Interval and Possibilistic Methods for Constraint-Based Metabolic Models

PhD Dissertation by

Francisco Llaneras Estrada

Supervisor

Jesús Picó i Marco

Valencia, March 15, 2010
Instituto Universitario de Automática e Informática Industrial
Universidad Politécnica de Valencia

UNIVERSIDAD
POLITECNICA
DE VALENCIA

This work is partially supported by the Spanish Government (Program CICYT-FEDER DPI2005-01180 and DPI2008-06880-C03-01). The author is also recipient of a fellowship from the Spanish Ministry of Education (FPU AP2005-1442).

In the following site the reader will find updated information regarding this thesis. This includes corrections and clarifications, connections with future works, publications, software and tools, etc.

http://science.ensilicio.com/Thesis/Thesis.html

Agradecimientos

La tesis doctoral es un viaje emocionante, exigente y no exento de riesgos. Por suerte, resulta mucho mejor, más sencillo y más divertido, cuando uno tiene una lista de agradecimientos tan extensa como la mía.

Estoy en deuda con mi director, Jesús Picó, por darme la oportunidad de trabajar en lo que me gusta, por descubrirme un tema fascinante y por dirigirme sabiamente en el desarrollo de esta tesis. También con mis supervisores durante mis estancias en el extranjero, Georges Bastin y Vassily Hatzimanikatis, porque me permitieron aprender de su trabajo, conocer sus instituciones y explorar dos países.

Le agradezco a Antonio Sala que se acercase un buen día con una propuesta para colaborar (haciendo así fructíferas las reuniones en la biblioteca); y a Marta Tortajada que fuese mi coautora en varias ocasiones y que sea una de las personas cuyas ideas más me han ayudado. He aprendido mil cosas de otras personas —profesores, compañeros, colegas, revisores, etc.— y a todos les estoy agradecido.

A los compañeros que encontré camino al doctorado les agradezco, sobretodo, que hicieran mi trabajo divertido. Agradezco cada café y cada conversación. Mención especial merecen mis principales compañeros de viaje, Pepe y Sergio, por su camaradería en el trabajo y en la vida.

Es probable que sin el apoyo de otras personas no hubiese terminado esta tesis... y es seguro que de haberlo hecho sería irrelevante. A mi familia —Juan, María, Juan— les agradezco que sean mi refugio, porque la existencia de ese «lugar» me permite el lujo de ser temerario. A mis amigos en la vecindad les agradezco su interés por las victorias de cada día; y tanto a ellos como al resto, les doy las gracias por creer incondicionalmente que lo que hacía, fuese lo que fuese, era digno de admirar. Mi deuda es especial con Lucia porque ella fue la recompensa en los días de trabajo y sacrificio.

Por último, y por encima de todo, quiero agradecer a mis padres todo lo que hicieron y el modo preciso en que lo hicieron. Sé que eso excede el ámbito de esta tesis, pero este párrafo es un buen lugar para mostrarles mi gratitud.

Gracias.

Abstract

This thesis is devoted to the study and application of constraint-based metabolic models. The objective was to find simple ways to handle the difficulties that arise in practice due to uncertainty (knowledge is incomplete, there is a lack of measurable variables, and those available are imprecise). With this purpose, tools have been developed to model, analyse, estimate and predict the metabolic behaviour of cells.

The document is structured in three parts. First, related literature is revised and summarised. This results in a unified perspective of several methodologies that use constraint-based representations of the cell metabolism. Three outstanding methods are discussed in detail, network-based pathways analysis (NPA), metabolic flux analysis (MFA), and flux balance analysis (FBA). Four types of metabolic pathways are also compared to clarify the subtle differences among them.

The second part is devoted to interval methods for constraint-based models. The first contribution is an interval approach to traditional MFA, particularly useful to estimate the metabolic fluxes under data scarcity (FS-MFA). These estimates provide insight on the internal state of cells, which determines the behaviour they exhibit at given conditions. The second contribution is a procedure for monitoring the metabolic fluxes during a cultivation process that uses FS-MFA to handle uncertainty.

The third part of the document addresses the use of possibility theory. The main contribution is a possibilistic framework to (a) evaluate model and measurements consistency, and (b) perform flux estimations (Poss-MFA). It combines flexibility on the assumptions and computational efficiency. Poss-MFA is also applied to monitoring fluxes and metabolite concentrations during a cultivation, information of great use for fault-detection and control of industrial processes. Afterwards, the FBA problem is addressed. A possibilistic approach is derived to get predictions under the assumption that cells have evolved to be optimal (Poss-FBA). It captures alternate optima and grades sub-optimality, thus relaxing the original assumption. The last contribution is a procedure to validate constraint-based models when data are scarce. This procedure mitigates validation problems with small metabolic networks.

This thesis highlights the importance of accounting for uncertainty when modelling living cells and promotes a constraint-based perspective: if we cannot exactly model how cells operate, use the knowledge available to distinguish what is possible from what is not. Following this idea, methods are proposed that start by representing the available knowledge and its uncertainty, and then exploit this representation to generate reliable new information.

Resumen

Esta tesis se ha centrado en el estudio y aplicación de modelos del metabolismo celular basados en restricciones. El objetivo era encontrar formas sencillas de afrontar los problemas que surgen en la práctica como consecuencia de la incertidumbre (los organismos modelados no son bien conocidos, faltan variables medibles y las disponibles son imprecisas). Con este propósito se han desarrollado herramientas para modelar, analizar, estimar y predecir el comportamiento metabólico de células vivas.

El documento se ha estructurado en tres partes. Primero, se revisó y resumió la literatura relacionada con el tema. Como resultado se ofrece una perspectiva unificada de metodologías que emplean modelos basados en restricciones para representar el metabolismo celular. Tres metodologías se discuten detalladamente: *network-based pathways analysis* (NPA), *metabolic flux analysis* (MFA), y *flux balance analysis* (FBA). También se comparan cuatro definiciones de rutas metabólicas para aclarar sus diferencias.

La segunda parte se dedicó al estudio de métodos intervalares para modelos basados en restricciones. La primera contribución es una aproximación intervalar al MFA tradicional particularmente útil al estimar flujos metabólicos en escenarios de escasez de datos (FS-MFA). Esta estimación informa sobre el estado interno de las células, el cual determina el comportamiento que éstas exhiben. La segunda contribución es un procedimiento para monitorizar los flujos metabólicos durante un proceso de cultivo en escenarios de escasez de datos.

La tercera parte del documento aborda el uso de teoría de posibilidad. La principal contribución es un marco posibilístico para (a) evaluar la consistencia de un conjunto de medidas experimentales, y (b) estimar flujos metabólicos (Poss-MFA). Esta aproximación combina flexibilidad en las hipótesis y eficiencia computacional. Poss-MFA se aplica después en la monitorización de flujos y concentración de metabolitos externos, información de utilidad para la detección de fallos y el control de procesos industriales. A continuación, se propone un enfoque posibilístico para FBA que permite obtener predicciones asumiendo que las células han evolucionado para mostrar un comportamiento óptimo (Poss-FBA). El método propuesto es capaz de capturar múltiples óptimos y gradar la optimalidad de distintas predicciones, relajando así la hipótesis original. La última contribución es un procedimiento para validar modelos cuando los datos disponibles son escasos. Este procedimiento mitiga los problemas de validación con redes metabólicas de pequeño tamaño.

En resumen, esta tesis subraya la importancia de considerar incertidumbre al modelar células vivas y promueve un enfoque basado en restricciones. Siguiendo esta idea, se han propuesto métodos que comienzan representando el conocimiento disponible y su incertidumbre para luego explotar dicha representación y generar nueva información de forma fiable.

Resum

Esta tesi s'ha centrat en l'estudi i aplicació de models del metabolisme cel·lular basats en restriccions. L'objectiu era trobar formes senzilles d'afrontar els problemes que sorgixen en la pràctica com a conseqüència de la incertesa (els organismes modelats no són ben coneguts, falten variables mesurables i les disponibles són imprecisas). Amb este propòsit s'han desenrotllat ferramentes per a modelar, analitzar, estimar i predir el comportament metabòlic de cèl·lules vives.

El document s'ha estructurat en tres parts. Primer es va revisar i resumir la literatura relacionada amb el tema. Com resultat s'oferix una perspectiva unificada de metodologies que fan ús de models basats en restriccions per a representar el metabolisme cel·lular. Tres metodologies es discutixen en detall: *network-based pathways analysis* (NPA), *metabolic flux analysis* (MFA), i *flux balance analysis* (FBA). També es comparen quatre definicions de rutes metabòliques per a aclarir les seues diferències.

La segona part es va dedicar a l'estudi de mètodes intervalares per a models basats en restriccions. La primera contribució és una aproximació intervalar al MFA tradicional particularment útil per estimar els fluxos metabòlics en escenaris d'escassetat de dades (FS-MFA). Esta estimació informa sobre l'estat intern de la cèl·lules, el qual determina el comportament que estes exhibixen. La segona contribució és un procediment per a monitoritzar els fluxos metabòlics durant un procés de cultiu en escenaris d'escassetat de dades.

La tercera part del document aborda l'ús de teoria de possibilitat. La principal contribució és un marc posibilístic per a (a) avaluar la consistència d'un conjunt de mesures experimentals, i (b) estimar els fluxos metabòlics (Poss-MFA). Esta aproximació combina flexibilitat en les hipòtesis i eficiència computacional. Poss-MFA s'aplica després en la monitorització dels fluxos i les concentracións dels metabòlits externs, informació d'utilitat per a la detecció de problemes i el control de processos industrials. A continuació, es proposa un enfocament posibilístic per a FBA que permet obtindre prediccions assumint que les cèl·lules han evolucionat per a mostrar un comportament òptim (Poss-FBA). El mètode proposat és capaç de capturar múltiples òptims i avaluar l'optimitat de distintes prediccions, relaxant així la hipòtesi original. L'última contribució és un procediment per a validar models quan les dades disponibles són escassos. Este procediment mitiga els problemes de validació amb xarxes metabòliques de dimensió reduïda.

En resum, esta tesi subratlla la importància de considerar la incertesa al modelar cèl·lules vives, i promou un enfocament basat en restriccions. Seguint esta idea, s'han proposat mètodes que comencen representant el coneixement disponible i la seua incertesa, per a després explotar aquesta representació i generar nova informació de forma fiable.

Table of contents

Part I: state of the art

Part II: Interval methods

Part III: Possibilistic methods

[T]he point of making models is to be able to bring a measure of order to our experience and observations, as well as to make specific predictions about certain aspects of the world we experience

(Casti, 1992)

Justification, Objectives and Contributions

Living organisms are complex. Even the simplest living cell is composed of an incredibly large number of multifunctional elements, which interact selectively and nonlinearly to produce the observed behaviour. This confers a crucial role to *mathematical models* in biology, they can mimic these interactions to help us understand how cells operate and predict their behaviour.

Models are thus a tool to improve our knowledge. They organise disparate information into a coherent whole; they enable studying properties that emerge from the whole cell and are not properties of individual parts. The modelling process itself results in hypothesis to be experimentally tested, thereby iteratively producing refined models and insight about cellular mechanisms. Mathematical models have also several applications in industries that involve biological processes, such as biomedicine, food industry or biotechnology. Models are used, for instance, to perform simulations, optimise variables, design experiments, and implement on-line quality control. Models are also a promising tool for metabolic engineering, allowing for directed manipulation of the gene content of an organism to obtain the desired behaviour.

Although other processes operate within cells, such us regulation and signalling, this thesis is focused on models of the *cellular metabolism*. The metabolism can be viewed as a chemical "factory" that converts available raw materials into energy as well as building blocks needed to produce biological structures, maintain cells alive, and carry out various cellular functions. This process can be represented with a metabolic network that encodes a set of biochemical reactions taking place within the cell. The nodes represent the involved metabolites and the edges represent the reaction rates or *metabolic fluxes*. Internal fluxes correspond to reactions occurring within cells and exchange fluxes to exchanges between the cells and their environment (uptake of substrates and formation of products). The set of flux values defines the metabolic state of cells or its phenotype, i.e., the behaviour they exhibit at a given time.

However, these networks of metabolic reactions are difficult to model. Considering all the mechanisms operating in metabolism will lead to detailed, quantitative predictions on cellular dynamics. Yet, lack of knowledge on the intracellular reactions and its parameters complicates this approach. As an alternative, classical *Stoichiometric Models* disregard the dynamics of the (fast) intracellular reaction and assumes that most internal metabolites rapidly reach their steady-state. This way, the state of the cells is represented without any information on the kinetics of the reactions.

Constraint-based Models appear as an extension of stoichiometric models. Along with the stoichiometric mass balances at steady-state, cells are subject to other constraints that limit their behaviour, such us thermodynamics or enzyme capacities. Imposing these constraints that operate at given circumstances it is possible to determine which functional states can and cannot be achieved by a cell. The imposition of constraints leads to a space of cellular phenotypes that, to the best of our knowledge, are feasible. Constraint-based models are thus conservative, but they do not require a particular type or amount of data to be useful. They are also scalable; new and better knowledge can be easily incorporated, just adding constraints, to improve the models.

Several methodologies employing constraint-based models can be found in literature. There are methods to *analyse* properties of the modelled organisms (e.g., identify optimal pathways), to *simulate* genetic modifications (e.g., gene deletions), and to *estimate* or *predict* the state exhibit by cells at given conditions. This thesis is devoted to study and improve these methodologies.

Objectives

The principal objectives pursued in this work are the following:

a) Survey methods that use constraint-based models to analyse, estimate or predict the metabolic behaviour of cells.

Several methods employ mathematical representations of cells that can be considered a constraint-based model, even if this is not always explicit. For this reason, it is worthy to do some efforts to present these methodologies with a unified perspective. This may allow to develop general solutions for related problems.

b) Identify the limitations of the studied methodologies.

The second objective is to identify the limitations that may arise when applying the standard methodologies to analyse, estimate or predict the metabolic behaviour of cells. In particular, the interest herein is on those difficulties that arise in scenarios of data scarcity, common in industry and research laboratories. In practice, uncertainty is often widely present: (i) there are no detailed models of the organism of interest, (ii) first-principles knowledge is incomplete, (iii) there is a lack of measurable variables, or (iv) the available measurements are imprecise.

c) Propose new methods to overcome the limitations found.

Once limitations have been identified, the next objective is to propose solutions for them, having the practical applicability in mind. These solutions should be kept simple and be justified theoretically.

d) Apply these methods in different real case studies.

All the contributions proposed in the preceding step should be tested experimentally when presented. Real data from different organisms will be used to show that the proposed methods are able to analyse, estimate or predict the metabolic behaviour of cells. Advantages over standard approaches should be illustrated.

Thesis outline

The first chapter reviews different kinds of mathematical models built to represent living cells in two fields: Bioprocess Engineering and Systems Biology. Chapter II is devoted to constraint-based models; there, the methodologies that are the context for the contributions of this thesis are presented with a unified perspective. Three methodologies are discussed in detail: Network-based Pathways Analysis (NPA), Metabolic Flux Analysis (MFA), and Flux Balance Analysis (FBA). Chapter III compares different proposals of Network-based pathways to clarify the intricate relationship among them.

The second part of the document is devoted to develop interval methods for constraint-based models. First, we address the MFA problem, the exercise of estimating the metabolic fluxes shown by cells by combination of a model and experimental measurements. Traditional MFA requires a large number of accurate measurements to be of use, but these are often not available. In chapter IV we propose an interval variant of MFA well suited for scenarios of data scarcity, the so-called flux-spectrum (FS-MFA). Representing fluxes with intervals allows accounting for uncertainty both in measurements and estimates; so the estimates are more reliable even if data is scarce (they are only as precise as allowed by the uncertainty). This enables using MFA in two common situations: when there is a lack of measurable fluxes, and when the measurements are highly imprecise. FS-MFA uses a linear programming formulation, so it is also simple and computational efficient. Using the same approach, chapter V discusses how to translate a given flux state into a pattern of pathway activities. Chapter VI describes a procedure for monitoring the metabolic fluxes during a cultivation process. The procedure employs FS-MFA to handle uncertainty and be of use in scenarios of data scarcity. It can be used to analyse collected data or to monitor a running process, mitigating the common absence of reliable on-line sensors in industry. Experimental data from cultivations of CHO cells and *C. glutamicum* illustrate the benefits of these proposals against traditional MFA approaches.

The third part of the document is devoted to the use of possibility theory in the context of constraint-based models. In chapter VII we introduce a possibilistic framework to (a) evaluate model and measurements consistency and (b) perform MFA flux estimations. The approach, called Poss-MFA, follows the original philosophy of constraint-based models, in the sense that it does not attempt necessarily to predict the actual fluxes with precision, but rather to distinguish "most possible" from "impossible" flux states. Poss-MFA gives possibility distributions as estimates that are more informative than point-wise ones when multiple values are reasonably possible. Besides, Poss-MFA considers measurements uncertainty and model imprecision in a flexible way (e.g., non-symmetric error), and is reliable in scenarios of data scarcity. The combination of flexibility of the assumptions and computational efficiency is a distinctive advantage of Poss-MFA over other approaches which either may rely on stronger assumptions (chi-squared distributions of errors, absence of irreversibility), or be only data-based (so they do not incorporate a model), or provide only point-wise estimates, or be computationally intensive (multi-variate integration in a general Bayesian estimation problem). In chapter VIII the possibilistic framework is adapted to account for extracellular dynamics. Poss-MFA is extended for monitoring time-varying fluxes and metabolite concentrations during a cultivation process. Then we approach dynamic FBA, a methodology to get predictions during a cultivation based on the assumption that cells have evolved to be optimal. A possibilistic variant, called Poss-FBA, allows to account for alternate optima and sub-optimality. These extensions are illustrated with real data from CHO cells and *Escherichia coli*. Finally, chapter IX presents a systematic, yet simple, procedure that employs Poss-MFA to validate constraint-based models when experimental data is scarce. The procedure has been applied with a model of *P. pastoris*, a yeast used in industry for the expression of recombinant proteins.

The last part of the thesis draws some general conclusions.

Contributions

The main contributions of this work are the following:

- **A unified perspective of methodologies that employ constraint-based models of the cell metabolism.** These methodologies have different purposes, use different mathematical tools, and rely on different assumptions; but they all exploit the properties of similar representations. Embracing these methodologies within the same framework makes it easy to extrapolate solutions from ones to others and develop common improvements.

- **An interval method to estimate the metabolic fluxes under data scarcity (FS-MFA).** The method is a simple and powerful improvement of traditional MFA. It is particularly useful to handle uncertainty: interval estimates are only as precise as allowed by the available knowledge. The benefits of FS-MFA have been illustrated with two real case studies.

- **A procedure for monitoring the metabolic fluxes during a cultivation process.** The procedure employs FS-MFA to handle uncertainty and lack of measurements. It has been tested with data from a cultivation of CHO cells.

- **A comparison of four definitions of network-based metabolic pathways.** This clarifies the relationship among four types of pathways, which subtle differences had been a source of misunderstanding in the literature.

- **A possibilistic framework to (a) evaluate measurements consistency and (b) perform flux estimations (Poss-MFA).** The combination of flexibility of the assumptions and computational efficiency is a distinctive advantage of Poss-MFA over other approaches. These advantages have been illustrated with several examples and a real case study.

- **A method based on Poss-MFA for monitoring the metabolic fluxes and the metabolite concentrations during a cultivation process.** The method can be also useful to fault detection in industrial processes. This method has been tested with data from a cultivation of CHO cells.

- **A possibilistic method to get dynamic FBA predictions of fluxes and metabolite concentrations (Poss-FBA).** The use of possibility theory allows to account for alternate optima and sub-optimality. The method has been illustrated with a simple model of *E. coli* and real data.

- **A simple procedure to validate constraint-based models in scenarios where experimental data is scarce.** The procedure mitigates the frequent lack of validation of small and medium metabolic networks (models).

Summarising, this thesis has been devoted to constraint-based models and the methodologies using them. We were interested in mitigating the difficulties that arise in practice due to uncertainty (model incompleteness, lack of measurable variables, and measurement errors). With this purpose in mind, we have developed interval and possibilistic methods that employ constraint-based models to analyse, estimate or predict the metabolic behaviour of cells. All these methods are able to represent our knowledge accounting for uncertainty, and then exploit this knowledge to generate reliable new information.

Publications

The results of this thesis have been published in:

Refereed Journal Papers

1. Llaneras F, Picó J (2007). An interval approach for dealing with flux distributions and elementary modes activity patterns. *Journal of Theoretical Biology*, 246(2).

2. Llaneras F, Picó J (2007). A procedure for the estimation over time of metabolic fluxes in scenarios where measurements are uncertain and/or insufficient. *BMC Bioinformatics*, 8:42.

3. Llaneras F, Picó J (2008). Stoichiometric modelling of the cell metabolism. *Journal of Bioscience and Bioengineering*, 1:1-12.

4. Llaneras F, Sala A, Picó J (2009). A possibilistic framework for metabolic flux analysis. *BMC Systems Biology*, 3:73.

5. Llaneras F, Picó J (2010). Which metabolic pathways generate and characterise the flux space? A comparison among elementary modes, extreme pathways and minimal generators. *J. Biomedicine and biotechnology*, vol. 2010.

6. Llaneras F, Tortajada M, Picó J (2010). Validation of a constraint-based model of *Pichia pastoris* growth under data scarcity. *BMC Systems Biology*, 4:115.

There is also a paper in preparation with the contents of chapter VIII.

Conference Presentations and Posters

7. Llaneras F, Picó J (2006). The linkage between flux distributions and elementary modes activity patterns: An interval Approach. *International Symposium on Systems Biology.*

8. Llaneras F, Bastin G, Picó J (2007). On metabolic flux analysis when measurements are insufficient and/or uncertain. *IAP Dysco workshop*.

9. Llaneras F, Tortajada M, Picó J (2007). Structural analysis of metabolic pathways applied to heterologous protein production in *P. pastoris*. *European Congress on Biotechnology*, *Journal of Biotechnology*, 131(2):S209.

10. Llaneras F, Sala A, Picó J (2008). A possibilistic framework for metabolic flux analysis. *Reunión de la red Española de Biología de Sistemas*.

11. Tortajada M, Llaneras F, Picó J (2008). Constraint-based modelling applied to heterologous protein production with *P. pastoris*. *Reunión de la red Española de Biología de Sistemas*.

12. Llaneras F, Sala A, Picó J (2009). Applications of possibilistic reasoning to intelligent system monitoring: a case study. *IEEE Multi-conference on Systems and Control.*

13. Llaneras F, Sala A, Picó J (2010). Dynamic flux balance analysis: a possibilistic approach. *Systems Biology of Microorganisms*.

14. Llaneras F, Sala A, Picó J (2010). Possibilistic estimation of metabolic fluxes during a batch process accounting for extracellular dynamics. *IFAC International Conference Computer applications in biotechnology*.

15. Tortajada M, Llaneras F, Picó J (2010). Possibilistic validation of a constraint-based model for *P. pastoris* under data scarcity. *IFAC International Conference Computer Applications in Biotechnology*.

Part I: state of the art

I

Mathematical models of cells

In this chapter we review the kind of mathematical models built to represent living cells in two fields, Bioprocess Engineering and Systems Biology. Both perspectives are addressed, and their goals and characteristics discussed. Then we show a non-exhaustive list of the most outstanding modelling methodologies in both domains and give criteria to classify them.

The last part of the chapter is devoted to a large family of models, called kinetic models. These are addressed here as opposite to the constraint-based models that will receive attention in chapter II.

1.1 Introduction

A model is a simplified or idealised representation of reality capable of representing an actual phenomenon; if it uses mathematical language it is called a mathematical model. Models are simplifications, because refer only to certain, user-defined aspects of reality. Bailey (1998) emphasises this relationship between models and its intended application by quoting Casti (1992):

> *«Basically, the point of making models is to be able to bring a measure of order to our experience and observations, as well as to make specific predictions about certain aspects of the world we experience»*

A model has to be constructed with a specific purpose, which determines what factors are relevant and what factors can be de-emphasised. Thereby, we restrict the model scope to represent only certain aspects of reality—those we are interested in—under certain conditions, and with a certain degree of detail. There are three reasons to proceed in this way: (1) to limit the need of experimental knowledge and quantitative data, (2) to reduce the model complexity, and (3) keep it amenable to formal analysis.

At this respect, cells and biological systems are somehow paradoxical. Although it is obvious that even the simplest living cell has a very complex molecular composition, the number of distinct behaviours that they display is much fewer. A large number of sets of multifunctional elements interact selectively and non-linearly to produce coherent rather than complex behaviours (Kitano, 2002).

Bellgardt and Schügerl give two possible reasons to this phenomenon, at least in the context of the cell metabolism (Schügerl, 2000):

> *«One reason is that the functional blocks of metabolism operate together—coordinated by a network of metabolic regulation and of exchange of mass, charge and energy—to ensure the survival and reproduction of the organism. Another reason is the tremendous number of cells in the population in the bioreactor that hides individual variations in their growth and leads to a smoothed average behavior»*

This important principle of simplicity from complexity differentiates the biological processes from others complex systems (Kitano, 2002; Palsson, 2000).

Moreover, this fact is connected with the two modelling strategies that one can follow to build models of cells. If we want to understand how the cell works, we have to deal with all this complexity: a network of multifunctional elements highly interconnected. However, if we are only interested in the global behaviour of cell populations, we can disregard most of its complexity and build simple representations capable of representing this behaviour. The first approach is the one typically followed in Systems Biology, whereas Bioprocess Engineering follows the second one. Anyway, this coexistence just strengthens our initial words about models: models have purposes, which determine what factors are relevant and what factors can be de-emphasised, and thus different applications require different models.

Along this chapter the wide field of mathematical modelling of cells and cell populations will be briefly reviewed from these two perspectives, Bioprocess Engineering and Systems Biology. Different modelling methodologies will be classified, and kinetic models will be discussed in more deep. To complete this review, the next chapter will be devoted to constraint-based models.

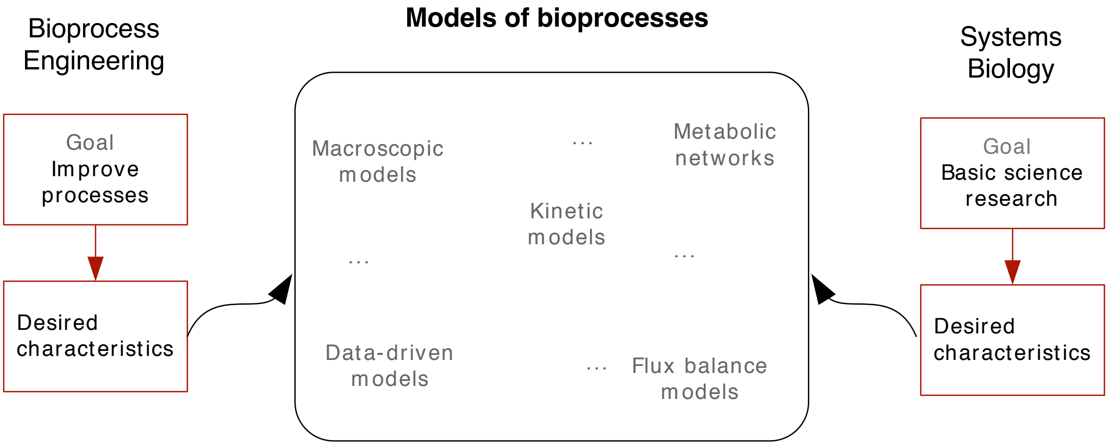


Figure 1.1. Two perspectives for mathematical modelling of bioprocesses.

1.2 Models for Bioprocess Engineering

Bioprocess Engineering concerns the improvement of industrial processes involving living organisms, usually cell populations. These processes are typically animal cell and microorganism cultures, and are employed for the production of enzymes, proteins, value-added chemicals, etc. In recent times, there has been a great emphasis on the use of biotechnological approaches, i.e., in the use of genetically modified microorganisms. Some applications, such as the production of pharmaceutical products or the production of chemicals avoiding the use of fossil fuels, are becoming increasingly important.

All these biological (biotechnological) processes are typically carried on vessels, or bioreactors, to keep cells under controlled conditions. Manipulating these conditions one can force cells to display the desired behaviour. Cells are typically maintained at appropriate environmental conditions (e.g., temperature, gas mixture or pH) and grown adding the required nutrients.

Notice that mathematical modelling concerns not only biological, but also physical aspects, since physical factors that affect the environment of the bioreactor may be considered (e.g., air distribution efficiency, oxygen mass transfer rates, degree of mixing). These factors are affected by the bioreactor design (e.g., geometry, mixing equipment) and by physical properties (e.g., liquid viscosity, interfacial tension).

Applications of models in Bioprocess Engineering

In order to achieve its main goal—improve the process performance—mathematical models are used with different purposes in Bioprocess Engineering:

- *Predict by simulation the process evolution at different conditions.* This is probably the most important purpose, since it underlies the others.

- *Develop model-based monitoring systems.* Mathematical models can be used in conjunction with on-line measurements to estimate process variables that cannot be directly measured. This topic is covered in (Bastin, 1990; Komives, 2003), and a recent example is given in (Veloso, 2009).

- *Fault detection or on-line quality control.* A monitoring system can be improved to detect deviations from the expected behaviour; these deviations can be diagnosed and remedies attempted. For example, multivariate statistical procedures based on PCA and PLS are often applied to monitor the progress of batch processes and detect batch-to-batch variations (Nomikos, 1995; Wold, 1987; Wold, 1998).

- *Improve the process through experiment design.* Many processes in bioreactors are a sequence of phases which differ in the environmental conditions and the feed inflow of substrates. Simulations can be used to choose a preliminary list of promising profiles, which can then be tested experimentally. See, for example, the model-based design of cultivation processes proposed in (Galvanauskas, 1998).

- *Process optimisation.* The aim of process optimisation is to find values for the adjustment of the manipulable parameters (e.g., environmental conditions, feeding rates) in such way that the benefit and cost ratio—defined by means of a quantitative function—reaches a maximum. Model-based optimisation provides an alternative to the trial-and-error methods that prevail in industry, and leads to better performance within shorter development time intervals. Optimisation strategies for bioprocesses can be classified in categories: one-time optimisation (Banga, 2003; 2008b), run-to-run optimisation (Camacho, 2007) and on-line optimisation (Visser, 2000).

- *On-line process control.* The problem of how to achieve a desired performance, in a reproducible manner, and reject the actual process disturbances. This implies: online monitoring the process and introduce an automatic feedback control (Bastin, 1990). Traditionally, a feeding profile was defined a priori for the substrate inflows, but a control law can be defined to automatically manipulate the inflow to avoid the effect of disturbances and guide the process evolution as desired, e.g., maintaining a stable biomass growth rate (Picó-Marco, 2004; 2006; Lee, 1999). A review about control of bioreactors is given in (Rani, 1999).

Main characteristics of models in Bioprocess Engineering

The models used in Bioprocess Engineering share some characteristics, which can be summarised as follows:

- Models consider non-biological factors.
- Models are quantitative and dynamic.
- Models are kept simple (and complexity arise from bottom-up).

The last two requirements are why Bioprocess Engineering has historically worked with unstructured models. On the one hand, the available experimental data was insufficient to develop (validate) dynamic models considering the intracellular behaviour (Palsson, 2000). In fact, the kinetic parameters of many intracellular reactions are still unknown, although the information is growing (Buchhold, 2002; Mashego, 2007). On the other hand, a mechanistic description of intracellular processes may result in a complex model, incompatible with most theoretical frameworks used in bioprocess engineering. Moreover, since simple models where successful, it was reasonably to incorporate complexity using a bottom-up approach.

For a long time simple, unstructured models have proved to be efficient for solving many problems. For example, in the major part of a batch experiment, all cell components eventually grow at the same rate (the so-called balanced growth condition), and the use of an unstructured model is adequate (Nielsen, 1992). Another example is given by feed-back control strategies, which are often based on simple, unstructured models. In this case, the control algorithm and the real measurements compensate the inaccuracy of the model which is managed as a disturbance.

1.3 Models for Systems Biology

In recent years a new discipline has emerged in the biology field that emphasises the importance of studying the cell metabolism under a system-level approach: the so-called Systems Biology[1] (Palsson, 2000; Kitano, 2002; Ideker, 2003; Klipp, 2005). It is commonly accepted that the appearance of high-throughput technologies, such as genomics transcriptomics, metabolomics or proteomics, is the cause of this transfor-

[1] The essence of Systems Biology is probably not new, but recent progresses (experimental and computational) are fuelling a renewed interest in a system-level approach to biology.

mation of biology. These techniques are providing a considerable amount of data, which implies a transition from a data-poor to a data-rich environment. It has been suggested that further biological discovery will be limited, not by the availability of biological data, but by the lack of available tools to analyse and interpret the data (Palsson, 2000). This explains why the main goal of Systems Biology is transform system-level data into system-level understanding. To accomplish its goal, Systems Biology combines experimental and theoretical approaches and assigns a central role to mathematical modelling (Kitano, 2002; Ideker, 2003; Stelling, 2004; Klipp, 2005).

Cells, tissues, organs, organisms and ecological webs are examples of biological systems which can be approached with Systems Biology. However, herein we reduce our scope to cells, and mainly to the cell metabolism. Considering the mechanisms operating in metabolism will lead to detailed, quantitative predictions on cellular dynamics (Stelling, 2004). However, the complexity of cells and the lack of knowledge on these mechanisms and its associated parameters has limited the success of this approach (Palsson, 2000). In fact, even though biological information is growing rapidly, we still do not have enough information to describe the cellular metabolism in mathematical detail for a single cell (Palsson, 2006; Bailey, 2001). Thus, the mathematical models of cells used in Systems Biology range from global, yet coarse, views of cellular systems to very detailed descriptions with a more limited scope.

Applications of models in Systems Biology

A non-exhaustive list of the applications of mathematical models in the context of Systems Biology includes:

- *Organize disparate information into a coherent whole.* Probably the goal of Systems Biology: combining in a rational manner the information about each component involved in a biological system.

