' + \mathbf{F} \tag{2.91}$$

Then it would regress between the respective loadings $\mathbf{T}$ and $\mathbf{U}$. This algorithm individually considers the independent and dependent variables, without combining the information from the two blocks.
The actual NIPALS algorithm for PLS is far more complicated. It follows an iterative procedure which mixes the computation of the principal components for $\mathbf{X}$ and $\mathbf{Y}$.
The iterative algorithm which gives each block information about the other one improves the inner relation $\hat{\mathbf{u}} = \mathbf{b}_i\mathbf{t}$. Since the order used for PCA has changed, the scores have lost their orthogonality. Orthogonality is not mandatory, but is useful for the inversion of the covariance matrix. The algorithm can be slightly modified to recover this property. The modified algorithm *(i)* introduces the additional vector $\mathbf{w}$ of the *weights* ($\mathbf{w}$ does not have any index since it is just an intermediate variable which is overwritten at every iteration) and *(ii)* adds the following extra loop after convergence on $\mathbf{t}_i$:

$$\mathbf{p}_i' = \frac{\mathbf{t}_i'\mathbf{X}}{\mathbf{t}_i'\mathbf{t}_i} \tag{2.92}$$

$$\mathbf{p}_i' = \frac{\mathbf{p}_i'}{\| \mathbf{p}_i' \|} \tag{2.93}$$

$$\mathbf{t}_i = \mathbf{t}_i \| \mathbf{p}_i' \| \tag{2.94}$$

2.4.5 Cross-validation

The number of components obtained from PCA is at maximum equal to the number of terms of the $\mathbf{F}(\mathbf{X})$ function. Not all the components are important, and some of these just describe noise. Defining a criterion to select the number of significant components is fundamental. Since PCA is an iterative method which builds up one component a time, this means to interrupt the procedure after a while.

How to select an appropriate number of components? There are different stopping criteria, generally based on a measure of the prediction error, e.g., the *sum of squares error* (SSE) or the *prediction residual sum of squares* (PRESS). Both examples provide an estimation of the predictive power of the model.

SSE is defined as:

$$\text{SSE} = \sum_{j=1}^{n} \frac{(z_j - y_j)^2}{n - i} \tag{2.95}$$

where n is the number of experimental data and i is the number of components selected. A stopping criterion could be (Figure 2.9):

$$\text{STOP IF: } \frac{\parallel SSE_i - SSE_{i-1} \parallel}{SSE_{i-1}} \geq 0.95 \tag{2.96}$$

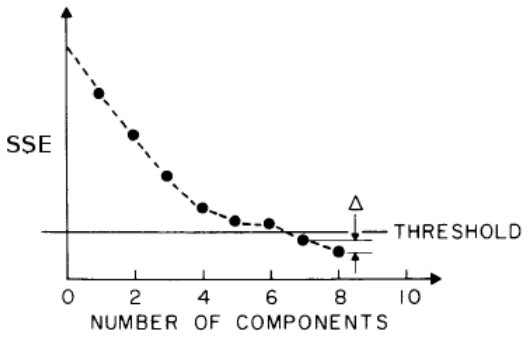


Figure 2.9: SSE vs number of components. The stop criterion requires to select a Δ value.

Another possible criterium is based on the specific property of the definition (2.95) not to monotonously decrease when increasing the number of components h (Figure 2.10). This is due to the choice of the denominator, which takes into account the number of components.

$$\text{STOP IF: } SSE_i - SSE_{i-1} \geq 0 \tag{2.97}$$

The previously presented procedure to select the number of components is useful during the construction of the model. For prediction instead, another method can be applied, called

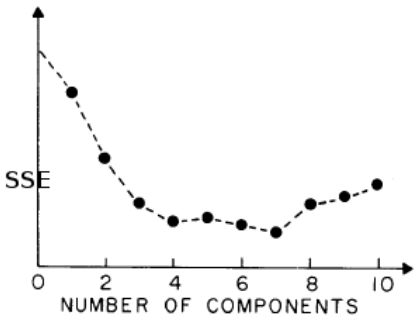


Figure 2.10: SSE vs number of components. The stop criterion is based on the existence of a local minimum in the profile.

cross-validation. Cross-validation does not just perform a simple regression for a selected model, but it also tests its predictive power. The regression is repeated many times, training each time on a different fraction $1-\alpha$ of the entire set of experiments and computing an SSE_i on the entire experimental range. In practise, the remaining fraction α is used to validate the obtained model. The final SSE of the model is an average of all the obtained SSE_i. How to properly select α? The choice depends on the structure and the characteristics of the particular set of data. This step is called *data folding.*
An additional step can be performed based on cross-validation. The PCA components result from a linear combination of the original variables, whose weights are selected by the NIPALS algorithm for PCA. If any of the weights is very small, the corresponding variable can be disregarded. Since each coefficient is estimated as many times as iterations are selected by data folding, a distribution of values is available for every component after cross-validation. A test of significance is performed on the mean value b of this distribution, for example a t-test. The t-test verifies the null hypothesis $\mathrm{H}_0 : b = 0$, returning a p-value. If this value is higher than a fixed one, the null hypothesis is assumed to be true. In this case the parameter can be disregarded, and the dimension of $\mathbf{X}$ is reduced.

2.4.6 Appendix: NIPALS algorithm for PCA and PLS

Algorithm: NIPALS for PCA

Input : $\mathbf{X}$

Output : $\mathbf{t}_i, \mathbf{p}_i$

for $i = 1$ **to** n

$\mathbf{t}_{h,i} = \mathbf{x}_j$

for $h = 1$ **to** k **do**

$$\mathbf{p}'_{h,i} = \frac{\mathbf{t}'_{h,i}\mathbf{X}}{\mathbf{t}'_{h,i}\mathbf{t}_{h,i}}$$

$$\mathbf{p}'_{h,i} = \frac{\mathbf{p}'_{h,i}}{\| \mathbf{p}'_{h,i} \|}$$

$$\mathbf{t}_{h+1,i} = \frac{\mathbf{X}\mathbf{p}'_{h,i}}{\mathbf{p}'_{h,i}\mathbf{p}_{h,i}}$$

if $\mathbf{t}_{h+1,i} \simeq \mathbf{t}_{h,i}$

BREAK

end if

end for h

end for i

where $\mathbf{x}_j$ is a random column of $\mathbf{X}$; i is the iteration on the PCA component; h is the iteration to reach convergence, and k is the maximum number of iterations.

Algorithm: NIPALS for PLS

Input : $\mathbf{X}, \mathbf{Y}$

Output : $\mathbf{t}_i, \mathbf{p}_i$

$\mathbf{E}_0 = \mathbf{X}$

$\mathbf{F}_0 = \mathbf{Y}$

for $i = 1$ **to** n

$\mathbf{u}_{h,i} = \mathbf{y}_j$

for $h = 1$ **to** k **do**

$$\mathbf{p}'_{h,i} = \frac{\mathbf{u}'_{h,i}\mathbf{X}}{\mathbf{u}'_{h,i}\mathbf{u}_{h,i}} \quad \left(\mathbf{w}' = \frac{\mathbf{u}'_{h,i}\mathbf{X}}{\mathbf{u}'_{h,i}\mathbf{u}_{h,i}}\right)$$

$$\mathbf{p}'_{h,i} = \frac{\mathbf{p}'_{h,i}}{\parallel \mathbf{p}'_{h,i} \parallel} \quad \left(\mathbf{w}' = \frac{\mathbf{w}'}{\parallel \mathbf{w}' \parallel}\right)$$

$$\mathbf{t}_{h,i} = \frac{\mathbf{X}\mathbf{p}'_{h,i}}{\mathbf{p}'_{h,i}\mathbf{p}_{h,i}} \quad \left(\mathbf{t}_{h,i} = \frac{\mathbf{X}\mathbf{w}'}{\mathbf{w}'\mathbf{w}}\right)$$

$$\mathbf{q}'_{h,i} = \frac{\mathbf{t}'_{h+1,i}\mathbf{Y}}{\mathbf{t}'_{h+1,i}\mathbf{t}_{h+1,i}}$$

$$\mathbf{q}'_{h,i} = \frac{\mathbf{q}'_{h,i}}{\parallel \mathbf{q}'_{h,i} \parallel}$$

$$\mathbf{u}'_{h+1,i} = \frac{\mathbf{Y}\mathbf{q}_{h,i}}{\mathbf{q}'_{h,i}\mathbf{q}_{h,i}}$$

if $h \geq 2$

if $\mathbf{t}_{h+1,i} \simeq \mathbf{t}_{h,i}$

BREAK

end if

end if

end for h

$\mathbf{E}_i = \mathbf{E}_{i-1} - \mathbf{t}_i\mathbf{p}'_i$ [OUTER RELATION (1)]

$\mathbf{F}_i = \mathbf{F}_{i-1} - \mathbf{u}_i\mathbf{q}'_i$ [OUTER RELATION (2)]

$\mathbf{b}_i = (\mathbf{t}'_i\mathbf{t}_i)^{-1}\mathbf{t}'_i\mathbf{Y}$

$\hat{\mathbf{u}} = \mathbf{b}_i\mathbf{t}$ [INNER RELATION]

$\mathbf{F}_i = \mathbf{F}_{i-1} - \hat{\mathbf{u}}_i\mathbf{q}'_i$

end for i

where $\mathbf{y}_j$ is a random column of $\mathbf{Y}$; i is the iteration on the PCA components; h is the iteration to reach convergence and and k is the maximum number of iterations.

Chapter 3

Results and Discussion

3.1 Introduction

The general aim of this study is to formulate a procedure which enables to predict the response of a micro-organism based only on the information available in real time during the fermentation, i.e, online. To be able to predict how a microbe and, for extension, an entire microbial population reacts to external stimuli that can be manipulated would allow to optimize the environmental conditions, pushing the culture in a determined direction. Considering an industrial fermentation, this would mean for example to optimize the production of the products of interest and, consequently, to increase the economical gain. This kind of study is classified as secondary model synthesis for micro-organisms, since secondary models introduce the effect of the environmental variables in the primary description of the system. The literature about this argument is quite recent, and considering the present state of knowledge the prospected potentiality is still far from the practical application. This study does not pretend to reach and fully explore this ambitious prospect, but it is a small step in the same direction. Which aspects of this wide field will be treated must be specified, and more in detail: what is meant for *response of the micro-organism*, i.e., what is the output of the process? Which information is available about the micro-organism, i.e., what is the input of the process? What does the procedure consist of?

Considering an industrial fermentation, the bacterial growth is promoted to produce some metabolite which can be valuable. The cell absorbs nutrients from the environment, and it expels the sub-products of its metabolism. In this context, the response of the micro-organism corresponds to the extracellular concentration of the metabolites, i.e., the concentration of the valuable products to be extracted from the medium and the concentration of the nutrients and/or inhibitors to maximize the productivity.

Since the process developed in this study is meant to be implemented in future as part of a

wider process control procedure, a necessary requisite is that it must work online. This means that it should make use of only information which is available during the process and it must require a computational time comparable with the dynamic evolution of the system. Considering a microbial population in a liquid mixed fermenter, such as a fed-batch bioreactor, the medium can be assumed as homogeneous. In this situation, the only measurements which can be collected online are about the environmental conditions. Applying a reductionist approach, among all the environmental conditions a few are selected to be relevant variables for the microbial system. Two kind of environmental variables can be distinguished: *(i)* the variables which are not affected by the bacterial growth, e.g., temperature, pH etc., and *(ii)* the variables which depend on the bacterial evolution, e.g., the extracellular concentrations of metabolites. In this study the variables pertaining to the first class will be always considered controlled, being kept constant, while the evolution of the second ones will be taken into account. Hence, both the starting and the final point of this procedure coincides with the extracellular concentrations. This condition perfectly adapts to an iterative online implementation.

The procedure requires to solve the primary model of the system (2.12), which is a dynamic model. Since this model contains the fluxes, how the fluxes continuously evolves in time must be known to solve it. *Isotopomer analysis* is a modern technique which allows to obtain time values for the fluxes. It measures the instant concentrations of labeled molecules by nuclear magnetic resonance (NMR) and/or gas chromatography/mass spectroscopy (GC/MS), and once the metabolic network has been completely characterized it provides the corresponding values of the other fluxes by solving stoichiometric balances. The expression of fluxes as function of the concentrations can then be obtained by regressing the experimental data collected in time. The problem, as was previously explained, is that isotopomer analysis is a discontinuous technique, and it cannot be performed online. Hence, the only way to include online information in the model is to express fluxes as function of the extracellular concentrations. The procedure's first step consists of finding a bond between fluxes and metabolite concentrations. The fluxes are used to solve the primary dynamic model. The solution is then slightly extrapolated in time, enabling to predict the reaction of the micro-organism and, in future applications, to optimize it. Since the growth and the dynamic evolution of a microbial population are quite slow, the time available for the solution of the procedure is in the order of the hour.

Different approaches were tested to complete this procedure. They distinguish each other in the way they relate the fluxes to the concentrations. In particular, a *(i)* grey-box (GB) approach, i.e., partially mechanistic, and a *(ii)* black-box (BB) approach, i.e., fully experimental, will be presented.

First, the two case studies used to test the procedure are introduced and described. Then every step of the procedure of both the approaches is analysed for both the case studies.

3.2 Data

3.2.1 Case studies

To perform a study about metabolic network-based modeling, a metabolic network is needed. Fortunately, many complete metabolic networks are inherited by the literature of the past decades. Metabolic networks can be very large systems, with many metabolites and reactions. Although the computational load proportionally increases with the system dimension, the approach here developed is general, and, at least in theory, it can be indifferently applied on small as on large networks. All the procedures of this study were tested on two different case studies: *(i)* a small-scale simulated network and *(ii)* the network which describes the metabolism of engineered *E. coli* for the production of 1,3-propanediol (Antoniewicz et al., 2007). The differences between these two systems, which will be here highlighted, allow also to investigate how the dimension of the network and the peculiarities of a real system influence the efficiency of the whole procedure.
The small-scale network is a *toy network*, i.e., a network whose data were not experimentally obtained but artificially simulated. It consists of 4 extracellular metabolites, 6 (m) intracellular metabolites and 9 (n) fluxes (Figure 3.1). Hence, the number of free fluxes is $n - m = 3$. The set of free fluxes is not univocally determined, since the basis of the null space of the **S** matrix is not unique. The fluxes 1, 5 and 7 were chosen. The measurements were simulated by computing the free fluxes as:

$$u_{1,ref} = \frac{c_{A,ext}}{1.5+c_{A,ext}} \tag{3.1}$$

$$u_{5,ref} = 0.2 \cdot \frac{c_{E,ext}}{3+c_{E,ext}} \tag{3.2}$$

$$u_{7,ref} = \frac{1}{1+c_{F,ext}} \tag{3.3}$$

The states, i.e., measurements of extracellular concentrations and fluxes, were then computed by solving the ODE system of the primary model, given the starting point for the extracellular concentrations at $[10, 15, 0, 0.1]$ and a time interval between 0 and 20 hours. A small amount of noise was added afterwards. The simulation provided 500 time points, equally distributed in the interval, for every extracellular concentration. The reference profiles for both extracellular concentrations and fluxes are reported (Figure 3.2, 3.3).

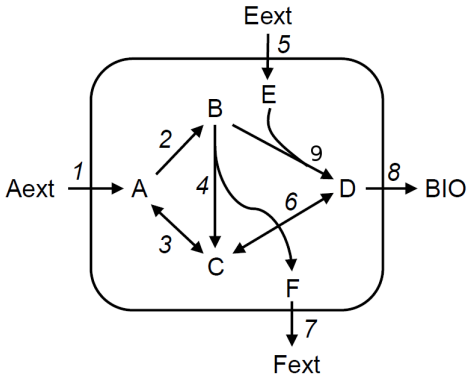
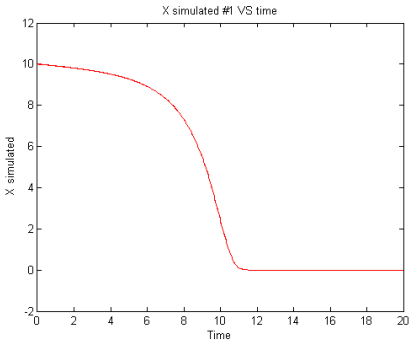
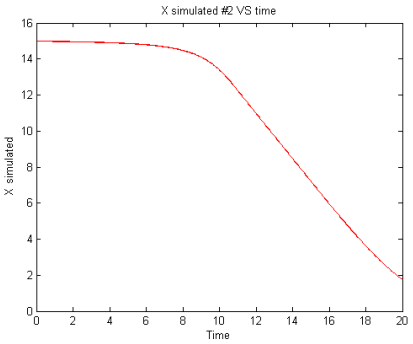


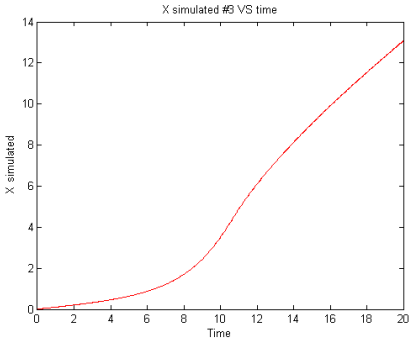
Figure 3.1: Small-scale network. A schematic representation of a microorganism, with the intracellular and extracellular metabolites and the relative fluxes.



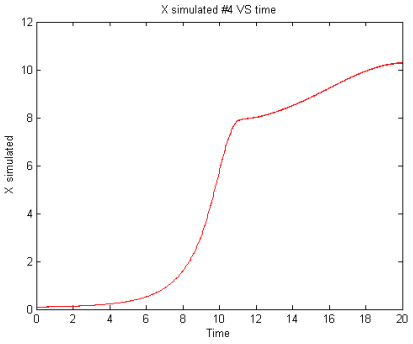
(a) Concentration A_{ext}



(b) Concentration E_{ext}



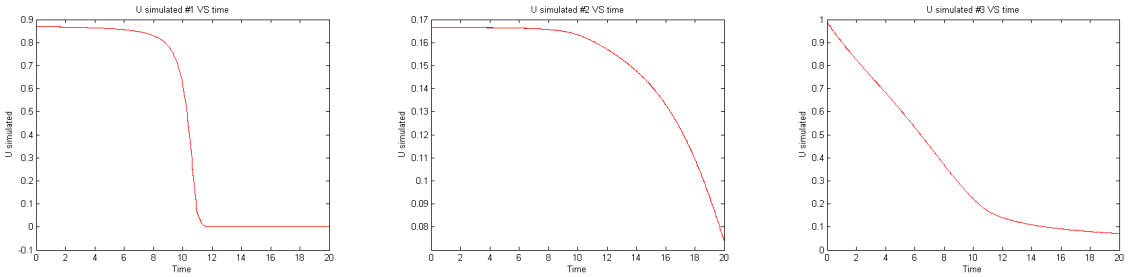
(c) Concentration F_{ext}



(d) Concentration BIO

Figure 3.2: Solution of the primary dynamic system for the toy network in the time experimental range.

The second case study is a real metabolic network instead. For all the details about the system this study refers to Antoniewicz et al. (Antoniewicz et al., 2007). In this article the



(a) u_1 (b) u_2 (c) u_3

Figure 3.3: Solution of the primary dynamic system for the *E.coli* network in the time experimental range.

author desired to make use of the detailed time profiles of *in vivo* fluxes to obtain additional insight in support of the assumed genotype of the organism. The data were collected from a fed-batch fermentation of *Escherichia coli* to produce 1,3-propanediol (PDO). PDO is a chemical which covers a role in many industrial productions: it is a building block in the production of polymers such as polytrimethylene terephthalate; it is added to the formulations of many products, e.g., composites, adhesives, laminates, coatings, moldings, aliphatic polyesters, copolyesters; it is a solvent. Consequently, this case study closely resembles an industrial fermentation. All the main environmental variables in the reactor, identified through a reductionist approach, are controlled, and they can be assumed to stay constant during the whole fermentation. The temperature was kept constant at 34^oC, pH at 6.8 ± 0.04 and the dissolved oxygen at $10\% \pm 0.7$ of saturation. The glucose was fed during the fermentation to keep its concentration in the medium more or less constant, around $45 \pm 5\, mM$. The use of fed-batch reactors to gradually feed the substrate during the fermentation is quite common, since, if all the substrate were provided at once from the beginning of the fermentation, the culture would know a very fast growth during the first hours, but then it would reach the stationary and the death phase before the end of the process. To maximize the productivity, to maintain the cellular population in the exponential growth phase as long as possible is convenient. The only environmental variables which are allowed to change are the extracellular concentration of metabolites. The network consists of 11 extracellular metabolites, 5 substrates, i.e., glucose ($Glc[e]$), citrate ($Cit[e]$), O_2, NH_3 and SO_4, 5 products, i.e., 1,3-propanediol ($PDO[e]$), biomass, CO_2 ($CO_2[e]$), acetate ($Ac[e]$) and ATP, and glycerol ($Glyc[e]$),which can be both a substrate or a product. Of these 11 extracellular metabolites, only 6 were not controlled and free to vary. The concentration of glucose in the feed of the fermenter ($GlucF[e]$) is considered as an additional extracellular metabolite, and it is correlated to the concentration

of feed in the medium adding a new flux to the network. The total number of extracellular concentrations results to be $k = 7$, while the total number of fluxes is $n = 69$. The number of intracellular metabolites is $m = 62$, and the number of free fluxes then is $n - m = 7$. Since the null space of the S_{int} matrix is not unique, many combinations of free fluxes can be selected. The study works well for every selected combination, but a proper one can influence some properties of the system making the calculation easier (3.3.3). The fluxes that were chosen as independent in this study are reported in Table 3.1.