- *Get insight on the modelled phenomena.* Generate experimentally testable hypotheses on underlying mechanisms as well as predictions of cellular behavior, thereby iteratively producing refined models and insight about the system (Stelling, 2004). This was called simulation-based analysis by Kitano (2002).

- *Explore questions not amenable to experimental inquiry.* An illustrative example is given by Bailey (1998), «*Now there is not need of dissect a genome […] since the entire palette of genes is accessible on the internet. Suddenly, and now inescapably, comes the question: what do these genes do, acting together?*».

- *Study systemic properties.* Those properties of the modelled system that emerge from the whole system and are not properties of individual parts. Examples are pathway redundancy in networks, or the coexistence of modules—units performing a particular function—in cellular systems.

- *Understand the essential qualitative features*. A mathematical model may be imprecise to provide quantitative predictions, however, its predictions can be valuable as a source of qualitative knowledge. See (Bailey, 1998) for a suggestive example.
- *Discover strategies for metabolic engineering*. Reliable models linking genotype and phenotype would allow for directed manipulation of the gene content of an organism to obtain a desired phenotype. This ability would provide a basis for the rational selection of drugs targets and metabolic engineering interventions to get strains with desired properties (Price, 2003).

Table 1.1. Comparison between models in Bioprocess Engineering and Systems Biology.

	Bioprocess Engineering models	Systems Biology models
Main goal	Improve industrial processes	Aid in basic science research
Modelled aspects	Biological and engineering	Purely biological
Characteristics	Quantitative	Quantitative or qualitative
(Typical)	Dynamic	Dynamic or static
	As simple as possible	Understandable
	Empirical	Highly knowledge based
	Often unstructured	Structured

Main characteristics of models in System Biology

The most important characteristics of models of cells and cell populations used in Systems Biology can be summarised as follows:

- Modelling is focused on the internal cell behaviour.
- Models are often complex.
- Models can be dynamic or static, quantitative or qualitative.

Most models in Systems Biology consider intracellular phenomena, to get insight into the operating mechanisms, or to exploit our knowledge about these mechanisms. The extracellular behaviour of cells, how they interact with its environment, is of course accounted for, but typically as consequence (outcome) of the internal processes.

Considering the intracellular processes often leads to complex models. This does not means that all aspects of the system need to be known, but due to the intrinsic complexity of cells, even a simple representation of its internal mechanisms results in complex models (e.g., with a considerable number of elements, non-linear relations, time-varying parameters, etc.).

Finally, the multiple objectives of Systems Biology imply that different kinds of models, dynamic or static, quantitative or not, may be of use.

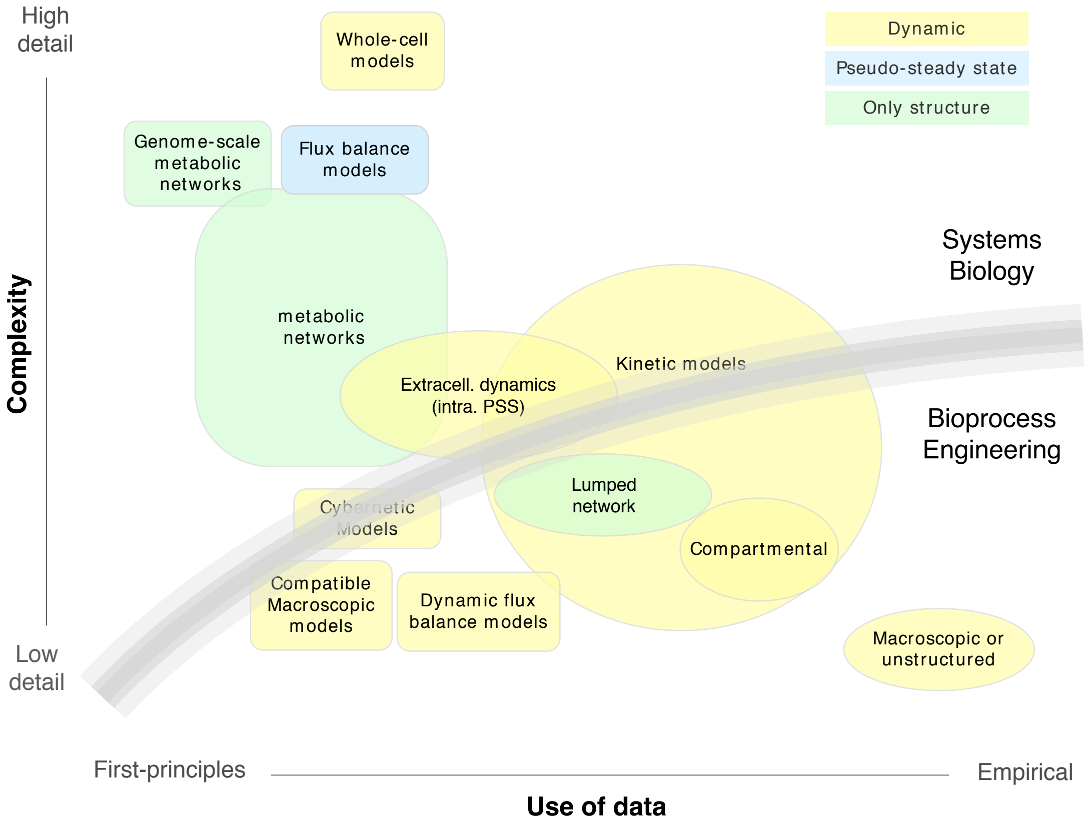


Figure 1.2. Different types of models classified by complexity and use of experimental data.

1.4 Classification of models of cells

In this section we describe different ways of classifying models, both in Bioprocess Engineering and Systems Biology. On the one hand, the purpose of the model determines which kind of model is desirable. On the other, very often the available knowledge (or data), constraints the kind of models that can be built. A non-exhaustive list of modelling approaches is given in Table 1.2 and Figure 1.2.

Data-driven or knowledge-based. A model is said to be data-driven when it is based on relationships between data. Some typical data-driven models are neural networks, fuzzy logic models and multivariate statistical models (e.g., those based on principal components analysis). On the contrary, a model is knowledge-based if its mathematical structure is derived (or inspired) from first principles knowledge about the modelled phenomena. Notice, however, that many knowledge-based models use some data, e.g., to fit parameters. In fact, the term data-driven is often reserved to purely data-driven models. Most models of cells and cell population systems are knowledge-based. This seems reasonably since there is a huge amount of qualitative knowledge available.

Parametric or non parametric. A model is called parametric if includes parameters requiring to be fitted with experimental data. Otherwise, the model is called non parametric. This classification is closely related with the previous one: a data-driven model is always parametric, but knowledge-based models can be parametric or non parametric.

Dynamic or static. A dynamic model represents changes of variables over time, while a static model does not. A static model describes the steady-state of the process at specific time instants that correspond to particular environmental conditions. A dynamic model represents the temporal evolution of the variables in the system, usually by means of ordinary (or partial) differential equations.

Structured or unstructured. The term unstructured designates models derived without an explicit consideration of processes operating inside the cells (Fredrickson, 1970). Basically, the cell is regarded as a black-box, or a catalyst for the conversion of substrates into products. Instead, a structured model accounts for (some) processes that operate inside the cells. A structured model may inform about the physiological state of the cells, its composition or its regulatory adaptation to the environmental changes; thus, structured models range from crude representations to highly detailed ones.[1]

Structured models typically arise: (a) As a way of improving the predictive capacity of an unstructured model (limited if the biological activity is characterised simply by the total biomass). This bottom-up approach—followed by Bioprocess Engineering—leads to low-complexity, compartmental models. (b) With the solely purpose of modelling the processes operating within cells. This top-down strategy—followed by Systems Biology—leads to highly detailed representations of the cell.

Segregated or non segregated. The majority of models of cell populations consider a homogeneous population. However, there are important phenomena which cannot be described under this assumption (Schügerl, 2000): alterations and disturbances in physiology and cell metabolism, morphological differentiation of the cells, mutations in the genome, spatial segregation, aggregation of cells or growth of more than one species, etc. To address this situation, simple, segregated models discriminating several classes of cells can be found in literature (Schüegerl, 2000; Henson, 2003). More complex models consider a continuous variation in cells properties by means of partial differential equations.

[1] In one extreme, a structured model may consider mechanisms operating in metabolism, signal processing and gene regulation; in the other, it might represent cells as a two-compartmental system.

Table 1.2. Knowledge based types of models for cells and cell populations.

Types	Methodology		Other characteristics	Ref.
Dynamic Unstructured	Macroscopic models	BE	Predictive Parametric	(Bastin, 1990)
	Compatible macroscopic models	BE	Predictive Parametric Derived from a structure	(Provost, 2006) (Teixeira, 2007)
	Dynamic flux balance models (Constraint-based models)	Both	Predictive Parametric Assumes optimality	(Mahadevan, 2003)
Dynamic Structured	Compartmental models	BE	Predictive Parametric	(Schügerl, 2000)
	Kinetic models	Both	Predictive Parametric	(Gombert, 2000)
	Cybernetic models	Both	Predictive Parametric Assumes optimality	(Ramakrishna, 1996)
	Whole-cell models	SB	Predictive Parametric Considers regulation, etc.	(Tomita, 2001)
Static Structured	Lumped metabolic networks	Both	Non-predictive Non-parametric	(Nielsen, 1992)
	(genome)-scale networks[a]	Both	Non-predictive Non-parametric	(Chassagnole, 2002) (Forster, 2003)
	Interaction-based models	SB	Non-predictive Non-parametric	(Stelling, 2004)
	Constraint-based models	SB	Non-predictive	(Gombert, 2000
	Flux balance models[b]	SB	Predictive Assumes optimality	(Price, 2003)

[a] Here we refer just to networks; the most common genome-scale models are classified as a particular type of (large) constraint-based models. [b] We consider "Flux balance models" as a subclass of "Constraint based models" that incorporate an assumption of optimal cell behavior.

Kinetic and constraint-based models. Beside these classifications, most models of cells, and particularly models of the cell metabolism, can be enclosed within two categories: kinetic model and stoichiometric (structured, or constraint-based) models.

Kinetic models are dynamic models accounting for the kinetics of intracellular cellular processes (e.g., enzyme-catalysed reactions, protein-protein interactions, or protein-DNA bindings). These models are typically formulated by means of ordinary differential equations. These models include reaction rates and other kinetic parameters that must be fitted using (dynamic) experimental measurements of inner processes, information that is often lacking. To avoid the need of kinetic data, constraint-based models can be build under the assumption that (most) intracellular processes are at steady-state. Notice that constraint-based models disregard intracellular dynamics, but are not necessarily static because extracellular dynamics (typically slower) can still be accounted for.

Both approaches will be discussed hereinafter. The rest of this chapter is devoted to kinetic models, and constraint-based models, which are those used along this thesis, will be reviewed in more deep in chapter II.

1.5 Kinetic models

The rest of this chapter will review different kinetic models of cells, starting form the simplest ones, and going on towards increasing levels of complexity.

Unstructured, kinetic models

Unstructured, kinetic models, often called macroscopic models, are the simplest ones: those that do not consider the internal structure of cells. The only biological variable considered in these models is the cell mass concentration or biomass, which is regarded as a black-box that converts substrates into products. Generally, biomass is linked with the extracellular species—substrates and products—by means of macro-reactions. Each macro-reaction has an elementary kinetic expression, such as Monod or Haldane, which describe the influence of substrates and product concentration or other variables, such as pH or temperature (Bastin, 1990; Dunn, 2000). Dynamical mass balances are then established from these macro-reactions identifying appropriate kinetic parameters from the available experimental data. This overall view represents an oversimplification of the reality. However, unstructured models have been successfully applied for long time in the field of bioprocess engineering.

The main characteristics of most unstructured, kinetic models are the following:

- They are knowledge-based.[1]
- They have parameters to be fitted with experimental data.
- They are dynamic.
- They are non-segregated.

The main advantage of unstructured models is its simplicity. This simplicity implies that unstructured models can be built without a huge amount of data and knowledge, because the number of variables of the model is kept at a minimum. Moreover, although experimental data is necessary, it can easily be obtained because only extracellular variables are included in the model (i.e., there is not need for intracellular measurements). These measurements can be acquired with a low sample rate to capture the dynamic behaviour, and then be used to fit the parameters of the model and to validate its predictions. This simplicity is also useful in those applications, such as process control and monitoring, where measurements are needed on-line.

Macroscopic models can be validated to guarantee that they emulate the actual process with accuracy under certain conditions; the environmental ones, which are under control, but also the intracellular state of cells, which is assumed to be constant.

Unstructured models fail whenever they are used inappropriately to describe situations where cells regulation, composition or morphology are important variables, i.e., when the characterisation of biological activity only by means of the total biomass is not sufficient. That may happen, for example, when a gene is induced or repressed, or when a genetically modified microorganism losses the modification. Another drawback of unstructured models is that they are not easily scalable. Although it is possible to incorporate complexity to an unstructured model adding new empirical parameters, this approach may result in a non understandable model. Proceeding in this manner we are disregarding our knowledge about the cell, which can be useful not only to keep the model understandable, but also to suggest extensions, and to build a structure where new experimental data can be incorporated as it become available.

Two examples will be reviewed for the shake of illustration. For details about unstructured models, consult the references (Schüegerl, 2000; Bastin, 1990; Dunn, 2000).

Example: one macro-reaction. Several cell processes can be described with simple macro-reactions linking product, substrates and microbial growth. Consider, for instance, one macro-reaction: substrate (s) $\xrightarrow{x}$ biomass (x) + product (p). Mass balances can be derived, resulting in the following ordinary differential equations:

[1] Its structure relies on our knowledge about cells, substrates needed to growth, excreted products, influence of pH, etc.

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t} &= \mu \cdot x - \mathrm{D} \cdot x \\ \frac{\mathrm{d}s}{\mathrm{d}t} &= v_s \cdot x + \mathrm{D} \cdot (s_i - s) \\ \frac{\mathrm{d}p}{\mathrm{d}t} &= v_p \cdot x - \mathrm{D} \cdot p\end{aligned} \tag{1a}$$

where x, s and p denote concentrations, s_i the substrate concentration in the inflow, and D the dilution rate (i.e., inflow per volume).

To complete the model, kinetic expressions should be given, such as:

$$\mu = \frac{\mu_{\max} s}{k_2 + s}, \quad v_s = \mathrm{Y}_{\mathrm{xs}} \cdot \mu + \mathrm{m}_{\mathrm{s}}, \quad v_p = \mathrm{Y}_{\mathrm{xp}} \cdot \mu + \mathrm{m}_{\mathrm{p}} \tag{1b}$$

where the specific rates of product formation and substrate utilisation are considered proportional to growth rate, and the growth rate is particular function of the substrate concentration (a so-called Monod kinetics).

This is a simple model, yet useful in many contexts. It includes the most fundamental observations concerning growth processes: (i) that the rate of cell mass production is proportional to biomass concentration; (ii) that there is an upper limit for growth rate on each substrate; and (iii) that the cells need substrate to survive. The model can be extended to include other phenomena, such as growth inhibition by the product, and similar macro-reaction schemes can be also stated (Bastin, 1990).

Example: *S. cerevisiae* model. A classic unstructured model of *S. cerevisiae* is the one developed by Sonnleitner and Käppeli (Sonnleitner, 1986). It is based on experimental observations and the hypothesis of a limitation in the oxidative capacity to explain the shift to ethanol formation observed in *S. cerevisiae*. The model also describes the decrease in the oxidative capacity with decreasing oxygen concentration—the so-called Pasteur effect. The model fits steady-state experiments very well, but it gives a poor description of transient operating conditions. To overcome this and other limitations several extensions of the model have been proposed (Nielsen, 1992).

Structured, kinetic models

Structured, kinetic models are the natural extension of unstructured models. Typically, the cell is structured into several intracellular compounds which are connected to each other and to the environment by fluxes, on the basis of the knowledge of fundamental biochemistry (Nielsen, 1992). The kinetic model is then built of balances of intracellular compounds represented with ordinary differential equations, which include reaction rates and other parameters. This formulation leads to quantitative predictions of the temporal evolution of the intracellular metabolites.

Structured, kinetic models are potentially more powerful than unstructured ones: (1) they may provide a realistic description of inner cell processes, (2) give more accurate predictions, and (3) be valid in a wider range of conditions. Nevertheless, these advantages do not come without cost: their development is a demanding task that requires better knowledge and more experimental data.

The degree of detail of structured, kinetic models varies within a wide range. In principle, the genome-scale reaction network that represents the whole-cell metabolism may be used as basis for a kinetic model (if available). However, major difficulties arise when trying to build a kinetic model based on a detailed reaction network (Gerdtzen, 2004):

(i) Changes in environmental conditions may cause cellular changes at many levels: transcription, translation and metabolic reactions.

(ii) Intracellular reactions are very complex and there are pathways for which detailed reactions have not yet been elucidated.

(iii) There is a lack of knowledge on kinetic mechanisms.

(iv) It is difficult to obtain experimentally the kinetic parameters for all intracellular reaction, due to the lack of available measurements (particularly at a sampling rate sufficient to capture intracellular dynamics).

The last one is probably the most critical: it implies that, even if it were possible to identify the kinetic mechanism for each intracellular reaction, the model will involve an extremely large number of equations for which many kinetic parameters are still unknown. The parameters estimation requires a special care to avoid a lack of identificability: the model complexity may lead to a parameter estimation that fits the experimental data very well, but that, in fact, is not capturing a physically valid behaviour (Dunn, 2000). In general, the experimental verification of model becomes increasingly difficult as the model complexity is increased (Schügerl, 2000).

To avoid this difficulties, smaller networks can be formulated by grouping many intracellular reactions into a reduced number of global reactions, or by including only the reactions which constitute the central metabolism. In this way, the highly detailed networks can be the basis for reasonably small kinetic models.

To close this section several examples of structured, kinetic models that can be found in the literature will be briefly described. Most of these examples came from the context of Bioprocess Engineering, but some of them were developed in the context of Systems Biology; more of these last ones can be found in (Klipp, 2005).

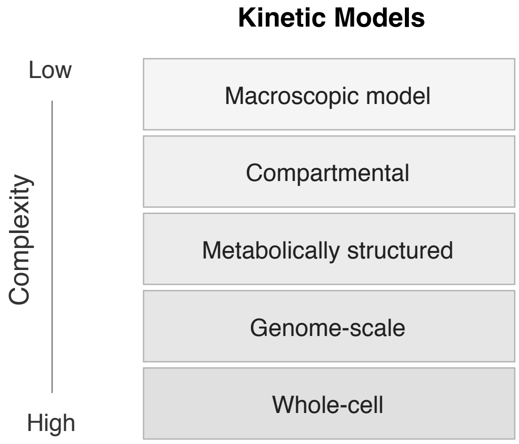


Figure 1.3. Kinetic models with increasing complexity.

Example: simple structured kinetic model. Consider the toy metabolic reaction network taken from (Provost, 2004) and depicted in Figure 1.4. There are two extracellular substrates (s_1 and s_2) and only one extracellular product (p_1). The cell is structured in 6 metabolites (e), one of them accumulated. The following mass balances can be stated:

$$\frac{\mathrm{d}x}{\mathrm{dt}} = \mu \cdot x + \mathrm{D} \cdot (x_i - x) \tag{2a}$$

$$\frac{\mathrm{d}\mathbf{e}}{\mathrm{dt}} = \mathbf{N_e} \cdot \mathbf{v} \cdot x + \mathrm{D} \cdot (\mathbf{e_i} - \mathbf{e}) \tag{2b}$$

$$\frac{\mathrm{d}\mathbf{c}}{\mathrm{dt}} = \mathbf{N} \cdot \mathbf{v} - \mu \cdot \mathbf{c} \tag{2c}$$

where **e** denotes the vector of extracellular metabolites concentrations (both substrates and products), $\mathbf{e_i}$ the inflow concentrations, **c** is the vector of intracellular metabolites, and $\mathbf{N_e}$ and **N** are stoichiometric matrices linking metabolites and fluxes.

In this particular example, considering that there is no inflow through the system boundaries (D=0) and taking into account the reactions in the network, the mass balances are the following:

$$\frac{\mathrm{d}x}{\mathrm{dt}} = \mu \cdot x = \begin{pmatrix} 1 & 1 & 1 & 1 & -1 & 0 & 0 & -2 \end{pmatrix} \cdot \mathbf{v} \cdot x \tag{3a}$$

$$\frac{\mathrm{d}\mathbf{e}}{\mathrm{dt}} = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \cdot \mathbf{v} \cdot x \tag{3b}$$

$$\frac{d\mathbf{c}}{dt} = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & -1 & 0 \end{pmatrix} \cdot \mathbf{v} - \mu \cdot \mathbf{c} \qquad (3c)$$

Note: the expression for the growth rate (μ) is based on the formation of all internal metabolites, however, many other approaches can be found in literature.

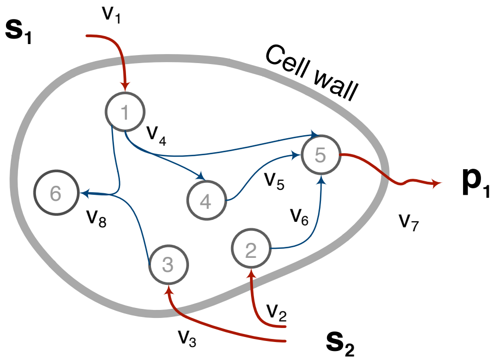


Figure 1.4. Simple reaction network extracted from (Provost, 2004).

Two-compartmental Model. The simplest approach to improve an unstructured model consists in dividing the cell into two compartments. A two-compartmental model to represent the diauxic growth of *Klebsiella terrigena* on the substrates glucose and maltose is described in (Schügerl, 2000). The first substrate is the preferred one and inhibits and represses the uptake of the second one. The enzymes of the maltose are represented with a compartment and another compartment stands for the remaining metabolism. The model fits with experimental data after identify its parameters (Schügerl, 2000).

Multi-compartment Model. By combination of the compartmental model concept with an intracellular ATP balance, Villadsen and Nielsen (1992) derived a kinetic model for *S. cerevisiae*. The model includes the shift to ethanol formation observed in the metabolism of *S. cerevisiae* during an aerobic glucose-limited chemostat, considering six macro reactions and three compartments. It was applied for simulation of a diauxic batch experiment and showed a good agreement with experimental measurements. Unfortunately, the large number of parameters needs a big amount of experimental data to be estimated, and, even with intracellular measurements, it is difficult to quantify the biomass compartments due to its difficult interpretation.

Biochemically Structured Model. Lei et al. (2001) developed a biochemically structured, kinetic model for the aerobic growth of *S. cerevisiae* on glucose and ethanol.

The model defines two compartments showing some similarities with the Nielsen and Villedsen model (1992). It provides a new interpretation of the shift in yeast metabolism based on the pyruvate and acetaldehyde branch points.

The model considers the following 12 reactions:

$$\mathrm{s_{glu}} \rightarrow \mathrm{s_{pyr}} + 0.33 \cdot \mathrm{NADH}$$

$$\mathrm{s_{pyr}} \rightarrow \mathrm{CO_2} + 1.67 \cdot \mathrm{NADH}$$

$$\mathrm{s_{pyr}} \rightarrow 0.67 \cdot \mathrm{s_{acetald}} + 0.33 \cdot \mathrm{CO_2}$$

$$\mathrm{s_{acetald}} + 0.5 \cdot \mathrm{NADH} \rightarrow \mathrm{s_{etOH}}$$

$$\mathrm{X_a} \rightarrow \mathrm{X_{Acdh}}$$

$$\mathrm{X_a} \rightarrow \mathrm{degrad.}$$

$$\mathrm{NADH} + 0.5 \cdot \mathrm{O_2} \rightarrow \mathrm{ATP}$$

$$\mathrm{s_{acetald}} \rightarrow \mathrm{s_{acetate}} + 0.5 \cdot \mathrm{NADH}$$

$$\mathrm{s_{acetald}} + 0.5 \cdot \mathrm{NADH} \rightarrow \mathrm{s_{etOH}}$$

$$\mathrm{s_{acetate}} \rightarrow \mathrm{CO_2} + 2 \cdot \mathrm{NADH}$$

$$\mathrm{s_{glu}} \rightarrow 0.91 \cdot \mathrm{X_a} + 0.08 \cdot \mathrm{CO_2} + 0.12 \cdot \mathrm{NADH}$$

$$\mathrm{s_{acetate}} \rightarrow 0.78 \cdot \mathrm{X_a} + 0.22 \cdot \mathrm{CO_2} + 0.4 \cdot \mathrm{NADH}$$

$$\mathrm{X_{Acdh}} \rightarrow \mathrm{degrad.}$$

The 12 reactions rates are formulated with Michaelis-Menten kinetics, and extended based on physiological knowledge. For the shake of brevity, only three of them are shown here:

$$v_1 = k_{1l} \cdot \frac{s_{glu}}{K_{1l} + s_{glu}} \cdot x_a + k_{1h} \cdot \frac{s_{glu}}{K_{1h} + s_{glu}} \cdot x_a + k_{1e} \cdot \frac{s_{glu}}{K_{1e} + s_{glu} \cdot \left(K_{1i} \cdot s_{actald} + 1\right)} \cdot s_{actald} \cdot x_a \tag{4a}$$

$$v_3 = k_3 \cdot \frac{s_{pyr}^4}{K_3 + s_{pyr}^4} \cdot x_a \tag{4b}$$

$$v_9 = \left(k_9 \cdot \frac{s_{glu}}{K_9 + s_{glu}} + k_{9e} \cdot \frac{s_{etOH}}{K_{9e} + s_{etOH}} \right) \cdot \frac{1}{K_{9i} \cdot s_{glu} + 1} \cdot x_a + k_{9c} \cdot \frac{s_{glu}}{K_9 + s_{glu}} \cdot x_a \tag{4c}$$

At this point, the mass balances can be formulated as follows:

$$\frac{\mathrm{d}x}{\mathrm{dt}} = \mu \cdot x, \qquad \text{where } x = x_a + x_{acdh} \tag{5a}$$

$$\frac{\mathrm{d}}{\mathrm{dt}} \begin{pmatrix} s_{glu} \\ s_{pyr} \\ s_{acetald} \\ s_{acetate} \\ s_{etOH} \end{pmatrix} = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0.98 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.36 & -1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.04 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \cdot \mathbf{v} \cdot x - \begin{pmatrix} s_{glu} - s_f \\ s_{pyr} \\ s_{acetald} \\ s_{acetate} \\ s_{etOH} \end{pmatrix} \cdot \mathrm{D} \tag{5b}$$

$$\frac{\mathrm{d}}{\mathrm{dt}}\begin{pmatrix} O_2 \\ CO_2 \\ x_a \\ x_{Acdh} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0.732 & 0.619 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.732 & 0.619 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix} \cdot \mathbf{v} - \mu \cdot \begin{pmatrix} 0 \\ 0 \\ x_a \\ x_{Acdh} \end{pmatrix} \qquad (5c)$$

This illustrates the difficulties that arise when a kinetic model gains in details: kinetic expressions are complex and fitting its parameters require more data than is often available.[1] For this particular work, Lei et al. developed a five-step procedure for parameters fitting (2001). The model was then validated on different experimental data. During a batch process, the model describes the glucose and ethanol profiles, and a reasonable prediction for pyruvate and acetate. However, the dynamic fed-batch experiments showed the limitations of the model.

Dynamic Model of S. Cerevisiae. In (Rizzi, 1997) an extensive kinetic model of glycolysis in *S. Cerevisiae* was introduced. The model is based on material balance equations of the key metabolites in the extracellular environment, the cytoplasm and the mitochondria. The model includes 22 compounds (for extracellular variables, intracellular metabolites and co-metabolites), 23 reactions and 23 kinetic reaction rates. It was verified by *in vivo* diagnosis of intracellular enzymes, and it was proved that its predictions fit reasonably well with experimental measurements.

Dynamic Model of *Escherichia coli*. In (Chassagnole, 2002) a detailed, dynamic model of the central carbon metabolism of *E. coli* was described. This was the first dynamic model linking the sugar transport system with the reactions of glycolysis and the pentose-phosphate pathway. It includes 18 compounds (for extracellular compounds and intracellular metabolites), 29 reactions and 29 kinetic reaction rates. Experimental measurements of intracellular metabolites at transient conditions were used to validate the structure of the model and to estimate the kinetic parameters.

Other structured, kinetic models

To close this chapter, we discuss two particular classes of structured, kinetic models: cybernetic models and whole-cell models.

Cybernetic models consider metabolic regulation as mediated through the control of enzyme synthesis and enzyme activity (Kompala, 1986). It is well-known that when faced with environmental changes, cells have more than one possible response in their metabolic machinery. The cybernetic approach assumes that metabolic systems have evolved optimal goal oriented strategies as a result of evolutionary pressures. Hence, cells switch their metabolism in response to changes in their environment in a manner

[3] Notice, moreover, that the representation used herein is still a huge simplification: there are thousands of metabolic reactions occurring within cells.

consistent with its optimal strategies. The outcome of these strategies modifies the intrinsic process kinetics. Interestingly, this assumption reduces the need of kinetic parameters. The applications of the cybernetic approach include models of diauxic growth of microorganisms (Kompala, 1986), the sequential and the simultaneous utilisation of substitutable substrates (Ramakrishna, 1996) and the growth of mammalian cell cultures (Guardia, 2000). It has been also used to aid in metabolic engineering tasks (Varner, 1999).

Whole-cell models are the first attempts to construct comprehensive, kinetic models of a complete cell (Tomita, 2001). The canonical whole-cell model consisted of a "virtual cell" with 127 essential genes selected from the genome set of *Mycoplasma genitalium* (Tomita, 1999). Ishii et al. illustrate the integrative nature of the whole-cell modelling approach when they explain the features of this first model (Ishii, 2004): «This virtual cell could transport extracellular glucose across the cell membrane, metabolize it through the glycolytic pathway and produce ATP molecules. These ATP molecules could in turn be utilized for the biosynthesis of phospholipids or the maintenance of the transcription/translation system». In contrast with other large-scale representations, the whole-cell modelling approach is focused on modelling the dynamic behaviour of cells. This is a very ambitious task because a great amount of quantitative data is needed—concentrations of metabolites and enzymes, flux rate, kinetic terms, etc. The development of high-throughput technologies to measure intracellular variables is one of the keys for the success of whole-cell models.

1.6 Conclusions

This chapter has been devoted to review the classes of models of cells and cell populations that are typically used in the fields of Bioprocess Engineering and System Biology. We have seen that there is a wide range of models, different in purpose and characteristics.

However, it seems that the models used in both domains are becoming more similar because models used in Bioprocess Engineering are gaining in detail to improve its predictive capacity—thanks to new measurement techniques that enable validation. At the same time, quantitative predictive models receive more attention form biologists due to the emergence of Systems Biology.

It is expected that the increasing availability of biological data, in conjunction with the currently available qualitative knowledge, may result in new (and better) models in future years. This will be particularly significant for basic science research, but bioprocess industries will be also fuelled by these advances.

Main references

- Bastin G, Dochain D (1990). *On-line Estimation and Adaptative Control of Bioreactors*. Amsterdam, Netherlands: Elsevier.

- Dunn IJ, Heinzle E, Ingham J, Prenosil E (2000). *Biological Reaction Engineering: Dynamic Modelling Fundamentals with Simulation Examples*. Wiley, Zürich.

- Schügerl K, Bellgardt KH (2000). *Bioreaction Engineering: Modelling and Control*. Heidelberg, Germany: Springer-Verlag.

- Bailey JE (1998). Mathematical modeling and analysis in biochemical engineering: past accomplishments and future opportunities. *Biotechnology Progress*, 14:8-20.

- Gombert AK, Nielsen J (2000). Mathematical modelling of metabolism. *Current Opinion in Biotechnology*, 11:180-186.

- Kitano H (2002). Computational systems biology. *Nature*, 420:206-210.

- Stelling J (2004). Mathematical models in microbial Systems Biology. *Current Opinion in Microbiology*, 7:513-518.

- Klipp E, Herwig R, Kowald A, Wierling C, Lehrach H. (2005). *Systems biology in practice: concepts, implementation and application*. Weinheim, Germany: Wiley-VCH.

- Nielsen J, Villadsen J (1992). Modelling of microbial kinetics. *Chemical Engineering Science*, 47:4225-4270.

II

Constraint-based models of the cell metabolism

Different methodologies use models of the cell metabolism that share two characteristics: (i) are derived from a metabolic network and (ii) assume steady-state for the intracellular metabolites. These methodologies have different purpose, employ different mathematical tools, and rely on different assumptions; but they all exploit the properties of a constraint-based description of cells.

In this chapter, we show that all these methodologies can be presented with a unified perspective under the label of constraint-based models. Next, three outstanding methodologies that use these kind of models are described: Network-based Pathways Analysis, Metabolic Flux Analysis, and Flux Balance Analysis.

Constraint-based modelling, and these three methodologies in particular, are the context for the contributions of this thesis that will described in subsequent chapters.

Part of the contents of this chapter appeared in the following journal article:

- Llaneras F, Picó J (2008). Stoichiometric Modelling of Cell Metabolism. *Journal of Bioscience and Bioengineering*, 105:1.

2.1 Introduction

An observed cellular behaviour may be explained by considering the constitutive elements of cells. However, to define the cell capabilities and predict its behaviour, the interactions between elements need to be considered. This confers a crucial role to networks because they embed these interactions, and thus they are responsible for observable cellular behaviour (Palsson, 2006). Examples of networks used in biology include regulatory and signaling networks; however, in terms of its biochemistry, kinetics, and thermodynamics, metabolism is the best characterized cellular network.

In this chapter, the terms *Stoichiometric modelling* and *Constraint-based modelling* are used to encompass methodologies based on representations of the cell metabolism that share two characteristics, the use of a metabolic network and the pseudo steady-state assumption (Figure 2.1):

- Stoichiometric models are derived from a metabolic network of the organism being modelled. The reaction stoichiometry embedded in these networks is the starting point, but the models are not limited to stoichiometry. A constraint-based perspective will be used to highlight this fact.

- Stoichiometric models disregard the dynamic intracellular behaviour, based on an assumption of steady state for (some) internal metabolites (Stephanopoulos, 1998).[1]

This way, stoichiometric or constraint-based modelling provides structurally detailed models, at the cost of disregarding the intracellular kinetics. Notice that two different notions of *model* will coexist hereinafter. We consider constraint-based representations as *models*, because they are mathematical descriptions of cell capabilities, even if they are unable to predict the behaviour shown at particular conditions. To avoid confusion regarding this, a model with predictive capacity is always explicitly named as a *predictive model* hereinafter.

The rest of the chapter is organised as follows. In sections 2.2 and 2.3 the classical principles of stoichiometric modelling are summarised. In section 2.4 we review this principles from a constraint-based perspective. The methodologies within the framework are briefly classified in section 2.5. Sections 2.6 to 2.8 are devoted to three methodologies of particular interest for the rest of this thesis: metabolic flux analysis, flux balance analysis, and network-based pathway analysis. Finally, the main conclusions are outlined.

[1] The dynamics of the extracellular reactions, the exchanges between cells and its environment, can still be taken into account.

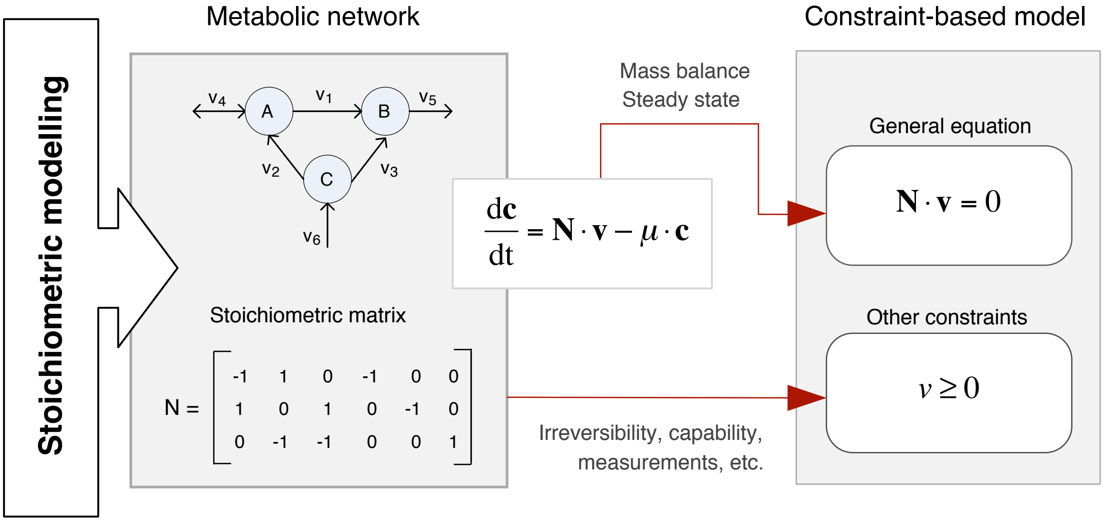

Metabolic network
Constraint-based model
Stoichiometric modelling
v_4
v_1
v_5
A
B
C
v_2
v_3
v_6
Stoichiometric matrix
$N = \begin{bmatrix} -1 & 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 & -1 & 0 \\ 0 & -1 & -1 & 0 & 0 & 1 \end{bmatrix}$
Mass balance
Steady state
$\frac{d\mathbf{c}}{dt} = \mathbf{N} \cdot \mathbf{v} - \mu \cdot \mathbf{c}$
General equation
$\mathbf{N} \cdot \mathbf{v} = 0$
Other constraints
$v \geq 0$
Irreversibility, capability,
measurements, etc.

Figure 2.1. Principles of the stoichiometric modeling framework. Given a metabolic network, the mass balance around each intracellular metabolite can be mathematically represented with an ordinary differential equation. If we do not consider intracellular dynamics, the mass balances can be described by a homogeneous system of linear equations: the so-called general equation. Other constraints can be also incorporated to further restrict the space of feasible flux states of cells.

2.2 Preliminaries: metabolic networks

Providing a comprehensive discussion of the importance of the metabolism is out of the scope of this chapter, but the following lines from Palsson (2006) will serve the purpose of motivation.

> *«Intermediate metabolism can be viewed as a chemical 'engine' that converts available raw materials into energy as well as building block needed to produce biological structures, maintain cells, and carry out various cellular functions.*
>
> *This chemical engine is highly dynamic, obeys the laws of physics and chemistry, and is thus limited by various physicochemical constraints. It also has an elaborate regulatory structure that allows it to respond to a variety of external perturbations.*
>
> *Metabolism comprises two types of chemical transformations: catabolic pathways that break down various substrates into common metabolites and anabolic pathways that collectively synthesize amino acids, fatty acids, nucleic acids, and other needed building blocks.*
>
> *During these processes, an intricate exchange of various chemical groups and reductionoxidation potentials takes place through a set of carrier molecules [e.g., ATP, NADH]. These carrier molecules and the properties that they transfer thus tie the metabolic network tightly together».*

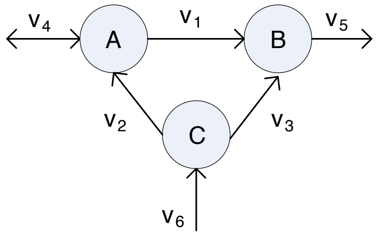
Traditionally, the metabolism was divided into individual metabolic pathways, which are indeed a central paradigm in biology. A metabolic pathway is a series of chemical reactions occurring within a cell, catalyzed by enzymes, resulting in either the formation of a metabolic product to be used or stored by the cell, or the initiation of another metabolic pathway. These pathways were defined on the basis of their step-by-step discovery, but this procedure is being substituted by a systemic approach.

With the arrival of genomics and proteomics and the increment in available quantitative data, the set of metabolic reactions (and pathways) involved in the whole cell metabolism are now assembled in networks (Palsson, 2006; Cornish-Bowden, 2000). Many metabolic networks are highly detailed to provide a comprehensive representation of the metabolism of a particular organism. However, smaller networks are also formulated, sometimes grouping sets of reactions or considering only parts of the metabolism, such as the central metabolism.

Metabolic networks and the stoichiometric matrix

The metabolism of living cells can be represented with a metabolic network under the form of a directed hyper-graph that encodes a set of elementary biochemical reactions taking place within the cell. In this hyper-graph the nodes represent the involved metabolites and the edges represent the metabolic fluxes or reaction rates. Two groups of fluxes can be defined: exchange fluxes and internal fluxes. Exchange fluxes represent an exchange with the environment outside the cells (uptake of substrates or formation of products). Internal fluxes represent metabolic reactions occurring within cells. A simple example is given in Figure 2.2.

The stoichiometric information embedded in a metabolic network with *m* metabolites and *n* reactions can be represented by a stoichiometric *m*×*n* matrix **N**, in which rows correspond to metabolites and columns to reactions.



$$\mathbf{N} = \begin{bmatrix} -1 & 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 & -1 & 0 \\ 0 & -1 & -1 & 0 & 0 & 1 \end{bmatrix}$$

Figure 2.2. A toy metabolic network. Nodes represent internal metabolites, edges the metabolic fluxes **v**, and arrows the reversibility of the reactions. Fluxes v_4, v_5 and v_6 correspond to exchanges with the environment.

2.3 Classical principles of stoichiometric modelling

Let us consider a cells population in an aqueous medium[1] and establish a set of mass balances to obtain a dynamic model (Bastin, 1990). The medium volume variation is given by: $\mathrm{dV/dt} = \mathrm{F_{in}} - \mathrm{F_{out}}$, where $\mathrm{F_{in}}$ and $\mathrm{F_{out}}$ are the inflow/outflow rates.

The growth rate of biomass (cells) can then be represented as follows:

$$\frac{\mathrm{d}x}{\mathrm{dt}} = \mu \cdot x - \mathrm{D} \cdot x + \mathrm{F}_x \qquad (1)$$

where x denotes the biomass concentration, μ its specific growth, D the dilution rate ($\mathrm{F_{in}}$/V), and $\mathrm{F_x}$ the biomass inflow rate which typically has value zero (because typically there is no biomass x_{in} in the inflow $\mathrm{F_{in}}$, and $\mathrm{F}_x = \mathrm{F}_{in} \cdot x_{in}/\mathrm{V}$).

Mass balances around the extracellular metabolites can be established as follows:

$$\frac{\mathrm{d}\mathbf{e}}{\mathrm{dt}} = \mathbf{v_e} \cdot x - \mathrm{D} \cdot \mathbf{e} + \mathbf{F_e} \qquad (2)$$

where **e** is the vector formed with the concentration of the extracellular metabolites (substrates and products), $\mathbf{v_e}$ the vector of specific, extracellular fluxes (uptakes and product formations), and $\mathbf{F_e}$ the net inflow/outflow of the extracellular metabolites.

Intracellular behavior

Given a metabolic network of the modelled cells, and extracting its stoichiometric matrix, the mass balances around the intracellular metabolites can be also represented by a set of ordinary differential equations (Provost, 2004),

$$\frac{\mathrm{d}(\mathbf{c} \cdot x)}{\mathrm{dt}} = \mathbf{N} \cdot \mathbf{v} \cdot x - \mathrm{D} \cdot \mathbf{c} \cdot x + \mathbf{c} \cdot \mathrm{F_x} \qquad (3)$$

where $\mathbf{c} = (c_1, c_2, ..., c_m)^{\mathrm{T}}$ is the vector of intracellular metabolites concentrations, and $\mathbf{v} = (v_1, v_2, \dots, v_n)^{\mathrm{T}}$ the vector of specific fluxes through each reaction, and **N** is the stoichiometric matrix linking fluxes and internal metabolites.

[1] In industrial processes the environment is usually a bioreactor, but herein we consider a more general situation; for instance, one could model the behaviour of cells on a natural environment, or model those processes that occur in a waste-water treatment plant.

To obtain a more operative expression we expand the derivatives,

$$\frac{\mathrm{d}(\mathbf{c}\cdot x)}{\mathrm{dt}}=\frac{\mathrm{d}\mathbf{c}}{\mathrm{dt}}\cdot x+\mathbf{c}\cdot\frac{\mathrm{d}x}{\mathrm{dt}} \tag{4}$$

By substituting (1) and (3) in (4), the mass balance equation around intracellular metabolites can be rewritten as follows:

$$\frac{\mathrm{d}\mathbf{c}}{\mathrm{dt}}=\mathbf{N}\cdot\mathbf{v}-\mu\cdot\mathbf{c} \tag{5}$$

This is the *dynamic mass balance equation*, which describes the evolution over time of the concentration of each metabolite. This equation implies that to model the dynamic evolution of intracellular metabolites we need information about stoichiometry (**N**), biomass growth (μ), and intracellular reaction fluxes (**v**).

Steady-state assumption

Unfortunately, the mechanisms of intracellular reactions are complex and still not very well understood. This, together with the lack of intracellular dynamic measurements, makes it difficult to build structured, kinetic models (Bailey, 1998; Palsson, 2000). This is why stoichiometric models disregard the dynamics of the intracellular reactions in (5) and assumes that (most) internal metabolites are at steady-state (Stephanopoulos, 1998).

This assumption is supported by the observation that intracellular dynamics are much faster than extracellular dynamics. Therefore, it is sensible to disregard its transient behaviour and consider that they rapidly reach the steady state.[1] The dilution term μ·c is also disregarded because it is generally much smaller than the fluxes affecting the same metabolite (Stephanopoulos, 1998).

Under these assumptions, the mass balances (5) can be described by a homogeneous system of linear equations, the so-called general equation:

$$\mathbf{N}\cdot\mathbf{v}=0 \tag{6}$$

In this way each stoichiometrically feasible steady-state is represented by a flux vector **v**. Notice, however, that this equation does not predict the actual state of cells. If **N** has full row rank, there are *m* independent equations. As *n* is typically larger than *m*,

[1] The steady-state assumption does not imply that the dynamic nature of the entire process is disregarded because extracellular dynamics (substrates uptake, product formation and biomass growth) can still be considered.

the system is underdetermined with *n-m* degrees of freedom. There is a whole space of feasible flux vectors, or flux states, that cells can shown.

The existence of multiple solutions makes sense, since cells show different behaviours depending on the environmental conditions, such as the availability of substrates or the temperature. Equation (6) must be seen as a representation of feasible states, or capabilities, of the metabolic network being modelled.

Equation (6) is the base of many tools to investigate the metabolism of living cells, some of which will be discussed in subsequent sections. First, let us discuss how additional constraints can be imposed to get richer representations of the cell metabolism.

2.4 Constraint-based modelling perspective

Constraint-based modelling is based on the fact that cells are subject to constraints that limit their behaviour (Palsson, 2006). In principle, if all constraints operating under a given set of circumstances were known, the actual state of a metabolic network could be elucidated; but most likely we will not be able to reach this state of knowledge soon (Palsson, 2000; Kitano, 2002). Nevertheless, imposing the known constraints, it is possible to determine which functional states can and cannot be achieved by a cell. The imposition of constraints leads to a space of feasible flux states, as it happens with the general equation (6), where every feasible flux vector lives (Wiback, 2004). Since a metabolic phenotype can be defined in terms of fluxes, this space represents, or at least contains, all the feasible phenotypes of cells (Edwards, 2002).

Now the general equation (6) can be seen as a set of stoichiometric constraints. In this way, the classical stoichiometric models can be seen as a particular kind of constraint-based models that only consider stoichiometric information.

Different types of constraints

Constraints can be divided in two main types: non adjustable (invariant) and adjustable ones (Table 2.1). The former are time-invariant restrictions of possible cell behaviour, whereas the latter depend on environmental conditions, may change through evolution, and may vary from one individual cell to another. Examples of non adjustable constraints are those imposed by thermodynamics (e.g, irreversibility of fluxes) and enzyme or transport capacities (e.g, maximum flux values). Enzyme kinetics, regulation, and experimental measurements are examples of adjustable constraints.

To study the invariant properties of a network, only invariant constraints can be used, because they are those that are always satisfied (i.e, they limit the cell capabilities). If adjustable constraints are used, the elucidated cell states will be only valid under the particular set of circumstances in which these constraints operate.

Table 2.1. Most common types of constraint.

Constraints	Type	Mathematical formulation
Systemic stoichiometry	Non-adjustable	$\mathbf{N} \cdot \mathbf{v} = 0$
Irreversibility of fluxes	Non-adjustable	$v \geq 0$
Enzyme/transporters capacities	Non-adjustable	$v^m \leq v \leq v^M$
Measured fluxes	Adjustable	$v = w \quad \text{or} \quad w^m \leq v \leq w^M$
Regulatory constraints	Adjustable	$v_o = 0, \quad \text{if} \quad v_i \neq 0$
Kinetic constants	Adjustable	$v = k \cdot \mathrm{C}_{\mathrm{m}}$, (C_m is a concentration)

Space of feasible flux states

The general equation (6) provides a set of stoichiometric constraints that link some fluxes with others, thus restricting the space of feasible flux vectors to a hyper-plane, a subspace of $\mathbf{R}^{\mathrm{n}}$ (Figure 2.3). An a second step, certain reactions are often considered irreversible, that is, able to operate only in one direction.

In this way, taking into account intracellular mass balances (6) and irreversibility constraints, a space of feasible steady state flux vectors b, the so-called *flux space*, can be defined as follows:

$$\mathbf{P} = \left\{ \mathbf{v} \in \mathbf{R}^{\mathrm{n}} : \begin{array}{c} \mathbf{N} \cdot \mathbf{v} = 0 \\ \mathbf{D} \cdot \mathbf{v} \geq 0 \end{array} \right\} \tag{7}$$

where **D** is a diagonal *nxn*-matrix with $\mathbf{D_{ii}} = 1$ if the flux *i* is irreversible, otherwise 0.

It is also very common to impose maximum flux values, derived from enzyme or transport capacities. In this way, one can add constraints of the form:

$$\mathbf{v^m} < \mathbf{v} < \mathbf{v^M} \tag{8}$$

If this data is available for every flux in the network, the flux space becomes a bounded space.[1] In mathematical terms, the convex polyhedral cone **P** is transformed into a bounded convex polyhedral cone (Figure 2.3).

Equations (6-8) represent most common non-adjustable constraints. These constraint define a space wherein every feasible flux vector always lives. They form a *constraint-based model* that describes in mathematical terms the capabilities of the metabolism under study. Other common non-adjustable thermodynamic constraint (Henry, 2006; Kümmel, 2006; Feist, 2007; Hoppe, 2007; Soh, 2010).

[1] In fact, capacity constraints for a subset of fluxes may be sufficient to get a bounded flux space.

Adding adjustable constraints

Adjustable constraints can also be incorporated to further restrict the space of feasible flux state or even to predict the actual fluxes. For example, regulatory constraints have been successfully imposed using Boolean logic operators (Covert, 2001; 2003), correlated reactions (Schilling, 2002), and control-effective fluxes (Stelling, 2002). There are also many methodologies that incorporate experimentally measured flux values as adjustable constraints. Details will be given below and in subsequent chapters.

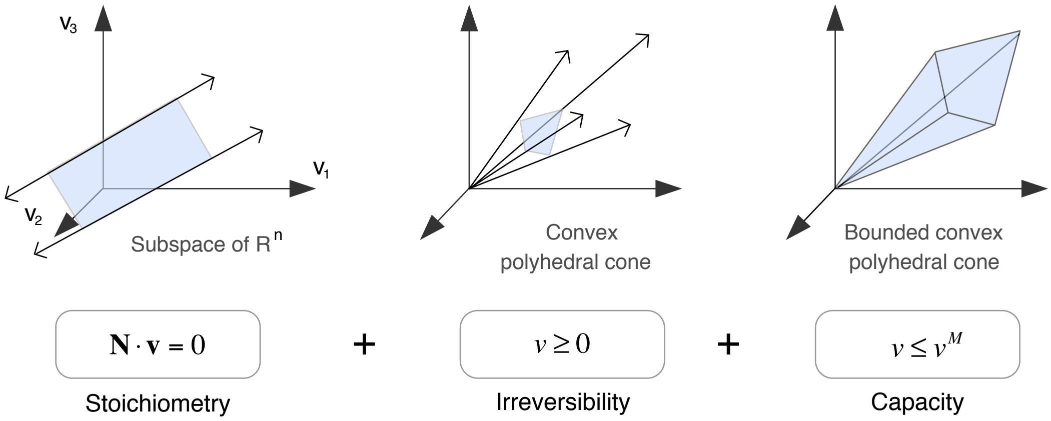


Figure 2.3. Space of feasible steady-state flux vectors by non-adjustable constraints.

2.5 Classification of constraint-based methodologies

There are several methods and techniques that exploit a constraint-based model. There is, however, a wide range of methodologies, they have particular purposes (e.g, analyse redundancy), employ a different mathematical frameworks (e.g, linear algebra), and are supported by particular assumptions (e.g, optimal behaviour).

A simple way to classify these methodologies is dividing them in two categories: those focused on analysing the entire flux space, and those that look for particular flux states within this space (see Figure 2.4).

Methodologies to systemic analysis

There are several approaches to study the modelled metabolism by means of the analysis of the flux space defined with (6-8). The objective of this approaches is elucidating systemic and emergent properties of the organism under investigation, those which do not derive from the elements that constitute the metabolic network, but which emerge from the interactions between those elements.

Pathway analysis with linear algebra

The equation (6) defines a homogeneous linear system of equalities, and therefore it can be analyzed using tools from linear algebra. For instance, the space of the solutions of (6) is defined by the null space (or kernel) of **N**. This is the space of stoichiometrically feasible (steady-state) flux vectors **v**.

The null space can be described by a $n \times (n\text{-}m)$ matrix,

$$\mathbf{K}(\mathbf{N}) \tag{9}$$

The columns of **K** are linear independent vectors that span the null space. These vectors form a basis of the space (6), and the dimension of **K** represents the degrees of freedom of this space.

Since the kernel is a basis of (6), every solution **v** in (6) can be expressed as a linear combination of these column vectors:

$$\mathbf{v} = \mathbf{K} \cdot \lambda \tag{10}$$

Note that if **K** exists, there are infinitely representations of **K**, because its columns can be linearly combined with each other.

Different analytical tools based on the null space **K** have been successfully applied in recent years. For instance, biochemically meaningful basis vectors have been used to get insight into pathway structures in a metabolic network (Schilling 1999). The null space has been also useful in the context of Metabolic Control Analysis (Reder, 1988; Heinrich, 1996).

However, the use of linear algebra to analyze the underlying metabolic networks has two main limitations: (i) inequalities cannot be used to represent well-known constraints, such as reactions irreversibility, and (ii) the obtained basis are not unique, and therefore they are not an invariant property. Ideas and tools from convex analysis has been used to overcome these limitations.

Pathway analysis with convex analysis

Convex analysis enables the analysis of linear systems of inequalities, thus making it possible to consider the irreversibility of the reactions, as given in equation (7). Using convex analysis, different concepts of network-based pathways have been proposed, such as elementary modes and extreme pathways (Papin, 2003; Papin, 2004). These pathways characterise, to some extent, the flux space defined in (7), and are being used to elucidate systemic properties, such as pathway length, network redundancy, enzyme subsets, or knockouts. These tools will be described in the next sections. Chapter III is devoted to compare some of them.

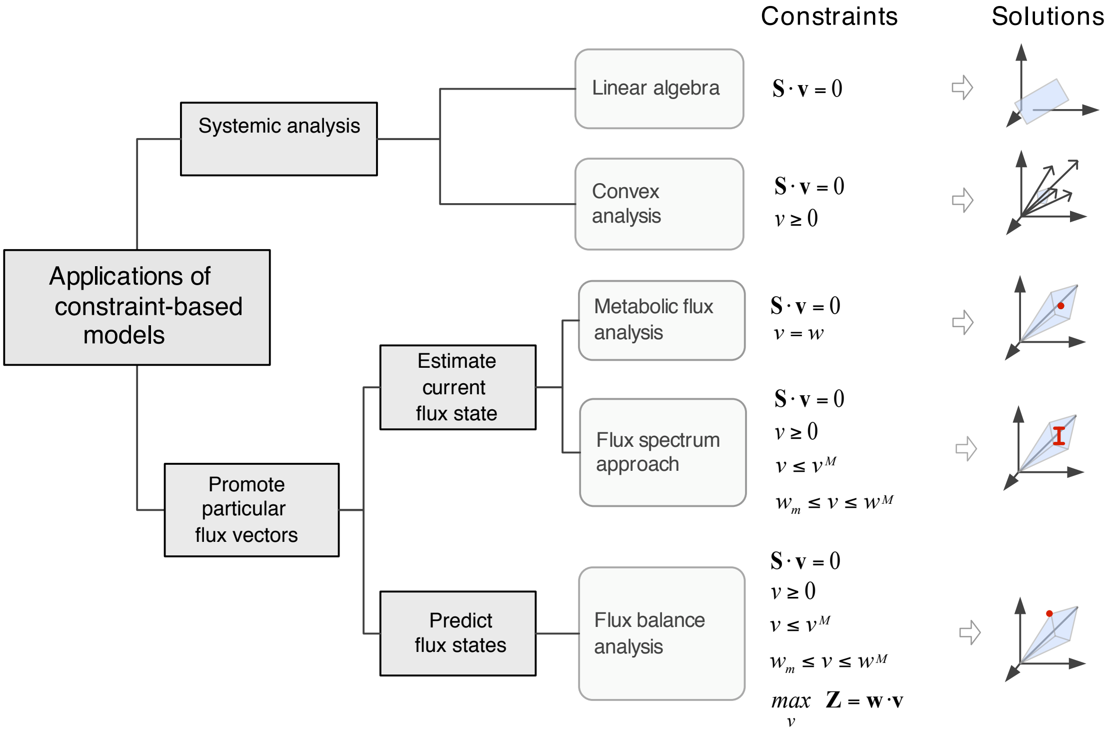

Constraints
Solutions
Applications of constraint-based models
Systemic analysis
Linear algebra
$\mathbf{S}\cdot\mathbf{v}=0$
Convex analysis
$\mathbf{S}\cdot\mathbf{v}=0$
$v\geq 0$
Promote particular flux vectors
Estimate current flux state
Metabolic flux analysis
$\mathbf{S}\cdot\mathbf{v}=0$
$v=w$
Flux spectrum approach
$\mathbf{S}\cdot\mathbf{v}=0$
$v\geq 0$
$v\leq v^M$
$w_m\leq v\leq w^M$
Predict flux states
Flux balance analysis
$\mathbf{S}\cdot\mathbf{v}=0$
$v\geq 0$
$v\leq v^M$
$w_m\leq v\leq w^M$
$\max_{v}\ \mathbf{Z}=\mathbf{w}\cdot\mathbf{v}$

Figure 2.4. Scheme of different methodologies that employ constraint-based models.

Methodologies to promote particular flux vectors

Several methodologies offset the under-determinacy of constraint-based models to promote particular flux vectors or metabolic states. This is achieved by adding adjustable constraints and making assumptions. This approach is mainly used (i) to estimate the flux state at given conditions, or (ii) build models capable of predicting the flux state that cells will exhibit at certain conditions.

Estimate the current flux state

A basic constraint-based model (6) can be coupled with *in vivo* experimental measurements of some fluxes to determine the complete flux state at the conditions where measurements were obtained. This is the approach used by metabolic flux analysis (Heijden, 1994; Stephanopoulos, 1998). Metabolic flux analysis has been extensively applied in recent years, and has been particularly successful in the fields of microbial production and animal cell culture.

Predict fluxes at given conditions

A predictive model is a mathematical representation of a system that predicts the outputs of the system given its inputs. In our context, certain constraints can be seen as inputs, such as substrates availabilities, and the flux vector as output. Since we do not know all the operating constraints, the imposition of the input constraints do not results in one unique prediction of the flux values; instead, a space of feasible steady state flux vectors is obtained. To determine which of these flux vectors is the actual one, further assumptions are needed. For instance, flux balance analysis gives point-wise predictions assuming that cells have evolved to be optimal respect to a (known) objective (Kauffman, 2003; Price, 2003).

In following sections three outstanding methodologies using constraint-based models will be presented: *Elementary modes analysis* (to discover pathways in a systematic way), *Metabolic flux analysis* (to estimate the fluxes exploiting the available measurements), and *Flux balance analysis* (to prediction fluxes assuming optimality).

2.6 Metabolic pathways analysis: identifying pathways

The purpose of network-based pathways analysis is twofold: first, identify a finite set of systemic pathways in a metabolic network; second, use these pathways to elucidate systemic properties and the capabilities of cell metabolism.

Generate the flux space

If we consider stoichiometry and reactions reversibility (7), the space of feasible flux vectors, or flux space **P**, is a convex polyhedral cone. Interestingly, convex analysis shows that any convex polyhedral cone can be generated by non-negative combination of a set of generating vectors $\mathbf{g_k}$ (Rockafellar, 1970):

$$\mathbf{P} = \left\{ \ \mathbf{v} \in \mathbf{R}^{\mathrm{n}} : \mathbf{v} = \sum_k w_k \cdot \mathbf{g_k} = \mathbf{G} \cdot \mathbf{w}, \quad w_k \geq 0 \ \right\} \tag{11}$$

Every feasible flux vector **v** in **P** can be represented as a non-negative combination of flux through these vectors **g**, which can be seen as pathways. In other words, all the flux states of a given metabolic network can be represented as an aggregation of fluxes through certain systemic pathways.

At least four related concepts of pathways fulfilling (11) have been proposed: extreme currents, elementary modes, extreme pathways and minimal generators. If one is only interested in generating the flux space fulfilling (11), the more reasonably choice is a minimal generating set, a smallest set of vectors holding the property (Urbanczik, 2005). Unfortunately this set is not unique in the general case. On the other hand, the set of elementary modes has another property that makes them a more powerful tool for the analysis of the underlying metabolism.

Elementary modes

The elementary modes are defined as the set of all the non-decomposable vectors in **P**. That is, all those vectors **e** in **P** that cannot be decomposed as a positive combination of two *simpler* [1] vectors in **P** (Schuster, 1999). This definition implies that:

(i) The elementary modes generate the flux space **P**, as in (11). Indeed, in chapter III we will show that they fulfil a more restrictive condition.

(ii) The set of elementary modes is unique. They are a systemic (and time-invariant) property of the given metabolic network.

(iii) Each **e** is non-decomposable. Each **e** represents a (stoichiometrically and thermodynamically) feasible route to the conversion of substrates into products and cannot be decomposed into simpler routes.

(iv) The elementary modes are all the routes consistent with property (iii).

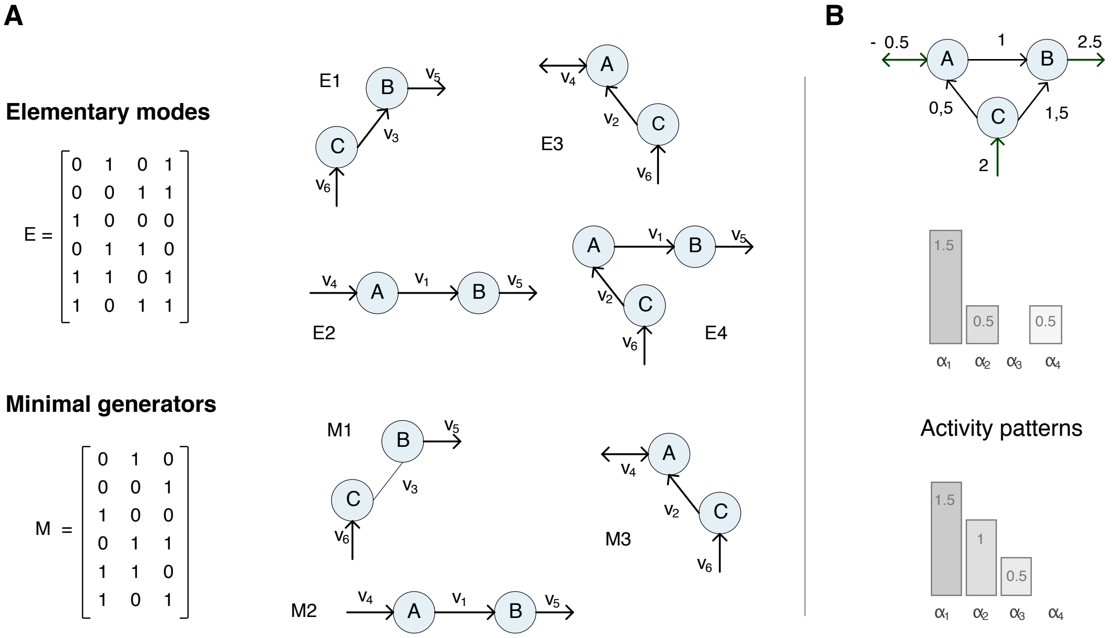


Figure 2.5. Example network-based pathways analysis. (A) The matrix and the pathways represent the 4 elementary modes and 3 minimal generators of the network depicted in Figure 2.2. Notice that E4 is not necessary to generate the flux space (so it is not a minimal generator). (B) Examples of possible translation of a given flux state into patters of pathway activities.

[1] Simpler vectors are vectors containing zero elements wherever vector **e** does, and include at least one additional zero component.

The fact that the set of elementary modes (EMS) comprises all the simple pathways in the network—its functional states—makes it possible to investigate the infinite behaviours that cells can show by simply inspecting them. This makes it easy to answer several questions: which reactions are essential to produce a certain compound, which will be the capabilities of the network if a reaction is knocked-out, etc. Answering these questions using the minimal generators or the extreme pathways may be difficult because one has to take into account the possible cancelations of reversible fluxes.

More details about pathway analysis will be given in chapter III, where four concepts of pathways are described and compared. The translation of a flux vector into a pattern of elementary modes activities will be addressed in chapter V.

Applications of network-based pathways analysis

Several applications of elementary modes and the closely related extreme pathways have been reported in the literature. Most of these applications are found in the context of microbial production, for the study of the metabolisms of *E. coli* (Schmidt, 1999), *Haemophilus influenzae* (Schilling, 2000; Papin, 2002), *Helicobacter pylori* (Price, 2002), and *Saccharomyces cerevisae* (Schwartz, 2006). However, elementary modes have also been used in botany (Poolman, 2003; Steuer, 2007) and in medicine (Zhong, 2002; Nolan, 2006).

2.7 Metabolic flux analysis: estimating fluxes

Generally speaking, metabolic flux analysis (MFA) combines a set measured fluxes (often extracellular ones) with a constraint-based model to get an estimate of all the fluxes. This results in a metabolic flux vector **v** that represents the steady state at which each reaction in the network occurs (Figure 2.6). This pattern of flux informs about the contribution of each reaction to the overall metabolic processes of substrate utilisation and product formation.

Consider a metabolic network with *m* internal metabolites and *n* reactions. Assuming that metabolites are at steady-state, mass balances can be formulated as follows:

$$\mathbf{N} \cdot \mathbf{v} = \mathbf{0} \tag{12}$$

Now, we consider that some fluxes in **v** have been measured, $\mathbf{v} = (\mathbf{v_u}\ \mathbf{v_m})$, keeping in mind that measurements are imprecise in practice, they can be represented as follows:

$$\mathbf{w_m} = \mathbf{v_m} + \mathbf{e_m} \tag{13}$$

where $\mathbf{e_m}$ represents measurement errors and $\mathbf{w_m}$ the measured values.

Hence, Traditional metabolic flux analysis (Heijden, 1994) can be defined as the exercise of determining the complete flux vector **v** that satisfies the balance equation (12) and is compatible with the measurements (13).