Free Fluxes	
u	**Reaction**
1	$GlucF[e] \rightarrow Gluc[e]$
2	$Cit[e] \rightarrow Cit$
3	$PDO \rightarrow PDO[e]$
4	$Ac \rightarrow Ac[e]$
5	$CO_2 \rightarrow CO_2[e]$
6	$O_2[e] \rightarrow O_2$
7	Biomass formation

Table 3.1: The fluxes which were chosen as independent for the *E. coli* network.

The simulation was led on a time interval between 15.4 and 44.6 hours. From literature 21 measurements per measured variable were provided, rendering a total of 189 measurements. This study was led using instead 500 measurements, equally distributed in the time interval, per measured variable. The time profiles which are obtained are almost continuous. The new data were generated by Vercammen with the dMFA approach briefly presented in section 2.2.3, using B-spline parametrization with incremental spline knot insertion (Figure 3.4, Figure 3.5, Figure 3.6).

3.2.2 Input data

As reported in subsection 3.2.1, the experimental data this study has worked on consist of 500 time points for each variable, concentrations and fluxes. Since the data are subject to an intrinsic variability, it is important to always consider how the error propagates in the different steps of the procedure. Additional data are required to estimate the uncertainty for each time point. A reasonable number of data to take into account the variability of the system could be about 100. Infinite sets of data can be generated for the toy network, since it is a simulated network. Also for the real system additional data can be easily obtained,

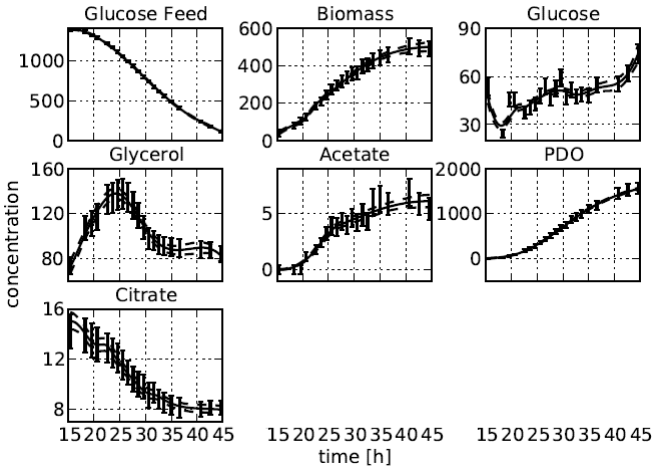


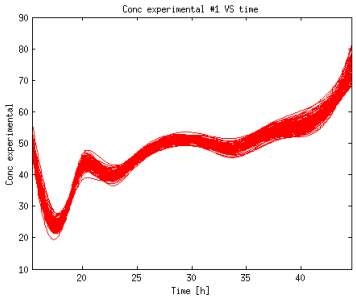
Figure 3.4: Experimental measurements and fitting profiles for the metabolite concentrations of the fed-batch case study.

since continuous functions of time were obtained for both fluxes and concentrations. For the fluxes these expressions were obtained by solving the dMFA problem (2.2.3), while for the concentrations a simple regression was performed on the experimental data. The additional data were obtained by normally perturbing 100 times the first experimental point in time. These 100 points were used as starting values to solve 100 times the primary dynamic system. Hence, 100 different time profiles were obtained, each one showing its own time evolution.
The way the data are organized and, consequently, how the results of the study are plotted is now explained and justified. The most common way to represent the variability for each time point is by using *error bars*, i.e., the confidence interval at $\alpha = 0.05$ centered around the mean value:

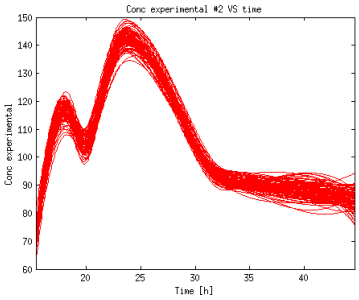
$$\text{ErrorBar} = \bar{X} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} \tag{3.4}$$

where n is the number of experiments used to compute the mean and the standard deviation. This representation of error individually considers each time point, and this way it does not take into account the dynamic nature of the different simulations. For this reason, it was chosen to always plot all the 100 time profiles. The dynamic correlation of data on the same profile is evident when looking at a typical graph used for regressions, which reports $y_{experimental}$ vs $y_{estimated}$ (Figure 3.7). From the plot is evident how the data are not randomly distributed, but they follow different profiles.

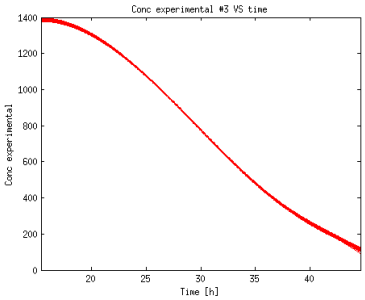
Having 100 curves for each variable being represented at once on a graph could be sometimes



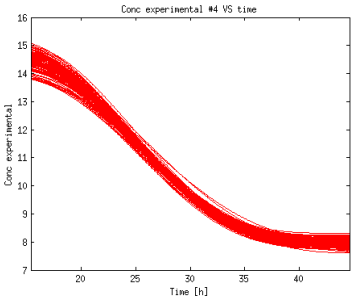
(a) Glucose concentration



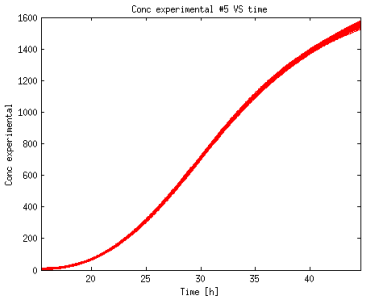
(b) Glycerol concentration



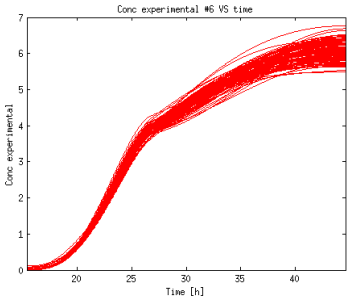
(c) Glucose feed concentration



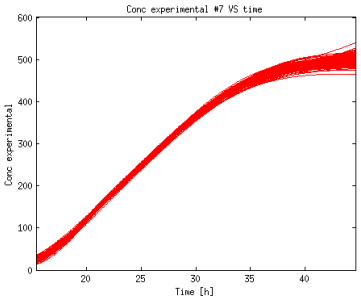
(d) Citrate concentration



(e) PDO concentration

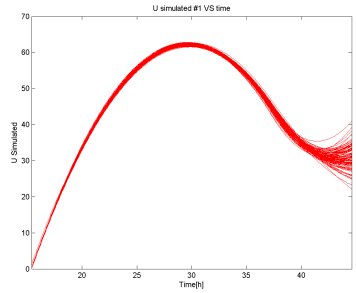


(f) Acetate concentration

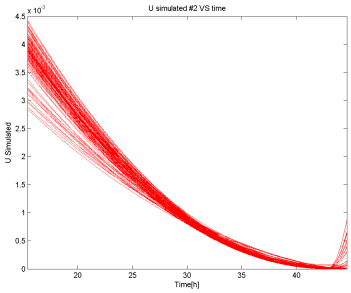


(g) Biomass concentration

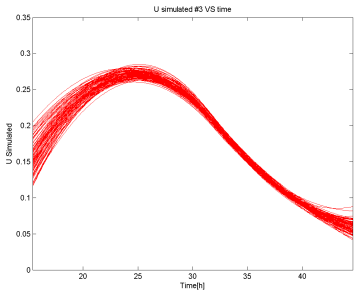
Figure 3.5: Experimental profiles of the extracellular concentrations for the *E.coli* network.



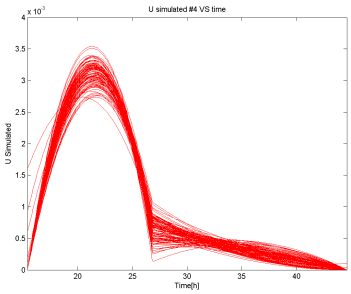
(a) u_1



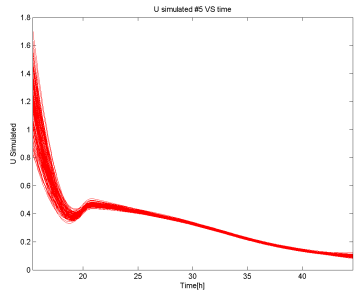
(b) u_2



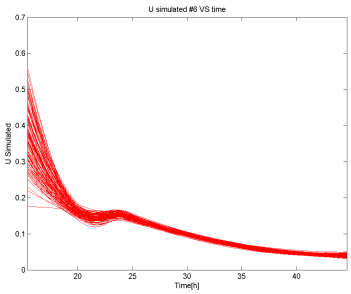
(c) u_3



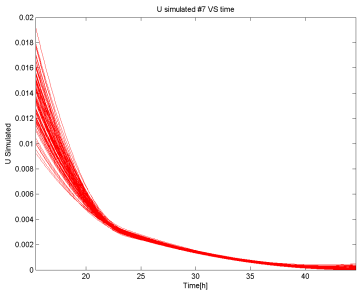
(d) u_4



(e) u_5



(f) u_6



(g) u_7

Figure 3.6: Experimental profiles of the fluxes for the *E.coli* network.

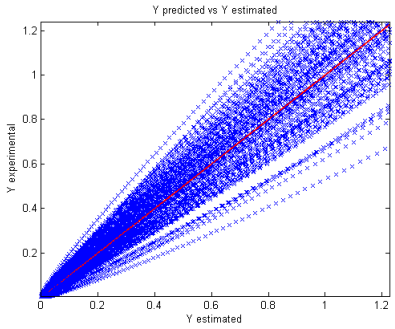
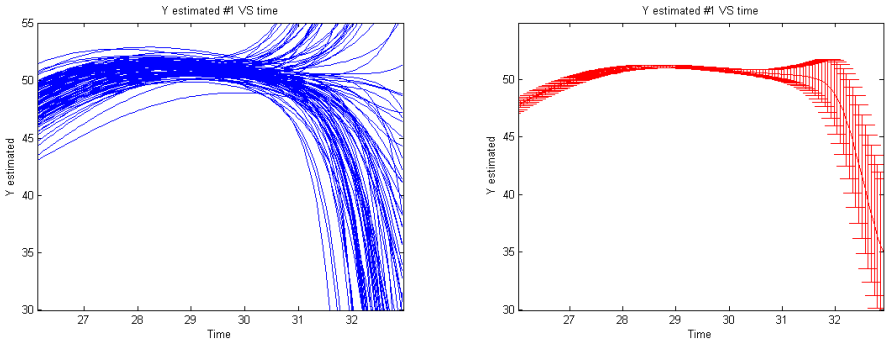


Figure 3.7: $Y_{estimated}$ versus $Y_{experimental}$.

confusing, but it allows to catch how the dynamic of each profile evolves (Figure 3.8). Looking at the error bars diagram, the plot appears to sharply deviate around $t = 31\,[h]$ and the confidence interval to drastically increase, but no evident reasons to explain this behavior can be identified. Instead, from the diagram which reports the 100 profiles it is evident how the problem lies in the solution of the dynamic system, which diverges, possibly showing an unstable behaviour of the system.



(a) 100 profiles (b) Error bars

Figure 3.8: Representation of the error: *a)* reporting all the 100 dynamic profiles; *b)* reporting the error bars for each time point.

The total amount of data for each variable, concentrations or fluxes, is $500\,\mathrm{x}\,100 = 5e5$. Since the number of variables is in the order of 10, and the number of parameters to correlate these variable will be slightly bigger, the system will be largely overdetermined. This is the best condition to perform a good regression. Such a huge amount of data is not comfortably manageable, and it must be properly organized. The input variables to all procedures are organized in a big matrix which concatenates 100 smaller matrices in vertical, each one having the variables on the rows and the time profile on the columns. The dimension of each

of these matrices is $n \times 500$, where n is the number of variables. The big matrix has dimension $100 \cdot n \times 500$.

3.3 Grey-box approach

3.3.1 Introduction

The first approach being presented is a *grey-box approach*. The name expresses the characteristic to be halfway between a *black-box model*, fully experimental, and a theoretical so called *white-box model*, fully mechanistic. What does fully experimental or fully mechanistic mean? A fully experimental model is a model generated by a regression on a particular set of data, being consistently able to properly describe just that set of data. Even if it does not include any comprehension of the system, it is generally a very flexible model, and it can be adapted to describe the system also in conditions slightly different from the training conditions. To obtain an experimental model can seem an immediate and basic task, but it is not. A good black-box model should represent a compromise between its description of the training set of data and its ability to be adapted to other data. In other words, the synthesis of a black-box model is a trade-off between its descriptive and predictive capability. On the contrary, a fully mechanistic model is a model which perfectly interprets the mechanisms that rule the system, a model which includes a complete knowledge and understanding of the system. While the first class of models is nowadays widely adopted in many fields and for many applications, including predictive microbiology and biotechnology, since it is the easiest and the most immediate kind of model which can be formulated about a system, it is not the same for the latter class. Fully mechanistic models are few, since finding models which do not contain adaptive parameters is quite difficult. What can be reached is a compromise, a grey-box model. The different grey-box models distinguish each other in their shade of grey, i.e., in how many empirical information they still contain and how deeply they penetrate the underlying mechanisms of the system. If the first models being adopted are generally black-box, since they don't require any knowledge about the system, the natural path would move toward mechanistic models, adding more and more biological information. In the case of this study the starting point was not a black-box model, but the wide number of studies published in literature about metabolic network-based modeling was used as a base to formulate a model which already includes mechanistic informations. Particularly, the grey-box approach here presented makes use of the primary dynamic model formulated by Van Impe (Van Impe et al., 2012) and of dynamic metabolic flux analysis (dMFA) and flux balance analysis (FBA) to estimate the flux distribution at each time point.

3.3.2 Formulation of the optimization problem

In section 2.2.4, two different approaches to determine the distributions of fluxes were presented. Since the main interest of this study is not in collecting biological information about the system, but in reproducing a set of fluxes as close as possible to the experimental values in large range of conditions, the ObjFind approach was chosen. The main difference between the ObjFind approach and the one proposed in this study is the final aim. While the original ObjFind estimates a set of coefficients with a linear objective function to formulate biological considerations or to test a previous hypothesis about the cell response, in this study the set of coefficients is used to solve the inverse problem, i.e., the determination of the fluxes by solving only the inner problem with the CoIs estimated with the bilevel problem. This problem is called *simulation problem*, and, theoretically, it should allow to obtain a distribution of fluxes close to the experimental one. In the original article about ObjFind (Burgard and Maranas, 2003), the *degeneracy* of the LP when facing the simulation problem was described, since it accepts many flux distributions as optimal solution. Using an LP formulation, the flux distributions through the network cannot be uniquely defined based solely on the identified CoIs. This is the first reason why the optimization problem was modified in the objective functions and in the constraints by introducing non linear terms. The second reason can be intuitively understood when considering the feasible area defined by the biological constraints of the inner problem. Most of the experimental flux distributions fall inside this area, not on its bounds. Since a linear objective function is monotonous, it will not be able to identify an optimal solution internal to the feasible area, but it will always find a solution on its bounds. The possibilities to overcome this problem just consists of adopting a non-linear objective function, which admits internal solutions to the feasible area, or to add further boundaries which include or at least approximate the experimental points. While the first alternative can find quite an easy biological explanation (see the list of objective functions proposed by Schuetz (Schuetz et al., 2007)), finding meaningful constraints which are close to the experimental point for every time instant is rather difficult. This would mean to leave aside part of the biological mechanistic considerations in favour of a precise description of the experimental data and a major flexibility of the model. Still the experimental characterization of the grey-box model remains strong.

For the procedure to succeed, the CoIs obtained from the direct optimization problem, i.e., the bilevel optimization problem, must satisfy an additional condition: they must vary continuously in time. The set of CoIs has to univocally correspond to a set of fluxes which are not constant in time. For example, a common problem to many objective functions is that they return a set of CoIs which is very unbalanced, i.e., with one CoI extremely larger than the

others. This fact is not a problem by itself. For example, the studies reported in literature had proven that one of the most common objective functions which returns a good description of the microbial response is the biomass growth. Since in the metabolic network one of the free fluxes is a combination of fluxes which expresses the biomass growth, it would be normal to expect the weight of this flux to be prevalent to the others. Nevertheless, if the different CoIs are too unbalanced, i.e., they differ in many orders of magnitude, the variation of the smaller CoIs could be disregarded with respect to the bigger one, and the set of CoIs would appear to be almost constant. Such a set of CoIs would not allow to solve the simulation problem. These and other empirical considerations can be done to define a proper formulation of the bilevel optimization problem, but at the end the best way to test the formulation is to solve the simulation problem time point by time point, outside the primary dynamic model, and to look at the results. The proper solution of the simulation problem is a necessary condition to the procedure of this study.

Considering the same problem as the original ObjFind method (2.25), but with a quadratic objective function as $f = \mathbf{u}^T \cdot \mathbf{C} \cdot \mathbf{u}$, the duality theorem cannot be applied anymore, and the solution of the inner problem is reformulated using the KKT conditions (2.3.3). In case of inequality constraints, the KKT conditions will produce a mathematical problem with complementarity constraints (MPCC), which is harder to solve.

$$\min_{\mathbf{c},\mathbf{u}\in\mathbb{R}^{n-m},\,\boldsymbol{\mu}_1\in\mathbb{R}^p,\,\boldsymbol{\mu}_2\in\mathbb{R}^n} \quad \sum_{i=1}^{n-m} (u_i - u_i^{exp})^2 \tag{3.5}$$

$$\text{s.t.} \quad \sum_{i=1}^{n-m} c_i = 1 \tag{3.6}$$

$$\mathbf{IR} \cdot \mathbf{K} \cdot \mathbf{u} \geq 0 \tag{3.7}$$

$$\mathbf{UB} - \mathbf{K} \cdot \mathbf{u} \geq 0 \tag{3.8}$$

$$(\mathbf{IR} \cdot \mathbf{K} \cdot \mathbf{u})^T \cdot \boldsymbol{\mu}_1 = 0 \tag{3.9}$$

$$(\mathbf{UB} - \mathbf{K} \cdot \mathbf{u})^T \cdot \boldsymbol{\mu}_2 = 0 \tag{3.10}$$

$$2 \cdot \mathbf{C} \cdot \mathbf{u} - (\mathbf{IR} \cdot \mathbf{K})^T \cdot \boldsymbol{\mu}_1 + \mathbf{K}^T \cdot \boldsymbol{\mu}_2 = 0 \tag{3.11}$$

$$(c_j \geq 0 \; \forall j \in 1, \cdots n-m)$$

$$(\mu_{1,j} \geq 0 \; \forall j \in 1, \cdots p)$$

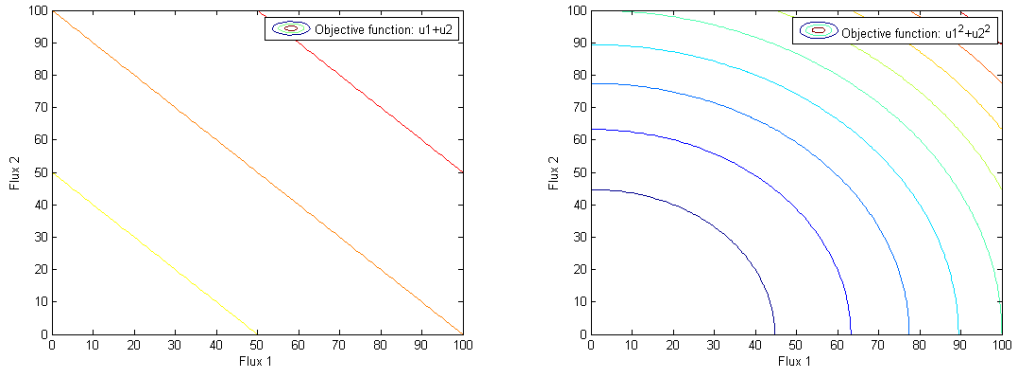
$$(\mu_{2,j} \geq 0 \; \forall j \in 1, \cdots n)$$

where μ_1 and μ_2 are lagrange multipliers for inequality constraints. In case of equality constraints, further additional lagrange multipliers λ should be added. In this case the problem is a NLP.

In this case, since physical information about the nature of the system were not available, the value of each element of **UB** can be set much bigger than the experimental values, remaining as an inactive constraint, or it can be omitted. Since the biggest flux is the one for glucose, which reaches a maximum around $70\,g/L \cdot h$, every element of the upper bound vector was set to 100. Furthermore, since the number of fluxes v is usually much bigger than the number of free fluxes, and during the reformulation of the problem a slack optimization variable must be added for every element of $\mathbf{v}$, to include this constraint means to notably increase the number of variables to optimize. Hence, in the following results, the upper bound constraint was considered only if it were proven not to badly affect the computational efficiency of the optimization problem, otherwise it was just disregarded.
As introduced in Section 2.3.1, two different frameworks are considered to solve the optimization problem: MATLAB® and CasADi. The peculiarities of each framework allow to solve the problem formulated in different ways. The performances of two different strategies, implemented in both the environments, were compared on the same problem, to establish the more efficient and precise strategy for this application. Choosing an optimization problem with a quadratic objective function, which admits an exact solution, 100 iterations were solved for each case. *(i)* First MATLAB® and CasADi were compared using for both the reformulated problem with KKT conditions. The MATLAB® solver was necessarily `fmincon`, since the MPCC problem introduce non linear constraints, and the interior point algorithm was chosen. While CasADi was able to find the exact solution, MATLAB® returns a different one. The difference between the solvers probably depends on the use of symbolic variables and automatic differentiation in CasADi, which allows to better solve the MPCC problem. *(ii)* Then the performance of the bilevel optimization problem in MATLAB® and the reformulated problem with KKT conditions in CasADi were compared. Using the interior point algorithm in MATLAB®, which was the only algorithm proven to properly work for this problem, both strategies return the exact solution. MATLAB® spent $420\,s \cong 7\,min$ to solve all the iterations, while CasADi $16\,s$. The total amount of iterations to complete in this study for both case studies was $5e5$, even without considering the eventual reformulation of the MPCC, which requires still more iterations. This consideration give a measure of how important it is to use a fast solver. Hence, all the following results were generated using the reformulation of the problem with KKT conditions using CasADi. MATLAB®, instead, was used afterwards to solve the simulation problem inside the primary dynamic system.

3.3.3 Optimization problem results

The interest is not only in solving the optimization problem, but also in checking how the whole procedure behaves when introducing the optimization as intermediate step. To do this,



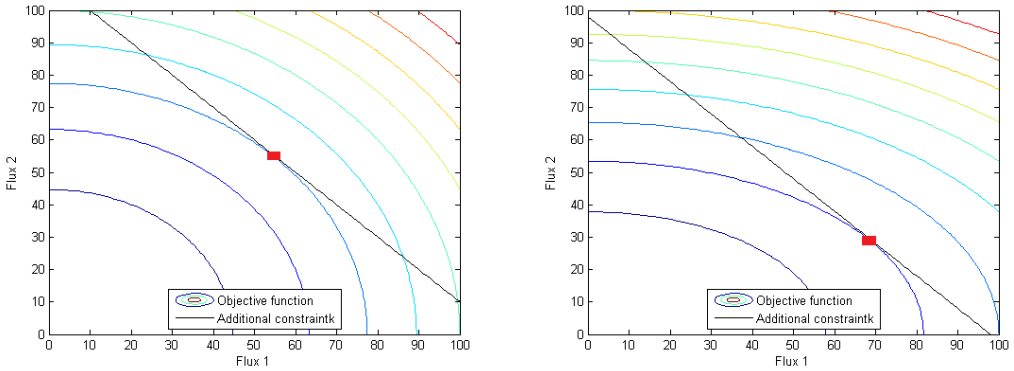
(a) $c_1 \cdot x_1 + c_2 \cdot x_2$ (b) $c_2 \cdot x_1^2 + c_1 \cdot x_2^2$

Figure 3.9: Plots of the level lines of a 3D function on the positive quarter: a) linear function; b) quadratic function. Both the linear and the quadratic functions monotonically increase, with a minimum in the origin and a not defined maximum, since they are unbounded.

the optimization problem has to be fully solved, which means both to obtain an objective function able to well fit the experimental data and to generate a set of CoIs whose time profiles vary smoothly and continuously in time and which univocally solve the simulation problem. Given the difficulties to find an objective function between the ones listed in literature (Schuetz et al., 2007) which satisfy all these requirements, the optimization problem was slightly modified. It was decided to renounce part of the mechanistic aspect of the model by introducing an experimental parameter to make the solution easier. This way of course the model loses in generality and in applicability, but it can be useful in the meanwhile to test the efficiency of the whole procedure, while a physical objective function still is lacking. The easiest way to find a formulation which provides a good solution was to visualize it on a bidimensional constrained space. In fact, to have a visual representation of a problem can always help its understanding.

As previously highlighted, simple linear and quadratic objective functions are not suitable for this problem since they are only able to find solutions which lie on the boundaries of the feasible area. The reason is that they monotonously increase or decrease (Figure 3.9).

When minimizing a quadratic function in the form $f = \mathbf{x}^T \cdot \mathbf{C} \cdot \mathbf{x}$ in the feasible region identified by the constraints of the problem (2.25), the minimum lays in the origin, far from a theoretical experimental point interior to this area. The summation of all the free fluxes is a line which crosses the feasible area. When imposing this summation to be bigger than some value, the minimum of the quadratic function would now lie on this constraint (Figure 3.10). Which point on the line, inside the feasible region, is the minimum depends on the shape of the objective function, i.e., on the CoIs. Since the CoIs are selected to minimize



(a) $x_1^2 + x_2^2$ (b) $c_2 \cdot x_1^2 + c_1 \cdot x_2^2$

Figure 3.10: Plots of the level lines of a surface: a) quadratic function $x_1^2 + x_2^2$ and constraint $x_1 + x_2 = c_b$; b) quadratic function $c_2 \cdot x_1^2 + c_1 \cdot x_2^2$ and constraint $x_1 + x_2 = c_b$. The addition of a linear constraint, which cut the positive quarter, allows to define a minimum of the quadratic objective function internal to the original feasible area. This minimum changes if the shape of the function will change, i.e., if the CoIs will change. The CoIs are optimized to make the solution always as close as possible to the experimental point.

the least squares error on the fluxes, if the parameter of the new constraint is set as a new optimization variable it would oblige the line to include the experimental point, and the other optimization variables would be estimated to make that point the minimum of the quadratic function in the new feasible area. Since the experimental time profiles for fluxes are smooth, the experimental point on this graph will move smoothly, and theoretically the set of CoIs should also vary continuously and smoothly. Furthermore, by introducing a new parameter in the outer optimization problem, the degrees of freedom of the problem are reduced, and the risk of multiple solutions diminishes.
The formulation of the bilevel optimization problem including the new constraint is:

$$\min_{\mathbf{c}\in\mathbb{R}^q} \quad \sum_{i=1}^{n-m} (u_i - u_i^{exp})^2 \tag{3.12}$$

$$\text{s.t.} \quad \sum_{i=1}^{q} c_i = 1 \tag{3.13}$$

$$\max_{\mathbf{u}\in\mathbb{R}^{n-m}} \quad f(\mathbf{u}, \mathbf{c}) \tag{3.14}$$

$$\text{s.t.} \quad \mathbf{IR} \cdot \mathbf{K} \cdot \mathbf{u} \geq 0 \tag{3.15}$$

$$\mathbf{UB} - \mathbf{K} \cdot \mathbf{u} \geq 0 \tag{3.16}$$

$$\mathbf{1}^T \cdot \mathbf{u} - c_b \geq 0 \tag{3.17}$$

where $q = (n - m) + 1$, c_b is the additional optimization variable and $\mathbf{1}$ is a vector of ones with dimension $[(n - m)\,\mathrm{x}\,1]$.
The problem reformulated with the KKT conditions, since the objective function is non linear, is :

$$\min_{\mathbf{c},\mathbf{u}\in\mathbb{R}^{n-m},\,\boldsymbol{\mu}_1\in\mathbb{R}^p,\,\boldsymbol{\mu}_2\in\mathbb{R}^n,\,\boldsymbol{\mu}_3,c_b\in\mathbb{R}} \sum_{i=1}^{n-m} (u_i - u_i^{exp})^2 \tag{3.18}$$

$$\text{s.t.} \sum_{i=1}^{n-m} c_i = 1 \tag{3.19}$$

$$\mathbf{IR}\cdot\mathbf{K}\cdot\mathbf{u} \geq 0 \tag{3.20}$$

$$\mathbf{UB} - \mathbf{K}\cdot\mathbf{u} \geq 0 \tag{3.21}$$

$$\mathbf{1}^T\cdot\mathbf{u} - c_b \geq 0 \tag{3.22}$$

$$(\mathbf{IR}\cdot\mathbf{K}\cdot\mathbf{u})^T\cdot\boldsymbol{\mu}_1 = 0 \tag{3.23}$$

$$(\mathbf{UB} - \mathbf{K}\cdot\mathbf{u})^T\cdot\boldsymbol{\mu}_2 = 0 \tag{3.24}$$

$$(\mathbf{1}^T\cdot\mathbf{u} - c_b)\cdot\mu_3 = 0 \tag{3.25}$$

$$2\cdot\mathbf{C}\cdot\mathbf{u} - (\mathbf{IR}\cdot\mathbf{K})^T\cdot\boldsymbol{\mu}_1 + \mathbf{K}^T\cdot\boldsymbol{\mu}_2 - (\mathbf{1}^T\cdot\mathbf{u})\cdot\mu_3 = 0 \tag{3.26}$$

$$(c_j \geq 0\ \forall j \in 1,\cdots n-m)$$

$$(\mu_{1,j} \geq 0\ \forall j \in 1,\cdots p)$$

$$(\mu_{2,j} \geq 0\ \forall j \in 1,\cdots n)$$

$$(\mu_3 \geq 0)$$

Since the variable c_b is always set by the optimization problem to the summation of the experimental values, it is possible to rewrite the inequality constraint as an equality one, which introduces a variable λ_1 instead of μ_3. Furthermore, it is also possible to disregard the upper bound constraints, because they are not active since the problem minimizes the objective function and the solution will always be far from the upper bounds. This way the number of complementarity constraints is reduced, and solving the optimization problem is easier. The reformulated problem with the last constraint expressed as an equality constraint and disregarding the upper bound constraints is:

$$\min_{\mathbf{c},\mathbf{u}\in\mathbb{R}^{n-m},\,\boldsymbol{\mu}_1\in\mathbb{R}^p,\,\lambda_1,c_b\in\mathbb{R}} \sum_{i=1}^{n-m}(u_i - u_i^{exp})^2 \tag{3.27}$$

$$\text{s.t.} \sum_{i=1}^{n-m} c_i = 1 \tag{3.28}$$

$$\mathbf{IR}\cdot\mathbf{K}\cdot\mathbf{u} \geq 0 \tag{3.29}$$

$$\mathbf{1}^T\cdot\mathbf{u} - c_b = 0 \tag{3.30}$$

$$(\mathbf{IR}\cdot\mathbf{K}\cdot\mathbf{u})^T\cdot\boldsymbol{\mu}_1 = 0 \tag{3.31}$$

$$2\cdot\mathbf{C}\cdot\mathbf{u} - (\mathbf{IR}\cdot\mathbf{K})^T\cdot\boldsymbol{\mu}_1 - (\mathbf{1}^T\cdot\mathbf{u})\cdot\lambda_1 = 0 \tag{3.32}$$

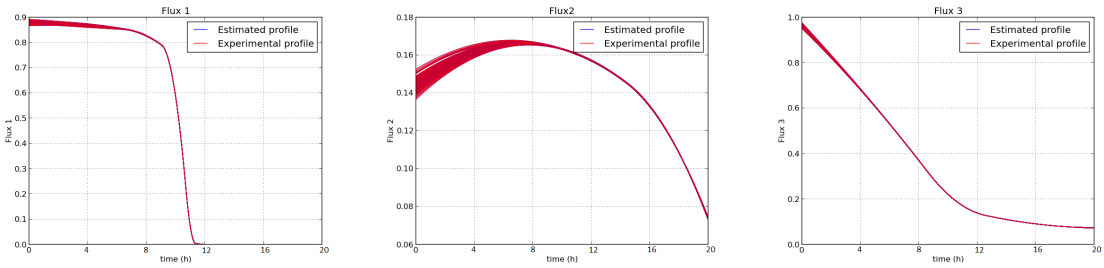
$$(c_j \geq 0\ \forall j \in 1,\cdots n-m)$$

$$(\mu_{1,j} \geq 0\ \forall j \in 1,\cdots p)$$

The set of free fluxes is not unique, but it depends on the selected $\mathbf{K}$ matrix, i.e., the null space of the $\mathbf{S}_{int}$ matrix. Although the procedure must work for every possible set of free fluxes, working with negative fluxes can introduce some additional difficulty. Considering the modified bilevel optimization problem (3.12), the problem is able to find the exact distribution of fluxes only if all the fluxes are positive. This is due to the particular form of the additional constraint which was chosen. To maintain the properties of a positive set of fluxes for a very general set, the free fluxes can be normalized between their maximal and minimum experimental value. Although the procedure was tested on a small example network and proven to work equally well with normalized and not-normalized fluxes, in this study the set of fluxes was chosen to be always positive on the experimental range, to avoid complications due to the normalization. Hence, the solution of the optimization can be visualized on the positive quadrant of the space of free fluxes.

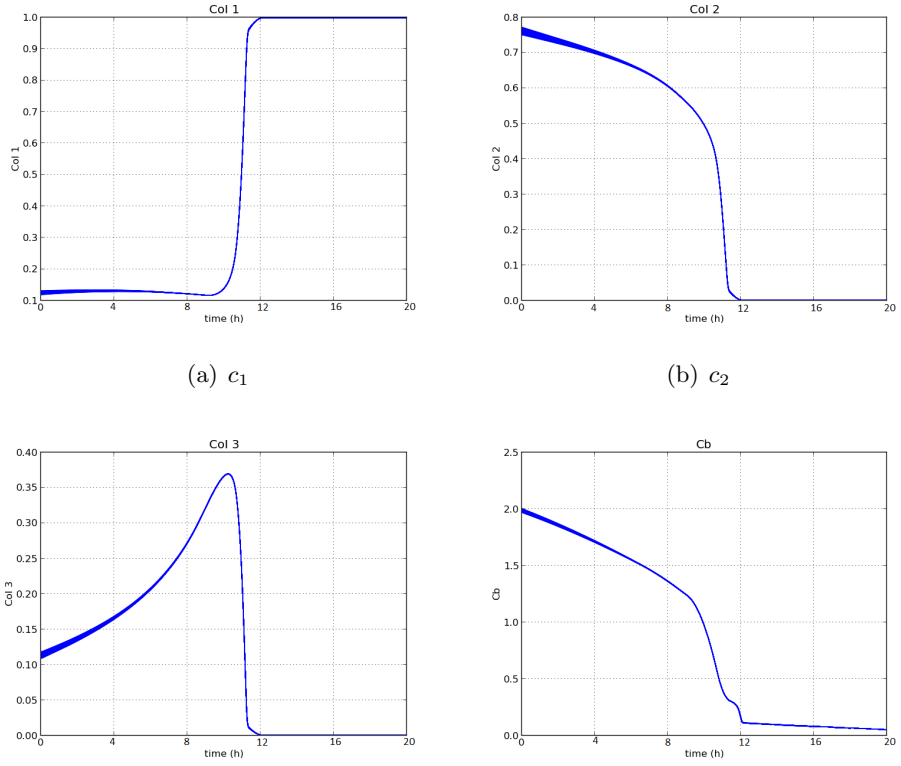
The solution obtained when solving the optimization problem (3.27) fits the experimental values perfectly. The time profiles of the simulated fluxes and CoIs obtained from the optimization problem are reported hereafter.

For the toy network, the number of free fluxes is 3, and the number of CoIs is consequently 4. The time profiles for each simulated free flux, compared with the corresponding experimental profiles, are reported in Figure 3.11. The time profiles for each CoIs are reported in Figure 3.12). For the *E.coli* network, the number of free fluxes is 7, and the number of CoIs is consequently 8. The time profiles for each simulated free flux, compared with the corresponding experimental profiles, are reported in Figure 3.14. It can be noted that the profiles are smooth enough to be regressed with a polynomial function, which is a necessary condition to solve the optimization and to go on in the procedure.



(a) u_1 (b) u_2 (c) u_3

Figure 3.11: Profiles for all the fluxes of the toy network: in red are the experimental data, in blue the trajectories obtained solving the bilevel optimization problem. The plot shows very good accordance between experimental and simulated profiles.



(a) c_1 (b) c_2

(c) c_3 (d) c_b

Figure 3.12: Profiles for all the optimization coefficients of the toy network, obtained solving the bilevel optimization problem. Considering the structure of the problem, which makes use of a quadratic objective function and of an additional experimental constraint, these profiles can not be easily interpreted from a biological point of view.

The time profiles for each CoI are reported in Figure 3.13.
Once each time point for each profile has been saved, the simulation problem must be solved, to check whether it is possible to return from the CoIs to the fluxes due to the degeneracy of the problem. The formulation of the inverse problem is:

$$\min_{\mathbf{u}\in\mathbb{R}^{n-m}} f = \mathbf{u}^T \cdot \mathbf{C} \cdot \mathbf{u} \tag{3.33}$$

$$\text{s.t. } \mathbf{IR} \cdot \mathbf{K} \cdot \mathbf{u} \geq 0 \tag{3.34}$$

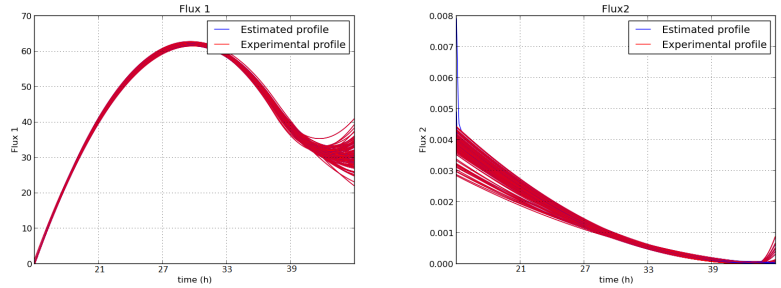
$$\mathbf{UB} - \mathbf{K} \cdot \mathbf{u} \geq 0 \tag{3.35}$$

$$\mathbf{1}^T \cdot \mathbf{u} - c_b \geq 0 \tag{3.36}$$

For the toy network, the time profiles for the fluxes obtained from the simulation problem (3.33) are reported in Figure 3.16. For the *E.coli* network, the time profiles for the fluxes obtained from the simulation problem (3.33) are reported in Figure 3.16.
As can be noticed, this particular formulation of the optimization problem not only allows to obtain smooth profiles for the CoIs, but it also permits the solution of the inverse problem, re-obtaining the original profiles. The new constraint added to this problem practically sets all simulated fluxes to the experimental value. The problem is of course that this experimental parameter has no mechanistic explanation, and it will become critical when the experimental information is not available. This will probably present a problem when trying to use the procedure to predict the micro-organism behavior. Still this modification allows to solve the optimization problem respecting all the requirements, and it is useful to test if the procedure can be completed.

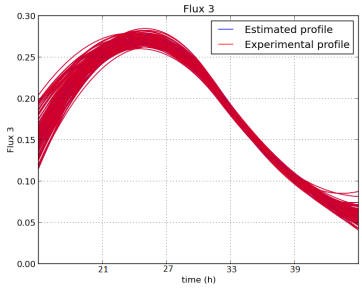
3.3.4 Regression for the grey-box approach

In this study the regression step has a fundamental role in enabling the solution of the dynamic system, since it has to link the variation of the fluxes to the variation of the extracellular concentrations. The first choice when a regression problem is approached is to select the kind of model to be used, i.e., linear or non linear with respect to the parameters. In fact algorithms and consequently software for the two categories are distinct. The shape of the model, and hence the kind of regression, since the two choices are linked, can be completely arbitrary or, in case the system under study is well known, it could be suggested by physical considerations. Considering the field of microbiology, since micro-organisms are extremely complex and highly non linear systems, the natural choice would be a non linear model. Nevertheless, in this study only linear regression was tested. The choice was sustained by a wide range of literature examples (Geeraerd et al., 2004; Gibson et al., 1988), which prove how linear mod-

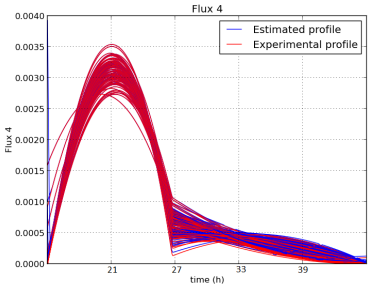


(a) u_1

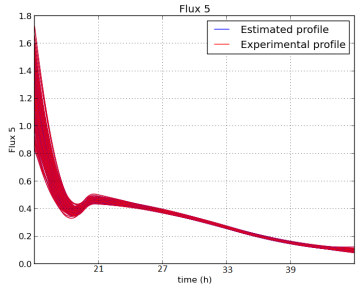
(b) u_2



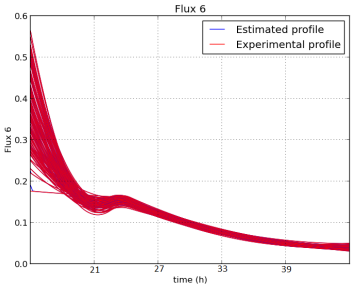
(c) u_3



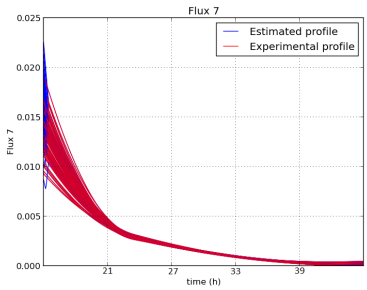
(d) u_4



(e) u_5

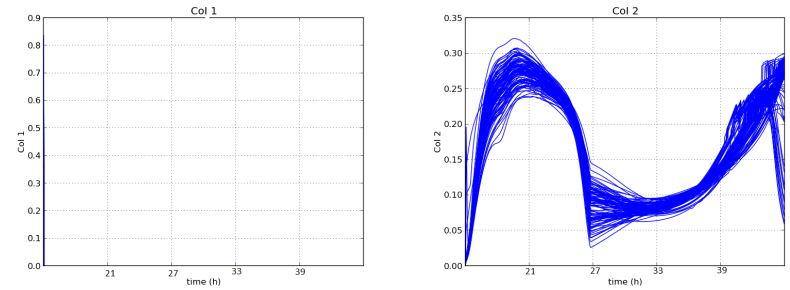


(f) u_6



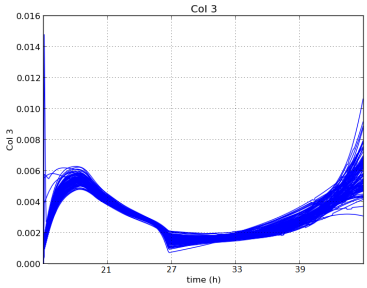
(g) u_7

Figure 3.13: Profiles for all the fluxes of the *E. coli* network: in red are the experimental data, in blue the trajectories obtained solving the bilevel optimization problem. The plot shows very good accordance between experimental and simulated profiles.