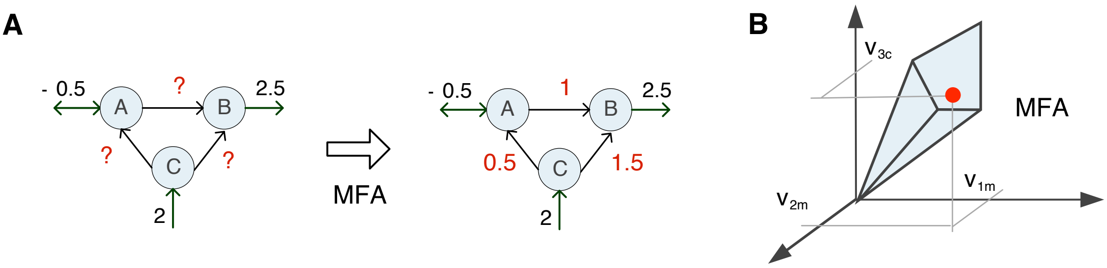


Figure 2.6. Traditional metabolic flux analysis. (A) Measured fluxes are coupled with the stoichiometric constraints to determine the remaining fluxes. (B) The under-determinacy of the flux space is offset by incorporating measurements (subindexes *m* denote measured fluxes, *c* determined ones).

Traditional MFA: problem determinacy and redundancy

If we define a dim{$\mathbf{v_m}$}×n selection matrix **Q** having exactly one “1” in each row and all other elements equal to zero, the system (12-13) can be rewritten as follows:

$$\begin{pmatrix} \mathbf{N} \\ \mathbf{Q} \end{pmatrix} \cdot \mathbf{v} + \begin{pmatrix} \mathbf{0} \\ \mathbf{e_m} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{w_m} \end{pmatrix} \tag{14a}$$

In practice $\mathbf{v_m}$ (and $\mathbf{e_m}$) is unknown due to noise, so one has to deal with the system:

$$\begin{pmatrix} \mathbf{N} \\ \mathbf{Q} \end{pmatrix} \cdot \mathbf{v} = \begin{pmatrix} \mathbf{0} \\ \mathbf{w_m} \end{pmatrix} \tag{14b}$$

With a classical classification of linear systems of equations, system (14) could be:

- *Underdetermined*. If less than *n-m* independent fluxes are measured, system (12) has infinite solutions. At least one flux, but probably most of them, cannot be determined.
- *Determined*. If exactly *n-m* independent fluxes are measured, system (14) has a unique solution. In this case, all fluxes can be uniquely determined.
- *Overdetermined*. If more than *n-m* independent fluxes are measured, system (14) probably has no solution—there are redundant measurements which are inconsistent.

This classification disregards the fact that (14) can be simultaneously underdetermined and redundant. A better classification was given by Klamt (2002).

Consider a partition in (12) between measured (*m*) and unknown fluxes (*u*):

$$\mathbf{N_u}\cdot\mathbf{v_u} = -\mathbf{N_m}\cdot\mathbf{w_m} \tag{15}$$

Then, determinacy and redundancy of the MFA problem can be defined as follows.

*System Determinacy and Calculability of Fluxe*s. System (15) is determined if rank($\mathbf{N_u}$) < *u* (*u* is the number of non-measured fluxes), i.e., if there are enough linearly independent constraints to uniquely calculate all non-measured fluxes $\mathbf{v_u}$. If the system is underdetermined, at least one flux in $\mathbf{v_u}$, and probably most of them, are non calculable.

System Redundancy and Consistency of Measurements. System (15) is redundant if rank($\mathbf{N_u}$) < *m,* if some rows in $\mathbf{N_u}$ can be expressed as linear combinations of other rows. This can lead to an inconsistent system if the vector $\mathbf{w_m}$ contains such values that no $\mathbf{v_u}$ exists that exactly solves (15). Redundancies can be exploited to analyse measurements consistency and adjust the measured values of the so-called balanceable fluxes.

Traditional MFA: calculation procedure

Traditional metabolic flux analysis (TMFA) is often performed with a two-step procedure (Heijden, 1994). First, consistency is analysed with a χ^2-test to ensure that measurements are free of gross errors (details below). Then, a weighted least squares problem is solved to get an estimate of $\mathbf{v}$:

$$\mathbf{v}^{\mathrm{mfa}} = \left(\mathbf{A}^{\mathrm{T}}\cdot\mathbf{F}^{-1}\cdot\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}\cdot\mathbf{F}^{-1}\cdot\mathbf{r}, \qquad \mathbf{A} = \begin{pmatrix} \mathbf{N} \\ \mathbf{Q} \end{pmatrix}, \mathbf{r} = \begin{pmatrix} \mathbf{0} \\ \mathbf{w_m} \end{pmatrix} \tag{16}$$

where it is assumed that errors $\mathbf{e_m}$ are distributed normally with a mean value of zero and a variance-covariance matrix $\mathbf{F}$.

Notice, however, that an equivalent (weighted) least squares problem can be formulated as a quadratic optimisation subject to linear constraints:

$$\mathbf{v}^{\mathrm{mfa}} = \min_{\mathbf{v}} \; \mathbf{e}_{\mathbf{m}}^{\mathrm{T}}\cdot\mathbf{F}^{-1}\cdot\mathbf{e_m} \quad \text{s.t.} \begin{cases} \mathbf{N}\cdot\mathbf{v} = \mathbf{0} \\ \mathbf{w_m} = \mathbf{v_m} + \mathbf{e_m} \end{cases} \tag{17}$$

Notice that, ideally, TMFA should be performed only when the system is determined and redundant. If it is not redundant, measurements consistency cannot be evaluated and the point-wise estimate given by (16) or (17) will be unreliable. If the system is underdetermined, the point-wise estimate given by (17) will be only one of multiple (infinite) possible values.

Evaluation of measurements consistency

Before applying TMFA, redundant measurements can be used to evaluate the consistency between measurements and model (Stephanopoulos, 1998). A redundant system will be consistent if it fulfils the consistency condition:

$$\mathbf{R} \cdot \mathbf{w_m} = 0, \quad \mathbf{R} = \mathbf{N_m} - \mathbf{N_u} \cdot \mathbf{N_u^{\#}} \cdot \mathbf{N_m} \qquad (18)$$

where **R** is the redundancy matrix and the operator ($^{\#}$) denotes the More-Penrose pseudo-inverse.

If inconsistency is detected, a χ^2-test can be used to evaluate its importance. The consistency analysis is based upon statistical hypothesis testing to determine if redundancies are satisfied to within expected experimental error.

The test is performed calculating a consistency index *h* as follows:

$$\begin{aligned} &\mathrm{h} = \varepsilon^{\mathrm{T}} \cdot \mathbf{W}^{\mathbf{-1}} \cdot \varepsilon \\ &\varepsilon = -\mathbf{R_r} \cdot \mathbf{v_m} \\ &\mathbf{W} = \mathbf{R_r} \cdot \mathbf{F_r} \cdot \mathbf{R_r^T} \end{aligned} \qquad (19)$$

where $\mathbf{R_r}$ is the reduced redundancy matrix (obtained by removing dependent rows in **R**) and $\mathbf{F_r}$ is the variance-covariance matrix of the measurements $\mathbf{w_m}$.

If a given $\mathbf{w_m}$ fails the consistency check ($h > \chi^2$), there is a (confidence level)% chance that either $\mathbf{w_m}$ contains gross errors or the stoichiometric matrix is incorrect.

Traditional MFA: calculation procedure by Klamt

It must be noticed that only some measured fluxes, the so-called balanceable, have an impact on the consistency analysis. These can be detected by inspection of **R**: exactly those $w_{m,j}$ for which the corresponding *j*-th column of **R** contains at least one nonzero value are balanceable (Klamt, 2002).

Interestingly, these balanceable measured fluxes can be adjusted (or balanced) if they are inconsistent. The adjusted values can be calculated as follows: [1]

[1] Expression (20) returns the original values of the non-balanceable fluxes and adjusted ones for the balanceable fluxes.

$$\mathbf{v}_{\mathbf{m}}^{\mathbf{mfa}} = \left(\mathbf{I} - \mathbf{F}_{\mathbf{r}} \cdot \mathbf{R}_{\mathbf{r}}^{\mathbf{T}} \cdot \mathbf{W}^{\mathbf{-1}} \cdot \mathbf{R}_{\mathbf{r}}^{\#}\right) \cdot \mathbf{w}_{\mathbf{m}} \tag{20}$$

After adjusting the measured fluxes, the non-measured fluxes can be calculated with equation (15). If the problem is determined and not redundant the unique $\mathbf{v_u}$ fulfilling (15) can be calculated using the inverse of $\mathbf{N_u}$. However, to get a solution in case (15) is determined and redundant, the pseudo-inverse should be used instead:

$$\mathbf{v}_{\mathbf{u}}^{\mathbf{mfa}} = -\mathbf{N}_{\mathbf{u}}^{\#} \cdot \mathbf{N}_{\mathbf{m}} \cdot \mathbf{v}_{\mathbf{m}}^{\mathbf{mfa}} \tag{21}$$

If the system is underdetermined, at least one non-measured flux, and probably most of them, are non uniquely determined and should not be calculated with (21). However, even in this case some fluxes may be calculable (Klamt, 2002).

Consider the general solution of system (15):

$$\mathbf{v}_{\mathbf{u}} = \mathbf{v}_{\mathbf{u}}^{\mathbf{p}} + \mathbf{K}_{\mathbf{u}} \cdot \lambda \tag{22}$$

where $\mathbf{v^p}$ is a particular solution, for instance, the one given by (20-21), λ is an arbitrary vector with *n-m* elements, and $\mathbf{K_u}$ is the kernel of $\mathbf{N_u}$. The product $\mathbf{K_u} \cdot \lambda$ spans the space of possible flux values and represents the underdeterminacy of the system.

We can define as calculable fluxes the elements $v_{u,j}$ in $\mathbf{v_u}$ which corresponding row in $\mathbf{K_u}$ is a null row. These elements are uniquely determined independently of λ; any particular solution $\mathbf{v^p}$ will assign the same value for those v_u classified as calculable. Therefore, their value can be taken from any solution, e.g., the one given by (10).

In summary, the procedure to apply TMFA proposed by Klamt et al. (2002), with slight changes, can be structured as follows:

Step 1 Balance the measured fluxes

1.1 Check if the system is redundant: rank($\mathbf{N_u}$) < *m*

1.2 If the system is redundant

- Evaluate consistency to detect gross errors with (19)
- Detect and adjust the balanceable fluxes with (20)

(thus obtaining a consistent set of measurements)

```
Step 2  Determine the non-measured fluxes
          2.1  Check if the system is determined: rank($\mathbf{N_u}$) = $u$
          2.2  It the system is determined, all fluxes are calculable
               If it is underdetermined, find calculable fluxes with (22)
          2.3  Get values of the calculable fluxes from (21)
```

More details about metabolic flux analysis will be given in chapters IV and VII. There, we propose alternative approaches to traditional MFA and illustrate their benefits with different cases of study.

Applications of metabolic flux analysis

Metabolic flux analysis has been widely used to characterise canonical states of cells, such as exponential batch growth or steady states in the continuous mode. In particular, animal cell cultures have received considerable attention (Bonarius, 1996; Follstad, 1999; Nyberg, 1999; Gambhir, 2003). Recently, there has been an increasing interest on the application of metabolic flux analysis to plant cell culture (Schwender, 2004; Ratcliffe, 2006). It has also been applied to study transient processes in microorganisms (Herwig, 2002), to on-line monitoring intracellular fluxes in mammalian cells (Henry, 2007), to determine the physiological state of a culture (Takiguchi, 1997), and to develop dynamic models (Teixeira, 2007). There are some medical applications of metabolic flux analysis, such as the generation of hypothesis for new therapeutical strategies (Calik, 2002), the optimisation of an extracorporeal bioartificial liver device (Sharma, 2005), and the investigation of metabolic responses of the rat liver to burn-injury-induced whole-body inflammation (Nolan, 2006). Metabolic flux analysis has been also combined with isotopic labelling experiments, allowing for a more reliable estimation of fluxes (Wiechert, 2001; Schmidt, 1999; Shirai, 2006).

2.8 Flux balance analysis: predicting fluxes

Flux balance analysis (FBA) is a methodology that uses optimisation to get predictions from a constraint-based model by invoking an assumption of optimal cell behaviour (Savinell, 1992; Varma, 1994; Edwards, 2002; Price, 2003; Palsson, 2006). Basically, one particular state among those that cells can show, accordingly to a constraint-based model, is chosen based on the assumption that cells have evolved to be optimal, i.e., that cells regulate its fluxes toward optimal flux states.

The procedure to build an FBA model can be summarised as follows (Figure 2.7):

Step 1 Define the flux space with (6-8). These constraints are the invariant structure of the model, and represent the capabilities of the cells.

$$\mathbf{N}\cdot\mathbf{v}=0 \;\text{ and }\; \mathbf{D}\cdot\mathbf{v}\geq 0$$

Step 2 Incorporate "input" constraints (adjustable ones), usually on a few uptake fluxes, based on capacities or availability of substrates.

$$\text{e.g., } \mathbf{v}_{\mathbf{u}}^{\mathbf{m}} \geq \mathbf{v}_{\mathbf{u}} \geq \mathbf{v}_{\mathbf{u}}^{\mathbf{M}}$$

Step 3 Assume that cells have evolved to achieve an optimal behaviour owing to evolutionary pressure. Then invoke an optimal use of resources (e.g., maximum growth), expressed by means of a (linear) cost index Z:

$$Z=\mathbf{d}\cdot\mathbf{v}$$

Step 4 Solve the formulated (linear) programming problem to obtain the flux vector that makes the best use of its resources to satisfy the stated objective function.

$$\mathbf{v}^{\mathbf{opt}} = \max_{v} Z \quad \text{s.t.} \left\{ \mathbf{N}\cdot\mathbf{v}=0 \quad \mathbf{D}\cdot\mathbf{v}\geq 0 \quad \mathbf{v}_{\mathbf{u}}^{\mathrm{m}} \geq \mathbf{v}_{\mathbf{u}} \geq \mathbf{v}_{\mathbf{u}}^{\mathrm{M}} \right\}$$

Notice that the aim of flux balance analysis is not to determine the flux vector that corresponds to a set of measurements (as in MFA), but to construct a model able to predict the phenotype that cells will show at certain conditions (those defined by the input constraints). Indeed, in most cases input constraints do not correspond to real measurements. Flux balance analysis is used to investigate hypothesis (e.g., test if a reduced uptake capacity can be the cause of an unexpected cell behaviour) and to evaluate a range of possibilities (e.g, find the best combination of substrates).

Metabolic objectives and optimization

It must be taken into account that FBA predictions, the optimal flux state, may not correspond to the actual fluxes exhibit by cells. To support the assumption of optimal behaviour, it must be hypothesised that: (i) cells, forced by evolutionary pressure, evolved to achieve an optimal behaviour with respect to certain objective, (ii) we know

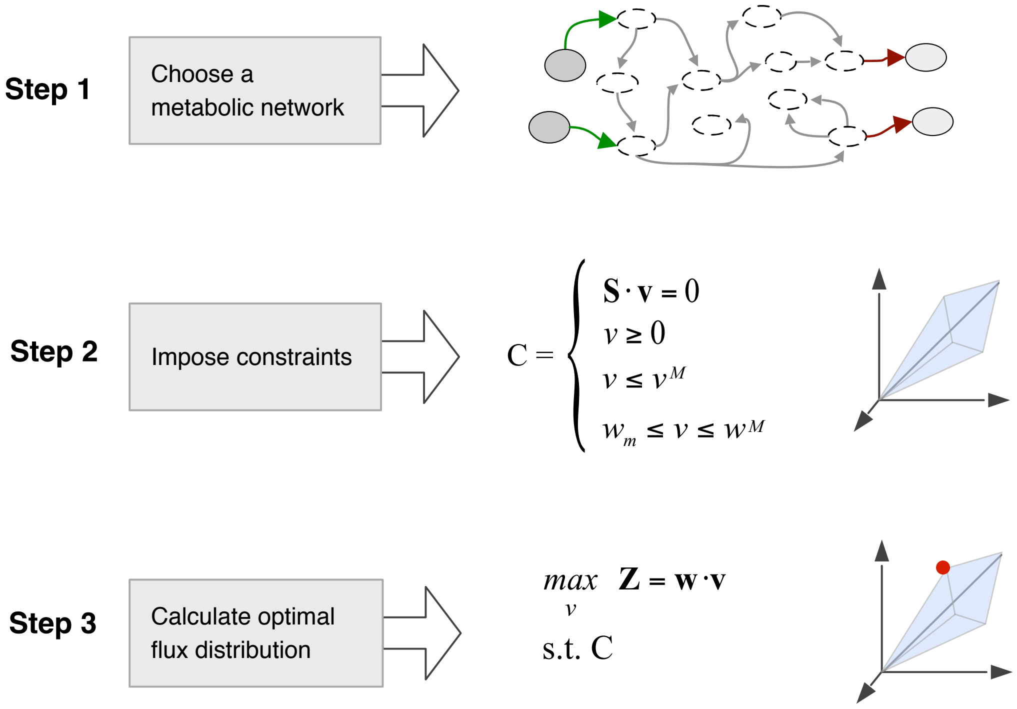

Step 1
Choose a metabolic network
Step 2
Impose constraints
$C = \begin{cases} \mathbf{S} \cdot \mathbf{v} = 0 \\ v \geq 0 \\ v \leq v^M \\ w_m \leq v \leq w^M \end{cases}$
Step 3
Calculate optimal flux distribution
$\max_{v} \ \mathbf{Z} = \mathbf{w} \cdot \mathbf{v}$
s.t. C

Figure 2.7. Procedure to develop a flux balance analysis model.

which this objective is, and (iii) the objective can be expressed, at least approximately, in convenient mathematical terms.

Clearly, predictions of flux balance analysis are dependent on the objective function being used. To date, the most commonly used objective function has been the maximisation of biomass, which leaded to predictions consistent with experimental data for different organisms, such as *Escherichia coli* (Varma, 1994b; Edwards, 2001) or *Helicobacter pylori* (Schilling, 2002). Other objective functions have been used, such as minimising ATP production, minimising nutrient uptake, or maximising metabolite production. Although linear functions are preferred (to keep the problem linear) non-linear functions have been also used. For example, a quadratic function is used in (Segre, 2002), and the authors suggest that genetically engineered knockout may undergo a minimal redistribution with respect to the flux configuration of the wild type cell. Remarkably, Schuetz et al. have shown the capacity of FBA to predict intracellular fluxes using different objective function (2007), but they pointed out that this requires to identify which are the relevant objectives of cells at different environmental conditions (Schuster, 2008; Schuetz, 2007).

Flux balance analysis will be also discussed in chapter VIII, where we introduce a possibilistic approach to FBA that provides rich predictions, accounts for sub-optimality, and considers (quasi) alternative optima solutions.

Applications of flux balance analysis

E. coli has been the most well studied microorganism due to the considerable amount of available data (Edwards, 2001; Reed, 2003; Feist, 2007). Other FBA models have been developed, for instance, for *Haemophilus influenzae* (Edwards, 1999), *H. pylori* (Schilling, 2002), *Saccharomyces cerevisae* (Forster, 2003), *Methanosarcina barkeri* (Feist, 2006) and *Synechocystis* (Montagud, 2010). As a result of these efforts, many applications of flux balance analysis have been investigated, some of which are summarised in Table 2.2 (Palsson, 2006; Edwards, 2002; Price, 2003). In recent years, the first medical applications of flux balance analysis have been carried out. Thiele et al. (2005) used FBA to investigate the metabolic network of human mitochondria and to evaluate the effect of potential disease treatments. FBA has been also applied to optimise the metabolism of cultured hepatocytes used in bioartificial liver devices (Sharma, 2005; Nolan, 2006). And there is a reasonable interest on the application of flux balance analysis of whole-plant models (Lange, 2006; Zhong, 2002).

Table 2.2. Applications and methods based on flux balance analysis.

Determine network properties	Yield of key cofactors and biosynthetic precursors
Redundancy studies	Detect alternate equivalent optima
	Analyze the flux variability of a given optimal
	Study the sensitivity of the optimal properties
Interpret experimental data	Analyze robustness against environmental perturbations
	Qualitatively classify metabolic state based on observations
Objective studies	predict optimal growth rates
	Elucidate which are the cell objectives (objective functions)
Simulated modifications	Study gene deletions/additions
	Predict the behavior of knockouts (MOMA)
	Gradual inhibition or enhancement of gene function
Accounting for regulation	Increase predictive power
Potential applications	Identify and prioritize candidate drug targets
	Direct strategies to engineer strains
	Evaluate the state of knowledge about the metabolism
	Design experimental programs
	Analyze enzyme deficiencies
	Evaluate genome annotations

2.9 Conclusions

Along this chapter classical stoichiometric models and the more powerful constraint-based models have been presented. We have also reviewed different methodologies that make use of these models. Particular attention has been paid to three outstanding methodologies that are the context for the contributions of this thesis: metabolic flux analysis, flux balance analysis, and metabolic pathways.

Metabolic flux analysis (MFA) uses experimentally measured data to estimate the metabolic state of cells at given conditions. It has been commonly used to study the exponential growth phase and steady states in continuous fermentation processes. There is also interest in the use of MFA for monitoring time-varying fluxes, particularly in industrial environments (Herwig, 2002; Henry, 2007; Takiguchi, 1997).

In chapter IV and VII, we introduce interval and possibilistic methods to perform MFA. These methods provide richer estimates, consider measurements uncertainty, and cope with scenarios of data scarcity. The estimation of fluxes over time will be discussed in chapters V and VIII.

Flux balance analysis (FBA) is a methodology to get predictions from a constraint-based model, so far, the only one applied in the genome-scale (Price, 2003). It is also of utility with simpler networks (Schuetz, 2007).[1]

A possibilistic approach to FBA will be discussed in chapter VIII. This approach provides rich predictions, accounts for sub-optimality, and considers (quasi) alternative optima.

Network-based pathways, such as elementary modes or extreme pathways, are tools to elucidate systemic properties and capabilities of cells. Despite being recent proposals, they have been used to improve our understanding of biological processes, guide metabolic engineering, and aid in the development of reduced models.

In chapter III, three definitions of network-based pathways will be compared. In chapter V, the translation of a flux state into a pattern of pathway activities will be addressed. Elementary modes will be also of use in the procedure to validate constraint-based models described in chapter IX.

In summary, constraint-based modelling is now a very active field—its applications increases steadily—and it is expected that this situation will continue. The results show that there is much valuable information that can be extracted for the reconstructed networks, even if intracellular kinetics are still unknown. Moreover, there is key advantage in the fact that the paradigm is scalable: new and better knowledge and data can be incorporated as additional constraints, thus improving the models in an iterative way.

[1] Indeed, the cybernetic approach, which is based on a similar assumption of optimal behaviour, has been used for the dynamic modelling of cells using very simple networks (Ramakrishna, 1996).

Main references

- Llaneras F. and Picó J. (2008) Stoichiometric modelling of cell metabolism *J Journal of Bioscience and Bioengineering*, 105 (1), 1-11.

- Bailey JE (2001). Complex biology with no parameters. *Nature biotechnology*, 19:503-504.

- Palsson BO (2006). *Systems biology: properties of reconstructed networks*. New York, USA: Cambridge University Press New York.

- Stephanopoulos GN, Aristidou AA (1998). *Metabolic Engineering: Principles and Methodologies*. San Diego, USA: Academic Press.

- Klamt S, Schuster S, Gilles ED (2002). Calculability analysis in underdetermined metabolic networks illustrated by a model of the central metabolism in purple non-sulfur bacteria. *Biotechnology & Bioengineering*, 77:734-751.

- Papin JA, Stelling J, Price ND, Klamt S, Schuster S, Palsson BO (2004). Comparison of network-based pathway analysis methods. *Trends in Biotechnology*, 22(8):400-405.

- Price ND, Papin JA, Schilling CH, Palsson BO (2003). Genome-scale microbial in silico models: the constraints-based approach. *Trends in Biotechnology*, 21(4):162-9.

- Edwards JS, Covert M, Palsson B (2002). Metabolic modelling of microbes: the flux-balance approach. *Environmental Microbiology*, 4:133-140.

III

Network-based metabolic pathways: a comparison

There is a great interest in systematically identifying the relevant pathways in a metabolic network, but, unsurprisingly, there is not a unique set of pathways to be tagged as relevant. At least four related concepts have been proposed: extreme currents, elementary modes, extreme pathways and minimal generators.

In this chapter, we will describe and compare these concepts. Basically, there are two properties that these sets of pathways can hold: they can generate the flux space—if every feasible flux vector can be represented as a non-negative combination of pathways activities—or they can comprise all the non-decomposable pathways in the network. The four concepts fulfil the first property, but only the elementary modes fulfil the second one. This subtle difference has been a source of errors and misunderstandings. This chapter attempts to clarify the intricate relationship among the different pathways by comparing them.

Part of the contents of this chapter appeared in the following journal article:

- Llaneras F, Picó J (2010). Which metabolic pathways generate and characterise the flux space? A comparison among elementary modes, extreme pathways and minimal generators. *Journal of Bioscience and Bioengineering*, vol. 2010.

3.1 Introduction

Recalling our standard notation, a metabolic network can be represented by a stoichiometric matrix **N**, where rows correspond to the *m* metabolites and columns to the *n* reactions. If one assumes that intracellular metabolites are at steady state, material balances can be formulated as follows (Stephanopoulos, 1998):

$$\mathbf{N}\cdot\mathbf{v}=0 \tag{1}$$

where $\mathbf{v} = (v_1, v_2, \ldots, v_n)^{\mathrm{T}}$ is the *n*-dimensional vector of flux through each reaction. Each feasible steady state is represented by a flux vector **v**.

Taking into account these mass balances and the irreversibility of certain reactions, the space of feasible steady state *flux vectors,* or *flux space*, can be defined as follows (see glossary at the end of the chapter for words in italics):

$$\mathbf{P}=\left\{\mathbf{v}\in\mathbf{R}^{n}:\begin{array}{c}\mathbf{N}\cdot\mathbf{v}=0\\ \mathbf{D}\cdot\mathbf{v}\geq 0\end{array}\right\} \tag{2}$$

where **D** is a diagonal *n*×*n*-matrix with $\mathbf{D_{ii}}$ = 1 if the flux *i* is irreversible (otherwise 0).

The flux space is the cornerstone of constraint-based modeling, as it was explained in chapter II. In this context, network-based pathways are used to investigate the modeled metabolism by the analysis of a finite set of relevant pathways, which ideally represent all the metabolic states that a cell can show. Some applications of this approach are enumerated in Table 3.1.

However, there is not a unique set of network-based pathways to be tagged as 'relevant' and different proposals have been applied with success: extreme currents, elementary modes, extreme pathways and minimal generators. These concepts are not equivalent, but closely related. There are three major properties that a set of network-based pathways may hold: (P1) they generate the flux space **P**, (P2) they are the minimal set of vectors fulfilling the first property, and (P3) they are all the non-decomposable pathways in the network. The fact that all the network-based pathways—elementary modes, extreme pathways, etc.—fulfil the first property but not the others has been a source of errors, imprecision and misunderstandings.

Along this chapter we discuss the relationship among the different network-based pathways from a theoretical point of view. We will start defining the four pathway concepts and then we will perform a comparison among them. Finally, we will present some examples and outline the major conclusions.

3.2 Different concepts of pathways

The first attempts to systemically extract a set of pathways from a given metabolic network were based on the assumption that all the fluxes were irreversible, or more precisely, that its dominant direction could be presumed. Convex algebra show that in this case the flux space **P** is a *pointed convex polyhedral cone* in the positive orthant R^n, which can be generated by non-negative combination of certain vectors, its edges or *extreme rays* (Rockafellar, 1970). See Figure 3.1 for a geometric illustration.

These extreme rays were flux vectors, or pathways, with a remarkable property (P1): the extreme rays generate the flux space **P**. That is, every flux vector **v** in **P** can be represented as a non-negative combination of fluxes through these pathways ($\mathbf{e_k}$ denotes the extreme rays):

$$\mathbf{P} = \left\{ \mathbf{v} \in \mathbf{R}^n : \mathbf{v} = \sum_k^e w_k \cdot \mathbf{e_k}, \quad w_k \geq 0 \right\} \tag{3}$$

An example illustrating this property was shown in chapter II, section 2.6. Notice also that in general a given **v** cannot be uniquely decomposed into an activity pattern **w**, but a space of valid solutions exists (Wiback, 2003).[1] This is also true for the rest of generating sets that will be introduced in subsequent sections.

Moreover, the set of extreme rays had two additional properties: (P2) it was the smallest (minimal) generating set of **P**, and (P3) the extreme rays were all the *non-decomposable* vectors in **P**, those that cannot be decomposed in simpler vectors (Gagneur, 2004). A *non-decomposable* vector is a minimal set of reactions that form a 'functional unit', if any of its participant reaction is not carrying flux, the others cannot

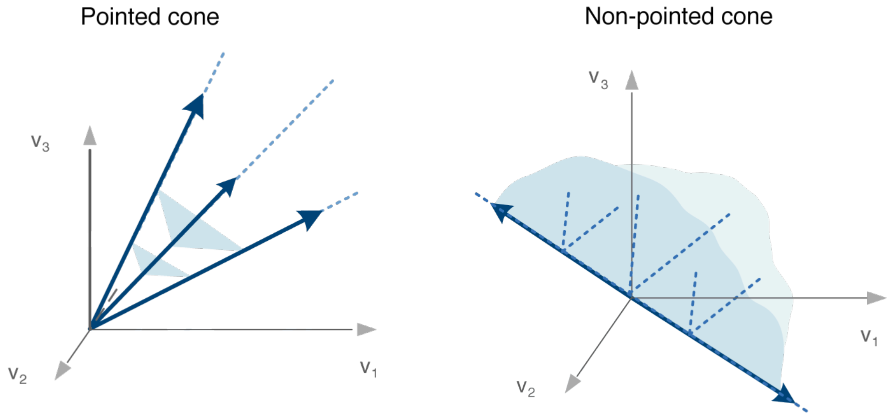


Figure 3.1. Extreme rays of two flux spaces.

[1] This issue will be discussed in more detail in chapter V.

operate alone. These functional units are the simplest steady state flux vectors that cells can show, and the rest of feasible states can be seen as the aggregated action of these units. This property makes it possible to investigate the infinite behaviors that cells can show by inspection of the finite set of non-decomposable vectors.

But what happens if not all fluxes can be assumed to be irreversible? If so, the extreme rays may lose these properties. Indeed, a set of vectors holding the three properties simultaneously (P1, P2 and P3) will not exist; there will be sets fulfilling P1 and P2, or P1 and P3, but not P2 and P3 in a general case.

Table 3.1. Applications of network-based pathways analysis.

Applications	References
Identification of pathways	(Schuster, 2000; Schuster, 1999)
Determination of minimal medium requirements	(Schilling, 2000)
Analysis of pathway redundancy and robustness	(Stelling, 2002; Price, 2002)
Linkage between structure and regulation…	
Correlated reactions (enzyme subsets)	(Papin, 2002; Pfeiffer, 1999)
Detect excluding reaction pairs	(Klamt, 2003)
Prediction of transcription ratios	(Stelling, 2002; Cakir, 2004)
Include regulatory rules	(Covert, 2003)
Support for metabolic engineering…	
Identification of pathways with optimal yields	(Schuster, 1999)
Evaluation of effect of addition/deletion of genes	(Carlson, 2002)
Inference of viability of mutants	(Stelling, 2002)
Detection of minimal cut sets	(Klamt, 2004)
Suggest operations to increase product yield	(Liao, 1996)
Translation of a flux vector into pathways activities…	
Particular solution methods	(Schwartz, 2006; Schwarz, 2005)
Alpha-spectrum	(Wiback, 2003; Llaneras, 2007)
Aid in the reconstruction of metabolic reaction networks…	
Assignment of function to orphan genes	(Forster, 2002)
Detection of infeasible circles	(Price, 2002; Beard, 2002)
Detection of network dead ends	(Schilling, 2002)
Support in the reconstruction of metabolic maps	(Cornish-Bowden, 2000)
Development of reduced, kinetic models	(Teixeira, 2007; Provost, 2004; 2006)

Extreme currents

Extreme currents are probably the first attempt to define a set of network-based pathways (Clarke, 1988). Their computation is based on splitting up each reversible reaction into two irreversible ones. If fluxes are reordered to separate the irreversible fluxes $\mathbf{v_I}$ and the reversible ones $\mathbf{v_R}$, the flux space (2) is augmented ($\mathbf{N} = [\mathbf{N_I}\ \mathbf{N_R}]$):

$$\mathbf{P_{rc}} = \left\{ \mathbf{v} \in \mathbf{R}^{n+r} : \begin{pmatrix} \mathbf{N_I} & \mathbf{N_R} & -\mathbf{N_R} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{v_I} \\ \mathbf{v}_R \\ \mathbf{v}'_R \end{pmatrix} = \mathbf{0} \ \text{ and } \begin{pmatrix} \mathbf{v_I} \\ \mathbf{v_R} \\ \mathbf{v'_R} \end{pmatrix} \geq \mathbf{0} \right\} \qquad (4)$$

The extreme rays of the cone $\mathbf{P_{rc}}$ are defined as the extreme currents of $\mathbf{P}$. Notice that $\mathbf{P_{rc}}$ is a pointed cone in the positive orthant R^{n+r}, so its extreme rays have all the properties mentioned above (P1-P3). However, $\mathbf{P_{rc}}$ lives in a higher-dimensional vector-space (augmented in one dimension for each split reversible reaction) and the extreme currents lose their properties when they are translated to the original vector-space.

In fact, it has been recently shown that the set of extreme currents (ECS) coincide with the set of elementary modes, which will be introduced below, when it is translated to the original vector-space (Wagner, 2005)—when computing the first a set of *r* spurious cycles appear (pathways formed by the forward and backward reaction of each reversible flux); however, these pathways are not considered meaningful (Schilling, 2000) and they disappear when the ECS are expressed in the original vector-space $\mathbf{R}^n$.

Elementary modes

The concept of elementary modes was introduced to extend the property of non-decomposability of the extreme rays (P3) to networks with reversible fluxes (Schuster, 1999; Schuster, 2000). A flux vector **e** is an elementary mode (EM) if and only if (Schuster, 2002):

C1) $\mathbf{e} \in \mathbf{P}$, and,

C2) there is no non-zero vector $\mathbf{v} \in \mathbf{P}$ such that the support of **v** supp(**v**) is a proper subset of the support of **e** supp(**e**).[1] In other words, **e** cannot be decomposed as a positive combination of two “simpler” vectors **v’** and **v’’** in **P** that contain zero elements wherever **e** does and include at least one additional zero

[1] The support of a vector **x** is the set of the indexes of the elements in **x** equal to zero. Examples: given **x**={4, 3, 0, 1, 0} and **y**={1, 3, 5, 2, 1}, its supports are supp(**x**)={3, 5} and supp(**y**)={∅}.

component each. This condition is the so-called non-decomposability, simplicity or genetic independence.

Thereby, the set of elementary modes (EMS) is defined as the set of all the non-decomposable vectors in the flux space (P3). This definition implies that the EMS fulfills property P1, as in (3), but also a more restrictive condition due to C2: a flux vector can always be represented as a non-negative combination of elementary modes without cancelations (Schuster, 2002):

$$\mathbf{P} = \left\{ \mathbf{v} \in \mathbf{R}^{\mathrm{n}} : \mathbf{v} = \sum_{k}^{e} w_k \cdot \mathbf{e_k}, \quad w_k \geq 0 \right\} \text{ without cancelations (*)} \tag{5}$$

*(*) if the sum runs over two or more indices **k**, all the **ek** have zero components wherever **v** has zero components and include at least one additional zero each.*

That means that the elementary modes are all the simple states (or functional units, or non-decomposable vectors) that a cell can show, and the rest of feasible states can be seen as its strictly aggregated action. That is, its aggregated action without cancelations. The "no cancelation rule" is relevant for several applications of network-based pathways; it makes it possible to investigate the infinite behaviors that cells can show by simply inspection of the finite set of elementary modes, because there is no possibility of cancelations of reversible fluxes. This allows to answer many interesting questions in an easy way, for example:

- Which reactions are essential to produce the compound Y? Those that participate in all the elementary modes producing Y.

- Is there a route connecting the educt A with the product Y? Only if there is an elementary mode connecting them.

- Which are the capabilities of the network if a reaction *r* is not carrying flux or has been knocked-out? The feasible states in these circumstances are only those that result from aggregating, with no cancelations, the elementary modes not involving *r* (i.e., the consequences of *r* not carrying flux can be directly predicted ignoring the elementary modes participated by *r*).

- Which is the optimal yield to produce Y from A? The (stoichiometrically) optimal pathway is the elementary mode consuming A and producing Y with the best yield.

As we will see in subsequent sections, the main difference among network-based pathways is that all of them satisfy (3), but only the elementary modes satisfy (5). This difference determines its applications.

Minimal generators

We have seen that the elementary modes generate the flux space, as in (3), but usually they are not the smallest set satisfying this condition because they have to fulfil the most stringent condition (5). Which is then the minimal set of vectors that generates **P** by non-negative combination? The term minimal generating set (MGS) has been recently coined to refer to this set (Wagner, 2005). Wagner et al. also shown how to obtain a MGS that is subset of EMS. However, there is not a unique minimal generating set in the general case: different MGS may exist within the EMS, and even vectors that are not EMs can be part of an MGS. Both cases will be discussed in following sections.

The idea of a minimal generating set also arises from a different point of view. It is well known that the elementary modes are not *systemically independent* because some modes can be represented as non-negative combination of others (Papin, 2004). These dependent modes are unnecessary to fulfil (3). Thus, any *irreducible subset* of the elementary modes, built by removing dependent modes, is a minimal generating set.

In summary, a set of minimal generators fulfils properties P1 and P2, whereas the elementary modes fulfil P1 and P3. The elementary modes include additional non-decomposable vectors to fulfil P3, which are redundant in (3) but necessary in (5). The fact that a MGS does not fulfils (5) reduces its utility for the analysis of the underlying metabolism. Remarkably, the questions mentioned in the previous section cannot be easily addressed using the MGS because the cancelation of reversible fluxes hides simple pathways. For example, the MGS has to be recalculated after a gene deletion, and similar difficulties arise in other applications. The advantage of the MGSs against the EMS is its reduced size: considering the central carbon metabolism of *E. coli*, the computation of the EMS returns more than 500000 EMs, whereas a MGS contains around 3000 MGs (Wagner, 2005). This also implies that obtaining the MGS is computationally more efficient. For these reasons, the MGS will be preferred in those applications that just require a set of vectors generating the flux-space. For instance, the MGS has been used to perform phenotype phase-plane analysis (Wagner, 2005) and it can be used to extract the minimal connections between extracellular compounds, information that can then be used to develop unstructured, kinetic models (Teixeira, 2007; Provost, 2006; Provost, 2004).

Extreme pathways

As it happens with the extreme currents, extreme pathways are obtained in an augmented vector-space (Schilling, 2000); however, only the internal fluxes are decomposed in both forward and backward directions.[1] Hence, if fluxes are reordered to

[1] The exchange fluxes, those that connect internal and external metabolites with one-to-one correspondence (Klamt, 2003), are kept as reversible.

separate the irreversible internal fluxes $\mathbf{v_I}$, the reversible ones $\mathbf{v_R}$ and the exchange fluxes $\mathbf{v_B}$, as $\mathbf{v} = [\mathbf{v_I}\ \mathbf{v_B}\ \mathbf{v_R}]^\mathbf{T}$, the flux space (2) can be reformulated as follows (where $\mathbf{N} = [\mathbf{N_I}\ \mathbf{N_B}\ \mathbf{N_R}]$):

$$\mathbf{P_{rc}} = \left\{ \mathbf{v} \in \mathbf{R}^{n+r} : \begin{pmatrix} \mathbf{N_I} & \mathbf{N_B} & \mathbf{N_R} & -\mathbf{N_R} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{v_I} \\ \mathbf{v_B} \\ \mathbf{v_R} \\ \mathbf{v'_R} \end{pmatrix} = \mathbf{0} \text{ and } \begin{pmatrix} \mathbf{v_I} \\ \mathbf{v_B} \\ \mathbf{v_R} \\ \mathbf{v'_R} \end{pmatrix} \geq \mathbf{0} \right\} \tag{6}$$

In this augmented vector-space, the set of extreme pathways (EPS) is a subset of the elementary modes that is systemically independent (Papin, 2004); however, the extreme pathways are not systemically independent in the original vector-space.[1] Therefore, they are not the irreducible subset of the elementary modes and they are not the minimal generating set (Wagner, 2005). Unfortunately, this notion was unclear in the literature until recently.

The extreme pathways fulfil property P1—they generate the cone as in (3) because only dependent elementary modes are discarded—, but not P2 nor P3 in the original vector-space. As it happens with the MGS, the fact that the EPS does not fulfil (5) reduces its utility in certain applications. Their advantage with respect to the EMS is its smaller size, but it must be kept in mind that very often the MGS will be smaller than the EPS (and never larger).

> *Example: two different vector-spaces*. Consider the small network depicted in Figure 3.2, case 2A. The 3 EPs of this network represented in the augmented vector-space {v_1, v_2, v_3, -v_3} are: E1=(1 0 1 0), E2=(0 1 0 1) and E3=(1 1 0 0). These 3 vectors are systemically independent. However, when translated to the original vector-space {v_1, v_2, v_3}, these vectors are: E1=(1 0 1), E2=(0 1 -1) and E3=(1 1 0), which are not longer systemically independent, since E1 = E2 + E3. Figure 3.2 also illustrates the systemic dependancy of the EPs.

3.3 Comparison of the different pathway concepts

This section is devoted to the comparison of the network-based pathways described above: extreme currents, minimal generators, elementary modes and extreme pathways. The case where all the fluxes are irreversible will be introduced first to contextualize the problem; then, the presence of reversible fluxes will be considered and the differences will become apparent (see Figure 3.2).

[1] Notice that even the ECs, which are equivalent to the EMs, are systemically independent in the augmented space where they are obtained.

Reference vector-space. Hereinafter we consider the original vector-space as the reference one: all the generating sets will be expressed as elements of the vector-space $\mathbf{R}^n$ where each flux corresponds to an axis. We choose $\mathbf{R}^n$ because it is the original space of the fluxes that connect the metabolites of the network, and thus it is the meaningful one. For instance, in the previous example the EPS expressed in the augmented vector-space were unable to capture the fact that pathway E1 can be seen as a combination of E2 and E3 (E1=E2+E3). Notice also that the relevant difference between equations (3) and (5), which depends on the cancelation of reversible fluxes, cannot be easily observed in the augmented vector-spaces. Since ECs and EPs are computed in augmented vector-spaces, once obtained, they have to be translated to $\mathbf{R}^n$, simply merging the decomposed reversible fluxes. This process also removes the spurious cycles (pathways formed only by the forward and backward reaction of each reversible flux) that appear as EPs and ECs in the augmented vector-spaces.

Case 1: All fluxes are irreversible

As explained in a previous section, when all the reactions are irreversible the flux space **P** is a convex cone that satisfies two conditions: (a) it is in the positive orthant $\mathbf{R}^+$ and (b) it is a *pointed cone*.

Condition (b) implies that **P** can be generated by non-negative combination of its extreme rays (3) (more details below). In fact, the extreme rays always belong to every generating set because by definition they cannot be generated by non-negative combination of other vectors within the cone. Thus, if the extreme rays are able to generate the cone, as it happens in this case, they are necessarily the minimal generating set. For the same reason, the extreme rays are always non-decomposable vectors of **P**. Moreover, condition (a) implies that the intersections of the cone with the (positive or negative) axis of the vector-space, which are potential non-decomposable vectors, cannot be interior points of **P**. Thus, in this particular case the extreme rays will be all the non-decomposable vectors in **P**.

These two conditions imply that when all fluxes are irreversible the extreme rays are the minimal generating set of the flux space (P1 and P2), but also the set of all non-decomposable vectors (P3). Since the ECs and the EPs are extreme rays of two cones defined in augmented vector-spaces where the reversible reactions are decomposed, it is obvious that, since there are no reactions to be decomposed, the ECs and EPs are the extreme rays of the original cone **P**. Therefore:

> *Rule 1. If all fluxes are irreversible, all the generating sets are equivalent, EMS = ECS = EPS = MGS, and coincide with the extreme rays of the flux space* ***P****.*

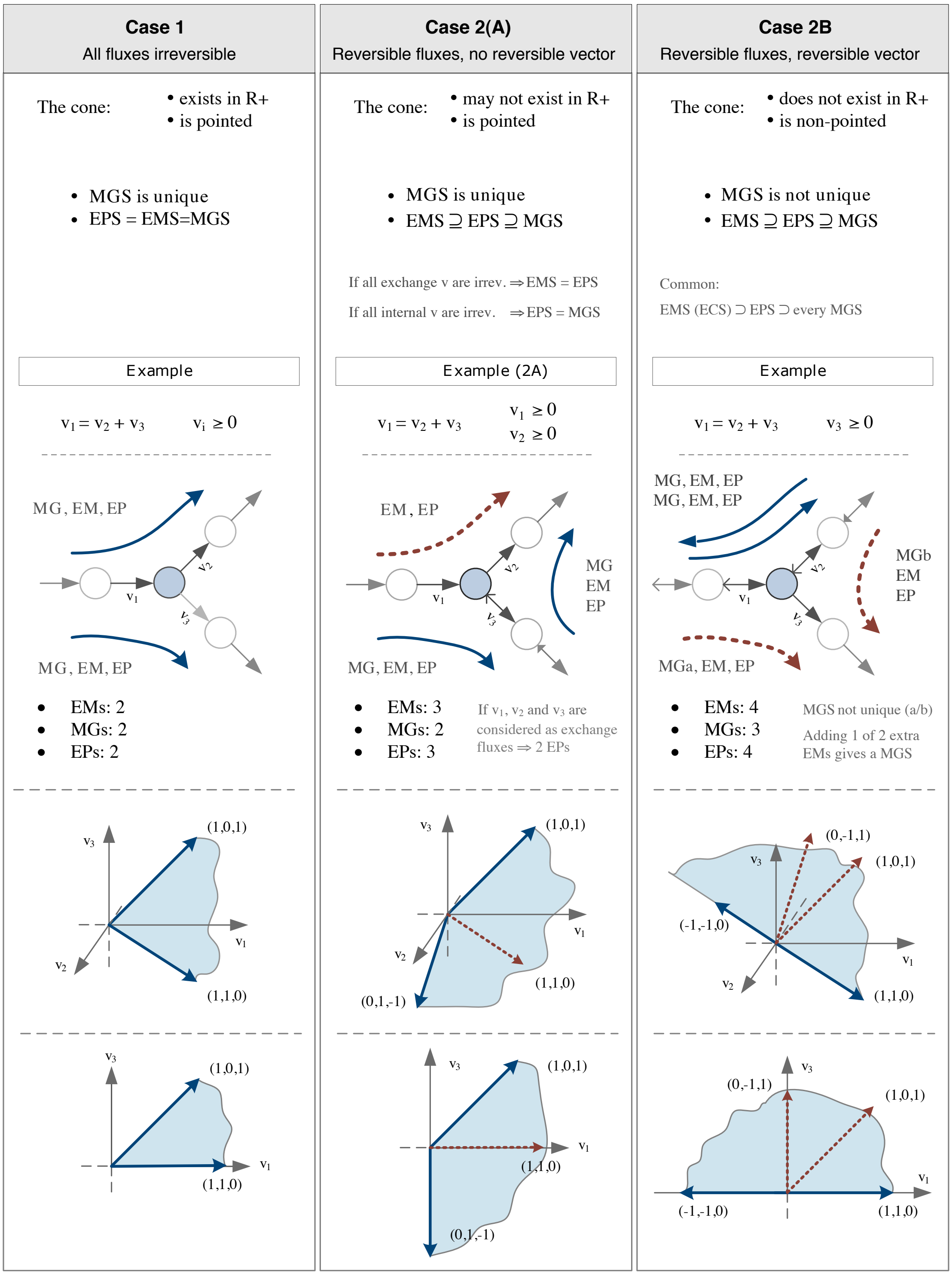


Figure 3.2. Case-based scheme of the different network-based pathways. Metabolites are represented with circles, and thin arrows represent the fluxes (reversible fluxes are double arrowed, and solid arrowhead defines the sign criteria). The axis at the bottom represent the flux-space over $\{v_1, v_2, v_3\}$, blue area, and its generating vectors. The blue thick arrows denote generating vectors that correspond to extreme rays of the cone, and the red dashed ones to other generating vectors.

Case 2: There are reversible fluxes

Now we consider the situation where certain fluxes are reversible. The flux space **P** is still a convex cone, but it is not necessarily in the positive orthant $\mathbf{R}^{+}$ and it may be non-pointed. If one reversible reactions is effectively reversible—i.e., both forward and backward directions can be followed by flux vectors—the cone will not be in the positive orthant (otherwise **P** would remain a pointed one in $\mathbf{R}^{+}$, as in case 1). Two situations are possible: case 2A, the cone is pointed, and case 2B, it is not.

Consider the *lineality space* of **P**, which represents the linear subspace contained in the cone, and defined as (details are given in the glossary at the end of the chapter):

$$\text{lin.space}(\mathbf{P}) := \{\mathbf{x} \in \mathbf{R}^{n} \mid \mathbf{A} \cdot \mathbf{x} = \mathbf{0}\}$$

The lineality space allows to characterise the cone as follows: ***P** is pointed if lin.space(**P**) = {0}; otherwise non-pointed*. Hence, **P** will be a non-pointed cone if a vector **x** and its opposite −**x** exist in **P**. These vectors would involve only reversible fluxes and represent *reversible vectors* that can operate in both directions. Thus, **P** is non-pointed cone if and only if it contains a reversible vector. It is also possible to check if a cone is pointed inspecting **K**, the kernel of **N**, arranged in a suitable way (Wagner, 2005).

The more important consequence of this classification is the following: a pointed cone **P** can be generated by non-negative combination of its extreme rays, which constitute its unique MGS, but this not longer true for a non-pointed one. A non-pointed cone still can be generated by non-negative combination, but a unique MGS will not exist.

Case 2A: reversible fluxes but not reversible vectors

If there are reversible fluxes but not a reversible vector, the flux-space **P** is still a pointed cone and it can be generated by its extreme rays (Schrijver, 1998). As explained above, if the extreme rays generate the cone, they are necessarily the minimal generating set because they belong to every generating set by definition.

> *Rule 2. If the flux space **P** does not contain a reversible vector, a unique MGS exists and it coincides with the extreme rays of **P**.*

However, if there are reversible fluxes, and they are effectively used in both directions, the cone is not restricted to the positive orthant $\mathbf{R}^{+}$. This implies that the intersections of vector-space axis with the cone will be non-decomposable vectors of **P**. That is, there are non-decomposable vectors in **P** that are not extreme rays. The EMS sill contains the extreme rays, which are always non-decomposable, but also other non-decomposable vectors. Notice that these extra EMs are necessary to generate the flux space **P** without cancelations (5), but can be redundant to fulfil (3).

> *Rule 3. The EMS (ECS) is always a superset of the extreme rays of the flux space **P**. If there are reversible fluxes, more EMS than extreme rays may exist.*

By rules 2 and 3 it follows that if the flux space **P** does not contain a reversible vector, the unique MGS is a subset of the EMS. Moreover, those EMs not belonging to the MGS will be systemically dependent and the MGS is the unique irreducible subset of the EMS.

> *Rule 4. If the flux space **P** does not contain a reversible vector, the unique MGS is the irreducible subset of the EMS. It can be extracted from the EMS selecting the systemically independent vectors (see appendix A).*

This property was incorrectly assigned to the extreme pathways in the past, but these are systemically independent only in an augmented vector-space and not in the original one (see example below). The EPs are the extreme rays of the cone obtained when the internal and reversible reactions are split, whereas the EMs (ECs) are the extreme rays of the cone obtained when all the reversible reactions are split. This differences determine the relationship among the concepts (Figure 3.3):

> *Rule 5. If the flux space **P** does not contain a reversible vector, the EPS can be a subset of the EMS, but in general it is not the MGS. That is, EMS (ECS) $\supseteq$ EPS $\supseteq$ MGS, and two particular cases exist:*
>
> *a. If all exchange fluxes are irreversible, EMS (ECS) = EPS*
>
> *b. If all internal fluxes are irreversible, EPS = MGS*
>
> *The two rules can be rephrased as follows:*
>
> *a. EPS can be a proper subset of the EMS ⇔ there are reversible exchange fluxes*
>
> *b. MGS can be a proper subset of the EPS ⇔ there are reversible internal fluxes*

Proof outline. (a) If all the reversible fluxes are internal, the EPs and the ECs (EMs) are the extreme rays of the same cone. (b) If all the internal fluxes are irreversible, the EPs are the extreme rays of the original cone, which coincide with the extreme rays due to rule 2.

Case 2B: Reversible fluxes and a reversible vector

If the reversible fluxes form a reversible vector, the flux space is now a non-pointed cone. A non-pointed cone can be represented as, $\mathbf{P_r} = \mathbf{H} + \mathbf{Q}$, where **H** is the linear

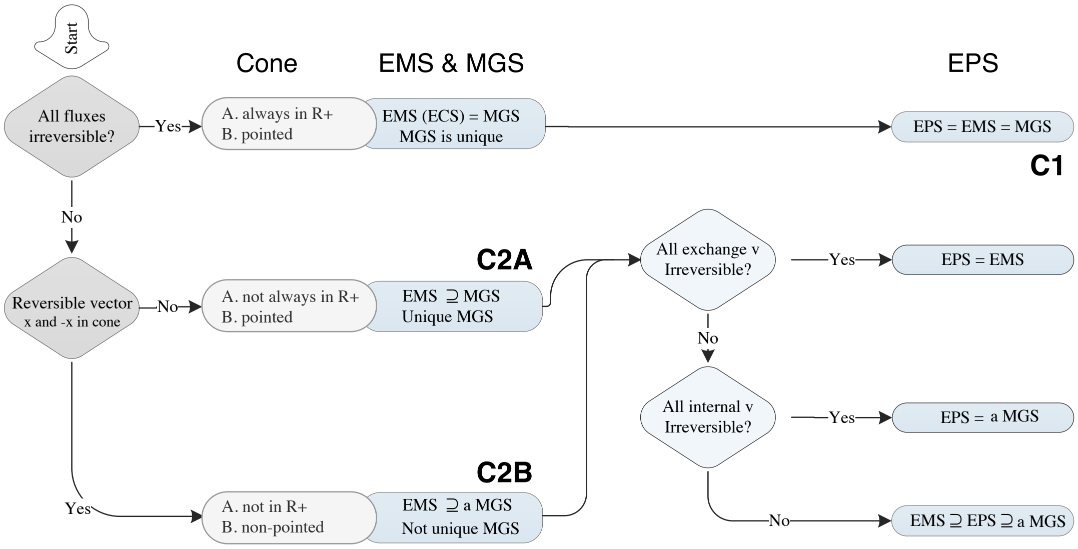


Figure 3.3. Relationship between different network-based pathways.

space lin.space($\mathbf{P_r}$), and $\mathbf{Q}$ is a pointed sub-cone, with $\mathbf{Q} \subseteq \mathbf{H}^{\perp}$ ($\mathbf{H}^{\perp}$ denotes the orthogonal complement of $\mathbf{H}$). This is indeed the general representation of a convex polyhedral cone, cases 1 and 2A were particular cases with $\mathbf{H} = \{0\}$. Thus, a non-pointed cone can be generated as follows (Schrijver, 1998):

$$\mathbf{P_r} = \left\{ \mathbf{v} \in \mathbf{R}^{n+r} : \quad \mathbf{v} = \sum_{k}^{nf} \lambda_k \cdot \mathbf{f_k} + \sum_{j}^{nb} \beta_j \cdot \mathbf{x_j}, \quad \lambda_k \geq 0 \right\} \tag{7}$$

where $\mathbf{f_k}$ are the 'irreversible' generating vectors, for which its opposite are not contained in $\mathbf{P_r}$, and $\mathbf{x_j}$ are the 'reversible' ones, for which its opposite $\mathbf{-x_j}$ is also contained in $\mathbf{P_r}$. Vectors $\mathbf{x_j}$ must form a base of $\mathbf{H}$, whereas vectors $\mathbf{f_k}$ generate the sub-cone $\mathbf{Q}$. Notice that $\mathbf{P_r}$ can still be generated by non-negative combination, as in (3), using $\mathbf{f_k}$, $\mathbf{x_j}$ and $\mathbf{-x_j}$ as generating vectors. Unfortunately, there is a price to pay for the cone being non-pointed: the set of minimal generating vectors is not unique anymore.

In fact, a minimal generating set of $\mathbf{P_r}$ can be obtained choosing an arbitrary base $\{\mathbf{x_j}\}$ of $\mathbf{H}$, and taking one arbitrary ray $\mathbf{f_k}$ from each minimal proper face of the cone (Schrijver, 1998). When the cone is pointed, there are no vectors $\{\mathbf{x_j}\}$ and the the minimal proper faces are the extreme rays, so they are uniquely defined.

The extreme rays of $\mathbf{P_r}$ will be present in any generating set because they cannot be represented as non-negative combination of other vectors in $\mathbf{P_r}$. However, they are insufficient to generate a non-pointed cone, they could even not exist.[1] Additional vectors $\{\mathbf{x_j}\}$ and $\{\mathbf{f_k}\}$ must be combined with the extreme rays to form a MGS, but the choice is not unique.

> *Rule 6. If the flux space $\boldsymbol{P_r}$ contains a reversible vector, its extreme rays are not a complete generating set and there is not a unique MGS.*

However, it is still possible to find a MGS containing only non-decomposable vectors, and thus being a subset of the EMS. This kind of MGS can be obtained with a lexico-smallest representation (Larhlimi, 2009) or extracted from the set of EMs as explained in below.

> *Rule 7. If the flux space $\boldsymbol{P_r}$ contains a reversible vector, a irreducible subset of the EMS constitutes a MGS formed only with non-decomposable vectors.*

Notice that other MGS will exist. Indeed, even more than one MGS formed with different non-decomposable vectors may exist, since there is not necessarily a unique irreducible subset of EMS. Both situations will be illustrated in next examples.

Regarding the EPS, rule 5 should be rephrased recalling that the MGS is not longer unique. Moreover, since reversible vectors are typically participated both by internal and external fluxes (except if they are futile cycles), a common situation arise where, EMS (ECS) $\supset$ EPS $\supset$ a MGS.

> *Rule 8. If the flux-space $\boldsymbol{P_r}$ contains a reversible vector, the EPS can be a subset of the EMS, but in general the EPS is not a MGS. The most common case will be EMS (ECS) $\supset$ EPS $\supset$ a MGS.*

Computing the different pathways

The elementary modes can be computed with *Metatool* (Pfeiffer, 1999) and *cellNetAnalyzer* (Klamt, 2003), both running under MATLAB, and with *OptFlux* (Rocha, 2010). The extreme pathways can be computed using *expa* (Bell, 2005). Minimal generating sets can be obtained using SNA (Urbanczik, 2006), a software package running under Mathematica, or using ccd (Fukuda, 1996) as reported in (Larhlimi, 2009).

[1] For instance, this is the case when all the fluxes are reversible, the cone is a *n*-dimensional vector-space generated only by vectors $\mathbf{x_j}$ and $-\mathbf{x_j}$.

In addition, we describe a simple method to get a MGS from the EMS extracting a irreducible subset. The procedure can be outlined with the following pseudo-code:

```
for each elementary mode $\mathbf{e_i}$ in $\mathbf{E}$
    define $\mathbf{A} = [\mathbf{M}\ \mathbf{E}_{\mathrm{r}}]$
    if (there is no $\mathbf{w} \geq \mathbf{0} \mid \mathbf{A} \cdot \mathbf{w} = \mathbf{e}$) then: add $\mathbf{e_i}$ to $\mathbf{M}$
end
```

where **E** is the matrix formed with EMs as columns, $\mathbf{E_r}$ is the submatrix of **E** only with columns after *i*, and **M** is a matrix collecting the MGs (and thus empty at the first iteration).

If the cone is pointed, the resultant set is the unique MGS (the extreme rays of the cone). Otherwise, it is one MGS (of many) formed with non-decomposable vectors.

3.4 Illustrative examples

Some examples will be used to illustrate the different cases described above. The first examples (1 to 5) use a simple network taken from Papin et al. (2004). The network has 6 reactions—3 internal and 3 exchanges—and three metabolites, so it has 3 degrees of freedom. If all the reactions were reversible, the kernel of **N** would provide a basis of the flux space formed by 3 reversible vectors. Herein we consider 5 examples where different reactions are irreversible (results are depicted in Figure 3.4).

Example 1. In the first example all fluxes are assumed to be irreversible (case 1). In this case, the flux space is a pointed cone in $\mathbf{R}^+$ and ECS, EMS, EPS and MGS are equivalent.

Example 2. Now the exchange flux v_4 is assumed to be reversible. This example corresponds to case 2A (the flux space is a pointed cone not in $\mathbf{R}^+$). In this case the EMS can be a superset of the MGS, as indeed happens in this example: EM4 is systemically dependent (EM4 = MG1 + MG2), so it is an EM but not a MG. On the other hand, the EPS is equal to the MGS because the internal fluxes are all irreversible. EM4 is not an EP because the reversible flux being cancelled in MG1 + MG2 is an exchange, so EM4 is systemically dependent in the vector-space where EPs are computed.

Example 3. In this third example the exchange flux v_4 and the internal flux v_2 are reversible. This is a general case and therefore, EMS ⊇ EPS ⊇ MGS. EM5 is neither an EP nor a MG (EM5 = MG1 + MG2). EM4 is not a MG (EM4 = MG3 + MG2), but it is an EP; one of the fluxes cancelled in MG3 + MG2 is an internal flux, so this cancelation cannot be done in the augmented vector-space where the EPs are computed.

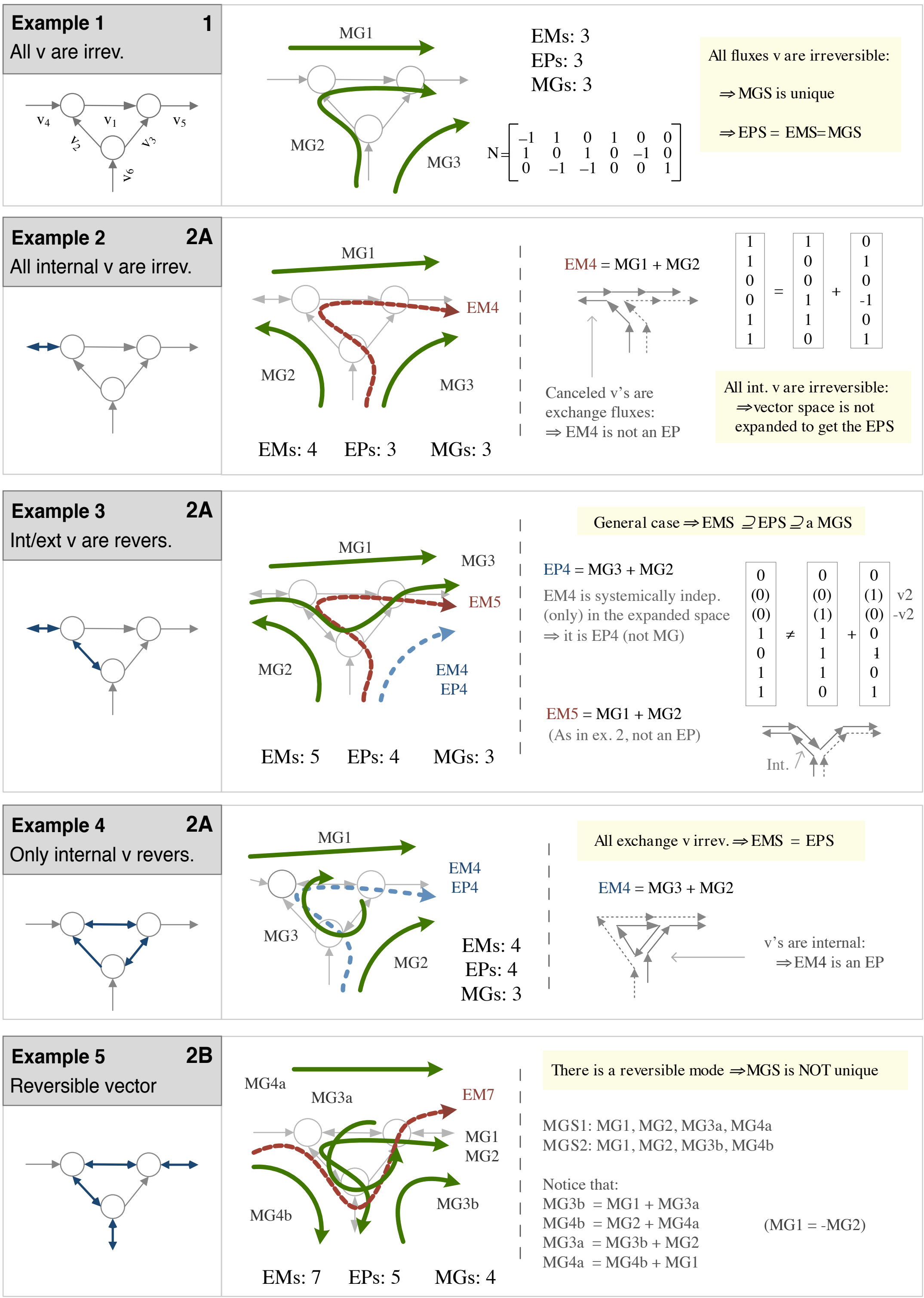


Figure 3.4. Illustrative examples of the differences among network-based pathways.

Example 4. In this example only two internal fluxes, v_1 and v_3, are reversible. Again, the EMS is a superset of the MGS: EM4 is not a MG because it is systemically dependent (EM4 = MG3 + MG2). On the other hand, as all the reversible fluxes are internal, the EPs and the EMs are necessarily equivalent.

Example 5. Now there are four reversible fluxes—v_1, v_2, v_5 and v_6—that define a reversible vector. This corresponds to case 2B, where the flux space is a non-pointed cone. There are 7 EMs and 5 of them are also EPs. The two vectors that form the reversible vector are extreme rays in this example. To form a MGS they need to be combined with 2 other vectors, but the choice is not unique. For instance, 2 subsets of EMs are minimal generating sets, MGS1 and MGS2.

Example 6. Klamt et al. uses a simple example, referred as N2 in their article, to investigate the relationship between the EMS and the EPS (Klamt, 2003). This network has 9 reaction (3 exchanges) and 6 metabolites. After computing the EMS, the EPS and the MGS, it turns out that there are 8 EMs and 5 EPs (the extra EM9/EP6 in (Klamt, 2003) disappears in the original vector-space because it is a spurious cycle caused by decomposing the reversible fluxes). Yet, the MGS contains only 4 vectors, indicating that there is an EP that is not systemically independent: it can be checked by simple inspection that EP1 = EP2 + EP4 (when they are represented in the original vector-space).

Example 7. Another example to be analyzed is the small network used by Schilling et al. (2000). We obtained 7 EMs and the 5 relevant EPs given in the paper. Again, the EPs are not systemically independent when translated to the original vector-space (EP2 = EP3 + EP5) and 4 vectors are sufficient to form a MGS. It turns out that the MGS is not unique because there is a reversible vector in the flux-space (in fact, the reversible vector defines two EPs: EP3 and EP4 use the same reactions but in opposite directions).

Example 8. We have also analyzed the metabolic network of CHO cells given in (Provost, 2006a). The network has 24 reactions (9 reversible) and 18 internal metabolites, so it has 6 degrees of freedom. There are 18 EMs and 8 EPs, but only 6 vectors form the unique MGS. More details about this model will be given in chapter IV, where it is used as a case study.

3.5 Conclusions

The purpose of network-based pathways analysis is to identify a finite set of systemic pathways in a metabolic network, and use these pathways to study the cell metabolism. In this chapter four similar concepts of network-based pathways have been described and compared.