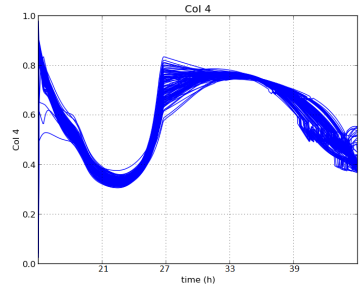


(a) c_1

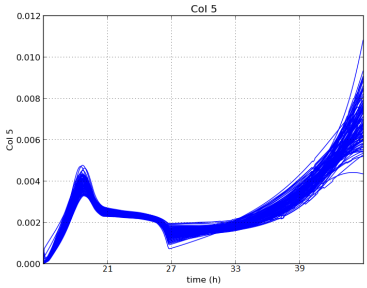
(b) c_2



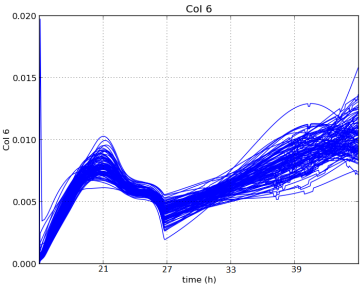
(c) c_3



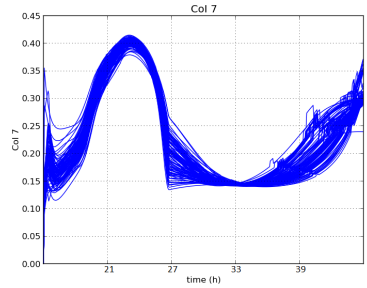
(d) c_4



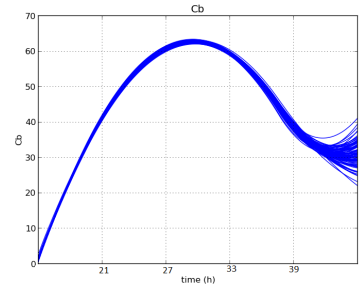
(e) c_5



(f) c_6

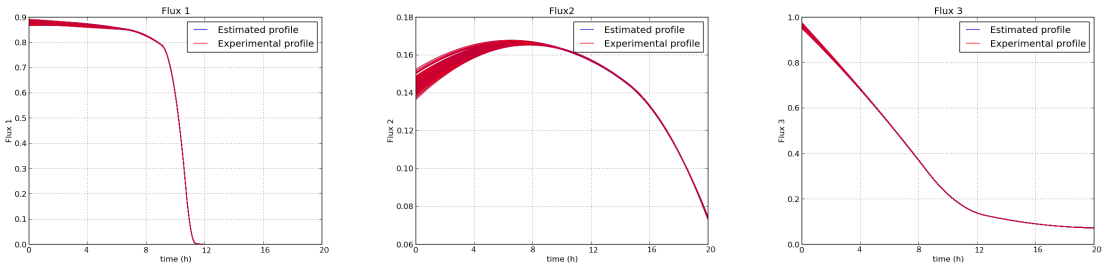


(g) c_7



(h) c_b

Figure 3.14: Simulation profiles for all the optimization coefficients of the *E. coli* network, obtained solving the bilevel optimization problem. Considering the structure of the problem, which makes use of a quadratic objective function and of an additional experimental constraint, these profiles can not be easily interpreted from a biological point of view.



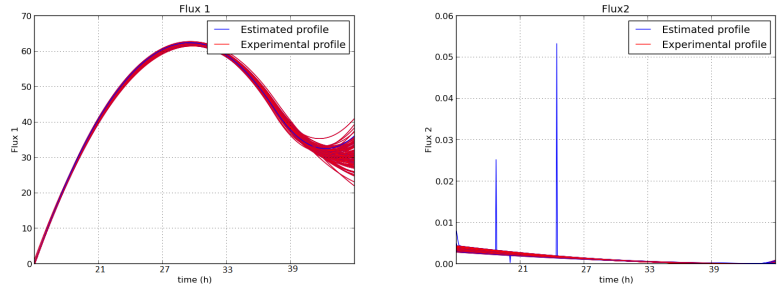
(a) u_1 (b) u_2 (c) u_3

Figure 3.15: Flux profiles of the toy network obtained from the simulation problem (blue) vs experimental flux profiles (red). The formulation of the simulation problem allows to reobtain almost perfectly the experimental profiles.

els, probably at the cost of a higher but still reasonable number of parameters, can provide good enough fitting of the biological experimental data. The results of the study confirmed this observation. Moreover, additional motivations were the easier implementation of a linear regression and the major number of software available to perform it compared to the non linear case. Stressing the concept, there were no mechanistic reasons to exclude the use of a non linear regression, which could theoretically better interpret the inherent mechanisms of a biological systems. The development of non linear models for microorganisms could surely represent a promising topic for future studies.

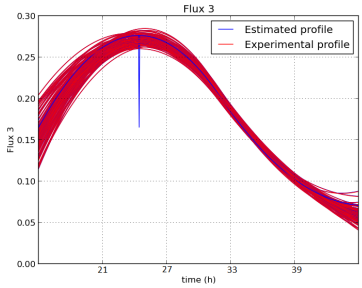
To obtain a regression curve which well fits the data is quite easy, but to obtain a good black-box model is far more difficult. The general rules and principles to formulate a proper black-box model are explained in subsection 1.5.3. Much more factors than the mere descriptive power of the model have to be considered, and to look at the data from different perspectives can help in the selection of the model. Different kinds of graphs are normally used to support the choice. Although some criterion based on these graphs can be implemented, it is difficult to formulate general rules. These criteria in fact could reveal to not be always effective, given the extreme variability of the input data and the lack of other information. If the user has some experience, the best criterion is always to look at the graphs and select case-by-case. The manual selection is suggested, but if the regression is inserted in a wider procedure, such as in this case, it can make the solution slower and non-automatic. In the following steps, different situations in which a delicate decision is required will occur. Whether to ask the user a manual choice or to leave the selection to the automatic criterion implemented will be decided case-by-case, depending on the respective improvements and on the loss of computation efficiency.

The common structure of the input data to a regression problem is different from the one

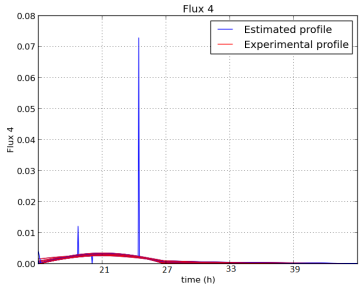


(a) u_1

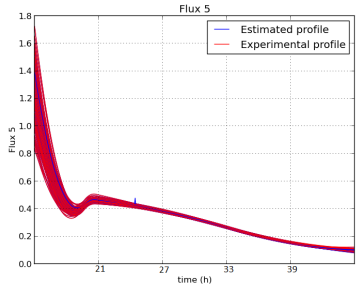
(b) u_2



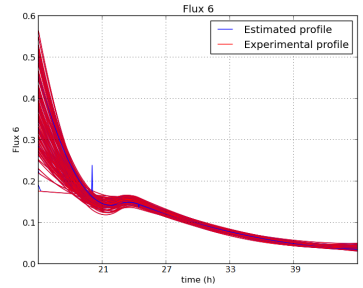
(c) u_3



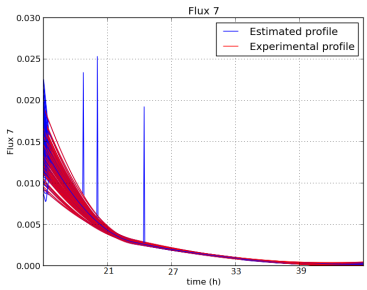
(d) u_4



(e) u_5



(f) u_6



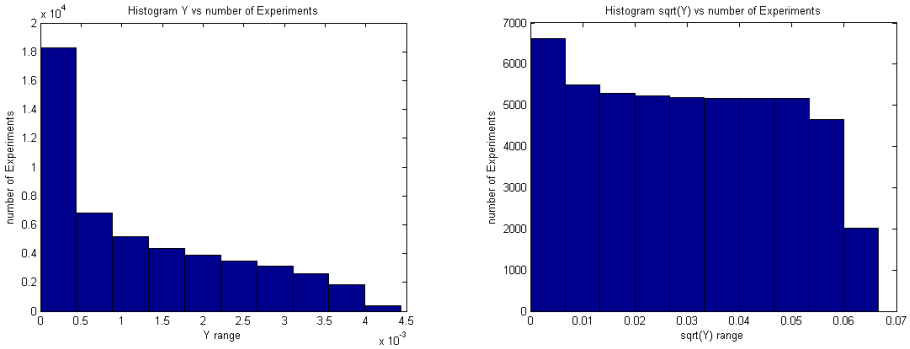
(g) u_7

Figure 3.16: Flux profiles of the *E.coli* network obtained from the simulation problem (blue) vs experimental flux profiles (red). The formulation of the simulation problem allows to reobtain almost perfectly the experimental profiles. The abrupt deviations show the existence of alternate optima.

previously presented (3.2.2). There is no longer need to visually preserve the dynamic behavior of the data. The regression considers for each variable a unique vector containing all the points available, listed without any specific order and independently whether they are correlated or not. The matrix containing the input data to this problem has the independent variables on the columns, and all the experimental values for each variable on the rows. Since the total amount of data for each variable is 500 time points x 100 iterations = $5e5$ points,the input matrix is shaped $5e5 \times m$, where m is the number of independent variables.
Before starting the regression, data are treated according with data pre-processing operations (2.4.1). Considering an experimental range, the best condition to perform a regression is generally to have data uniformly distributed on the whole range. During the parameter estimation step in fact, the best model is selected by trying to reduce the total distance between the points of the estimated model and the corresponding experimental values. Consequently, without a uniform distribution of data, the model would better fit the area where there are more data available, providing instead a worse fit where there is lack of data. It should be better, before performing regression, to check the distribution of data, and eventually to modify it. Since a black-box model is being considered, the regression operates in the same way on every possible set of data given as input. The substitution of a variable provides a regression curve which fits the new data, but it does not affect the global procedure. The estimated variables should then be re-substituted to be compared with the corresponding experimental profiles. An immediate way to visualize how the data are distributed on the experimental range is to plot an histogram. The experimental range of each variable is equally subdivided in intervals on the abscissa x and the number of experiments which falls in each interval is reported on the ordinate y (Figure 3.17). It can be seen how the distribution of data is much more uniform on the experimental range using the square root of the original variable.

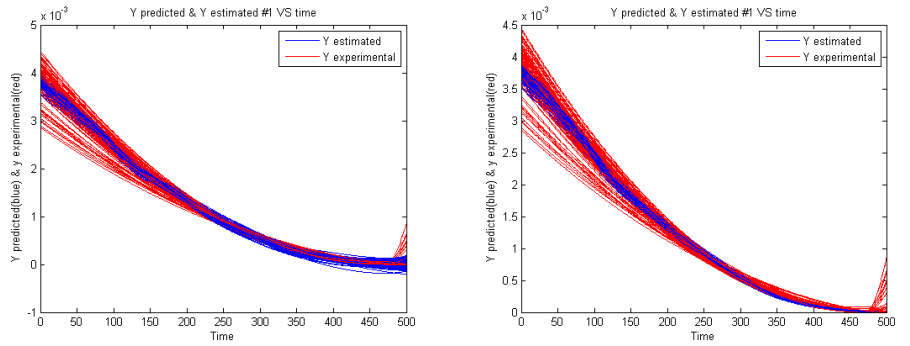
To generate the final results of this study an unmodified set of data for the dependent variables was used, since no relevant improvements in the fitting were registered with substituted variables (Figure 3.18). Furthermore, there are no general criteria to modify the data, and the substitution should be made by the user looking at the histogram of each variable. Consequently, it represents a limit to the possibility of making the procedure fast and automatic.

The regression method chosen for this study was Partial Least Squares (PLS). This method combines the advantages of the Principal Component Analysis (PCA), which reorganizes the original set of data, to good fitting properties. The algorithm was implemented in a software by Geert Gins, and it is a slightly modified version of the NIPALS algorithm (2.4.4). Mean-centering and variance-scaling are automatically performed on the input data by the software.



(a) u_2 (b) $\sqrt{u_2}$

Figure 3.17: Histograms for the flux 2 of the *E. coli* network: a) Y range vs number of experiments for every interval of Y values; b) $\sqrt{Y}$ range vs number of experiments for every interval of $\sqrt{Y}$ values. The distribution of data in plot b) is much more uniform than in plot a).

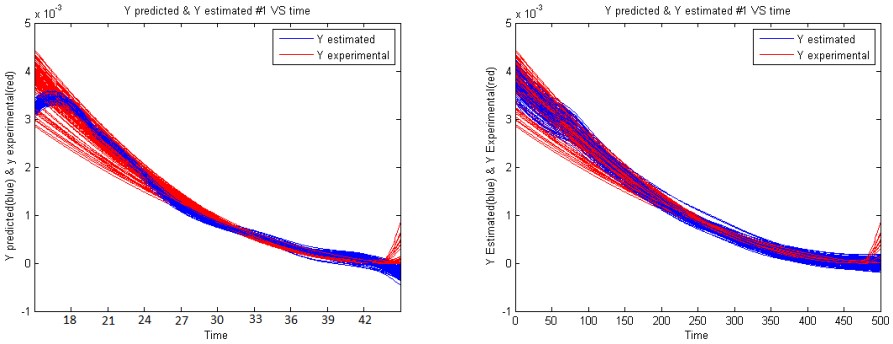


(a) u_2 (b) $\sqrt{u_2}$

Figure 3.18: Estimated model and experimental profiles comparison for the flux 2 of the *E. coli* network: *a)* the regression is performed using the variable u_2; *b)* the regression is performed using the variable $\sqrt{u_2}$. The substitution of the variable gives no evident improvements.

The PLS regression is not uninfluenced when providing the entire matrix containing all the dependent variables as input or just one column a time. If the regression problem is run with more than one dependent variable, the PCA axes slightly rotate, and the fitting of each variable gets worse. In this study a regression problem for each dependent variable was solved, since there are no particular reasons why the dependent variables should adapt to each other (Figure 3.19).

Previously to regression, a cross-validation is trained. Cross-validation, in this situation, is



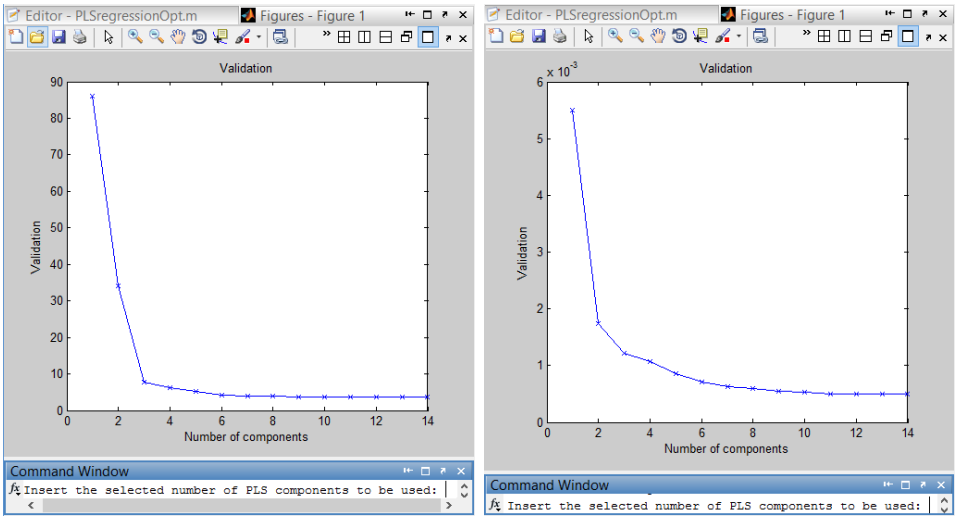
(a) Regression of $\mathbf{Y}$ (b) Regression of $\mathbf{y}$

Figure 3.19: Estimated model and experimental profiles comparison for the flux 2 of the *E. coli* network: *a)* the PLS regression giving input Y as a matrix; *b)* PLS regression providing input Y as a vector. The fitting is better when giving as input only one dependent variable a time.

meant to select the minimum number of terms which guarantees a good fitting. Limiting the number of terms is in line with *parsimony*, which is one of the fundamental principles of regression (1.5.3). The cross-validation step consists of comparing many kinds of models according with a defined criterion. The comparison does not only take into account the model obtained when training on the whole set of data, which tests only the descriptive capability of the model. The whole time range is divided into uniform intervals, and the model is trained as many times as there are intervals on a set of data which excludes one interval at a time. The criterion chosen in this study for the comparison is to compute the global *sum of squared errors* (SSE) for each model, weighted with a term which takes into account the number of terms considered.

$$\text{SSE} = \sum_{j=1}^{n} \frac{(z_j - y_j)^2}{n - i} \tag{3.37}$$

where n is the number or experimental data and i is the number of components selected. The SSE includes not only the training data of each cross-validation iteration, but the whole time range. Thanks to the denominator, the graph of SSE vs number of terms can even show a non-monotonic decrease, allowing a more correct choice, which takes into account not only the fitting but also the generalization properties of the model. Although some general criteria exist in literature to select the best number of terms on this graph (2.4.5), they were proven not to always accurately work. After various attempts, it was preferred to leave the choice to the user. During the cross-validation the solver stops as many times as there are Y variables,



(a) (b)

Figure 3.20: Interface for cross-validation step. The y axis called *validation* is the SSE value, while the x axis indicates the corresponding number of terms. *a)* There is a sharp change in the derivative of the curve, and the choice easily falls on 3 terms. *b)* In this case the SSE continues to gradually decrease, and the choice is less evident: a good value in this situation could be around 6 or 7 terms.

and it asks for a numerical value which is then saved as the number of terms to describe that variable. The number of terms to be selected is generally the one which guarantees the lowest SSE value, i.e., the best description of the data, according with a sufficient differential decrease of the SSE compared with the previous terms. Sometimes the choice is easy, since there is a visible gap in the derivative of the SSE curve between two successive terms. Usually, in these situations, an almost asymptotic value is reached for a very low number of terms. In other cases the choice is more delicate, and the number of terms must be chosen higher (Figure 3.20).

Generally, the use of principle components analysis represents a great aid for regressions. This is even more true for cross-validation. Cross-validation, in fact, compares the performance of many different models, starting from the easiest one and increasing gradually its complexity by adding new terms. If the set of input variables is not reorganized, there is no way to decide which variable should be added at each iteration. All the possible permutations of terms must be explored, causing a notable increase of the computational time. PCA instead generates the so called *components*, which are linear combinations of the original set of variables, and it orders the components for *importance*, i.e., which percentage of the original variance each component can explain. Hence, the cross-validation starts exploring from the most important

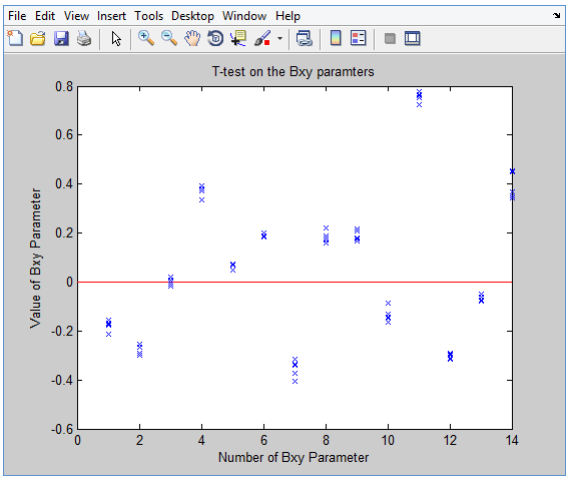


Figure 3.21: Graph to select the number of parameters, which reports the number of the parameters on the abscissa and the value of the parameters on the ordinate for the different cross-validation iterations. In this case, parameter n°3 can be disregarded based on the t-test.

component, and for every iteration it just has to add the successive component in order of importance. PCA is naturally part of the PLS algorithm (2.4.4), since PLS briefly consists of two combined PCAs on the dependent and independent variables, which are able to influence each other by making the final components slightly rotate.