We have seen that all the flux states of a given metabolic network can be represented as an aggregation of fluxes through its elementary modes, which are all the simple, or non-decomposable, pathways in the network. Nevertheless, the set of elementary modes is not the smallest set of pathways fulfilling this property. This role corresponds to the so-called minimal generating sets. In certain cases there is a unique minimal generating set, but often there are many of them. Interestingly, the set of elementary modes can be reduced by eliminating modes that are systemically dependent, resulting in a minimal generating set formed only with elementary modes. We have also highlighted that, contrarily to what has sometimes been stated, the extreme pathways are not a minimal generating set, because they are usually systemically dependent in the original vector-space.

The minimal generating sets can be of use in applications where a set of generating vectors is required. In these cases they will be preferred due to its reduced size and because their computation is more efficient. For instance, minimal generators are suitable for extracting the fundamental connections between extracellular compounds, information that can be used to develop unstructured, kinetic models (Teixeira, 2007; Provost, 2004; 2006). However, the analysis of the elementary modes is more powerful. The fact that the set of elementary modes comprises all the simple pathways in the network—its functional states—makes it possible to investigate the infinite behaviors that cells can show by simply inspecting them. This makes it easy to answer several questions: which reactions are essential to produce a certain compound, which will be the capabilities of the network if a reaction is knockout, etc. Answering these questions using the minimal generators or the extreme pathways may be difficult because one has to take into account the possible cancelations of reversible fluxes.

Significant efforts are being done to improve network-based pathways analysis, particularly in the context of genome-scale metabolic networks, where their more critical limitation appears. When the number of reactions in the network grows, the number of pathways dramatically increases, reducing understandability and even becoming not computable (Papin, 2004; Gagneur, 2004). Recent works have improved the computation algorithms (Klamt, 2005; Terzer, 2008), and proposed methods to get particular subsets of pathways (Figueiredo, 2009) or decompose large networks in modules (Schuster, 2002b). New concepts of pathways have been also recently introduced. Kaleta et al. have introduced *Elementary flux patterns*, which explicitly takes into account possible steady-states fluxes through a genome-scale network when analyzing pathways through a subsystem, thus allowing the application of many (not all) elementary-mode-based tools to genome-scale networks (Kaleta, 2009). Barrett et al. have used Monte Carlo sampling in conjunction with principal component analysis to obtain a low-dimensional set of pathways generating the flux space of genome-scale networks (Barrett, 2009).

Most applications of network-based pathway analysis are currently found in the context of microbial production (e.g., Schilling, 2000; Price, 2002; Schwartz, 2006), but also in botany (Poolman, 2003; Steuer, 2007) or in biomedicine (Zhong, 2002; Nolan, 2006).

Main references

- Llaneras F, Picó J (2010). Which metabolic pathways generate and characterise the flux space? A comparison among elementary modes, extreme pathways and minimal generators. *J. Biomedicine and biotechnology*, 1:2010.

- Rockafellar RT (1996). *Convex analysis*. Princeton, USA: Princeton University Press.

- Schrijver A (1988). *Theory of linear and integer programming*. Amsterdam, Netherlands: Wiley.

- Papin JA, Stelling J, Price ND, Klamt S, Schuster S, Palsson BO (2004). Comparison of network-based pathway analysis methods. *Trends in Biotechnology*, 22(8):400-405.

- Schuster S, Dandekar T, Fell DA (1999). Detection of elementary flux modes in biochemical networks: A promising tool for pathway analysis and metabolic engineering. *Trends in Biotechnology*, 17(2):53-60.

- Schilling CH, Letscher D, Palsson BO (2000). Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology*, 203 (3) :229-248.

- Wagner C, Urbanczik R (2005). The geometry of the flux cone of a metabolic network. *Biophysics Journal*, 89(6):3837-3845.

Part II: Interval methods

IV

Interval estimates of metabolic fluxes under data scarcity

This chapter describes an interval approach to perform flux estimations, a variant of metabolic flux analysis particularly well suited to scenarios of data scarcity. This approach exploits the available measurements, coupled with a constraint-based model, to estimate each metabolic flux. The approach is based on a linear programming formulation, so it is simple and computationally efficient.

We use two real cases studies to illustrate the limitations of traditional MFA and show the benefits of the interval approach.

Part of the contents of this chapter appeared in the following journal articles:

- Llaneras F, Picó J (2007). An interval approach for dealing with flux distributions and elementary modes activity patterns. *Journal of Theoretical Biology*, 246(2):290-308.
- Llaneras F, Picó J (2007). A procedure for the estimation over time of metabolic fluxes in scenarios where measurements are uncertain and/or insufficient. *BMC Bioinformatics*, 8:421.

4.1 Introduction

As we saw in chapter II, constraint-based models define the space of metabolic states that cells can exhibit, but do not predict which of these are likely to take place at given circumstances. These predictions can be obtained with flux balance analysis (FBA), which assumes that cells have evolved to be optimal (Price, 2003). FBA is able to predict the actual fluxes, but requires to identify the objectives relevant at different conditions (Schuster, 2008; Schuetz, 2007). As an alternative, one can perform a metabolic flux analysis (MFA), which, generally speaking, is the exercise of estimating the fluxes shown by cells combining the model with experimental measurements.

Traditional MFA uses only measurements of uptake and production rates (i.e., fluxes in and out cells) that are balanced around the intracellular metabolites (see chapter II). This purely stoichiometric approach has some limitations, particularly in scenarios lacking data and where measurements are imprecise. Traditional MFA requires a large number of accurate measurement to be of use, but these are often not available.

In this chapter we follow a constraint-based approach to address this problem. We present a variant of MFA that exploits an interval representation of fluxes.[1] The proposed method, called flux-spectrum (FS-MFA) is particularly well suited for scenarios of data scarcity, scenarios where: (a) isotope experiments are not available, (b) there is often a lack of measurable fluxes, and (c) the available measurements may be imprecise or inaccurate.

The benefits of an interval approach to MFA can be summarised as follows:

- It considers reactions irreversibility and other inequality constraints, and represents the measured fluxes with intervals, thus capturing its uncertainty.
- It provides interval estimates, instead of point-wise ones.
- Intervals estimates are more reliable (their uncertainty is explicit) and richer (more informative).
- Intervals estimates enable the use of MFA in two new cases: (a) when there is a lack of measurements, and (b) when these are highly inconsistent. Point-wise estimates fail in both cases, but intervals may be valuable.

The chapter is structured as follows: we first review traditional MFA and discuss its limitations. In section 4.3, flux-spectrum MFA is introduced as an alternative. Afterwards, two cases of study are used to illustrate the limitations of traditional MFA—some of them not well-know—and show the advantages of the flux-spectrum.

[1] Instead of representing a flux with a single value v, an interval is used. This is useful in two common situations: (a) if fluxes are uncertain ([0.9, 1.1]), and (b) if they are partially unknown ([0, ∞], or [0, 5]).

4.2 Preliminaries on metabolic flux analysis

Recalling ideas from chapter II, let us consider a metabolic network with *m* internal metabolites and *n* reactions. Thus, assuming that internal metabolites are at steady-state, mass balances around internal metabolites can be formulated as follows:

$$\mathbf{N}\cdot\mathbf{v} = \mathbf{0} \tag{1}$$

where $\mathbf{v} = (v_1, v_2, ..., v_3)^T$ is the *n*-dimensional vector of metabolic fluxes, and **N** is a stoichiometric matrix.

A flux vector **v** represents the metabolic state of the cells at a given time, without any information on the kinetics of the reactions. Notice that as typically *n* is larger than *m*, the system (1) is underdetermined, i.e., there is a wide range of stoichiometrically-feasible flux vectors.

Now, we consider that some fluxes in **v** have been measured, $\mathbf{v} = (\mathbf{v_u}\ \mathbf{v_m})$. Keeping in mind that measurements are imprecise, they can be represented as follows:

$$\mathbf{w_m} = \mathbf{v_m} + \mathbf{e_m} \tag{2}$$

where $\mathbf{e_m}$ represents measurement errors and $\mathbf{w_m}$ the measured values.

Traditional metabolic flux analysis (TMFA) can be defined as the estimation of a flux vector **v** satisfying (1-2) for a "reasonably small" measurement error. TMFA is often formulated as a two-step procedure: (1) analyse the consistency of the measurements to detect gross errors, and (2) solve a weighted least squares problem to estimate **v**. Details about TMFA calculations can be found in chapter II (section 2.8).

Let us recall the concepts of determinacy and redundancy. If we split (1) between measured (*m*) and unknown fluxes (*u*), we obtain the equation:

$$\mathbf{N_u}\cdot\mathbf{v_u} = -\mathbf{N_m}\cdot\mathbf{w_m} \tag{3}$$

This equation allows us to classify any MFA problem as follows:

Determinacy. If the system (3) is determined,[1] there are enough linearly independent constraints to uniquely calculate all non-measured fluxes $\mathbf{v_u}$. If it is underdetermined, at least one flux in $\mathbf{v_u}$, probably most of them, are non calculable.

[1] If rank($\mathbf{N_u}$) = *u* (*u* is the number of non-measured fluxes).

Redundancy. If the system (3) is redundant,[1] some rows in $\mathbf{N_u}$ are linear combinations of other rows, this can lead to an inconsistent system if $\mathbf{w_m}$ contains such values that no $\mathbf{v_u}$ exists that exactly solves (3). These redundancies can be exploited to analyse measurements consistency and adjust some measured fluxes.

Remember that, ideally, traditional TMFA should be performed only when the system is determined and redundant: (a) if it is not redundant, measurements consistency cannot be evaluated, and the point-wise estimate given by TMFA will be unreliable, and (b) if the system is underdetermined, a point-wise estimate will be only one of multiple (infinite) possible values (Klamt et al., 2002).

Limitations of MFA

Although it has been successfully applied for many years (see chapter II for examples), this traditional formulation of MFA has some limitations:

(i) It only considers equality constraints. (For example, reversibility constraints or maximum flux values cannot be taken into account.)

(ii) It provides only point-wise estimates, uninformative and unreliable when the uncertainty is significant.

(iii) It cannot be used if measurements are (highly) inconsistent, because point-wise estimates cannot reflect their obvious high uncertainty.

(iv) It requires a large number of measured fluxes to be of use: the system (3) has to be determined and redundant. Otherwise, the given estimate will be one of many possible ones.

Several alternatives have been suggested to face these limitations (Bonarius, 1997). For instance, quadratic programming allows to get estimates considering irreversibility constraints (but inherits the rest of drawbacks and the χ^2 tests lose validity). There are also proposals to incorporate assumptions to overcome the lack of measurements. Nookaew et al. have proposed to get estimates based on the assumption that cells are likely to use as many pathways as possible to maintain robustness and redundancy (2007). Related hypotheses have been formulated using the concept of elementary modes (Poolman, 2004; Schwartz, 2006). The FBA assumption of optimal cell behaviour could be also invoked. Another option is incorporate intracellular data obtained from stable isotope tracer experiments (Sauer, 2006; Szyperski, 1998; Wiechert, 2001). Yet, data from isotope tracer experiments will not be considered in this work because they are seldom available.

[1] If $\text{rank}(\mathbf{N_u}) < m$.

Instead, we follow a constraint-based approach to introduce a variant of MFA. We do not attempt necessarily to predict the actual fluxes with accuracy, but to obtain candidate flux values by means of intervals. We will show that this approach overcomes the limitations of traditional metabolic flux analysis described above, providing reasonably estimates in scenarios lacking data and where measurements are imprecise, without new hypothesis and without data from isotopic experiments.

4.3 Flux-spectrum MFA: an interval approach

Let us approach metabolic flux analysis with a constraint-based perspective. First, along with the mass balances at steady-state (1), we consider the irreversibility of certain reactions:

$$\mathbf{D}\cdot\mathbf{v}\geq\mathbf{0} \tag{4}$$

where **D** is a diagonal *n*×*n*-matrix with $\mathbf{D_{ii}}$ = 1 if the flux *i* is irreversible (otherwise 0).

Hence, the *flux space* of feasible (steady) state flux vectors is defined as:

$$\mathbf{P}=\left\{\mathbf{v}\in\mathbf{R}^{\mathrm{n}}:\begin{array}{c}\mathbf{N}\cdot\mathbf{v}=\mathbf{0}\\ \mathbf{D}\cdot\mathbf{v}\geq\mathbf{0}\end{array}\right\} \tag{5}$$

The flux space can be seen as a simple *constraint-based model.* which can be easily extended adding adjustable constraints for measured fluxes at given circumstances. To account for the uncertainty of the measurements, each measured flux can be represented with an interval $v_{m,i}=\left[v_{m,i}^{m},v_{m,i}^{M}\right]$, and then (2) is substituted by inequalities:

$$\mathbf{v_m^m}\leq\mathbf{v_m}\leq\mathbf{v_m^M} \tag{6}$$

At this point, the constraints given by mass balances and irreversibility constraints (5), together with the measured fluxes (6), define the so-called *current flux space* **F**:

$$\mathbf{F}=\left\{\mathbf{v}\in\mathbf{R}^{\mathrm{n}}:\begin{array}{c}\mathbf{N}\cdot\mathbf{v}=\mathbf{0}\\ \mathbf{D}\cdot\mathbf{v}\geq\mathbf{0}\\ \mathbf{v_m^m}\leq\mathbf{Q}\cdot\mathbf{v}\leq\mathbf{v_m^M}\end{array}\right\} \tag{7}$$

where **Q** is a matrix that selects the measured fluxes having exactly one "1" per row, other elements zero. The space **F** contains the flux vectors $\mathbf{v}\in\mathbf{P}$ compatible with the measurements.

At this point, flux-spectrum MFA (FS-MFA) can be defined as the exercise of estimating the flux vectors **v** that fulfil the constraints (7).

Classifying FS-MFA problems

Before addressing the flux estimations, the FS-MFA problems defined with **F** in (7) should be classified in analogy to traditional MFA problems, accounting for its consistency, closure and determinacy.

Consistency. A FS-MFA problem is *consistent* if there is at least one vector $\mathbf{v} \in \mathbf{F}$; otherwise the FS-MFA problem is inconsistent (Figure 4.1A). Notice that the consistency of a TMFA problem and the consistency of the correspondent FS-MFA problem are not equivalent: FS-MFA considers reactions irreversibility to detect new inconsistencies and considers measurements uncertainty, so that the problem can be consistent even if the original measurements are not (Figure 4.1B).

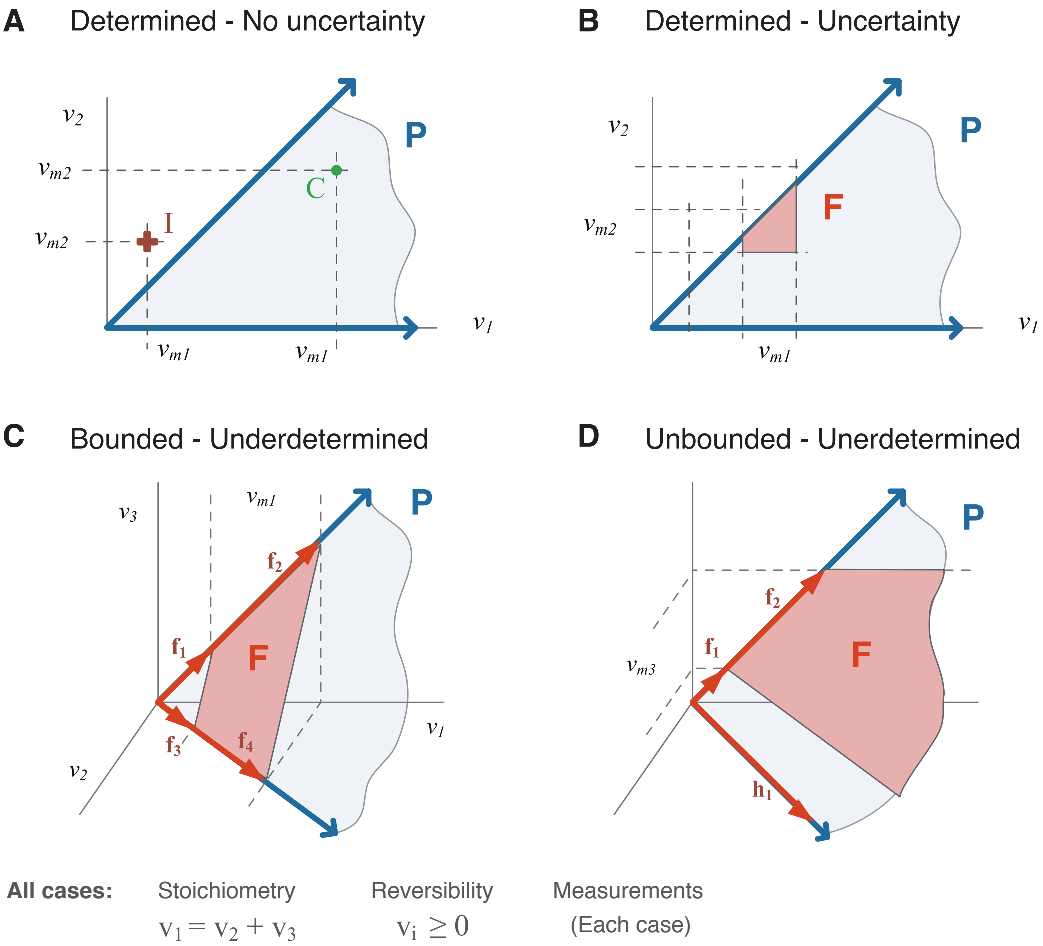


Figure 4.1. Projections of the flux space P and the current flux space F. Minimal generators of F are depicted with red arrows (unbounded and bounded generators are hi and fi, respectively). Subindexes m denote measured fluxes, and letters I and C inconsistent and consistent sets of fluxes, respectively.

Closure. A FS-MFA problem is bounded (or closed) if and only if **F** is bounded. In other words, the problem is said to be bounded in if the range of possible values for each flux is bounded, if $\forall i$ = 1...*n*, values v_i^m and v_i^M exist so that $v_i^m \leq v_i \leq v_i^M$. A bounded FS-MFA problem can be considered *solvable*, in the sense that all the fluxes can be estimated (Figure 4.1D).

Determinacy. Using the classical definition given in the introduction, a FS-MFA problem is said to be determined if the measurements impose enough linearly independent constraints to uniquely determine all the fluxes.[1] Although a FS-MFA problem can be solvable even if it is underdetermined, the notion of determinacy is still useful. A *determined* FS-MFA may have multiple solutions, but it is always bounded and all fluxes can be estimated (Figure 4.1C); on the contrary, if a problem is *underdetermined*, the flux estimation will be always non unique, and maybe unbounded.

The flux-spectrum

Once the admissible flux space **F** has been defined (7), interval estimates can be easily obtained for each measured and non-measured fluxes. These intervals are obtained solving two linear programming (LP) problems for each flux v_i as follows:

$$\forall v_i,\ \ i = 1\ldots n \quad \begin{cases} v_i^m = \ \min v_i \quad \text{s.t. } \mathbf{v} \in \mathbf{F} \\ v_i^M = \ \max v_i \quad \text{s.t. } \mathbf{v} \in \mathbf{F} \end{cases} \tag{8}$$

In this way we get an interval estimate for each flux, an interval bracketing its possible values, $v_i \in \left[v_i^m, v_i^M \right]$.

The flux-spectrum $\mathcal{S}$ can be defined as the set of these intervals:

$$\mathcal{S} = \left\{ \mathbf{v} \in \mathbf{R}^{\mathrm{n}} : v_i^m \leq v_i \leq v_i^M \right\} \tag{9}$$

The flux-spectrum $\mathcal{S}$ is the smallest "plane-parallel" set that encloses the flux space **F**. $\mathcal{S}$ encloses **F** but contains other flux vectors that do not fulfil (7). However, this overestimation is unavoidable if one wants to give an independent estimation for each flux.

The width of the intervals reflects the precision of the estimate, which depends on the number of non-measured fluxes, the irreversible reactions, the available measurements and the considered degree of uncertainty. Of course, the further constraints are available the tighter intervals are obtained.

[1] Reactions irreversibility and measurements uncertainty are not considered in this analysis.

As it will be shown in next sections, the interval approach of the flux-spectrum provides several advantages over traditional MFA.[1]

Simple example of FS-MFA

We now apply FS-MFA to a simple example. A toy network and its stoichiometric matrix **N** are given in Figure 4.1. All fluxes except v_4 are irreversible, so matrix **D** is defined as, **D** = diag(1 1 1 0 1 1).

Three fluxes in the network are measured at successive time instants {v_3, v_5 and v_6}, but its uncertainty is initially not taken into account. The MFA problem is determined and not redundant, so there is a unique flux vector **v** fulfilling (7). In this case, FS-MFA provides the same point-wise estimate that TMFA (Figure 4.2).

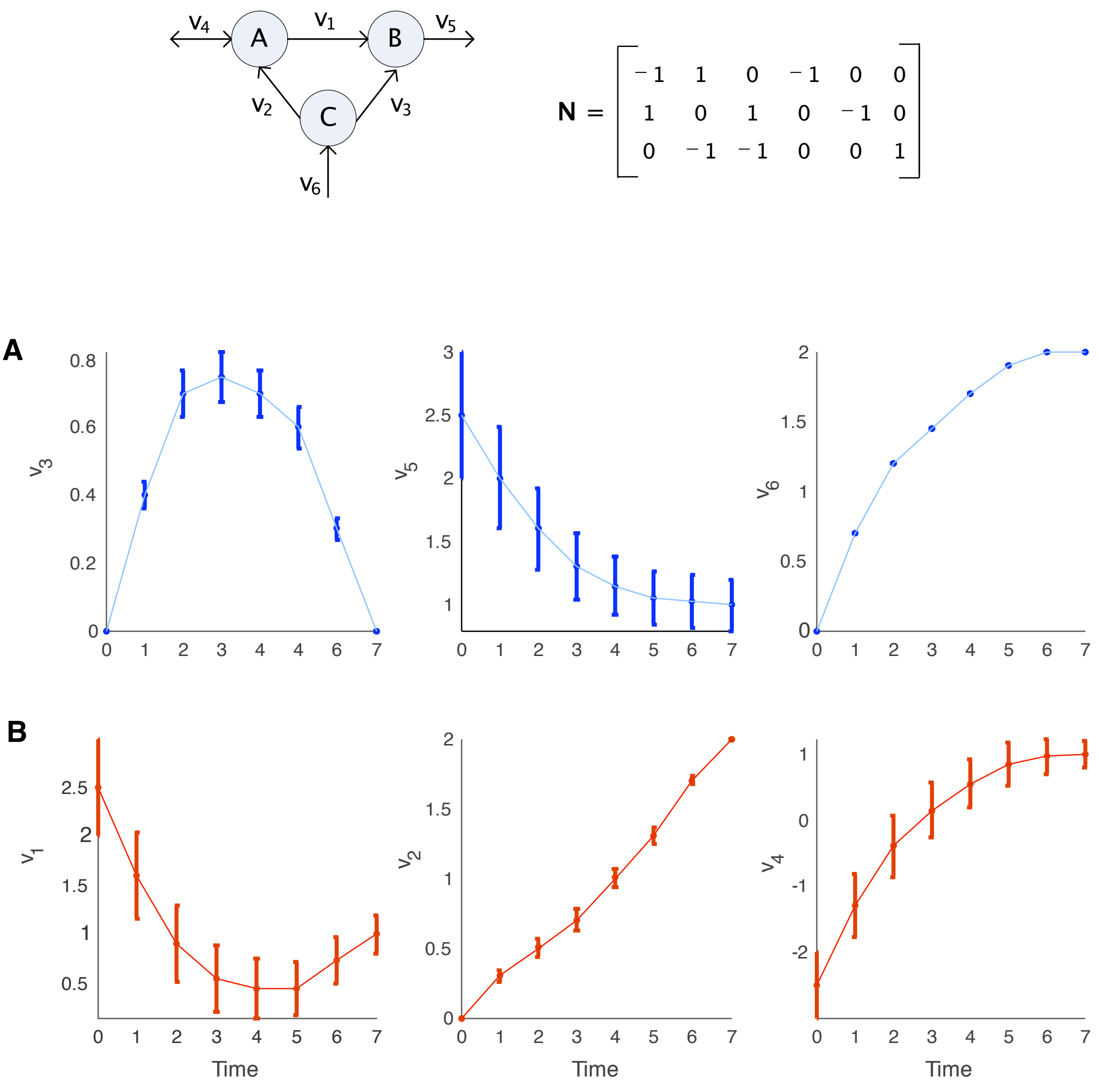
However, we should consider that measurements are in practice imprecise. For instance, we can assume an uncertainty around the measured values of ±10% for v_3, ±20% for v_5 and 0% for v_6, and define the constraints in (6) accordingly. The, solving the linear programming problems in (8), FS-MFA provides interval estimates that have into account the uncertainty of the original measurements (Figure 4.2).

Benefits of the FS-MFA

As shown in the previous example, if uncertainty is not considered, all fluxes are reversible, and the MFA problem is determined, FS-MFA gives the same point-wise estimate that traditional MFA. In addition, FS-MFA bring several advantages that can be summarised as follows (table 4.1):

- *FS-MFA considers reactions irreversibility*. These and other inequality constraints further restrict the interval estimates (Figure 4.3). This will be useful to handle uncertainty and the lack of measurements. Moreover, these constraints can detect inconsistencies even if there are not redundant measurements.
- *FS-MFA represents the measured fluxes with intervals to capture its uncertainty*. This also allows one to incorporate other knowledge, such us capacity constraints, or measurements that are highly uncertain.
- *FS-MFA provides interval estimates, instead of point-wise ones*. Intervals estimates are more reliable (their uncertainty is explicit) and richer (more informative).

[1] Mahadevan and Schilling used a similar approach to analyse alternate optimal solutions in constraint-based metabolic models (Mahadevan, 2003). Their proposal, called *Flux Variability Analysis*, solves a similar set of LP problems, but with a different purpose. Flux variability analysis follows is an FBA approach: it incorporates the assumption of optimal cell behaviour as constraint to predict the cell behaviour. However, the flux-spectrum is an MFA-wise method: it incorporates a set of experimentally measured fluxes, instead of the optimality assumption, to estimate the cell behaviour at a given moment.



$$\mathbf{N} = \begin{bmatrix} -1 & 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 & -1 & 0 \\ 0 & -1 & -1 & 0 & 0 & 1 \end{bmatrix}$$

Figure 4.2. FS-MFA example. A toy metabolic network and its stoichiometric matrix. (A) Three measured fluxes. Intervals represent their uncertainty (±10% for v_3, and ±20% for v_5). (B) Fluxes estimated with FS-MFA and TMFA are denoted with red intervals and red lines, respectively.

- *FS-MFA interval estimates enable MFA when there is a lack of measurements*, i.e., when the FS-MFA problem is underdetermined. TMFA point-wise estimates fail in this situation because they provide only one of multiple solutions, while FS-MFA intervals capture all of them. These interval estimates are typically narrow enough to be valuable and informative.

- *FS-MFA intervals estimates enable MFA when measurements are highly inconsistent.* A point-wise estimate cannot be chosen in this situation—because the measurements have been proved to be highly uncertain—, but we can define a band of uncertainty around the measurements to enclose nearby consistent measurements, and get interval estimates from it (Figure 4.3). Furthermore, the band size needed to find the first solution provides an indication of the degree of inconsistency.

These benefits will be illustrated with two cases studies in subsequent sections.

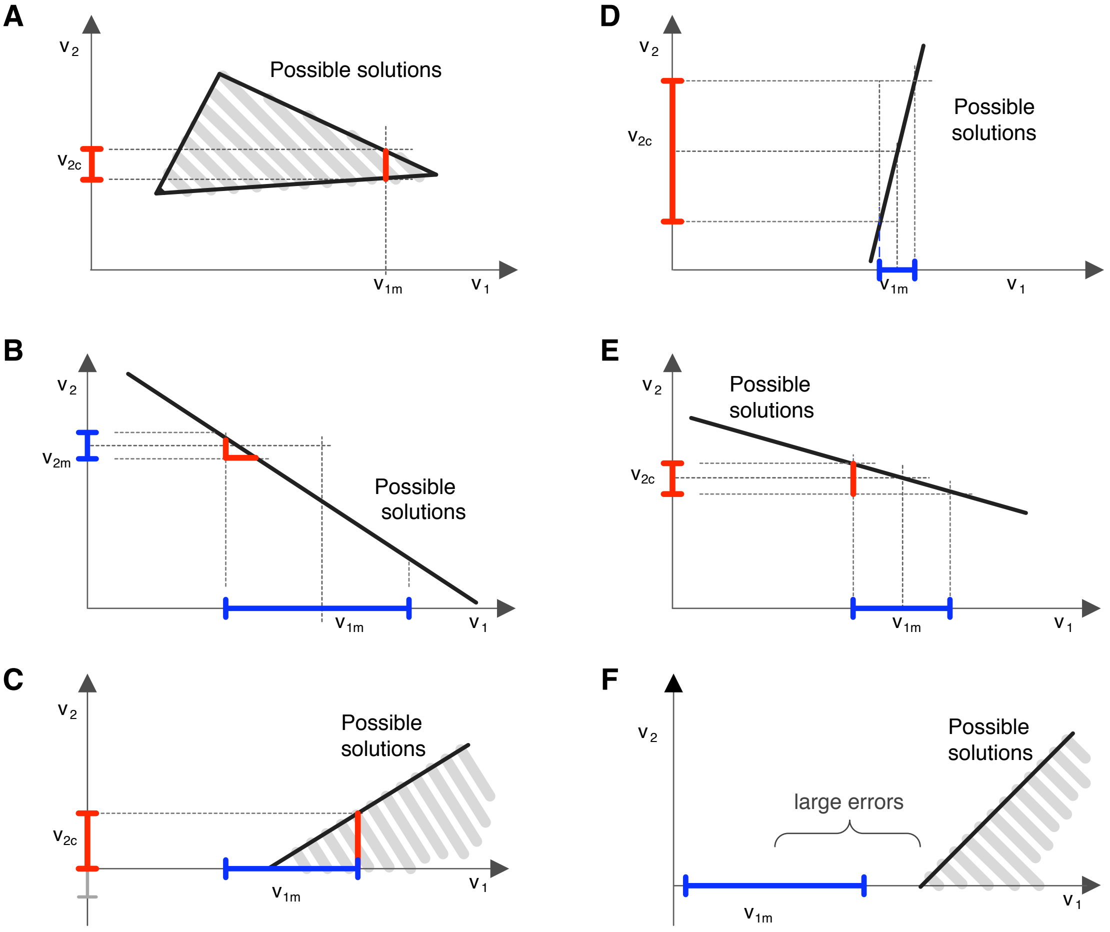


Figure 4.3. FS-MFA in use. Each figure shows a schematic projection of a high-dimensional flux space. (A) Underdetermined case. (B) Determined and redundant case. (C) Adding reversibility constraints. (D) Detection of sensitivity problems. (E) Considering uncertainty. (F) Consistency analysis with reversibility constraints. In all cases, the space of possible solutions before taken into account v_1 and v_2 is represented with a black line or a polygon; the uncertainty of the measured fluxes is represented with blue intervals, and the interval estimates with red intervals; subindex *m* and *c* denote measured and calculated fluxes, respectively.

Limitations of FS-MFA

We have seen that FS-MFA brings some interesting advantages over traditional MFA, but it still has some limitations.

- *The flux-spectrum is an overestimation.* Since each individual flux cannot be varied independently, there are combinations of fluxes within the flux-spectrum that are unfeasible flux vectors, i.e., that do not fulfil (7). Unfortunately, this overestimation is unavoidable if one wants to give an independent estimation for each flux. Notice also that it is guaranteed that all the feasible solutions of (7) are captured by the flux-spectrum intervals.
- *MFA is still limited to small metabolic networks.* TMFA can only be applied with relatively small networks, otherwise the available measurements (even if 13C data are available) are insufficient to offset the network under-determinacy. FS-MFA reduces this difficulty, thanks to the irreversibility constraints and the use of intervals, being able to get estimates in underdetermined cases. However, if the under-determinacy is large, the interval estimates will be wide and even unbounded.
- *Interval estimates tend to be conservative.* To enclose all values that are reasonably possible, the interval description of the measurements tend to be conservative, and this is translated to the estimates. A single interval cannot distinguish highly possible values from those which are only reasonably possible. In other words, an interval is more informative than a single value, but it is still limited. This problem will be addressed in chapter VII using a possibilistic framework.

Table 4.1. Comparison between MFA and FS-MFA.

	Rich data	Data Scarcity		
	Determined *Redundant*	*Determined* *Not Red.*	*Underdet.* *Redundant*	*Underdet.* *Not Red.*
Traditional MFA (TMFA)				
Flux estimation	o	o		
... with high uncertainty				
Evaluates consistency (χ^2)	o		o	
Evaluates consistency (irreversibility)				
Flux-spectrum MFA (FS-MFA)				
Flux estimation	o	o	(o)	(o)
... with high uncertainty	o	o	(o)	(o)
Evaluates consistency (χ^2)	o		o	
Evaluates consistency (irreversibility)		o		o

The symbol “o” denotes a feature, and “(o)” a potential feature.

Parametric description of the current flux space

The current flux space **F** can be defined with a set of constraints (7), but this description is not operative. The flux-spectrum is a more useful description, but at the cost of overestimating the space of feasible flux states. To provide a third alternative, this section introduces an exact and parametric description of **F**.

From a geometric perspective, the current flux space **F** is a convex polyhedron of the form $\{\mathbf{x} | \mathbf{A} \cdot \mathbf{x} \geq \mathbf{b}\}$, where **A** is a matrix and **b** a column vector.[1] Interestingly, any convex polyhedron can be decomposed as the sum of a convex hull and a convex polyhedral cone (Le Verge, 1994; Schrijver, 1999). Therefore:

$$\mathbf{F} = \underbrace{\left\{ \sum_{j}^{q} \omega_j \cdot \mathbf{x_j} : \quad \omega_j \geq 0, \quad \sum_{j}^{q} \omega_j = 1 \right\}}_{\text{Convex hull}} + \underbrace{\left\{ \sum_{k}^{p} \sigma_k \cdot \mathbf{h_k} : \quad \sigma_k \geq 0 \right\}}_{\text{Convex Polyhedral Cone}} \tag{10}$$

where ω_j are σ_k are weights, vectors $\{\mathbf{x_j}\}$ are the vertices of the hull and vectors $\{\mathbf{h_k}\}$ are a generating set of the cone. (The first are sometimes called *bounded generators* and the second *unbounded generators*.)