A further step can be added to cross-validation before the final parameter estimation. The parameters obtained from the cross-validation can be checked to decide whether they can be assumed to be zero. For every training set of data and every parameter a value is now available from cross-validation. A t-test with significance level of 5% is performed on the distribution of parameter values, to determine if they can be assumed to come from a distribution with null mean and unknown variance (Figure 3.21). If the t-test accepts the null hypothesis, that term is disregarded in the following regression.
After performing cross-validation, which selects the number of components and eliminates the ones whose parameter is assumed to be null, a new PLS regression is performed, introducing one column of the Y matrix a time and already providing the number of components to be used. This is the proper regression step, which estimates the parameters of the model. Three kinds of linear model structures were tested, gradually increasing the complexity: *(i)* a linear model both in the parameters and the variables, *(ii)* a linear model in the parameters and quadratic in the variables, with only quadratic terms; *(iii)* a linear model in the parameters and quadratic in the variables, with both quadratic and cross-product terms. The number of total terms increases from one model structure to the other, but thanks to cross-validation, the

final number of parameters of the model is limited, depending on the number of components selected. In case of number of independent variables $m = 4$, the three model structures would appear as:

$$\mathbf{Y} = a_1 \cdot \mathbf{x}_1 + a_2 \cdot \mathbf{x}_2 + a_3 \cdot \mathbf{x}_3 + a_4 \cdot \mathbf{x}_4 \tag{3.38}$$

$$\mathbf{Y} = a_1 \cdot \mathbf{x}_1 + a_2 \cdot \mathbf{x}_2 + a_3 \cdot \mathbf{x}_3 + a_4 \cdot \mathbf{x}_4 + a_5 \cdot (\mathbf{x}_1)^2 + a_6 \cdot (\mathbf{x}_2)^2 + a_7 \cdot (\mathbf{x}_3)^2 + a_8 \cdot (\mathbf{x}_4)^2 \tag{3.39}$$

$$\begin{aligned}\mathbf{Y} = {} & a_1 \cdot \mathbf{x}_1 + a_2 \cdot \mathbf{x}_2 + a_3 \cdot \mathbf{x}_3 + a_4 \cdot \mathbf{x}_4 + a_5 \cdot (\mathbf{x}_1)^2 + a_6 \cdot (\mathbf{x}_2)^2 + a_7 \cdot (\mathbf{x}_3)^2 + a_8 \cdot (\mathbf{x}_4)^2 + \\ & + a_9 \cdot \mathbf{x}_1 \cdot \mathbf{x}_2 + a_{10} \cdot \mathbf{x}_1 \cdot \mathbf{x}_3 + a_{11} \cdot \mathbf{x}_1 \cdot \mathbf{x}_4 + a_{12} \cdot \mathbf{x}_2 \cdot \mathbf{x}_3 + a_{13} \cdot \mathbf{x}_2 \cdot \mathbf{x}_4 + a_{14} \cdot \mathbf{x}_3 \cdot \mathbf{x}_4\end{aligned} \tag{3.40}$$

Generalizing the expression (3.40):

$$\mathbf{Y} = \mathbf{a}_1 \cdot \mathbf{X} + \mathbf{a}_2 \cdot \mathbf{X}^2 + \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} a_{3\,i,j} \cdot \mathbf{x}_i \cdot \mathbf{x}_j \tag{3.41}$$

Total number of terms		
Kind of model	**Toy network**	***E.coli* network**
Linear	4	7
Quadratic	8	14
Quadratic + cross-products	14	34

Table 3.2: Total amount of terms for PLS regression for the different model structures and case studies.

Although there exist many criteria to verify the goodness of a regression model, the final judgment should always be based on the fitting, by just looking at the estimated and the experimental profiles of the dependent variables over time. To compare the estimated model to the experimental values, all the pre-processing operations must be inverted. *(i)* For variance-scaling, each column must be multiplied by the respective variance; *(ii)* for mean-centering, to each column the mean of that column must be added; *(iii)* for substitution of variables, the original variable must be plotted.

3.3.5 Results of PLS regression for the grey-box approach

The aim is to express the CoIs as a continuous function of the extracellular concentrations. This regression model enables to obtain the fluxes starting from the experimental values of

the concentrations, by solving the optimization problem and the regression. The part of the procedure which includes the optimization and the regression can be written as:

$$\begin{cases} \mathbf{u} = g(\mathbf{c}) \text{ (Optimization)} \\ \mathbf{c} = f(\mathbf{x}(t)) \text{ (Regression)} \end{cases} \rightarrow \mathbf{u} = g(f(\mathbf{x}(t))) \tag{3.42}$$

The inputs to each problem will be listed first, and then the plots resulting from the regression will be shown and discussed for both the case studies.

The regression software requires as input just which kind of model to select from the proposed ones, i.e., (3.38),(3.39) or (3.40), and the number of cross-validation iterations, i.e., the number of selected training sets. Since the cross-validation explores each possible model which can be generated from the original terms, in theory the wider the range of original terms, the better will be the regression, without renouncing to the principle of parsimony. Having 500 time points, a proper number of training sets for cross-validation could be 5. This means that the whole time range is subdivided in 5 intervals of 100 points each. Every cross-validation iteration excludes one of these intervals, and it trains the PLS regression on the remaining 400 points. At every iteration the interval which is disregarded is switched, exploring all the possible permutations.

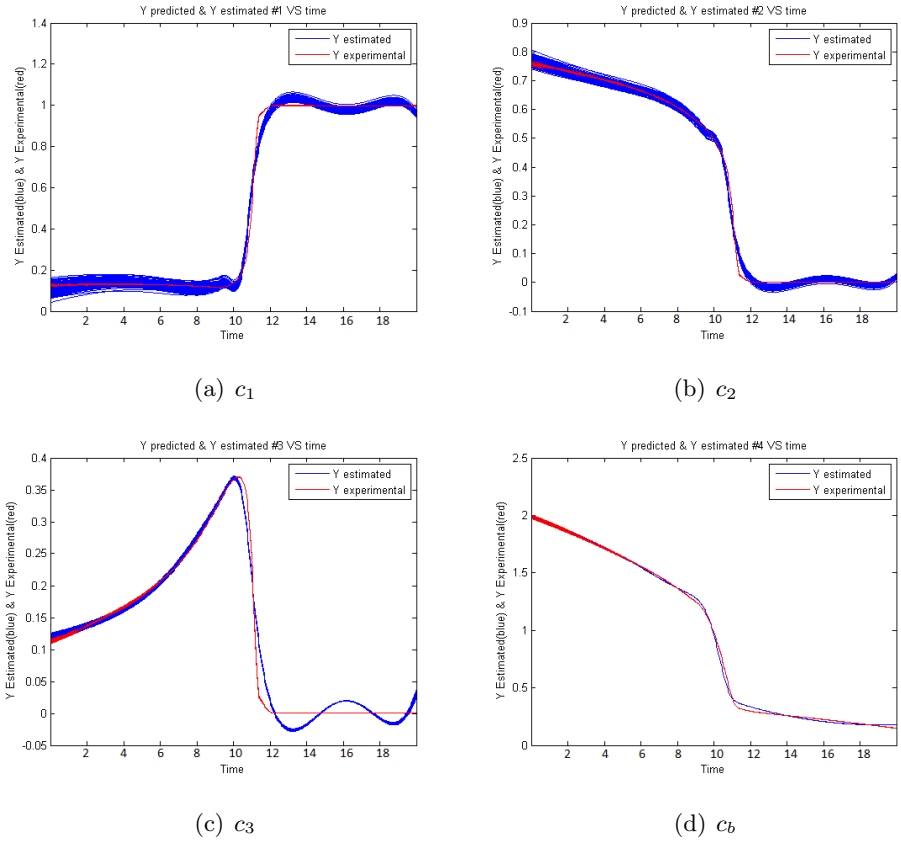
For the toy network, there are 4 extracellular concentrations and 4 coefficients, and the results of the regression are reported in Figure 3.22. The number of components used is reported in Table 3.3.

PLS for toy network, GB approach	
Dependent variable	N of components
CoI 1	8
CoI 2	8
CoI 3	7
Cb	5

Table 3.3: PLS components for the toy network.

For the *E.coli* network, there are 7 extracellular concentrations and 8 coefficients (Figure 3.23). The number of components used is reported in Table 3.4.

The resulting fit is very good, but the estimated curves oscillate much more than the experimental ones. This is probably due to the choice of a polynomial model. The polynomial model is useful because it is very flexible and it can adapt to almost every type of curve, but it does not catch the real mechanism which is behind the experimental profile. Hence, the



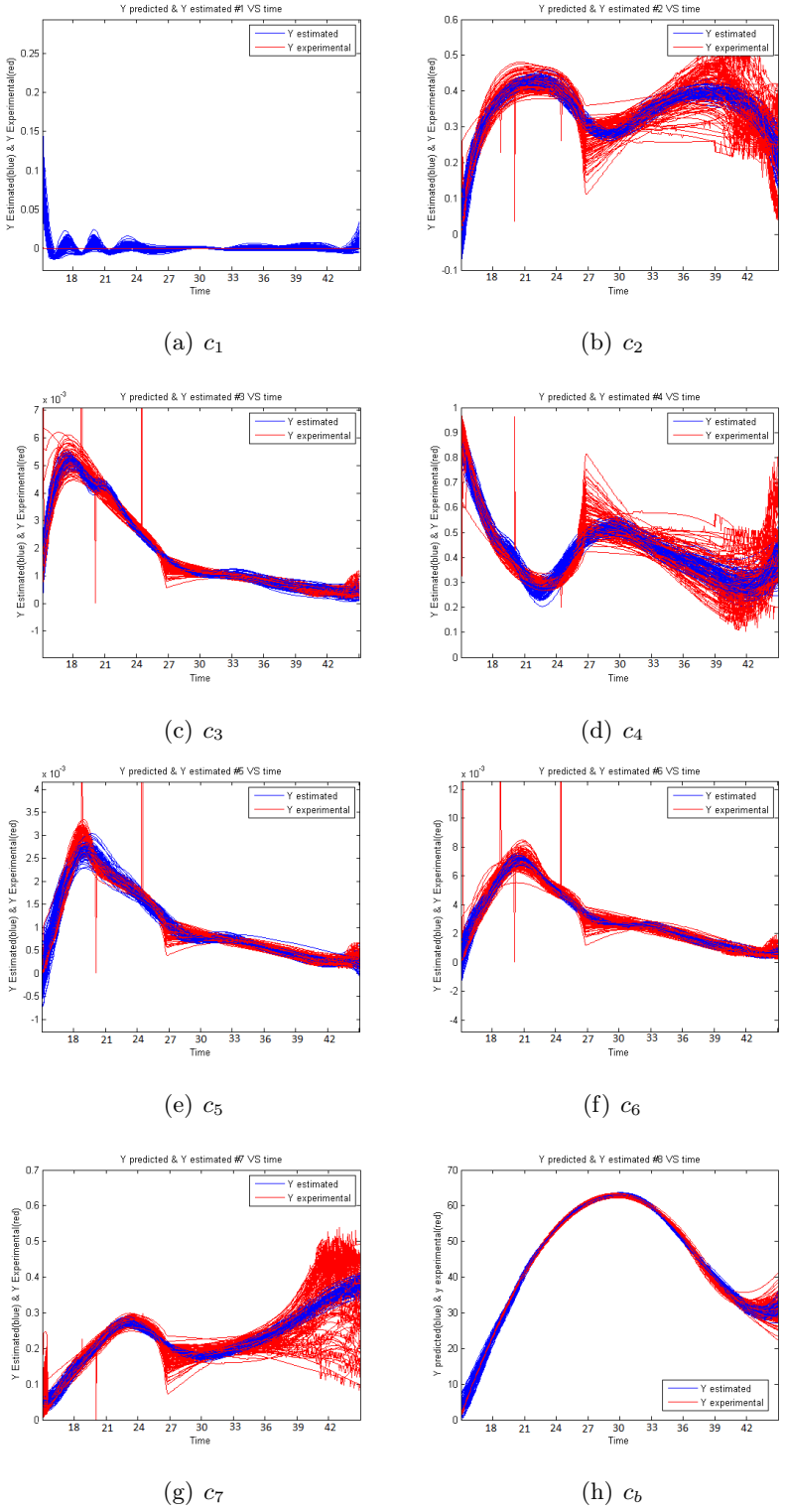
(a) c_1 (b) c_2

(c) c_3 (d) c_b

Figure 3.22: Coefficient profiles for the toy network: in red are the simulated profiles obtained from the bilevel optimization problem, in blue the profiles obtained using the PLS regression. The fit is quite good, but where the simulated profiles becomes asymptotic, the regression model oscillates around the asymptote.

number of terms is probably superior than a proper nonlinear model, causing more oscillations and the risk of over-fitting. This fact is particularly evident when the model reaches a stationary region: while the system asymptotically approaches the stationary value, the polynomial model oscillates around it. Polynomials in fact always go to infinity, and they are not able to simulate asymptotic behaviours. These oscillations have a notable consequence when the stationary value of the experimental profile corresponds to 0, since the model becomes negative.

What would happen in the grey-box approach if a regressed CoI turns negative? In the optimization problem the CoIs were normalized between 0 and 1, and their bounds were fixed. Considering the modified optimization problem (3.12) with the quadratic objective function $f = \mathbf{x}^T \cdot \mathbf{C} \cdot \mathbf{x}$, if all the CoIs were positive, the correspondent quadratic function in the bidimensional case would be $x_1^2 + x_2^2 = 0$. The 3D representation of the function is reported



(a) c_1 (b) c_2

(c) c_3 (d) c_4

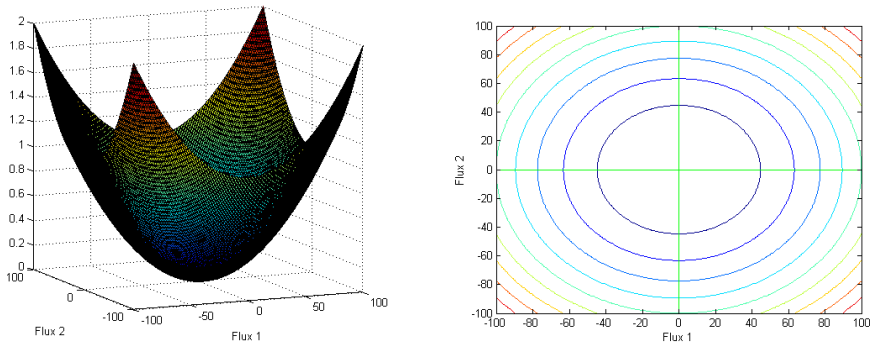
(e) c_5 (f) c_6

(g) c_7 (h) c_b

Figure 3.23: Coefficient profiles for the *E.coli* network: in red are the simulated profiles obtained from the bilevel optimization problem, in blue the profiles obtained using the PLS regression.

PLS for *E.coli* network, GB approach	
Dependent variable	N of components
CoI 1	12
CoI 2	15
CoI 3	12
CoI 4	7
CoI 5	15
CoI 6	15
CoI 7	8
Cb	5

Table 3.4: PLS components for the *E. coli* network.



(a) $x_1^2 + x_2^2$ `surf` (b) $x_1^2 + x_2^2$ `contour`

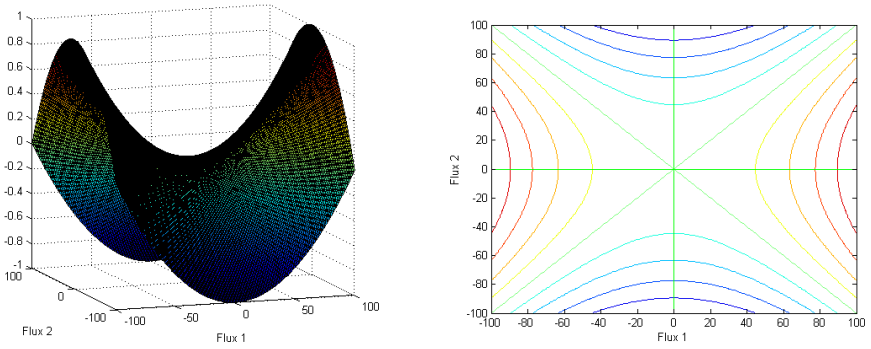
Figure 3.24: Surface plot and level line plot. The quadratic function shows a defined minimum in the origin.

in Figure 3.24).

The function shows a definite minimum in the origin, and the considerations made for the modified bilevel problem are perfectly valid. If one of the CoIs assumed a negative value instead, the bidimensional quadratic function would be $x_1^2 - x_2^2 = 0$, and its 3D plot would be completely different (Figure 3.25).

The function doesn't show a defined minimum anymore, but it goes to $-\infty$, and the optimization problem is able to reach only solutions which lie on the bounds.

The previous consideration shows the importance of bounding the regression for the CoIs to positive values. The easiest way to do this would just be to set the model to zero if its effective value turns negative. Nevertheless, this expedient causes a non smooth profile for the optimization parameters, which could create problems during the solution of the simulation



(a) $x_1^2 - x_2^2$ `surf` (b) $x_1^2 - x_2^2$ `contour`

Figure 3.25: Surface plot and level line plot. The shape of the quadratic function has changed, and it does not have a defined minimum anymore.

problem. The alternative is to use multiple linear regression instead of PLS, imposing each point to be positive.

3.3.6 Multiple Linear Regression for the grey-box approach

The importance of bounding the coefficients to positive values during the regression step was previously highlighted. Since no software for regression exists in MATLAB® which allows to set upper and lower bound to the variables, the problem must be formulated as a constrained regression.

The problem with PLS is that, given the complexity of the algorithm, it is difficult to modify some step and still maintain good fitting results. How to impose constraints to the parameter estimation? A normal least squares method finds the parameter values as analytic solution of an optimization problem. A least squares problem linear in the parameters $\boldsymbol{\beta}$ appears as:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mid y_i^{exp} - \sum_{j=1}^{m} \beta_j \cdot x_j(t_i) \mid^2 = \parallel \mathbf{y} - \mathbf{X} \cdot \boldsymbol{\beta} \parallel^2 \tag{3.43}$$

where $n > m$, i.e., the system is overdetermined. This means that there are more equations than unknowns. The matrix $\boldsymbol{\beta}$ of the coefficients is determined by minimizing the SSE. This optimization problem has an analytic solution by computing the derivative of the least squares with respect to the parameters and imposing it to be 0.

$$\hat{\boldsymbol{\beta}} = arg \min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) \tag{3.44}$$

$$\frac{\partial S}{\partial \hat{\boldsymbol{\beta}}} = \mathbf{X}^T \cdot \mathbf{y} - \mathbf{X}^T \cdot \mathbf{X} \cdot \hat{\boldsymbol{\beta}} = 0 \tag{3.45}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y} \tag{3.46}$$

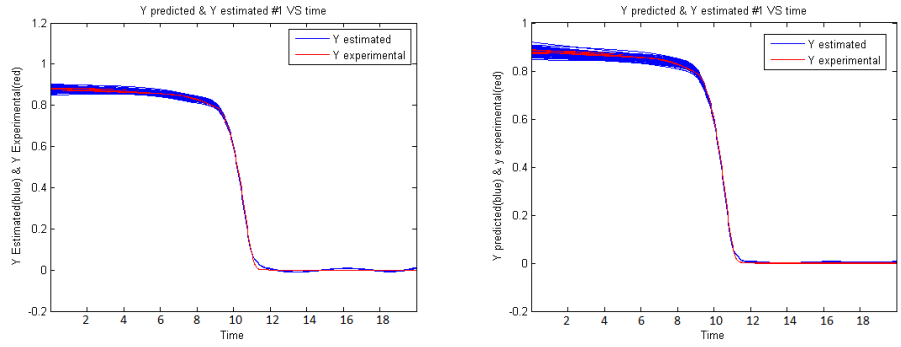
where the term $(\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T$ is called *pseudoinverse*.
A way to introduce constraints in the regression could be to substitute the analytic expression of the solution of this problem with a solver for constrained least squares method such as `lsqlin` in MATLAB®, which requires constraints to be linear in the parameters. This strategy was first tested on a PLS algorithm. Nevertheless, the complexity of the PLS algorithm makes very difficult to modify some passage, and the final fitting is not good any more.
An alternative chance is to renounce to the advantages of the PLS implementing this strategy on a normal *multiple linear regression problem* (MLR) (Figure 3.26, Figure 3.27). The input dependent variables were provided as a matrix, regressing the variables in the same problem. Differently from PLS, this does not affect the final results, since every column of the input matrix is individually considered. It was necessary to implement a non-negativity constraint for each variable and for each time point, bounding the entire estimated profile to be positive. This means to add a big number of constraints, but the solution is quite fast since the problem is QP:

$$\min_{\boldsymbol{\beta}} \quad \| \mathbf{Y} - \mathbf{X} \cdot \boldsymbol{\beta} \|^2 \tag{3.47}$$

$$\text{s.t.} \quad \boldsymbol{\beta} \cdot \frac{(\mathbf{X} - \bar{\mathbf{X}})}{\boldsymbol{\sigma}_{\mathbf{X}}} \cdot \boldsymbol{\sigma}_{\mathbf{Y}} + \bar{\mathbf{Y}} \geq 0 \tag{3.48}$$

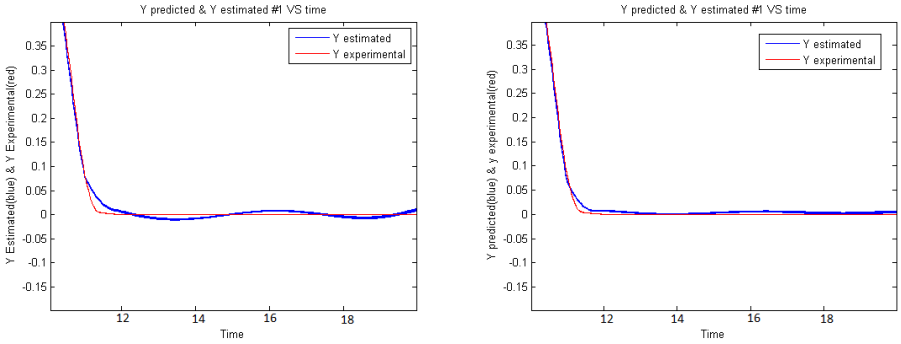
The particular way to express the constraint stems from the necessity of comparing the non mean-centered and non variance-scaled estimated values with the experimental data. Nevertheless, the expression is still linear in the parameters, and the optimization problem remains QP.
The results of the regression using the MLR algorithm for the toy network are reported in Figure 3.28. For the *E. coli* network instead, the results are reported in Figure 3.29. When looking at the graphs, the constrained MLR seems to provide a very good fitting, solving the problem of the negative values for the fluxes.

For the MLR regression the PCA was not implemented, and the cross-validation is not applied then. The model is always the most complex one, built using all the terms of the quadratic



(a) PLS (b) Constrained MLR

Figure 3.26: Comparison of the estimated models obtained with the two different regression algorithms. In red are the experimental data, in blue the profiles obtained with PLS regression (a) and constrained MLR (b). The curve from constrained MLR never becomes negative.



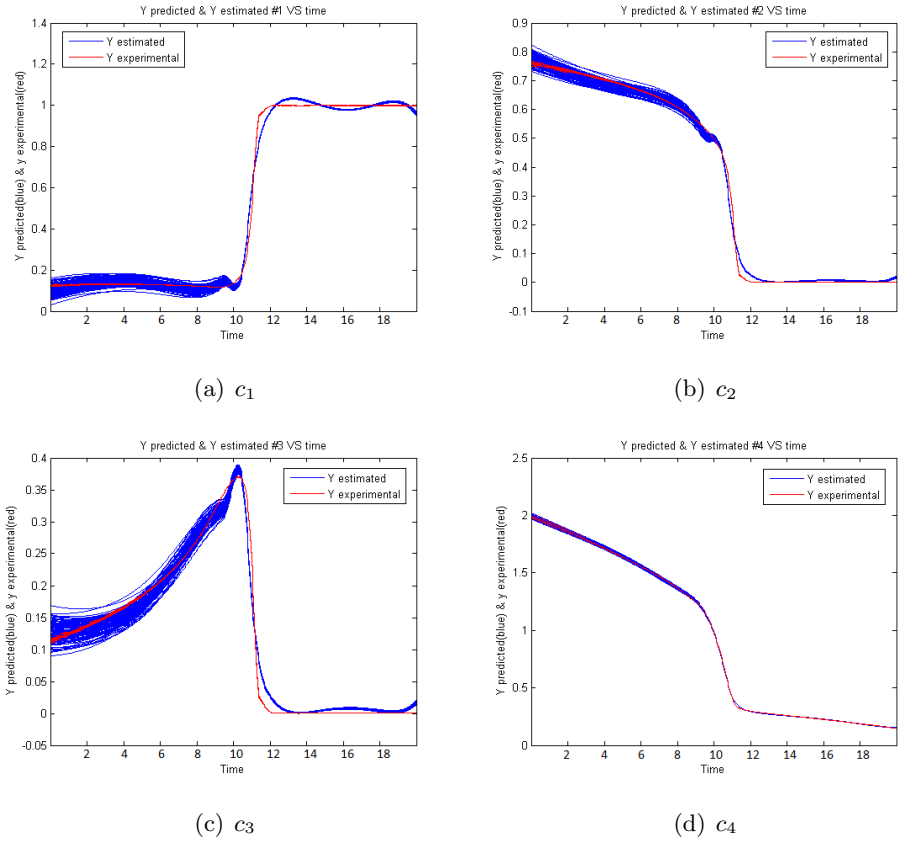
(a) PLS (b) Constrained MLR

Figure 3.27: Comparison of the estimated models obtained with the two different regression algorithms. The graphs are the same as the previous ones, but zoomed around the asymptotic region to highlight the difference between PLS and constrained MLR regression.

form with cross-product terms. The number of used terms is 14 for the toy network and 34 for the *E.coli* network. This is such a high number of terms, much higher of what is necessary to obtain a good fit. Since part of the terms gives just a small contribution to the fitting, the fact of using so much terms does not notably influence the regression, but it can cause a loss of generalization properties for the model. For the future use of the procedure, PCA should always be implemented in case of using a linear model for regression.

3.3.7 Dynamic system solution for the grey-box approach

The grey-box approach is a succession of many steps, very different between each other and inspired from different studies. Each step implies a certain computational time, but it also



(a) c_1 (b) c_2

(c) c_3 (d) c_4

Figure 3.28: Coefficient profiles for the toy network using the constrained MLR regression: in red are the experimental data, in blue the profiles obtained with regression.

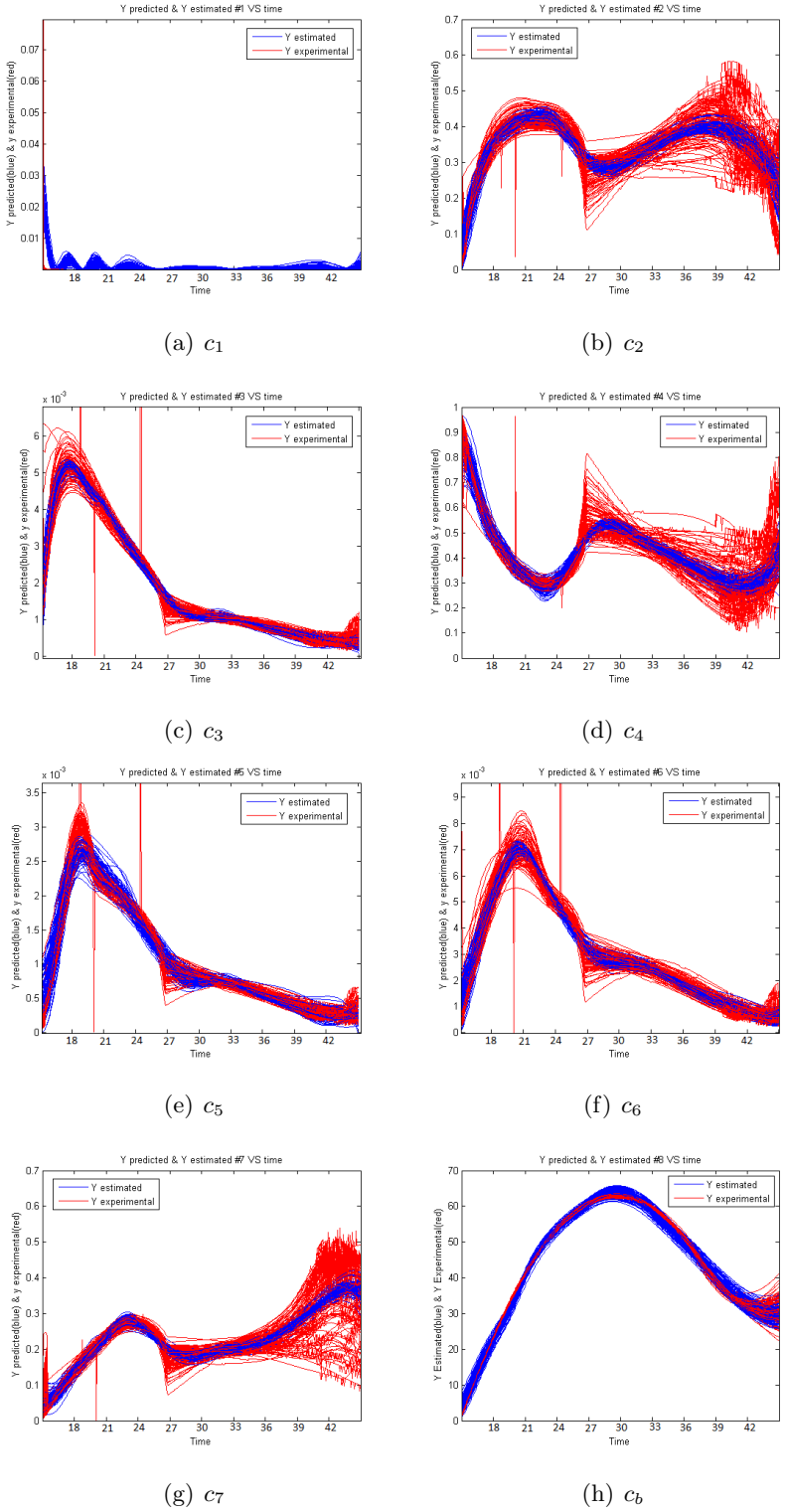
introduces a component of error with respect to the exact solution. Depending on how much the system under study is sensible to errors with respect to the experimental data, the increase of the error percentage could make the solution of the dynamic system difficult. The optimization problem in particular, derived from the FBA method, was extrapolated from its original context and slightly adapted for the aim of this application. The uncertainty introduced by this step makes the solution of the dynamic system very delicate.

Once an analytic expression for the fluxes has been found, the primary dynamic system can be solved. It is a system of as many ordinary differential equations as there are extracellular concentrations. The system appears as:

$$\frac{d\mathbf{C}_{macro}}{dt} = \mathbf{S}_{ext} \cdot \mathbf{K} \cdot \mathbf{u} \cdot N_{macro} \tag{3.49}$$

$$\mathbf{u} = f(\mathbf{C}_{macro}, T, pH, \ldots, \mathbf{\Phi}) \tag{3.50}$$

The extracellular concentrations $\mathbf{C}_{macro}$ will be generically called $\mathbf{X}(t)$ in the following trea-



(a) c_1 (b) c_2

(c) c_3 (d) c_4

(e) c_5 (f) c_6

(g) c_7 (h) c_b

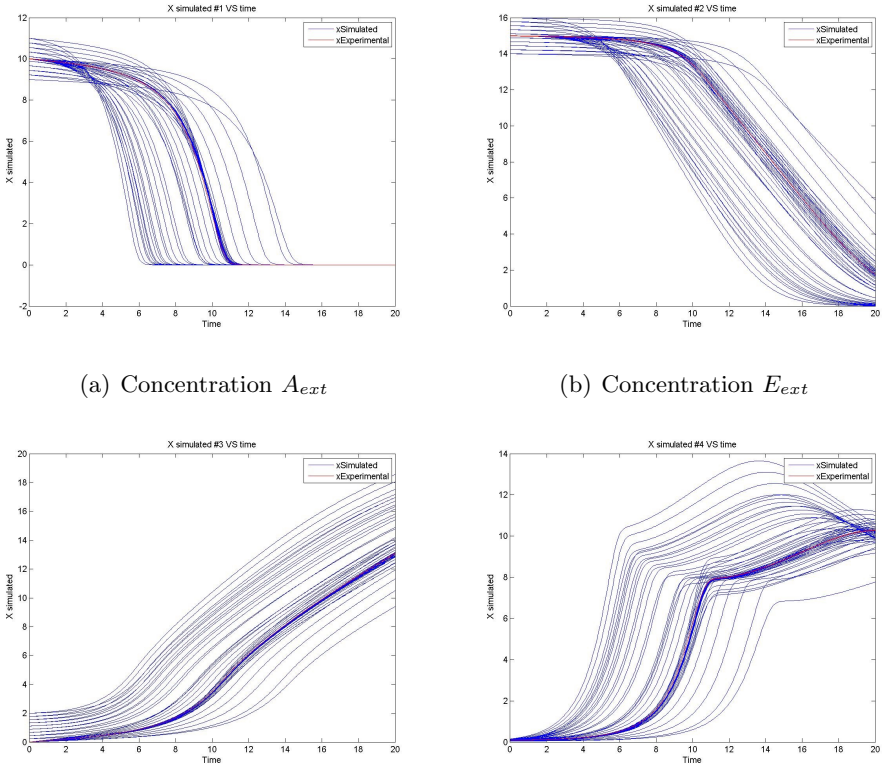
Figure 3.29: Coefficient profiles for the *E.coli* network using the constrained MLR regression: in red are the experimental data, in blue the profiles obtained with regression.

tise. Since in this problem the macroscopic number of cells is the last element of the vector of extracellular concentrations, N_{macro} will be substituted by $\mathbf{xx} \cdot \mathbf{X}(t)$, where $\mathbf{xx}$ is a row vector of zeros with only a 1 which corresponds to the extracellular concentration of cells. Working in an environment with constant temperature, pressure, pH etc., the only variables affected by the growing culture are the metabolite concentrations. The final system is:

$$\frac{d\mathbf{X}(t)}{dt} = \mathbf{S}_{ext} \cdot \mathbf{K} \cdot \mathbf{u} \cdot \mathbf{xx} \cdot \mathbf{X}(t) \tag{3.51}$$

$$\mathbf{u} = f(\boldsymbol{\beta}, \mathbf{X}(t)) \tag{3.52}$$

where $\boldsymbol{\beta}$ are the parameters estimated during the regression step.
Without extrapolating the results of the dynamic system in time, which is considered a risky operation, the model can be tested on different training sets obtained by perturbing the initial point for $t = 0\,h$ and solving a different simulation from each of these starting point. This effect was included in the procedure by reporting for each plot always 100 profiles, which take into account the variability of the system. Consider the toy network, whose equations are known and they can be used to generate infinite additional data. The 100 profiles of each concentration were already reported (Figure 3.3). Even if the variance between the different profiles, obtained by normally perturbing the starting point of each variable inside its 95% confidence interval, is quite limited in these graphs, it must be taken into account that the dynamic system solution internal to the procedure has to deal with much larger variance, due to the introduction of the optimization and regression steps. An analysis of sensitivity was performed on the dynamic system solution for the toy network by increasing the perturbation of the initial points. For systems of differential equations, a fundamental property is the *stability*. A dynamic system is stable if small perturbations of the input data lead to small perturbations of the outputs, i.e., if a trajectory, which starts from an initial point near the original profile, indefinitely stays inside a defined neighbourhood of this one. A trajectory, even called *orbit*, can be attracted from the original profile, converging on it after a transient, or it can be repelled, diverging from it. This concept is similar to the *condition number* of a function with respect to a variable, which indicates how sensitive the function is to errors in that variable. If a perturbation on that variable will be propagated or softened is a property of the system. The plots in Figure 3.30 report the profiles obtained by perturbing the starting point both of one concentration at a time and of more concentrations at the same time.
How much the variance increases along the time evolution between the different profiles is evident. This fact highlights that the system is very sensitive to errors, and it gives a measure of how difficult it could be to solve the entire grey-box procedure, where each step naturally implies its own error.



(a) Concentration A_{ext}

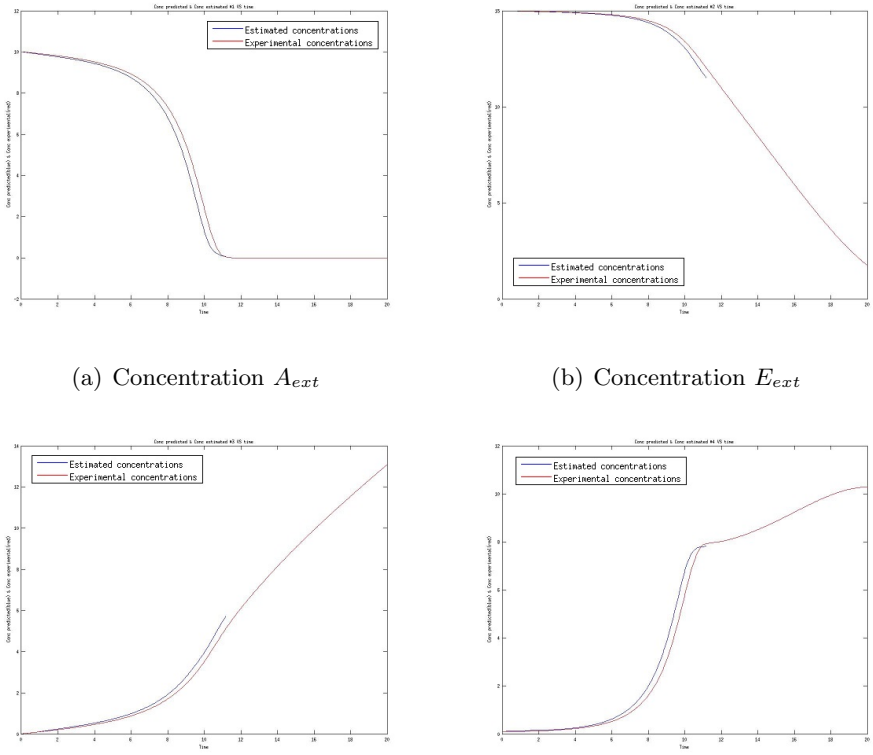
(b) Concentration E_{ext}

(c) Concentration F_{ext}

(d) Concentration BIO

Figure 3.30: Solution of the primary dynamic system for the toy network, artificially perturbing the starting point for each concentration: in red are the experimental data, in blue the simulated profiles. Although for a small perturbation of the initial points of the simulation, a much bigger increase of the variance is obtained, still, after a while, the different profiles seem to reconverge to the experimental trajectory.

The results of the dynamic system which includes all the steps of the grey-box procedure will now be presented. The information and the results of the optimization and the regression step are the ones presented in the subsection 3.3.3 and 3.3.6, respectively.
For the toy network, the number of extracellular concentrations is 4, and the results are reported in Figure 3.31. As can be noticed from the plots, the ODE solver is not able to complete the integration of the system over the whole time range. Both `ode45` and `ode15s`, for stiff problems, were tested, providing the same result. The default absolute and relative tolerances for these solvers are set to $1e-6$ and $1e-3$, respectively. Even the selection of a lower precision does not improve the results.
To understand where the problem lies, the procedure is split in many different steps, gradually complicating it and passing from the analytic solution towards the actual one.



(a) Concentration A_{ext}

(b) Concentration E_{ext}

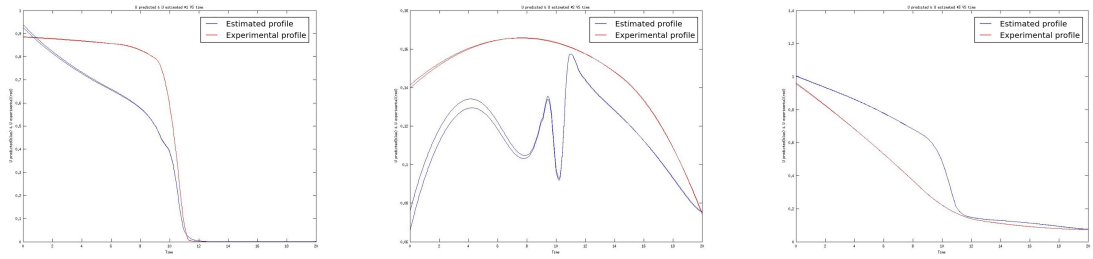
(c) Concentration F_{ext}

(d) Concentration BIO

Figure 3.31: Solution of the primary dynamic system for the toy network: in red are the experimental profiles, in blue a simulated one. The simulation stops before completing the integration.

The first step consists of solving the dynamic system using the profiles of the CoIs directly obtained from the simulation problem (Figure 3.12), excluding the regression. Since the simulation problem proved to properly work with these profiles, reproducing almost exactly the original fluxes, the solution of the dynamic problem should reproduce exactly the original profiles of the concentrations. The CoIs as continuous function of time are obtained with a simple linear interpolation between the adjacent values. This is necessary because the ODE solver makes use of a variable integration step, and it could require the values of the CoIs for different time points from the original 500 ones. As it was expected, the solution of the dynamic system with these profiles of the optimization coefficients returns the exact profiles of the concentration. The problem is not in the ODE solver or in the solution of the dynamic system. Further complications have to be added.