The generating vectors $\mathbf{h_k}$ and $\mathbf{x_j}$ provide a parametric (or explicit) description of the current flux space **F**. Any flux vector $\mathbf{v} \in \mathbf{F}$ can be represented as a non-negative combination of these generating vectors; and any combination of these vectors, satisfying the conditions for ω_j and σ_k, corresponds to a flux vector $\mathbf{v} \in \mathbf{F}$. Figure 4.4 provides a graphical representation.

Notice that vectors $\mathbf{x_j}$ correspond to vertices of the convex hull and are uniquely defined, but this is not necessarily true for the vectors $\mathbf{h_k}$ generating the cone: they will be unique if the cone is pointed, but not otherwise.[2]

Remark. If **F** is bounded (it is a polytope), it can be generated using vectors $\mathbf{x_j}$ only:

$$\mathbf{F} = \left\{ \sum_{j}^{q} \omega_j \cdot \mathbf{x_j} : \quad \omega_j \geq 0, \quad \sum_{j}^{q} \omega_j = 1 \right\} \tag{11}$$

[1] Remember that the flux space **P** is not a convex polyhedron, but a convex polyhedral cone of the form $\{\mathbf{x} | \mathbf{A} \cdot \mathbf{x} \geq 0\}$. The cone **P** was studied in chapter III.

[2] Any convex polyhedral cone **C** can be decomposed as the sum of a pointed cone and a linear space, **C** = **Cp** + lin.space(**C**). **C** can thus be represented as non-negative combination of a minimal set of vectors: $\{\mathbf{g_1}, ..., \mathbf{g_s}, \mathbf{b_1},...,\mathbf{b_p}\}$, with $\mathbf{gi} \in \mathbf{Cp} \setminus$lin.space(**C**) and $\mathbf{b_i} \in$ lin.space(**C**). This representation is in general not unique (Schrijver, 1999). But if **C** is pointed, then lin.space(**C**)={0} and the extreme rays form the unique, minimal generating set of **C**. This issue was discussed in chapter III.

In this case, the minimal generating set of **F** could be obtained solving a vertex enumeration problem.

Parametric description: finding generating vectors

A set of generating vectors for **F** can be obtained as follows: (1) encode the convex polyhedron **F** as a convex polyhedral cone **C**, (2) obtain a generating set for **C**, and (3) translate the generating set of **C** into a generating set of **F**.

The polyhedron **F** can be encoded as an auxiliary cone **C** introducing a scalar variable λ to transform the system of linear inequalities **A**·**x** ≥ **b** into an equivalent homogeneous one (Le Verge, 1999):

$$\mathbf{C}=\left\{\begin{pmatrix}\mathbf{v}\\ \lambda\end{pmatrix}\in\mathbf{R}^{n+1}\ \mathbf{:}\ \begin{array}{c}\begin{pmatrix}\mathbf{N} & 0\end{pmatrix}\cdot\begin{pmatrix}\mathbf{v}\\ \lambda\end{pmatrix}=\mathbf{0}\\ \begin{pmatrix}\mathbf{D} & \mathbf{0}\\ \mathbf{Q} & -\mathbf{v}_{\mathbf{m}}^{\mathbf{m}}\\ -\mathbf{Q} & \mathbf{v}_{\mathbf{m}}^{\mathbf{M}}\end{pmatrix}\cdot\begin{pmatrix}\mathbf{v}\\ \lambda\end{pmatrix}\geq\mathbf{0}\end{array}\right\} \qquad (12)$$

If a vector **v** satisfies (7), then (**v**, λ) satisfies (12) and, conversely, each solution (**v**, λ) of (12) yields the solution **v**/λ of (7), unless λ = 0.

Now we define the function Ψ(**v**, λ) that allows us to transform a given generating set of the cone **C** to one of the polyhedron **F**:

$$\Psi(\mathbf{v},\lambda)=\begin{cases}\mathbf{v} \quad \text{if} \quad \lambda=0\\ \mathbf{v}/\lambda \qquad \text{else}\end{cases} \qquad (13)$$

If $\{\mathbf{g_1},\ldots,\mathbf{g_s}\}$ is a generating set of **C**, then $\Psi(\{\mathbf{g_1},\ldots,\mathbf{g_s}\})$ is a generating set of **F**. For each, $\mathbf{g_i}=(\mathbf{v_i},\lambda_i)^T$, then $\Psi(\mathbf{g_i})$ is an unbounded generator if $\lambda_i = 0$. If $\lambda_i \neq 0$, $\Psi(\mathbf{g_i})$ is bounded.

In this way, the problem of finding a generating set for the polyhedron **F** has been transformed into that of finding a generating set for the cone **C**, a set of vectors $\{\mathbf{g_1}, \ldots, \mathbf{g_s}\}$ that fulfill the following:

$$\mathbf{C}=\left\{\sum_{\mathrm{j}}^{\mathrm{s}}\alpha_j\cdot\mathbf{g}_{\mathrm{j}},\quad \alpha_j\geq 0\right\} \qquad (14)$$

A minimal generating set of convex polyhedral cone can be found applying the Chernikova's algorithm (Chernikova, 1965; Le Verge, 1994) with the software ccd (Fukuda, 1996). The pathways described in chapter III are also generating sets (not always minimal) so they fulfil (14). Elementary modes can be computed with Metatool (Pfeiffer, 1999), cellNetAnalyzer (Klamt, 2007) and , and OptFlux (Rocha, 2010).[1]

Parametric description and the flux-spectrum

Notice that the flux spectrum $\boldsymbol{\mathcal{S}}$ described in a previous section can be obtained from the explicit description of **F**. The bounds v_i^m and v_i^M can be directly obtained from the set of generating vectors $\{\mathbf{h_k}\}$ and $\{\mathbf{x_j}\}$ as follows:

$$
\begin{aligned}
v_i^m &= \begin{cases} -\infty & \text{if} \quad \exists \mathbf{h_k} \ \forall h_{k,i} < 0 \\ \min\left\{x_{1,i} \ldots x_{s,i}\right\} & \text{else} \end{cases} \\
v_i^{\mathrm{M}} &= \begin{cases} \infty & \text{if} \quad \exists \mathbf{h_k} \ \forall h_{k,i} > 0 \\ \max\left\{x_{1,i} \ldots x_{s,i}\right\} & \text{else} \end{cases}
\end{aligned}
\qquad (15)
$$

where $x_{j,i}$ denotes the i-th element of $\mathbf{x_j}$ and $h_{k,i}$ denotes the i-th element of $\mathbf{h_k}$.

Notice, however, that computing $\boldsymbol{\mathcal{S}}$ using linear programming (8) is more efficient than using a parametric representation (10) and then obtaining $\boldsymbol{\mathcal{S}}$ by means of (15).

Parametric description and the centroid

Although interval estimates are more reliable and richer, sometimes it is useful (or necessary) to combine them with point-wise estimates. One option is calculate a weighted least squares solution (the one given by TMFA, but considering irreversibility constraints). Another sensible choice is the centroid of **F**, a flux vector "surrounded" by all the feasible states within **F**.[2]

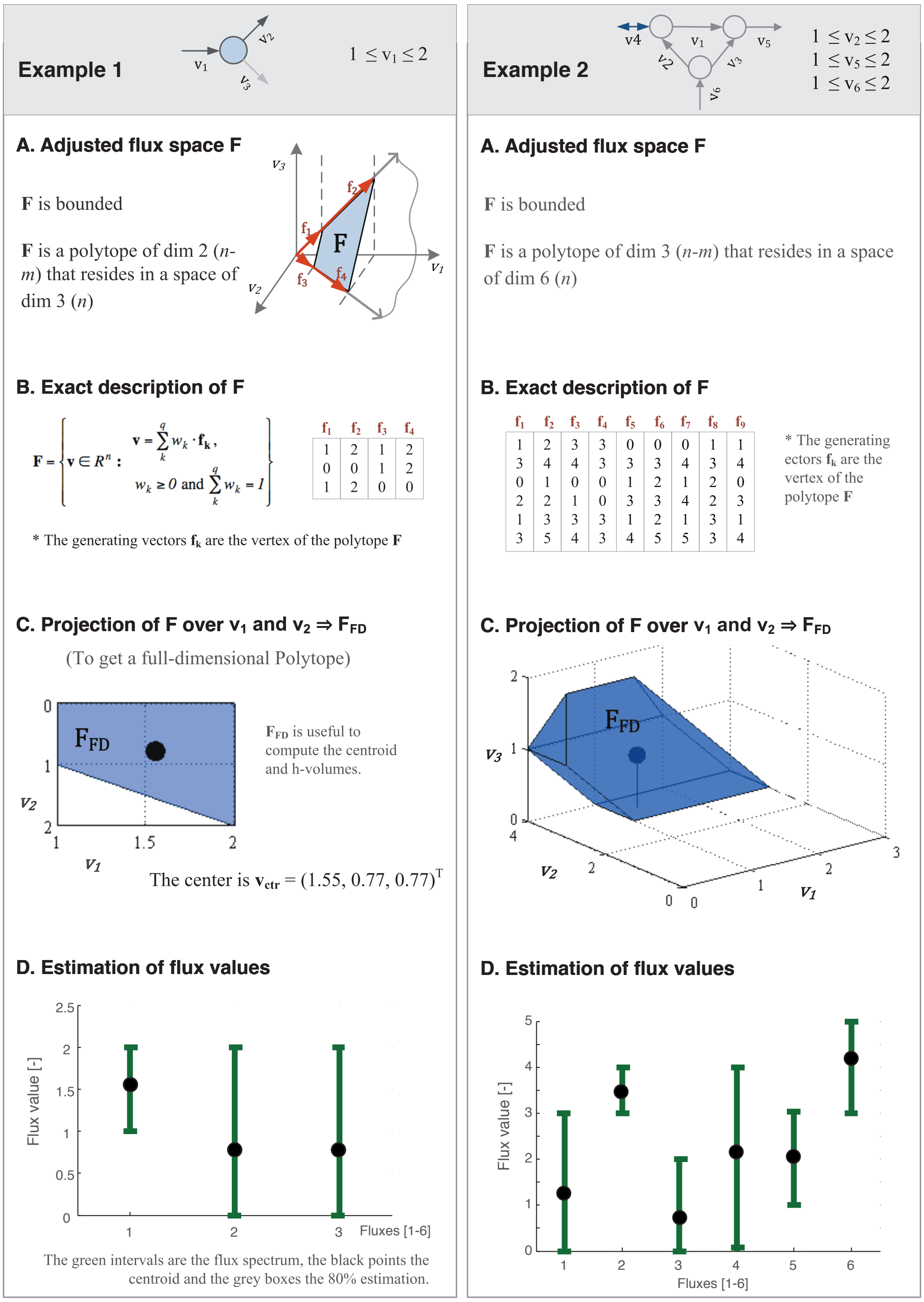
A three-step procedure can be used to obtain the centroid of a bounded flux space **F**, (1) project **F** into a full-dimensional polytope $\mathbf{F_{FD}}$, (2) compute the centroid of $\mathbf{F_{FD}}$, (3) and then recover the centroid of **F** from it.

[1] Some tools require that inequalities appear in (12) as diagonal n×n-matrix **H** with $\mathbf{H_{ii}} = 1$ or $\mathbf{H_{ii}} = 0$. To satisfy this condition, slack variables $\mathbf{s_1}$ and $\mathbf{s_2}$ as follows:

$$\mathbf{Q}\cdot\mathbf{v} \geq -\mathbf{v}_{\mathbf{m}}^{\mathbf{M}} \quad \rightarrow \quad \mathbf{Q}\cdot\mathbf{v} + \mathbf{v}_{\mathbf{m}}^{\mathbf{M}} = \mathbf{s}_1, \quad \mathbf{s}_1 \geq \mathbf{0} \qquad (\mathbf{s}_2 \text{ is defined in analogy, but to } \mathbf{v}_{\mathbf{m}}^{\mathbf{m}})$$

Hence, the cone **C** is reformulated as $\mathbf{C^*} = \{\mathbf{x} \,|\, \mathbf{A}\cdot\mathbf{x} = 0, \mathbf{H}\cdot\mathbf{x} \geq 0\}$, with $\mathbf{x} = (\mathbf{v} \ \ \mathbf{s_1} \ \ \mathbf{s_2} \ \ \lambda)^{\mathrm{T}}$.

[2] The flux-spectrum should be also computed to check that the centroid represents the whole current flux space reasonably well. If the intervals of the flux-spectrum are large—indicating that the estimation is imprecise—the centroid, or any other point-wise estimation, will be unreliable.



Example 1

$1 \leq v_1 \leq 2$

A. Adjusted flux space F

F is bounded

F is a polytope of dim 2 (*n*-*m*) that resides in a space of dim 3 (*n*)

B. Exact description of F

$$\mathbf{F} = \left\{ \mathbf{v} \in R^n : \begin{array}{l} \mathbf{v} = \sum_k^q w_k \cdot \mathbf{f_k}\,, \\ w_k \geq 0 \text{ and } \sum_k^q w_k = 1 \end{array} \right\}$$

f_1	f_2	f_3	f_4
1	2	1	2
0	0	1	2
1	2	0	0

* The generating vectors $\mathbf{f_k}$ are the vertex of the polytope **F**

C. Projection of F over v_1 and v_2 ⇒ F_{FD}

(To get a full-dimensional Polytope)

F_{FD} is useful to compute the centroid and h-volumes.

The center is $\mathbf{v_{ctr}} = (1.55, 0.77, 0.77)^T$

D. Estimation of flux values

The green intervals are the flux spectrum, the black points the centroid and the grey boxes the 80% estimation.

Example 2

$1 \leq v_2 \leq 2$
$1 \leq v_5 \leq 2$
$1 \leq v_6 \leq 2$

A. Adjusted flux space F

F is bounded

F is a polytope of dim 3 (*n*-*m*) that resides in a space of dim 6 (*n*)

B. Exact description of F

f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9
1	2	3	3	0	0	0	1	1
3	4	4	3	3	3	4	3	4
0	1	0	0	1	2	1	2	0
2	2	1	0	3	3	4	2	3
1	3	3	3	1	2	1	3	1
3	5	4	3	4	5	5	3	4

* The generating ectors $\mathbf{f_k}$ are the vertex of the polytope **F**

C. Projection of F over v_1 and v_2 ⇒ F_{FD}

D. Estimation of flux values

Figure 4.4. Parametric representation of the current flux space **F** and the centroid.

$\mathbf{F}$ is a (*n-m*)-dimensional polytope in an *n*-dimensional space (Figure 4.4A). The equalities in (1) imply that *m* fluxes can be considered as dependent ($\mathbf{v_D}$) of others ($\mathbf{v_I}$), so we can project $\mathbf{F}$ over (*n-m*) independent fluxes to obtain a full-dimensional polytope $\mathbf{F_{FD}}$ (Figure 4.4C). Notice that reordering rows, equation (1) can be reformulated as, $(\mathbf{N_I}\ \mathbf{N_D})*(\mathbf{v_I}\ \mathbf{v_D})^T = 0$, so that each $\mathbf{v}$ can be reconstructed from $\mathbf{v_I}$ (coordinates of $\mathbf{F_{FD}}$), as follows:

$$\mathbf{v} = \begin{pmatrix} \mathbf{v_I} \\ \mathbf{v_D} \end{pmatrix} = \Omega \cdot \mathbf{v_I}, \qquad \Omega = \begin{pmatrix} \mathbf{I} \\ -\mathbf{N_I^{-1}} \cdot \mathbf{N_D} \end{pmatrix} \tag{16}$$

Taking this into account, we can choose $\mathbf{v_I}$ to project $\mathbf{F}$ over the first *n-m* coordinates such that $\mathbf{N_D}$ is invertible (Braunstein, 2008). The vertexes that define $\mathbf{F_{FD}}$ can be extracted from the vertexes $\{\mathbf{x_1},..., \mathbf{x_s}\}$ of $\mathbf{F}$, by simply removing the rows that correspond to dependent fluxes (Figure 4.4B).

At this point, we can obtain the centroid $\mathbf{c_{FD}}$ of the polytope $\mathbf{F_{FD}}$, for example, dividing it into simplices and determining the weighted sum of their centroids[1]. Finally, the centroid $\mathbf{c}$ of $\mathbf{F}$ is recovered from $\mathbf{c_{FD}}$ from (16).

Remark on computation efficiency. The procedure outlined to compute the centroid requires a parametric description of $\mathbf{F}$, its vertexes $\{\mathbf{x_1},..., \mathbf{x_s}\}$, which computation is expensive. Indeed, it has been recently proved that computing the centroid is a NP-hard problem (Rademacher, 2007). As an alternative, the centroid can be approximated sampling random points from the polytope, the number of samples depending polynomially on the desired approximation (Elbassioni, 2009; Kannan, 1997).

4.4 Case study: cultivation of CHO cells

In this section we will use the example of Chinese hamster ovary (CHO) cells cultivated in batch mode in stirred flasks to illustrate the methods presented along the chapter.

- We introduce and describe the example of CHO cells, showing how to formulate the flux estimation problem.

- For the sake of comparison, we include the results given by traditional metabolic flux analysis (TMFA).

[1] We use a MATLAB implementation of this method written by Michael Kleder, which is based on the Quickhull algorithm (Bradford, 1995).

- We demonstrate that the flux-spectrum (FS-MFA) can be of use under data scarcity, both in scenarios lacking measurements (where TMFA cannot be applied), and in scenarios where measurements are uncertain.

After this, a second case study will be devoted to show the limitations of TMFA, provide further validation of FS-MFA, and discuss its benefits in other scenarios.

Preparation: metabolic network and constraint-based model

We will use the metabolic network depicted in Figure 4.5, which has been taken from (Provost, 2004), but with reactions for nucleotide synthesis taken from (Provost, 2006a). The network describes the metabolism concerned with the two main energetic nutrients, glucose and glutamine. The metabolism of the amino-acids provided by the culture medium is not included. Four pathways are considered: the glycolysis, the glutaminolysis, the TCA cycle and the nucleotides synthesis. The complete lists of compounds is given in tables 2 and 3, and the list of reactions in Table 4.4.

The stoichiometric matrix **Nᵢ** to define (1) or (7) is thus the following:

$$\mathbf{N_i} = \left(\begin{array}{cccccccccccccccccc}
1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 \\
0 & 0 & 0 & 0 & 1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1 & 1 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & -1
\end{array}\right)
\begin{array}{|c}
\text{G6P} \\ \text{DAP} \\ \text{G3P} \\ \text{R5P} \\ \text{Pyr} \\ \text{ACA} \\ \text{Cit} \\ \text{AKG} \\ \text{Mal} \\ \text{Glu} \\ \text{Osa} \\ \text{Asp}
\end{array} \qquad (17)$$

The information in the matrix **Nᵢ** defines the stoichiometric constraints (1):

$$\mathbf{N_i} \cdot \mathbf{v_i} = 0 \qquad (18)$$

The extracellular fluxes for glucose (v_G), lactate (v_L), and alanine (v_A) coincide with three fluxes of the network, and they need to be incorporated by inspection of the network. It is also natural to assume that the formation of purine and pyrimidine nucleotides is the same. As a result, four new equations are incorporated (Provost, 2004):

$$\begin{aligned} -v_G &: \ v_1 \\ v_L &: \ v_6 \\ v_A &: \ v_7 \end{aligned}$$

$$\begin{aligned} v_{NH4} &: \ v_{19} = v_{15} + v_{16} \\ -v_Q &: \ v_{20} = v_{16} + v_{17} + 2 \cdot v_{18} \\ v_{CO2} &: \ v_{21} = v_3 + v_8 + v_{10} + v_{11} + v_{13} - v_{18} \\ v_{Extra} &: \ v_{22} = 0 = v_{17} - v_{18} \end{aligned}$$

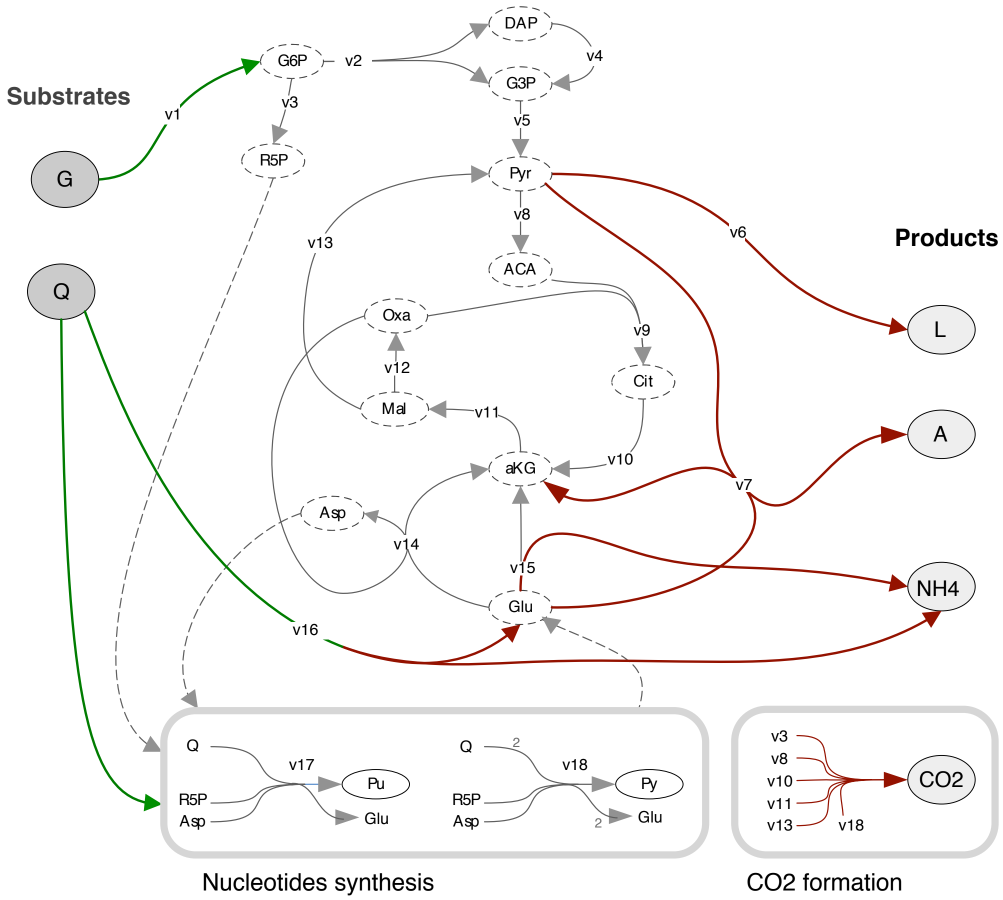


Figure 4.5. Simplified metabolic network of CHO cells metabolism (Provost, 2004).

Table 4.2. List of substrates and products.

G	Glucose	initial substrates	Q	Glutamine	initial substrates
L	Lactate	extracell. product	A	Alanine	extracell. product
NH4	Ammonia	extracell. product	CO2	Carbon dioxide	extracell. product
Pu	Purine	intracell. product	Py	Pyrimidine	intracell. product

Table 4.3. List of balanced metabolites.

G6P	Glucose-6-phosphate	G3P	Glyceraldehyde-3-phosphate
DAP	Dihydroxy-acetone Phosphate	Pyr	Pyruvate
R5P	Ribose-5-Phosphate	ACA	Acetyl-coenzyme A
Cit	Citrate	Oxa	Oxaloacetate
Mal	Malate	aKG	α-ketoglutarate
Glu	Glutamate	Asp	Aspartate

For convenience, these extracellular fluxes and the constraints regarding the nucleotides can be represented defining a 4×18 matrix $\mathbf{N_e}$ fulfilling the equation:

$$\mathbf{v_e} = \mathbf{N_e} \cdot \mathbf{v} \tag{19}$$

$$\text{with } \mathbf{N_e} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

In this way (18) and (19) can be joined to define an extended homogeneous system of linear equations where all the extracellular fluxes appear as a unique flux in **v**:

$$\mathbf{N} \cdot \mathbf{v} = \mathbf{0}, \quad \text{with} \quad \mathbf{N} = \begin{pmatrix} \mathbf{N_i} & \mathbf{0} \\ \mathbf{N_e} & \mathbf{I} \end{pmatrix}, \ \mathbf{v} = \begin{pmatrix} \mathbf{v_i} \\ \mathbf{v_e} \end{pmatrix} \tag{20}$$

The extended system has 16 metabolites (*mx*) and 22 reactions (*nx*). The system is underdetermined and has 6 degrees of freedom.

Then we assume that all the reactions are constrainted to be positive, so that the matrix **D** in $\mathbf{D} \cdot \mathbf{v} \geq 0$ is a *n*-dimensional diagonal matrix of "1". Many reactions in the network are indeed reversible (e.g., v_2, v_4, v_5, v_6 and v_7), but herein we consider only one possible direction, the one exhibited during the growth phase (Provost, 2004; Provost, 2006a, 2006b). Therefore, the model will be valid in this phase, but not under different conditions (e.g., when glucose is exhausted and lactate and alanine are consumed instead of produced).

This way, we have completely defined the *flux space* of admissible steady state flux vectors—as in equation (5)—that corresponds to the given network.

Fluxes estimated with Traditional MFA

In (Provost, 2004), experimentally measured values are given for 6 fluxes (in bold in table 4.4). In this case, the rank of $\mathbf{N_u}$, 16, is equal to the number of unknown fluxes, 22-6, so the MFA problem is determined and not redundant (see sections 4.2 and 4.3 for details). The unique flux vector fulfilling (3) has been estimated using traditional MFA as explained in chapter II. The results, given in tables 4 and 5 (reference column), are exactly those reported by Provost and Bastin (2004). To provide further validation of these data, experimental measurements and estimated fluxes from other studies with mammalian cells have been included in table 4.4.

Table 4.4. Production/consumption rates and reaction fluxes.

	dataset P		dataset G		dataset B
Production and uptake rates	[mM/(d·10^9 cells)]	[%]	32h [%]	24h [%]	[%]
G (*v_1*)	**4.05***	**100****	**100****	**100****	**100****
Q (*v_{20}*)	**1.18**	**29.14**	-	-	**17.00**
L (*v_6*)	**7.39**	**182.47**	**173.91**	**177.08**	**102.00**
A (*v_7*)	**0.26**	**6.42**	6.5217	10.41	**7.00**
NH4 (*v_{19}*)	**0.96**	**23.70**	-	-	-
CO2 (*v_{21}*)	2.61	64.44	-	-	**126.00**
Py-Pu (*v_{22}*)	**0**	0	-	-	.
Reaction fluxes					
1: G→G6P	4.05 [a]	**100.00**	**100**	**100**	**100.00**
2: G6P→G3P+DAP	3.76	92.84	-	-	-
3: G6P→R5P+CO2	0.28	6.91	-	-	76
4: DAP→G3P	3.76	92.84	-	-	-
5: G3P→Pyr	7.53	185.93	-	-	-
6: Pyr→L	7.39	**182.47**	**173.91**	**177.08**	**102**
7: Pyr+Glu→A+aKG	0.26	**6.42**	6.5217	10.42	**7**
8: Pyr→ACA+CO2	0.34	8.40	13.043	33.33	8
9: Oxa+ACA→Cit	0.34	8.40	-	-	27
10: Cit→aKG+CO2	0.34	8.40	-	-	6
11: aKG→Mal+CO2	1.10	27.16	23.913	62.5	-
12: Mal→Oxa	0.63	15.56	15.217	45.83	-
13: Mal→Pyr+CO2	0.47	11.60	2.17	25	-
14: Oxa+Glu→Asp+aKG	0.28	6.91	8.69	45.83	6
15: Glu→aKG+NH4	0.20	4.94	6.52	10.42	-1
16: Q→Glu+NH4	0.75	18.52	4.34	31.25	18
17: R5P+Asp+Q→Pu + Glu	0.14	3.46	-	-	-
18: R5P+Asp+2Q+CO2→Py+Glu+Mal	0.14	3.46	-	-	-

*Experimentally measured values are in bold. **Fluxes are represented as percentage of glucose uptake.

P: Provost & Bastin (2004). Data from the growth phase of CHO cells cultivated in batch mode (μ=0.69d^{-1}). Measurements and fluxes computed with traditional MFA.

G: Gambhir et al. (2003). Data from a cultivation of hybridoma cells in batch mode (μ=0.72d^{-1}) at two time instants of the growth phase. Measurements and fluxes calculated with a variant of traditional MFA based on carbon and nitrogen balances.

B: Bonarius et al. (1996). Data from a cultivation of hybridoma cells in continuous mode (μ=0.83d^{-1}).

Comments: The data correspond to experiments with differences in cultivation modes, medium, type of cells, bioreactor conditions, etc. Nevertheless, datasets P and G (32h) show a good agreement for all fluxes except 13 and 16. These two fluxes are closer lla to G (24h) suggesting that dataset P corresponds to cells at a state between the two time instants. Dataset B corresponds to an experiment in continuous mode, where cells exhibit a different metabolic state (the measured values diverge from P and G).

FS-MFA: scenarios lacking measurements

To illustrate one of the benefits of FS-MFA, we have estimated the fluxes in underdetermined scenarios that use different subsets of measurements. The results are given in table 4.5 (columns L1-L3) and compared with those obtained with TMFA in the previous section, where all the measurements were available.

Scenario L1. Let us consider that only 4 fluxes are measurable instead of 6: glucose, alanine, glutamine and CO_2. The MFA problem is underdetermined, so TMFA cannot be applied, and there are no calculable fluxes.[1] However, interval estimates can be obtained with FS-MFA solving the LP problems in (8). The results, depicted in Figure 4.6, show that the flux-spectrum intervals are accurate, similar to the point-wise estimates given by TMFA, but using 4 measurements instead of 6. Some interval estimates are wider (v_{15} and v_{19}), but most of them are precise and narrow.

Scenario L2. Now we consider that a different set of 4 fluxes are measured. In this case the estimates are slightly worse: the average interval size is 13% instead of 7.3%. This suggests that fixing the value of v_{21} (CO_2) impose a stronger constraint than fixing v_6 (L)—at least in combination with the other measurements. This is indeed reasonable, because CO_2 participates in 6 reactions and lactate just in 1.

Scenario L3. Although 5 fluxes are now measured, the MFA problem remains underdetermined, so TMFA cannot be used. However, as shown in Figure 4.6, the interval estimates given by FS-MFA are practically equivalent to those obtained with MFA.

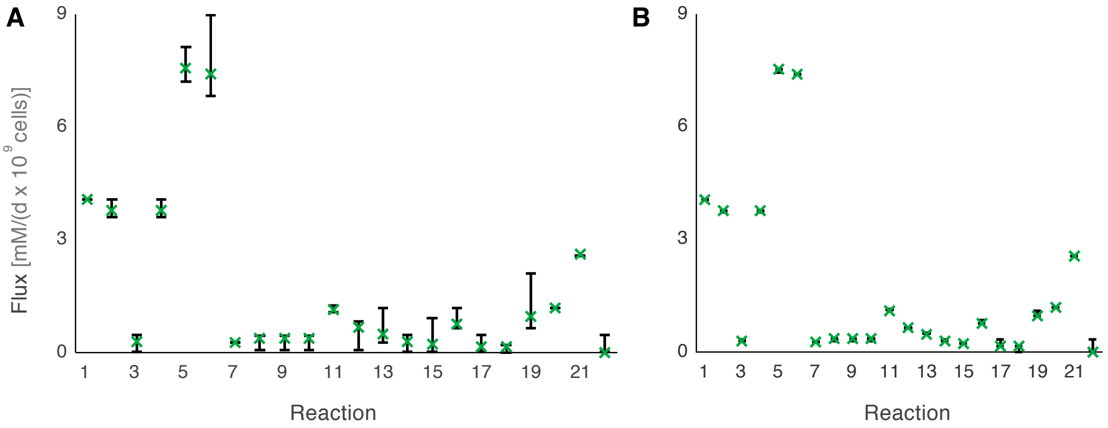


Figure 4.6. FS-MFA estimates in two underdetermined scenarios: L1 and L3. (A) FS-MFA applied from only 4 measured fluxes: v_1 (G), v_7 (A), v_{20} (Q), and v_{21} (CO_2). (B) FS-MFA applied from 5 measured fluxes v_1 (G), v_6 (L), v_7 (A), v_{20} (Q) and v_{21} (CO_2). In both figures the fluxes estimated with TMFA from 6 measured fluxes is included for the sake of comparison (crosses).

[1] The kernel of $\mathbf{N_u}$, has no null rows. See chapter II for details.

Table 4.5. Quantitative comparison of FS-MFA estimates for different cases.