Since adding the regression of the CoIs as a function of the extracellular concentrations to the dynamic system the problem would be the complete one, which was proven not to be

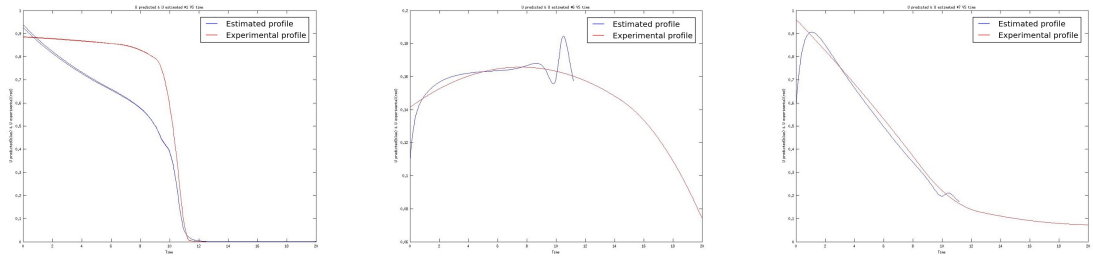


(a) u_1 (b) u_2 (c) u_3

Figure 3.32: Solution of the simulation problem with the regressed profiles of the CoIs point by point for the toy network. In red are the experimental data, in blue the simulated profiles. The simulated profiles do not closely follow the experimental ones, and they present abrupt oscillations.

able to complete the integration, the second step solves the simulation optimization problem with the regressed profiles of the CoIs for each of the 500 time points. The values of the fluxes are returned point by point, excluding the complications due to the integration. These profiles of the fluxes (Figure 3.32) are the same ones that the ODE solver uses to solve the dynamic problem. As can be seen, the simulated profiles of the fluxes are really far from the real ones, and they are even subject to notable and unmotivated oscillations. This fact does not find an explanation in the regressed CoIs (Figure 3.28, blue profiles), since these profiles, obtained with the MLR technique, don't show evident or big deviations from the training profiles (Figure 3.28, red profiles). This behaviour is probably due to the extreme sensitivity of the simulation problem to the CoIs provided as input. This is actually a problematic issue, since the profiles obtained from regression would never perfectly reproduce the original one. This kind of deviation cannot be easily avoided.

The confirmation that the problem of the dynamic system solution lies in the simulation problem can be obtained by checking the profiles of the fluxes during the solution of the dynamic system. Since the fluxes are not the variables of the dynamic system, the ODE solver does not return their value. This problem can be overcome by transforming the ordinary differential equation (ODE) system into a differential-algebraic equation system (DAE), and adding the algebraic equations which return the values of the fluxes: $u_i = 0,\ (\forall i \in 1, \cdots p)$, where n is the number of fluxes (or, expressing the fluxes as one vector of dimension $(n \,\mathrm{x}\, 1)$). A DAE system can be easily solved by `ode15s` by providing a singular mass matrix. The mass matrix $\mathbf{M}$ is a diagonal matrix of dimension $(m \,\mathrm{x}\, m)$, where m is the number of unknowns or equations, with 1 in correspondence of the differential equations and 0 in correspondence of the algebraic equations. The mass matrix simply multiplies the left argument of the



(a) u_1 (b) u_2 (c) u_3

Figure 3.33: Fluxes of the toy network obtained from the dynamic system with the regressed profiles of the CoI. In red are the experimental data, in blue are the simulated profiles. The integration stops when the simulated fluxes begin to oscillate.

dynamic system, not modifying the dynamic equations but setting to zero the derivatives of the stationary variables:

$$\mathbf{M} \cdot \frac{\partial \mathbf{x}}{\partial t} = f(t, \mathbf{x}) \tag{3.53}$$

The flux profiles obtained when solving the DAE system are reported in figure 3.33.
As it can be seen, the integration stops when the flux profiles begin to abruptly oscillate, confirming this fact to be the cause why the ODE solver cannot complete the integration.
The same considerations are valid for the *E. coli* network, but the negative effects caused by the extreme sensitivity of the simulation problem are much bigger, since the CoIs show profiles which are more difficult to fit, and the error introduced by the regression is consequently higher. Also for the *E. coli* network in fact, the ODE solver was not able to complete the integration of the dynamic system.

3.4 Black-box approach

3.4.1 Introduction

The results of the previous section (3.3) prove that the complications of the grey-box approach represent an insurmountable obstacle to overcome, and they make it impossible to complete the procedure of integrating the primary dynamic system. These complications derive from the attempt to insert biological information about the system for determining how the fluxes vary in time in the procedure, through the FBA method. The introduction of this information is fundamental to make the model more significant and representative of the real system.

Even if many models currently used in literature still maintain a large experimental characterization, the natural evolution of research consists of always inserting new and additional knowledge in the mathematical description of the system, increasing the mechanistic aspect of the model. Nevertheless, in the case of this study, since the grey-box approach showed to be too complicated and heavy to be solved, it was chosen to make the procedure lighter, eliminating the steps connected with the FBA method. This means of course to make a step back from the future of this kind of modeling, since part of the mechanistic information of the model is excluded and the experimental characterization of the model is increased. Still, considering the innovation of the proposed approach, it was decided to prefer the flexibility of the model, making it possible to complete the procedure. Stressing the concept, this was just a practical choice, and it does not exclude the future possibility of alternatively reformulating the whole procedure moving toward mechanistic models.

As was told in the introduction of the chapter (3.1), what distinguishes the two approaches is the way fluxes and extracellular concentrations are bonded. The black-box approach usually represents the first and the most direct attempt to approach modeling. In fact a simple black-box model is used, i.e., a fully experimental model. Disregarding any specific information about the system under study, this kind of model is just based on the experimental data available for the independent and dependent variables, i.e., extracellular metabolite concentrations and fluxes, respectively. A regression between fluxes and concentrations is performed to obtain a continuous function starting from isolated experimental points. Although to perform a regression could seem quite an immediate and simple task, many more aspects should be considered to obtain a model which not only fits well the data, but is also as much *significative* as a fully experimental model can be. In this context, for *significative* is meant a model which possesses good *generalization* properties (Geeraerd et al., 2004). This is of course difficult for a model which is built based on just a particular set of experimental data. The attempt to generalize the model can involve different sets of data inside the same experimental range, or data taken outside it. Testing the model according with the first case is called *validation*, in the second case *extrapolation*. In particular, the extrapolation of a model is usually an unrecommended practice. Nevertheless, considering a future application of the procedure as part of an online strategy of predictive control, the extrapolation required would be limited to a small time interval outside the experimental range. If the model would be able to catch just the direction in which the system is evolving, possibly it could already be considered useful information. This aspect of the black-box model will be tested at last after obtaining and validating the model.

3.4.2 Regression for the black-box approach

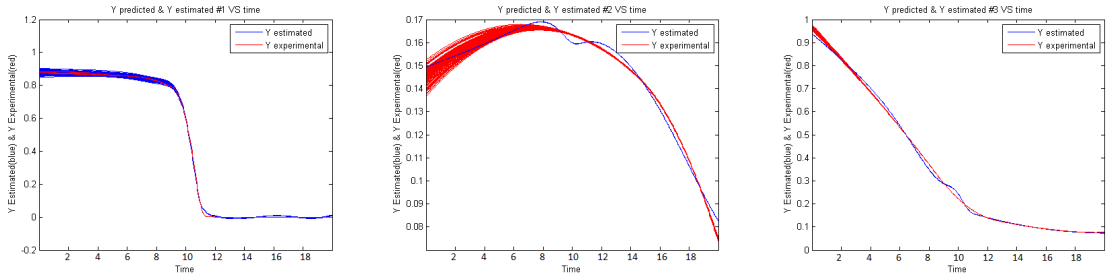
In the black-box approach, the distribution of the fluxes is directly linked to the variation of the extracellular concentrations. The fluxes are assumed as dependent variables, and regressed as function of the concentrations. Since the concentrations are the variables of the dynamic system, the model so formulated allows to compute the distribution of fluxes for every time point during the integration:

$$\mathbf{u} = f(\mathbf{x}(t)) \tag{3.54}$$

The considerations about the regression are the same as highlighted for the black-box approach (3.3.4). The only difference is about the negative values assumed by the dependent variables caused by the oscillations of the regression model. As was told, the black-box model is chosen as a polynomial linear in the regression parameters, due to its particular flexibility and adaptability. Since it does not reflect in any way the knowledge about the system, the number of parameter to be used is probably bigger than a proper non linear model, and this fact causes the model to oscillate more. Being a polynomial, the regression model is not able to approximate asymptotic behaviours, and when the system reaches a stationary region the polynomial tend to oscillate around the training profile. In particular, when the training profile asymptotically approaches 0, the regression model could assume negative values. While for the grey-box approach this fact could represent a problem (3.3.5), it is theoretically not the same for the black-box approach, where the dependent variables of the regression are no longer the optimization coefficients but the fluxes. There is no physical or biological reason why the fluxes should be bounded to positive values. Since the fluxes are just reaction rates, a negative value for a flux can be interpreted as an inverse reaction, i.e., a reaction which proceed consuming its products to produce the reactants. The only fluxes bounded to positive values are the ones which are selected by the irreversibility matrix **IR** and involved in the irreversibility constraint. Consequently, the negative values assumed by the fluxes should not cause problem to the solution of the dynamic system, and the PLS regression can be used instead of the MLR of the grey-box approach.

3.4.3 Results of PLS regression for the black-box approach

The results of the regression problem are presented for both case studies. The regression model is the one with cross-products (3.40), and the algorithm was PLS (2.4.4).
For the toy network, the number of extracellular concentrations is 4, and the number of free fluxes is 3. The regression graphs obtained are reported in Figure 3.34. The number of



(a) u_1 (b) u_2 (c) u_3

Figure 3.34: Flux profiles for the toy network: in red are the experimental data, in blue are the profiles obtained with PLS regression. The regression provides an accurate fit.

components manually selected for each free flux is reported in Table 3.5.

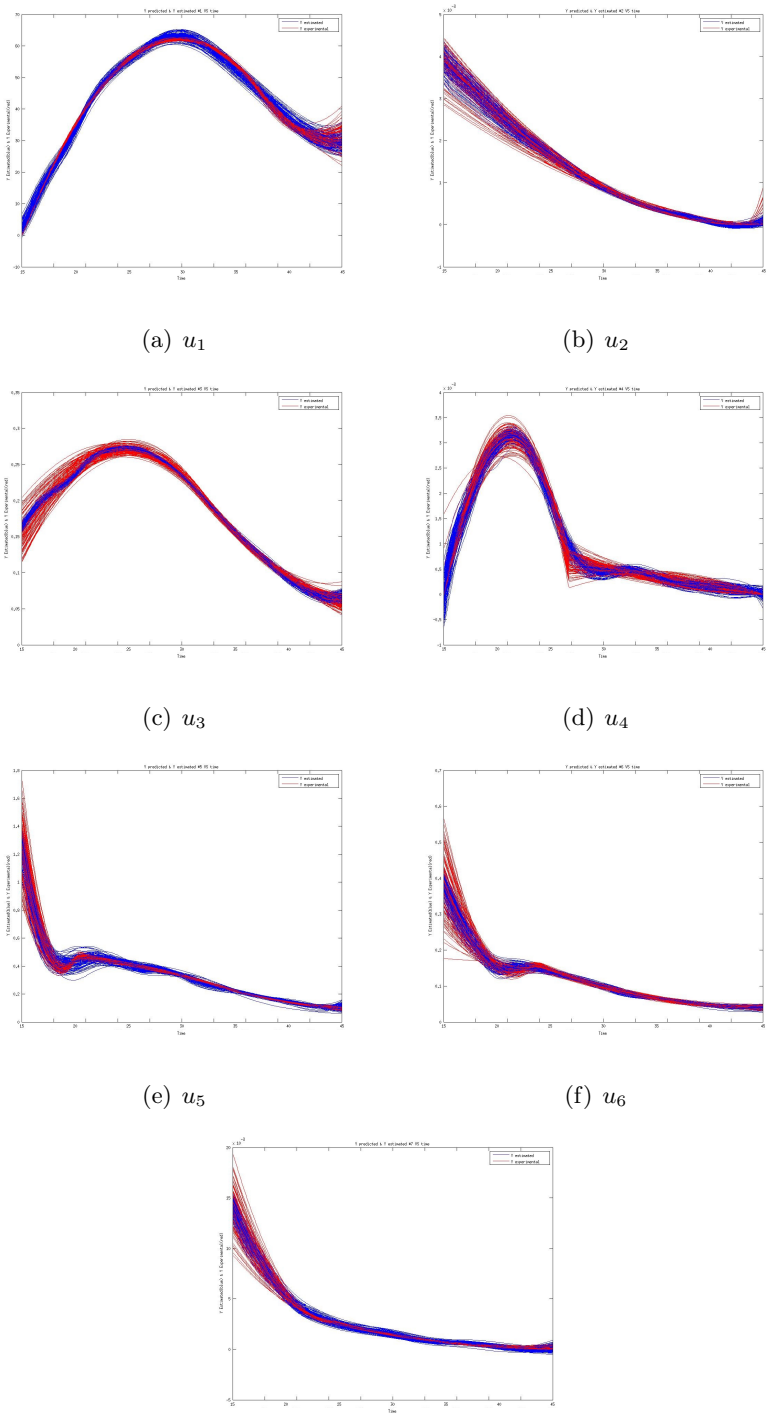
PLS for toy network, BB approach	
Dependent variable	N of components
Flux 1	8
Flux 2	4
Flux 3	5

Table 3.5: PLS components for the toy network.

The *E. coli* network is bigger than the toy network. The number of both extracellular concentrations and free fluxes is 7.The regression graphs obtained are reported in Figure 3.35. The number of selected components is reported in Table 3.6. The considerations are analogous to the grey-box approach.

PLS for *E.coli* network, BB approach	
Dependent variable	N of components
Flux 1	5
Flux 2	10
Flux 3	6
Flux 4	15
Flux 5	15
Flux 6	15
Flux 7	13

Table 3.6: PLS components for the *E. coli* network.



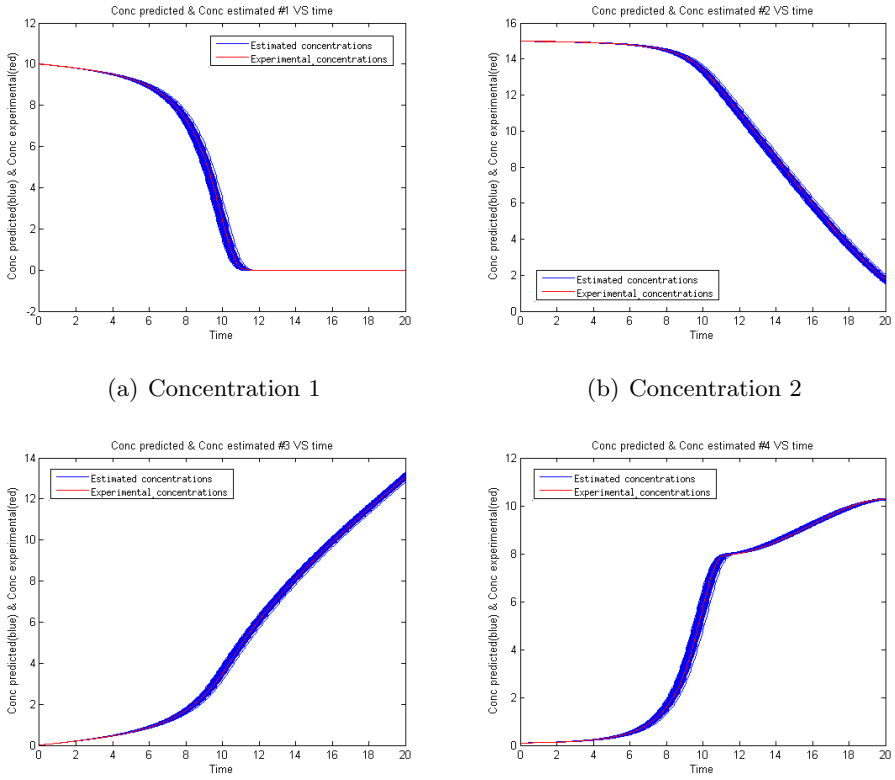
(a) u_1 (b) u_2

(c) u_3 (d) u_4

(e) u_5 (f) u_6

(g) u_7

Figure 3.35: Flux profiles for the *E.coli* network: in red are the experimental data, in blue are the profiles obtained with PLS regression. The regression provides an accurate fit.



(a) Concentration 1

(b) Concentration 2

(c) Concentration 3

(d) Concentration 4

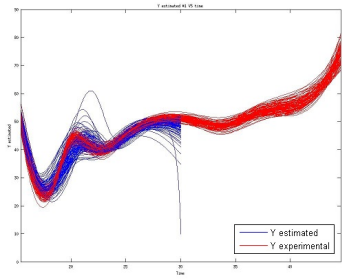
Figure 3.36: Solution of the primary dynamic system for the toy network: in red are the experimental data, in blue the simulated profiles. The integration in completed, and the simulated profiles closely follow the experimental trajectories.

3.4.4 Dynamic system solution for the black-box approach

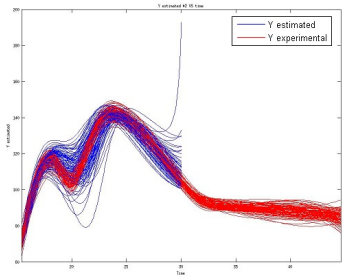
Once the fluxes have been expressed as continuous functions of time, the primary dynamic system can be solved (Equation 3.51).

For the toy network the number of extracellular concentration is 4, and the results are reported in Figure 3.36. The simulation closely follows the experimental profiles. The variance is slightly bigger, as it could be expected for the propagation of error. For the *E. coli* network the extracellular concentrations are 7. The results are reported in Figure 3.37. As it can be seen, the ODE solver cannot conclude the integration. Even when modifying the default tolerance the solution does not improve. The reasons which cause this kind of output must be investigated.

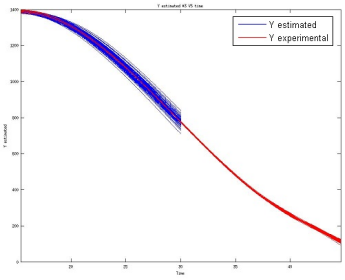
A possible explanation can be found when looking at the behaviour of the profiles near the critical point of the integration. The profiles of the extracellular concentrations 3, 4, 5, 6 and



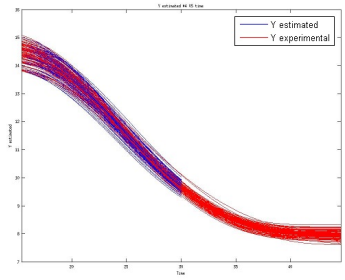
(a) Glucose concentration



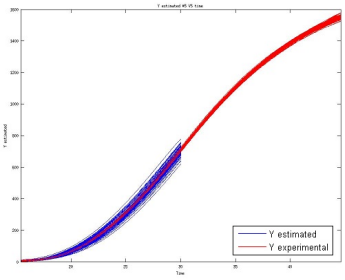
(b) Glycerol concentration



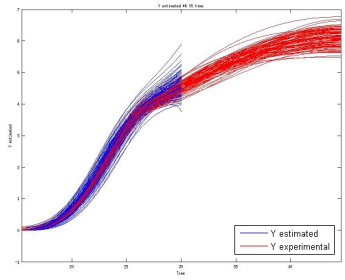
(c) Glucose Feed concentration



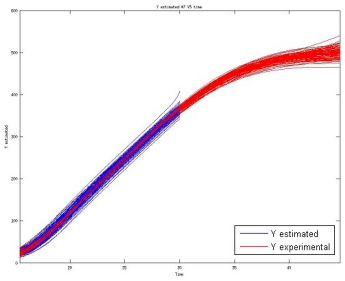
(d) Citrate concentration



(e) PDO concentration



(f) Acetate concentration



(g) Biomass concentration

Figure 3.37: Solution of the primary dynamic system for the *E.coli* network: in red are the experimental data, in blue the simulated profiles. The simulated profiles follow the experimental trajectories quite well in the first part, then they diverge, and the integration stops.

7 approximate quite well the experimental ones. Not the same appears for the concentrations 1 and 2, whose profiles show a big increase of the variance in the last hours of the simulation, and some profiles even diverge from the experimental trajectories. The different behaviour of the simulation can be due to the fact that the experimental profiles of the concentrations 1 and 2 are less regular than the others, presenting more oscillations in the initial part of the simulation. In this area, the simulated profiles seem to follow the experimental trend until the first oscillation peak is reached, then they gradually get far from the original trajectory. This is the typical behaviour of an unstable system. As was previously introduced (3.3.7), a stable system is able to lessen the error with respect to its input variable, while an unstable system exponentially increases it, so that a small change in the input produces a big change in the output. The concept of stability for systems of differential equations is analogous to the *conditioning* of a function. The criterion to verify whether a linear dynamic system $\mathbf{x(t)}' = \mathbf{A} \cdot \mathbf{x(t)}$ is stable is based on the eigenvalues of the A matrix. If all the eigenvalues of this matrix are positive numbers, the system is stable. If the eigenvalues are complex numbers, only the real part is significant for this criterion. If even only one eigenvalue has a positive real part, the system is unstable. For non-linear dynamic systems, an extension of this criterion does exist. The stability of the system in the neighbourhood of a point is assessed by linearizing the system with respect to all the variables around that point, computing the partial derivatives. The eigenvalues of the Jacobian of the dynamic system must be checked, according with the stability criterion.

The stability of the dynamic system (3.51) for the *E. coli* network was tested for different values of the input variable, the extracellular concentrations. Since the system under study is autonomous, i.e., it does not explicitly depend on time, the jacobian of the system is the same in every moment of the simulation. Then, a time near the critical value $t = 30h$, where the integration stops, was chosen. The range between the maximum and minimum values assumed by each extracellular concentration during the simulation was uniformly subdivided in $k = 10$ points, collecting all the points in a matrix of dimension (nxk), where n is the number of extracellular concentrations. The stability of the system was tested for each element of this matrix, numerically computing the Jacobian and the eigenvalues. The resulting values show the existence of at least one positive eigenvalue for most of the points inside the tested range. The system is then unstable.

The second consideration is focused on why the instability of the system is particularly evident in two of the seven extracellular concentrations, while the others seem to follow quite well the experimental profiles. A possible explanation was already given considering the oscillations that characterize the experimental profiles of concentrations 1 and 2. An alternative or additional explanation could be bonded to the structure of the $\mathbf{S}_{ext} \cdot \mathbf{K}$ matrix, which multiplies

the vector **u** of the fluxes in the dynamic system:

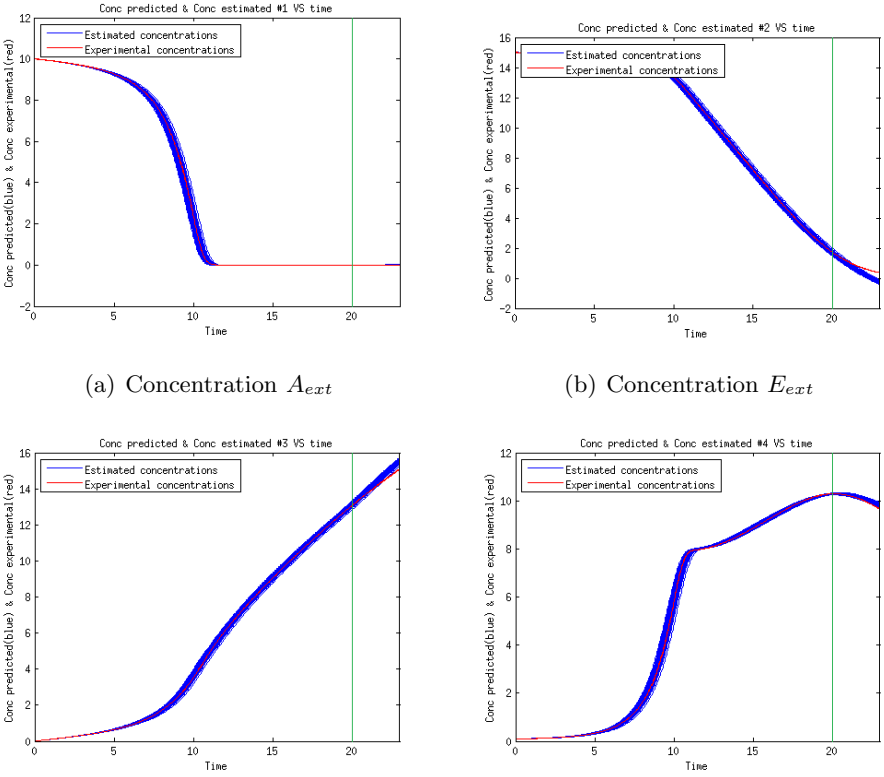
$$\mathbf{S}_{ext} \cdot \mathbf{K} = \begin{pmatrix} 1.000 & 2.500 & 5.000e-1 & -3.333e-1 & -1.167 & 1.000 & -4.312 \\ 0.000 & -3.000 & -2.000 & 0.000 & 2.000 & -2.000 & -4.603 \\ -1.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & -1.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 3.968e+1 \end{pmatrix} \tag{3.55}$$

This matrix has dimension $(m \times n)$, where m is the number of extracellular concentrations and n is the number of free fluxes. Each row corresponds to a concentration, and the columns describe which are the fluxes each concentration depends on. It can be noticed that the concentrations 1 and 2, corresponding to the first two rows of the $\mathbf{S}_{ext} \cdot \mathbf{K}$ matrix, depend on many fluxes, while the other concentrations just on one flux each. This observation means that the error intrinsically introduced by the regression on each flux is summed during the estimation of the concentrations 1 and 2, increasing the effects of the instability of the system. The stability is a property of the system, and not much can be done to improve it.

3.4.5 Testing the predictive capability

In the previous section (3.4.4), the black-box procedure was completed for the toy network. Since the procedure did not properly work for the *E.coli* network, this case study will not be included in this section. Furthermore, while for the toy network additional data can be generated at every moment, since it is a simulated network, not the same can be done for the *E.coli* network, whose data are experimental and were taken from literature (Antoniewicz et al., 2007). The only way to test the validity of the procedure for the *E.coli* network outside the training range would be to exclude part of the data during the training phase, and use these data to test the extrapolation.

Since the simulations for the toy network follow the experimental profiles quite well, the next step is to test the predictive power of the model, i.e., to check whether the model can be generalized to new data. The new data can be generated inside the original time range by just changing the starting points for the simulations, or outside it, by extrapolating the response of the system in time. Extrapolation is always a delicate operation, even if the model had been largely validated. In particular, for polynomial expressions, the higher the degree of the polynomial, the more it will oscillate, and the faster it will diverge outside the training range. This is even more critical particularly in case of asymptotic behaviour.



(a) Concentration A_{ext}

(b) Concentration E_{ext}

(c) Concentration F_{ext}

(d) Concentration BIO

Figure 3.38: Solution of the primary dynamic system for the toy network, extrapolating the model in time: in red are the simulated data, in blue the solutions of the dynamic system.

The simulations from different starting points are critical for the toy network, since the primary dynamic system was proven to be very sensitive to changes in the initial values (3.30). To try applying the regression obtained from one set of data on another one which starts from a different point does not provide useful results.

The extrapolation in time will now be tested, slightly increasing the time interval from $0 \div 20h$ to $0 \div 23h$ (Figure 3.38). The model results to catch the behavior of the system for a very short while, after which it will probably get far from the experimental data. No applications of predictive control are available until now for this type of systems.

Chapter 4

Conclusions

In this chapter some general conclusions are drawn about the approach proposed by this study, in light of the results presented in *Results and discussion* (Chapter 3).
This study proposes a new approach to solve a metabolic network-based primary model of the growth of cellular cultures online. Metabolic network-based modeling was chosen since it is one of the most promising strategies for biological systems. Many metabolic networks have already been fully characterized in the literature of the past, and the derived models showed to provide good accordance with many sets of experimental data. Furthermore, they do not have the necessity to estimate kinetic expressions, which, in a so articulated and nonlinear system as a living organism, would be too many and too complicated to determine. Resulting from some simplifying hypothesis (2.2.1), the model is quite easy, but still it requires to estimate at every time point the distribution of fluxes, which cannot be measured online during the process. At first, the FBA method (Burgard and Maranas, 2003; Gianchandani et al., 2008) was applied to estimate these fluxes as the ones which optimize an objective function of cellular metabolism, theoretically dependent on the condition of the cell, i.e., in which growth phase it is going through, and on the environmental conditions, i.e., just the abundance or lack of substrates and the accumulation of products, in this case. Successively, since the FBA method introduces difficulties that were shown to be practically insurmountable, the second approach was tried, which simply correlates the offline experimental measurements of the fluxes to the online variation of the extracellular metabolite concentrations. Considering that the procedure was tested for both approaches with two different case studies (3.2.1), at the end four formulations of the problem were faced. Between these formulations, only one, the black-box approach solved for the smaller case study, called *toy network*, could be successfully completed. All the others instead were proven to accumulate during the procedure a quantity of error which makes the simulation to diverge from the experimental behaviour. This is probably due in part to the excessive sophistication of the procedure, which extrapolates and

combines many strategies with different original aims, and in part on the formulated model itself, which was proven to be extremely sensitive to error in the inputs. The different approaches will now be considered, and for each one some conclusions both general and relating to a single step are drawn.

The first consideration is about the grey-box approach, whose critical step is the optimization problem. The optimization problem is solved in two different moments, i.e, *(i)* the bilevel problem, which is used off-line to estimate the coefficients of the objective function and *(ii)* the simulation problem, which is effectively solved during the integration. In other studies about flux balance analysis, the fluxes are left free to vary, comparing the results obtained with different objective functions to determine which is the one that better describes the behaviour of the organism in that particular situation. Many objective functions, which are nonlinear combinations of fluxes without experimental coefficients, can be formulated based on biological considerations, and a list of objective functions with their respective biological meaning is provided by Schuetz (Schuetz et al., 2007). This method works very well to increase the understanding of the cellular metabolism, but not for the aim of this study, since the simulated distribution of fluxes is not constrained in any way to closely follow the profiles of the fluxes obtained when regressing the measurements of off-line isotopomer analysis. Consequently, the addition of the experimental coefficients to the objective function and of the outer problem to estimate them was necessary. The outer problem is analogous to the determination of the coefficients of a regression problem. The system is much more bonded to the training data, since the fluxes are forced to follow the experimental profiles. The biological interpretation is then completely left to the optimization coefficients. The formulation of the bilevel problem was inspired by the ObjFind approach presented in literature (Burgard and Maranas, 2003), which uses a linear objective function simply made by the weighted summation of all the fluxes multiplied by the corresponding coefficients. The original method used the obtained set of coefficients to determine which fluxes were more important inside the objective function, i.e., which fluxes the cell seemed to optimize, obtaining information or validating hypotheses about the cellular metabolism. Burgard also says that the simulation problem, i.e., the optimization of the objective function with the set of coefficients estimated by the bilevel problem, cannot be univocally solved, since the same distribution of coefficients correspond to many sets of fluxes. In this study, to make the solution of the simulation problem, which is inserted in the integration, possible, the linear objective function was transformed into a quadratic function, and an experimental constraint was added, which bounds the summation of the simulated fluxes to be equal to the summation of the experimental fluxes. In this way the dynamic system can be correctly solved, but the set of coefficients is far more difficult to be interpreted, renouncing part of the biological information. Finally, the

model loses all its mechanistic characterization, and the reasons which justified the addition of the optimization problem to the procedure fall. The optimization step was completely excluded, and the black-box approach was adopted. Given the importance of considering the physics of the system for the model to better represent reality, it would be important to add new biological information to the model. Furthermore, the solution of the simulation problem was still very delicate, being easily subject to alternate optima and extremely sensitive to error in the input coefficients. In future, possible improvement to this study would be given by finding a biological objective function able to properly solve the simulation problem, or finding an alternative formulation of the optimization problem able to maintain both the experimental data fitting and the biological relevance.

The second consideration is about the primary dynamic system proposed in literature (Van Impe et al., 2012). The system proved to be very sensitive to error, i.e., for small changes in the initial starting points of the simulation, the estimated profiles drastically increase their variance with respect to the experimental trend. For dynamic systems, this behaviour is called *instability*. Actually, the behaviour which was observed was slightly different from instability, since the simulated profiles did not diverge, but after a while seem to reconverge toward the experimental one. The variance of the model appears to vary a lot during the integration, and this can represent a problem for the solution of the online procedure, since simulations from different starting points always diverge, being the error increased by the intermediate steps. The dynamic system inserted inside the online procedure was in fact definitely unstable, and also starting from the same initial point, the profile risks to diverge due to the error accumulated during the integration. The stability being a property of the system, probably there is not an intuitive or immediate way to improve the situation.

At the end, due to the critical points here presented, only the dynamic system of the black-box approach for the toy network was completely integrated, and even extrapolated in time to verify how much it is able to follow the response of the micro-organism. The model is still strongly bonded to the experimental training set of data, but its extrapolation seems to catch the trend of the experimental data, at least in a limited time near the training interval. It should be tested if this prediction can be useful to implement some online control strategy. Hopefully, future improvements in each step of the procedure will make it possible to increase the predictive power of the model, producing results which can be extended to cover a larger range of situations.

Finally, remaining on the case of black-box models, the regression which correlates the extracellular concentrations to the optimization coefficients in the grey box-approach, or the fluxes in the black-box approach, could be improved. Only linear regression was tested, due to practical reasons and justified by a wide range of previous studies which confirmed this

type of models to effectively work. Nevertheless, considering the extreme complication and the high nonlinearity of the systems which try to describe the living organisms, for sure a nonlinear model would be more appropriate. If a functional form would be found which better interprets the mechanisms that take place inside the cell, the number of experimental coefficients could be probably diminished, and the model would be more easily generalized. In future, nonlinear regression should be tested.

The modeling here presented is in general, not theoretically depending on the features of the particular metabolic network, and it can be applied to any fermentation. Although this study recognizes the primary importance of developing a comprehensive and quantitative understanding of biological systems, to allow for better control of the fermentations and improved process performance, it still proposes an innovative method, which can be considered a first attempt of online metabolic network based-modeling. This consideration justifies the choice of a model which maintains a strong experimental characterization, giving priority to its flexibility and adaptability more than to the biological interpretation. Still there are many steps of the proposed procedure which should be adjusted or re-invented to improve and extend its applicability, and only after guaranteeing a proper functioning of the experiment based model more mechanistic information could be included, with their respective complications. A natural evolution of this study will be to bond the flux variation to the state of the culture, using the experimental data to formulate and validate more significant models, not limited to a specific set of data.

Although being ambitious, the promise of a computationally efficient methodology which works online would enable real-time process control and monitoring, improving the efficiency and the productivity of biotechnological processes. The data from off-gas analysis and off-line metabolite measurements would be continuously updated, providing new inputs to the procedure. Since the performance of a fermentation depends on two different scales, the macroscopic scale of the bioreactor and the microscopic of cellular metabolism, information from both these levels is necessary to implement a *proactive* control strategy. The greatest aid to biological modeling would probably derive from enhancements in the experimental characterization of the microscopic scale, which is nowadays still problematic and limited. The expectation of new discoveries able to at least improve the present situation in predictive microbiology is not utopian, and much can still be expected from such a fresh and ongoing research area.

Bibliography

M. R. Antoniewicz, D. F. Kraynie, L. A. Laffend, J. González-Lergier, J. K. Kelleher, and G. Stephanopoulos. Metabolic flux analysis in a nonstationary system: Fed-batch fermentation of a high yielding strain of *E. coli* producing 1,3-propanediol. *Metabolic Engineering 9*, pages 277–292, 2007.

J. Baranyi and T. Roberts. A dynamic approach to predicting bacterial growth in food. *International Journal of Food Microbiology 23*, pages 277–294, 1994.

J. Baranyi and T. Roberts. Predictive microbiology - quantitative microbial ecology. *Culture, 25*, 2004.

L. T. Biegler. *NONLINEAR OPTIMIZATION, Concepts, Algorithms and Applications to Chemical Processes.* MOS-SIAM Series on Optimization. SIAM, New York, 2010.

J. Broughall, P. Anslow, and D. Kilsby. Hazard analysis applied to microbial growth in foods: development of mathematical models describing the effect of water activity. *Journal of Applied Bacteriology 55*, pages 101–110, 1983.

S. Brul and H. Westerhoff. *Systems biology and food science*, pages 250–288. Woodhead, Sawston, Cambridge, UK, 2007.

A. P. Burgard and C. D. Maranas. Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Wiley Periodicals, Inc. Biotechnol Bioeng 82*, pages 670–677, 2003.

G. Buzzi-Ferraris and F. Manenti. *Interpolation and Regression Models for the Chemical Engineer: Solving Numerical Problems.* Wiley-VCH, Boschstrasse 12, D-69469, Weinheim, Germany, 2010.

C. Darwin. *The origin of species by means of natural selection.* Crowell, New York, USA, 1899.

M. Diehl. *Script for Numerical Optimization Course.* KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium, 2009.

P. Doran. *Bioprocess Engineering Principles.* Academic Press, Waltham, Massachusetts, USA, 1995.

J. Edwards, R. Ibarra, and B. Palsson. In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature Biotechnology 19*, pages 125–130, 2001.

A. Feist and B. Palsson. The biomass objective function. *Current Opinion in Microbiology 2010, 13*, pages 344–349, 2010.

A. Geeraerd, V. Valdramidis, F. Devlieghere, H. Bernaert, J. Debevere, and J. Van Impe. Development of a novel approach for secondary modelling in predictive microbiology: incorporation of microbiological knowledge in black box polynomial modelling. *International Journal of Food Microbiology 91*, pages 229–244, 2004.

C. Genigeorgis. Factors affecting the probability of growth of pathogenic microorganisms in foods. *Journal of the American Veterinary Medical Association 179*, pages 1410–1417, 1981.

E. Gianchandani, M. Oberhardt, A. Burgard, C. Maranas, and J. Papin. Predicting biological system objectives de novo from internal state measurements. *BMC Bioinformatics 2008*, 2008.

A. Gibson, N. Bratchell, and T. Roberts. Predicting microbial growth: growth responses of salmonellae in a laboratory medium as affected by ph, sodium chloride and storage temperature. *International Journal of Food Microbiology 6*, pages 155–178, 1988.

J. Lederberg. Infectious disease as an evolutionary paradigm. *Emerging Infective Disease 3*, pages 417–423, 1997.

R. W. Leighty and M. R. Antoniewicz. Dynamic metabolic flux analysis (dmfa): A framework for determining fluxes at metabolic non- steady state. *Metabolic Engineering 13*, pages 745–755, 2011.

L. Leistner. *Hurdle effect and energy saving*, pages 553–557. W.K., Applied Science Publ., London, 1978.

L. Leistner. Food preservation by combined methods. *Food Res. Int. 25*, pages 151–158, 1992.

T. McMeekin, J. Brown, K. Krist, D. Miles, K. Nuemeyer, D. Nichols, J. Olley, K. Presser, D. Ratkowsky, T. Ross, M. Salter, and S. Soontranon. Quantitative microbiology: A basis for food safety. *Emerging Infect. Dis. 3*, page 541 549, 1997.

T. McMeekin, J. Olley, D. Ratkowsky, and T. Ross. Predictive microbiology: towards the interface and beyond. *International Journal of Food Microbiology, 73*, pages 395–407, 2002.

T. McMeekin, J. Bowman, O. McQuestin, L. Mellefont, T. Ross, and M. Tamplin. The future of predictive microbiology: Strategic research, innovative applications and great expectations. *International Journal of Food Microbiology 128*, pages 2–9, 2008.

T. A. McMeekin and T. Ross. Predictive microbiology: providing a knowledge based framework for change management. *International Journal of Food Microbiology, 78*, pages 133–153, 2002.

J. Monod. The growth of bacterial cultures. *Annual Review of Microbiology 3*, pages 371–394, 1949.

G. Najafpour. *Biochemical Engineering and Biotechnology.* Elsevier, Philadelphia, USA, 2007.

J. Niklas, E. Schräder, V. Sandig, T. Noll, and E. Heinzle. Quantitative characterization of metabolism and metabolic shifts during growth of the new human cell line age1.hn using time resolved metabolic flux analysis. *Bioprocess Biosyst Eng 34*, (1):533–545, 2011.

J. Olley and D. Ratkowsky. Temperature function integration and its importance in storage and distribution of flesh foods above the freezing point. *Food Technol. Aust. 25*, pages 66–73, 1973a.

J. Olley and D. Ratkowsky. The role of temperature integration in monitoring fish spoilage. *Food Technol. N. Z. 8*, pages 147–153, 1973b.

D. Ratkowsky. Principles of nonlinear regression modeling. *Journal of Industrial Microbiology, 12*, pages 195–199, 1993.

D. Ratkowsky, R. Lowry, T. Mcmeekin, A. Stokes, and R. Chandler. Model for bacterial culture growth rate throughout the entire biokinetic temperature range. *Journal of Bacteriology, vol. 154, No. 3*, pages 1222–1226, 1983.

T. Roberts, A. Gibson, and A. Robinson. Prediction of toxin production by clostridium botulinum in pasteurised pork slurry. *J. Food Technol. 16*, pages 337–355, 1981.

T. Ross, J. Baranyi, and T. McMeekin. *Predictive microbiology*, pages 1699–1710. Elsevier Inc., Philadelphia, USA, 1999.

L. Rosso, J. Lobry, S. Bajard, and J. Flandrois. Convenient model to describe the combined effects of temperature and ph on microbial growth. *Applied and Environmental Microbiology, Vol. 61, No.2*, pages 610–616, 1994.

R. Schuetz, L. Kuepfer, and U. Sauer. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Molecular Systems Biology 3*, pages 1–15, 2007.

W. Scott. The growth of microorganisms on ox muscle: ii. the influence of temperature. *Journal of the Council for Scientific and Industrial Research Australia 10*, pages 338–350, 1937.

J. Van Impe, D. Vercammen, and E. van Derlinden. Toward a next generation of predictive models: A systems biology primer. *Food Control 29*, pages 336–342, 2012.

M. Zwietering, I. Jongenburger, F. Rombouts, and K. van't Riet. Modelling of the bacterial growth curve. *Appl. Environ. Microbiol. 56*, pages 1875–1881, 1990.