	L1 *(v_1, v_6, v_{20} and v_{21})*		L2 (v_1, v_6, v_7 and v_{20})		L3 *($v_1, v_6, v_7, v_{20}, v_{21}$)*		U (L3 + uncertainty)		Ref.
Reactions	Flux [a]	IS [%c]	Flux [a]	IS [%c]	Flux [a]	IS [%c]	Flux [a]	IS [%]	Flux [a]
1: G→G6P	**4.0546**	-	**4.0546** [b]	-	**4.0546**	-	[3.85, 4.25]	5.49	***4.05***
2: G6P→G3P+DAP	[3.59, 4.05]	6.21	[3.64, 4.05]	5.52	[3.71, 3.76]	0.69	[3.48, 4.01]	7.06	*3.76*
3: G6P→R5P+CO2	[0, 0.4587]	6.21	[0, 0.40]	5.52	[0.28, 0.33]	0.69	[0.10, 0.49]	5.24	*0.28*
4: DAP→G3P	[3.59, 4.05]	6.21	[3.64, 4.05]	5.52	[3.71, 3.76]	0.69	[3.48, 4.00]	7.06	*3.76*
5: G3P→Pyr	[7.19, 8.10]	12.41	[7.29, 8.10]	11.04	[7.43, 7.53]	1.38	[6.96, 8.0]	14.11	*7.53*
6: Pyr→L	[6.82, 8.96]	28.97	**7.39**	-	**7.39**	-	[7.02, 7.76]	10.01	***7.39***
7: Pyr+Glu→A+aKG	**0.26**	-	**0.26**	-	**0.26**	-	[0.25, 0.28]	0.36	***0.26***
8: Pyr→ACA+CO2	[0.06, 0.42]	4.97	[0, 1.63]	22.08	[0.28, 0.34]	0.83	[0.09, 0.49]	5.41	*0.34*
9: Oxa+ACA→Cit	[0.06, 0.42]	4.97	[0, 1.63]	22.08	[0.28, 0.34]	0.83	[0.09, 0.49]	5.41	*0.34*
10: Cit→aKG+CO2	[0.06, 0.42]	4.97	[0, 1.63]	22.08	[0.28, 0.34]	0.83	[0.09, 0.49]	5.41	*0.34*
11: aKG→Mal+CO2	[1.06, 1.24]	2.48	[0.64, 2.81]	29.44	[1.10, 1.13]	0.41	[1.00, 1.25]	3.36	*1.10*
12: Mal→Oxa	[0.06, 0.82]	10.35	[0.36, 1.63]	17.17	[0.62, 0.63]	0.14	[0.30, 0.89]	7.89	*0.63*
13: Mal→Pyr+CO2	[0.26, 1.18]	12.41	[0.27, 1.18]	12.27	[0.47, 0.51]	0.55	[0.25, 0.89]	8.65	*0.47*
14: Oxa+Glu→Asp+aKG	[0, 0.45]	6.21	[0, 0.40]	5.52	[0.28, 0.33]	0.69	[0.10, 0.49]	5.24	*0.28*
15: Glu→aKG+NH4	[0, 0.91]	12.41	[0.01, 0.91]	12.27	[0.20, 0.24]	0.55	[0, 0.63]	8.54	*0.20*
16: Q→Glu+NH4	[0.63, 1.18]	7.45	[0.64, 1.18]	7.36	[0.75, 0.84]	1.24	[0.60, 1.07]	6.38	*0.75*
17: R5P+Asp+Q→Pu	[0, 0.45]	6.21	[0, 0.4079]	5.52	[0.14, 0.33]	2.62	[0.05, 0.49]	5.97	*0.14*
18: R5P+Asp+2Q→Py	[0, 0.18]	2.48	[0, 0.1813]	2.45	[0, 0.14]	1.93	[0, 0.19]	2.68	*0.14*
19:→ NH4	[0.63, 2.10]	19.86	0.65, 2.10]	19.63	[0.96, 1.09]	1.79	[0.60, 1.69]	14.82	***0.96***
20:→ Q	**1.18**	-	**1.186**	-	**1.18**	-	[1.12, 1.24]	1.60	***1.18***
21:→ CO2	**2.55**	-	[1.28, 7.26]	80.96	**2.55**	-	[2.42, 2.68]	3.46	*2.55*
22: Pu-Py (constraint)	[0, 0.45]	6.21	[0, 0.40]	5.52	[0, 0.33]	4.55	[0, 0.49]	6.70	***0.00***
Mean		7.31		13,2		0.93		6.40	
Standard Deviation		6.89		17,39		1.06		3.48	

[a] in mM/(d*10^9 cells). [b] measured values are in bold. [c] interval sizes w.r.t. biggest measured flux (v_7). A: Underdetermined case (2 dof). B: Underdetermined case (2 dof). C: Underdetermined case (1 dof). Column D: C with uncertainty band of ±5%. Reference: determined case, fluxes computed with traditional MFA.

FS-MFA: scenarios of measurements uncertainty

As explained in the previous section, one of the benefits of FS-MFA is that it considers the uncertainty of the measurements, which is also transferred to the estimates, and explicitly reflected in their intervals.

Scenario U. We consider the scenario L3, but we incorporate uncertainty adding a band of of ±5% around the measured values. After this, the flux-spectrum intervals are obtained as usual (8). The obtained interval estimates, given in table 4.5, are wider, but still useful: the interval sizes range between 1.6% and 14.82% instead of 0.14% and 4.5%, the average interval size is 6.4%, and only 8 intervals are wider than 10%.

FS-MFA: dealing with reversible reactions

In the previous examples all reactions have been assumed to be irreversible, but FS-MFA can be applied if this is not the case. For instance, let us consider that reactions 2, 4, 5, 6 and 7 are reversible. In this case, FS-MFA estimation in the scenario L3 gives exactly the same results. The same happens if the measured fluxes are $\{v_1, v_6, v_{20}, v_{21}\}$ or $\{v_1, v_6, v_7, v_{21}\}$. However, this is not always the case. For instance, when the measured fluxes are $\{v_1, v_{20}, v_{21}, v_{22}\}$, the intervals for fluxes v_7, v_8, v_{15} and v_{19} are unbounded. Indeed, there is a route involving the reactions 7, 8, 15, and 19 that transforms alanine into lactate through 7, and since none of the fluxes is measured, the flux through the route cannot be bounded.

4.5 Case study: *C. glutamicum*

In this section we will use a a classical model of *Corynebacterium glutamicum* and experimental data from a batch fermentation to illustrate some limitations of traditional metabolic flux analysis (TMFA), and show that it can be overcome using FS-MFA.

- We validate FS-MFA against experimentally measured fluxes.
- We show that TMFA is unreliable if there are not redundant measurements.
- We show that even with redundant measurements, TMFA point-wise estimates can be deviated due to uncertainty. Conversely, FS-MFA is more reliable because the interval estimates are only as precise as allowed by the uncertainty.
- We demonstrate that FS-MFA can provide good estimates even if only a few external metabolites measured—when the MFA problem is underdetermined and TMFA cannot be used.

Preparation: metabolic network and constraint-based model

Corynebacterium glutamicum is a glutamic acid bacteria used to produce lysine by microbial fermentation from glucose. A stoichiometric network for this bacteria has been taken from (Gayen, 2006), but it is a slight variation of the one constructed by Vallino et al. (1994). The reactions considered to describe the primary metabolism of *C. glutamicum* necessary to support lysine and biomass synthesis from glucose are thus included in the network. A reaction of ATP dissipation is included to allow variations of the maintenance related ATP consumption and the operation of futile cycles. A closed balance was assumed for NADPH in the works by Gayen et al. and Vallino et al. However, since assumption has been questioned (Yang, 2006; Wittmann, 2002; Marx, 1996), so we decided to remove the balance of NADPH from the network.

The network considers 41 reactions and 39 metabolites (tables 6, 7, 8 and 9). There are 4 redundant mass balances[1], and therefore the row-rank of the network is 36 and it has 5 degrees of freedom. The corresponding 36×41 stoichiometric matrix **N** is

Table 4.6. Extracellular metabolites.

BIOMASS	Glucose	LYSI	Lysine
CO2	Carbon dioxide	NH3	Ammonia
GLC	Glucose	TREHAL	Trehalose
H2O	Water		

Table 4.7. List of internal metabolites.

ADP	Adenosine diphosphate	GLUT	Glutamate
AKG	Kalpha-etoglutaric acid	OAA	Oxalate
AKP	2-Amino-6-ketopimelate	PEP	Phosphoenolpyruvate
ALA	Alanine	PYR	Pyruvate
ASP	Aspirate	RIB5P	Ribose-5-phosphate
ATP	Adenosine triphosphate	RIBU5P	Ribulose-5-phosphate
E4P	Erythrose--4-phosphate	SED7P	Sedoheptulose-7-phosphate
FAD	Flavin adenine dinucleotide (oxidized)	SUC	Succinate
FADH	Flavin adenine dinucleotide (reduced)	SUCCOA	Succinyl coenzyme A
FRU6P	Fructose-6-phosphate	VAL	Valine
G3P	3-Phosphoglycerate	XYL5P	Xylulose-5-phosphate
GAP	Glyceraldehydes-3-phosphate	ACCOA	Acetyl coenzyme A
GLC6P	Glucose-6-phosphate	COA	Coenzyme A
GLUM	Glutamine	MDAP	Meso-Diaminopimelate
NAD	Nicotinamide adenine dinucleotide (oxidized)		
NADH	Nicotinamide adenine dinucleotide (reduced)		
NADP	Nicotinamide adenine dinucleotide phosphate (oxidized)		
NADPH	Nicotinamide adenine dinucleotide phosphate (reduced)		

[1] Pairs of balanced metabolites that impose the same constraint, for instance ATP and ADP.

given in table 4.10. The vector of reactions irreversibility, which defines the diagonal of the matrix **D**, is also given in table 4.10. These two matrices define the flux space **P** in (5), the constraint-based model that we will use along this section.

Table 4.8. List of considered reactions of the central carbon metabolism of *C. glutamicum*.

System	Reaction
Glucose Phosphotransferase System	1. X_GLC + PEP -> GLC6P + PYR
Storage Compounds; Trehalose	2. 2 GLC6P + ATP <> TREHAL + ADP
EMP Pathway	3. GLC6P <> FRU6P
	4. FRU6P + ATP -> 2 GAP + ADP
	5. GAP + ADP + NAD <> NADH + G3P + ATP
	6. G3P <> PEP + H2O
	7. PEP + ADP -> ATP + PYR
	8. PYR + NADH <> LAC + NAD
Carboxylation reaction	9. PEP + CO2 -> OAA
TCA Cycle	10. PYR + COA + NAD -> ACCOA + CO2 + NADH
	11. ACCOA + OAA + H2O + NADP <> AKG + COA + NADPH + CO2
	12. AKG + COA + NAD -> SUCCOA + CO2 + NADH
	13. SUCCOA + ADP <> SUC + COA + ATP
	14. SUC + H2O + FAD +NAD <> FADH + OAA + NADH
Acetate Production or Consumption	15. ACCOA + ADP <> AC + COA + ATP
Glutamate, Glutamine, Alanine, and Valine Production	16. NH3 + AKG + NADPH <> GLUT + H2O + NADP
	17. GLUT + NH3 + ATP -> GLUM + ADP
	18. PYR + GLUT -> ALA + AKG
	19. 2 PYR + NADPH + GLUT -> VAL + CO2 + H2O + NADP + AKG
Pentose Phosphate Pathway	20. GLC6P + H2O + 2 NADP -> RIBU5P + CO2 + 2 NADPH
	21. RIBU5P <> RIB5P
	22. RIBU5P <> XYL5P
	23. XYL5P + RIB5P <> SED7P + GAP
	24. SED7P + GAP <> FRU6P + E4P
	25. XYL5P + E4P <> FRU6P + GAP
Oxidative Phosphorylation	26. 2 NADH + O2 + 4 ADP -> 2 H2O + 4 ATP + 2 NAD
	27. 2 FADH + O2 + 2 ADP -> 2 H2O + 2 ATP + 2 FAD
Asparate Amino Acid Family	28. OAA + GLUT <> ASP + AKG
	29. ASP + PYR + 2 NADPH + ATP -> AKP + 2 NADP + ADP + H2O
	30. AKP + SUCCOA + H2O + GLUT -> MDAP + COA + AKG + SUC
	31. MDAP -> LYSI + CO2
ATP Dissipation	32. ATP -> ADP
Biomass Synthesis	33. 30 PYR + 21 GLC6P + 7 FRU6P + 150 G3P + 52 PEP + 13 GAP + 332 ACCOA + 126 RIB5P + 80 ASP + 33 LYSI + 446 GLUT + 25 GLUM + 54 ALA + 40 VAL + 100 NADPH + 3000 ATP -> 1000 BIOMAS + 143 CO2 + 100 NADP + 332 COA + 364 AKG + 3000 ADP

Table 4.9. Stoichiometric matrix *C. glutamicum*.

	Irreversible	1	0	0	1	0	0	1	0	1	1	0	1	0	0	0	0	1	1	1	1	0	0	0	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0
	Reaction	GLU	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	O_2	NH_3	BIO	LYS	TRE	CO_2	H_2O	NADPH
1	AC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	AKP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0
	(ACCOA)	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-332	0	0	0	0	0	0	0	0
3	AKG	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	-1	0	1	1	0	0	0	0	0	0	0	0	1	0	1	0	0	364	0	0	0	0	0	0	0	0
4	ALA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-54	0	0	0	0	0	0	0	0
5	ASP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	-80	0	0	0	0	0	0	0	0
6	ATP (ADP)	0	-1	0	-1	1	0	1	0	0	0	0	0	1	0	1	0	-1	0	0	0	0	0	0	0	0	4	2	0	-1	0	0	-1	-3000	0	0	0	0	0	0	0	0
7	BIO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1000	0	0	-1	0	0	0	0	0
8	CO2	0	0	0	0	0	0	0	0	-1	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	143	0	0	0	0	0	-1	0	0
9	COA	0	0	0	0	0	0	0	0	0	-1	1	-1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	332	0	0	0	0	0	0	0	0
10	E4P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	FADH (FAD)	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	FRU6P	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	-7	0	0	0	0	0	0	0	0
13	G3P	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-150	0	0	0	0	0	0	0	0
14	GAP	0	0	0	2	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	1	0	0	0	0	0	0	0	-13	0	0	0	0	0	0	0	0
15	GLC6P	1	-2	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	-21	0	0	0	0	0	0	0	0
16	GLUM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-25	0	0	0	0	0	0	0	0
17	GLUT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	-1	-1	0	0	0	0	0	0	0	0	-1	0	-1	0	0	-446	0	0	0	0	0	0	0	0
18	H2O	0	0	0	0	0	1	0	0	0	0	-1	0	0	-1	0	1	0	0	1	-1	0	0	0	0	0	2	2	0	1	-1	0	0	0	0	0	0	0	0	0	-1	0
19	LAC	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	LYSI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	-33	0	0	0	-1	0	0	0	0
21	MDAP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0
22	NADPH (NADP)	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	-1	0	0	-1	2	0	0	0	0	0	0	0	0	-2	0	0	0	-100	0	0	0	0	0	0	0	1
23	NADH (NAD)	0	0	0	0	1	0	0	-1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	NH3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
25	O2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
26	OAA	0	0	0	0	0	0	0	0	1	0	-1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0
27	PEP	-1	0	0	0	0	1	-1	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-52	0	0	0	0	0	0	0	0
28	PYR	1	0	0	0	0	0	1	-1	0	-1	0	0	0	0	0	0	0	-1	-2	0	0	0	0	0	0	0	0	0	-1	0	0	0	-30	0	0	0	0	0	0	0	0
29	RIB5P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	-1	0	0	0	0	0	0	0	0	0	-126	0	0	0	0	0	0	0	0
30	RIBU5P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	SED7P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	SUC	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
33	SUCCOA	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0
34	TREHAL	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0
35	VAL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	-40	0	0	0	0	0	0	0	0
36	XIL5P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Preparation: measured fluxes of *C. glutamicum*

Experimental data of a batch fermentation of *C. glutamicum* cultured on minimal glucose medium was taken from (Vallino, 1994). There, the fluxes of biomass and several external metabolites (lactate, acetate, glucose, O_2, CO_2, NH_3, lysine, and trehalose) were experimentally measured. The accumulation of lactate and acetate were negligible, so their flux is always zero in this study. The rest of measured fluxes and its standard deviations are given in table 4.10. The high uncertainty of the measurements is illustrated by the 90% confidence intervals (MR90 in table 4.10).

As expected, the original measurements (M) are slightly inconsistent; they do not exactly satisfy the network stoichiometry. We follow two approaches to exploit these inconsistencies and obtain better (adjusted) measured fluxes: (a) Perform a consistency analysis using a χ^2-test, which do not detect gross errors (h = 1.23), and then adjust the measurements using weighted least squares (see chapter II for details). (b) Perform Monte Carlo simulations to compute the ranges that contain those values in M that satisfy the stoichiometry and the reactions irreversibility (5).

TMFA calculations were performed with the three-step procedure described in chapter II, accounting for the standard deviations given in table 4.10. FS-MFA estimates were performed representing the measured fluxes with the intervals of 90% confidence (MR90 in table 4.10).

TMFA and FS-MFA estimates against measurements

In this section we use the experimental measurements described above to validate FS-MFA estimates and to illustrate the limitations of TMFA. We perform 5 batteries (A-D) of estimations using different sets of measurements. In each run, a subset of the seven known fluxes are really used as measurements (as inputs), while the rest are estimated. These estimates are then compared with the experimental values, thus providing a cross-validation of FS-MFA and TMFA.

A. Leave-1-measurement-out. All the MFA problems on this battery are redundant[1], so we could checked if the measurements pass the χ^2-tests. This is the scenario where TMFA is supposed to be reliable, but the results show that in some cases (A1 and A7) its estimations are significantly deviated from the experimental values.

Conversely, FS-MFA intervals show a good agreement in all the cases (although they are sometimes slightly conservative). Notice also that the degree of overestimation of the flux-spectrum intervals varies from some cases to others, which indicates that some measurements impose stronger constraints. For instance, the overestimation in A1 is negligible, but important in A4. Notice also that the centroid, and even the centre of the intervals, are better point-wise estimates than the ones given by TMFA.

[1] There are 5 degrees of freedom and 6 independent measurements.

Table 4.10. Experimentally measured fluxes during a batch fermentation of *C. glutamicum*.

Metabolite (# reaction)		Production/consumption rates or fluxes (mM/h)			
		M *Measurements*	*MR90* *Meas. range 90%*	*CR* *Consistent range*	*WL* *WLSQ adjust.*
GLC (1)	Consump.	40.6±22	[-4.4, 76.8]	[20.1, 35.9]	25
O2 (34)	Consump.	59.2±5.9	[49.5, 68.9]	[49.5, 67.6]	592
NH3 (35)	Consump.	64.8±44	[-7.5, 137]	[8.3, 27.1]	17
LYSE (37)	Production	0.04±.01	[0.02, 0.06]	[0.02, 0.06]	23
TREHAL (38)	Production	0.4±2	[-2.9, 3.7]	[0.04, 3.7]	4
Biomass (36)	Production	21.9±5.4	[13, 30.8]	[13, 30.8]	66
CO2 (39)	Production	61.9±6.2	[51.4, 71.8]	[52.6, 71.8]	618

M: Original measurements and its standard deviation (Vallino, 1994). MR90: Values with 90% of confidence. CR: intervals bracketing the consistent values in M. WL: weighted least squares adjustment.

B. Leave-2-measurements-out. 12 out of 20 MFA problems of this battery are determined, but not redundant. TMFA is typically unreliable in this situation, because measurements consistency cannot be checked, but herein we know a priori that the measurements are consistent, so this problem is avoided. Yet, the results show that in most cases (e.g., B2, B7 or B8), TMFA estimations are deviated from the experimental values. The rest of MFA problems of the battery (B13 to B21) are underdetermined, so TMFA cannot be applied.

On the contrary, FS-MFA is still able to provide valuable estimates, even in those cases that are underdetermined. The centroid is within the measured intervals in 28 out of 42 cases, and always close to them. The flux-spectrum intervals are wider than in the previous case, but still informative.

C. Leave-1-measurement-out (balanced NADPH). At this point we modify the network to include the cofactor NADPH and assume that it is balanced, as done in the original work of Vallino et al. (1994). Again, FS-MFA estimates provide better results than TMFA, particularly in cases C4 and C6.

Notice also that the estimates show a good agreement with the experimental data, thus indicating that the assumption of a balanced NADPH is, at least, compatible with the extracellular behaviour of cells at the given conditions.

D. Leave-2-measurement-out (balanced NADPH). All the estimation problems are redundant, so this is again a scenario where TMFA is supposed to be reliable. However, the results show that TMFA estimates can be highly deviated from the experimental values. On the contrary, FS-MFA estimates fit quite well. Notice also that the flux-spectrum intervals are again precise, indicating that the assumption of a balanced NADPH may be valid.

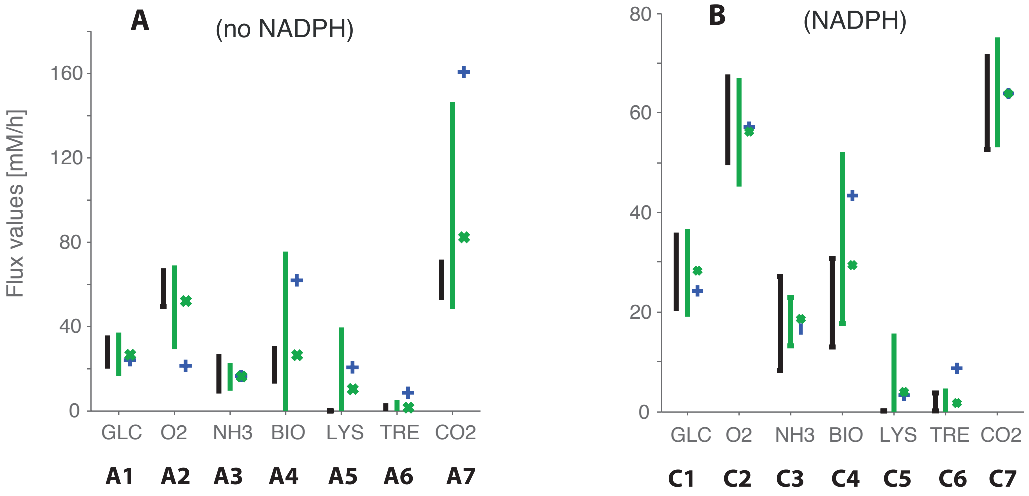


Figure 4.7. FS-MFA and TMFA estimations against experimental measurements. In batteries A and C, 6 out of the 7 measurements are inputs, the remaining one is used for validation. Only fluxes used for validation are depicted. Experimentally measured values (CR) are indicated with a black interval, the flux-spectrum with light green intervals, the centroid with an "x", and TMFA estimate with a "+".

Leave-3-measurement-out (balanced NADPH). Most of the MFA problems of this battery are determined, but not redundant (28), and as expected, TMFA estimates are deviated in many of them. There are also 7 underdetermined problems, where TMFA cannot be applied. FS-MFA estimates are remarkable good for the 35 cases, particularly if one takes into account that only 4 fluxes are measured.

In summary, (i) we have corroborated that TMFA is unreliable when there are not redundant measurements (batteries B and E), (ii) while FS-MFA provides a good estimation, only slightly more imprecise that the obtained when redundant measurements where available. Moreover, (iii) it has been shown that even if there are redundant measured fluxes and its inconsistency is low, TMFA can be unreliable due to the effect of measurements uncertainty (see A2, A4, A7, D10 and D15), but (iv) FS-MFA gave better results for all these batteries.

FS-MFA estimates in data scarce scenarios

This section shows the results given by FS-MFA in some scenarios where TMFA cannot be applied or is unreliable due to a lack of measurable fluxes.

Scenario 1: measuring seven fluxes. The case where all the measured fluxes given in table 4.6 are considered will be used as reference. Again, the results point out that even if there are redundant measurements, the uncertainty may have a significant effect over the estimation of certain fluxes (e.g., v_{32} or v_{40}). FS-MFA copes with this providing interval estimates that are only as precise as allowed by the uncertainty.

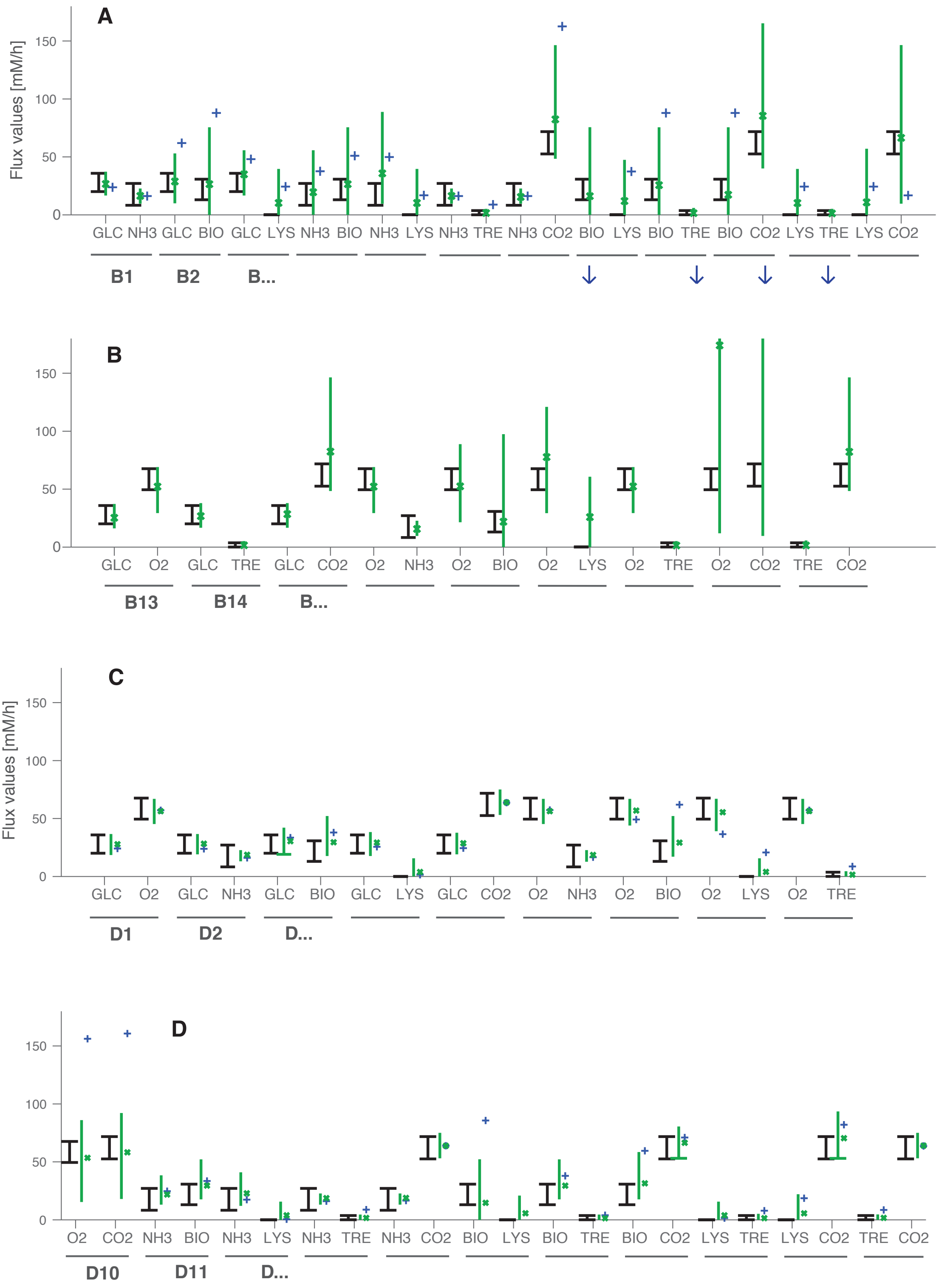


Figure 4.8. FS-MFA and TMFA estimations against measurements. In batteries B and D, 5 out of 7 measurements are inputs, the remaining one is used for validation. Both batteries are equivalent, but in D the cofactor NADPH is assumed to be balanced. Only fluxes used for validation are depicted. Experimentally measured values (CR) are indicated with a black interval, the flux-spectrum with light green intervals, the centroid with an “x”, and TMFA estimate with a “+”.

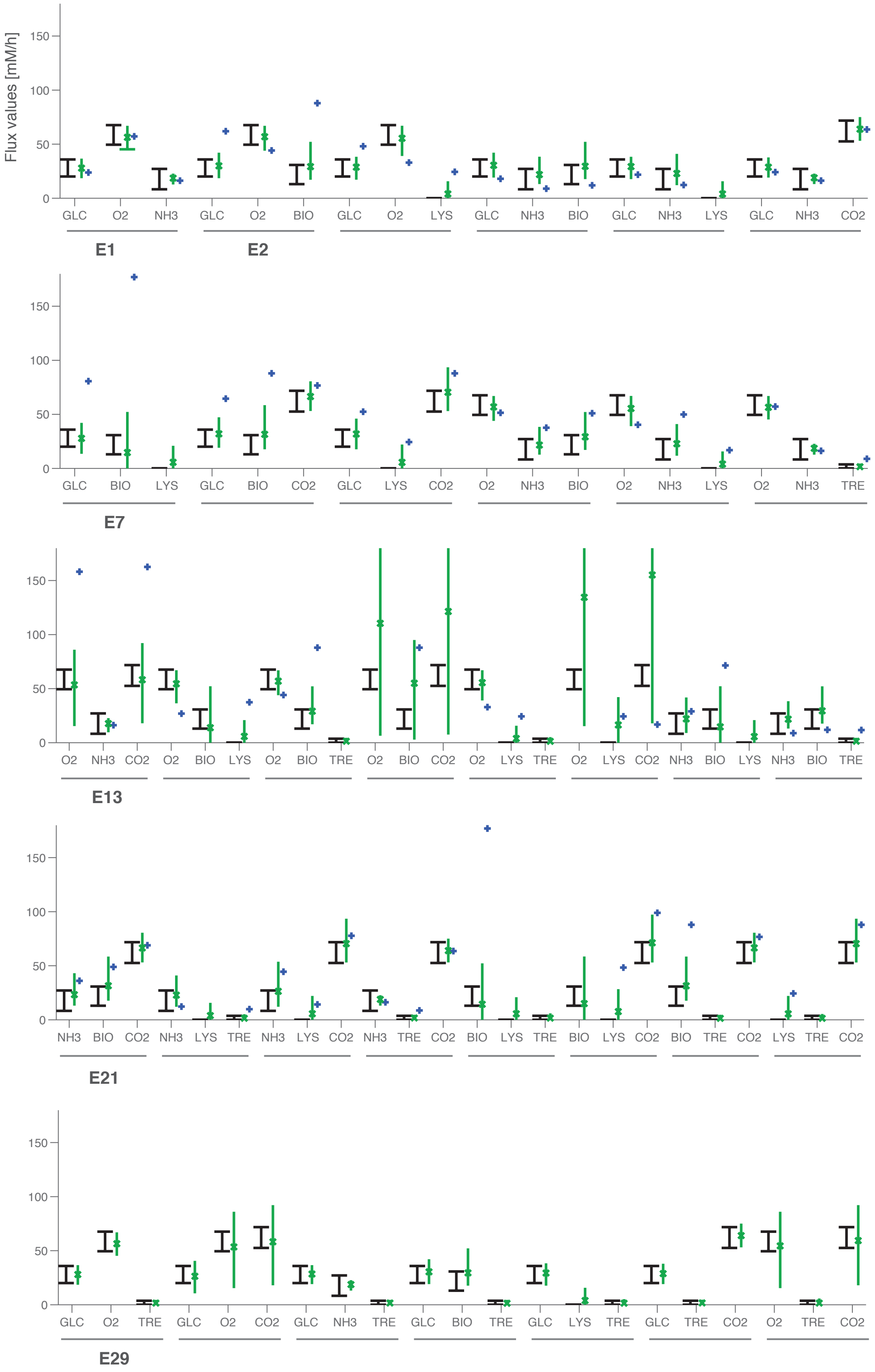


Figure 4.9. FS-MFA and TMFA estimations against experimental measurements. In each battery, 4 out of the 7 measurements are inputs, the remaining three are used for validation. Only fluxes used for validation are depicted. Experimentally measured values (CR) are indicated with a black interval, the flux-spectrum with light green intervals, the centroid with an “x”, and TMFA estimate with a “+”.

Scenario 2: measuring five fluxes. If 5 fluxes are measured {v_{GLC}, v_{O2}, v_{LYSE}, v_{Bio} and v_{CO2}}, the flux estimation problem is not redundant, so TMFA will be unreliable. Nevertheless, the results in Figure 4.10 show that FS-MFA provides a very good estimation. The results are practically equivalent to those obtained in the scenario 1 (the centroid has a mean deviation of 0.054 mM/h with respect S1, for a flux vector with a mean value of 19.85mM/h).

Scenario 3: measuring four fluxes. When 4 fluxes are measured {v_{GLC}, v_{O2}, v_{Bio} and v_{CO2}}, the flux estimation problem is underdetermined, so TMFA cannot be applied. Yet, FS-MFA provides a valuable estimation (see Figure 4.10).

In the scenarios 1 to 3 we have incorporated an artificial flux of NADPH to estimate the total amount being consumed or produced by cells at the given conditions. As it can be seen in Figure 4.10, the value of this flux (v_{41}) is between -39.4 and 13.6 mM/h, indicating that a balanced NADPH, even if it is not the only possibility, is compatible with the measurements and the model. At this point, we repeat the FS-MFA estimations in the same 3 scenarios, but now we assume that the cofactor NADPH is balanced, thus improving the accuracy of the results, if the assumption is indeed acceptable. The results are depicted in Figure 4.11.

Scenario 4: measuring five fluxes (balanced NADPH). The MFA problem is now redundant thanks to the added balance NADPH balance. Interestingly, even if 5 fluxes are measured instead of 7, the obtained flux estimates are similar to those in the reference scenario S1 (the centroid has a mean deviation of 1.4 mM/h).

Scenario 5: measuring four fluxes (balanced NADPH). The flux estimates are similar to those in S1, and much more precise than those obtained in S3 (the mean deviation of the centroid in S5 is 2.5 mM/h, significantly better than the 7.18 mM/h of S3). This suggests that, in this particular case, the assumption of a balanced NADPH is partially overcoming the lack of measurements.

Scenario 6: measuring two fluxes (balanced NADPH). In this case only two fluxes are measured {v_{Bio} and v_{CO2}}, so the MFA problem is underdetermined with two degrees of freedom. Yet, the estimates are similar to those in S1 and practically equivalent to those in S5.

The results show that the actual fluxes can be estimated even if only a few measurements are available. In this particular case, FS-MFA provides an estimate even if only two external fluxes—growth rate and CO2 production—were measured. Clearly, the structure of a highly simplified metabolic network restricts the flux states that cells can show. However, it must be keep in mind that a small network may be biased to fit a particular cell state, thus being valid only under certain conditions. The network model is used herein only as an example, so this problem has not been addressed. However, the validation of reduced networks like this one will be discussed in chapter IX.

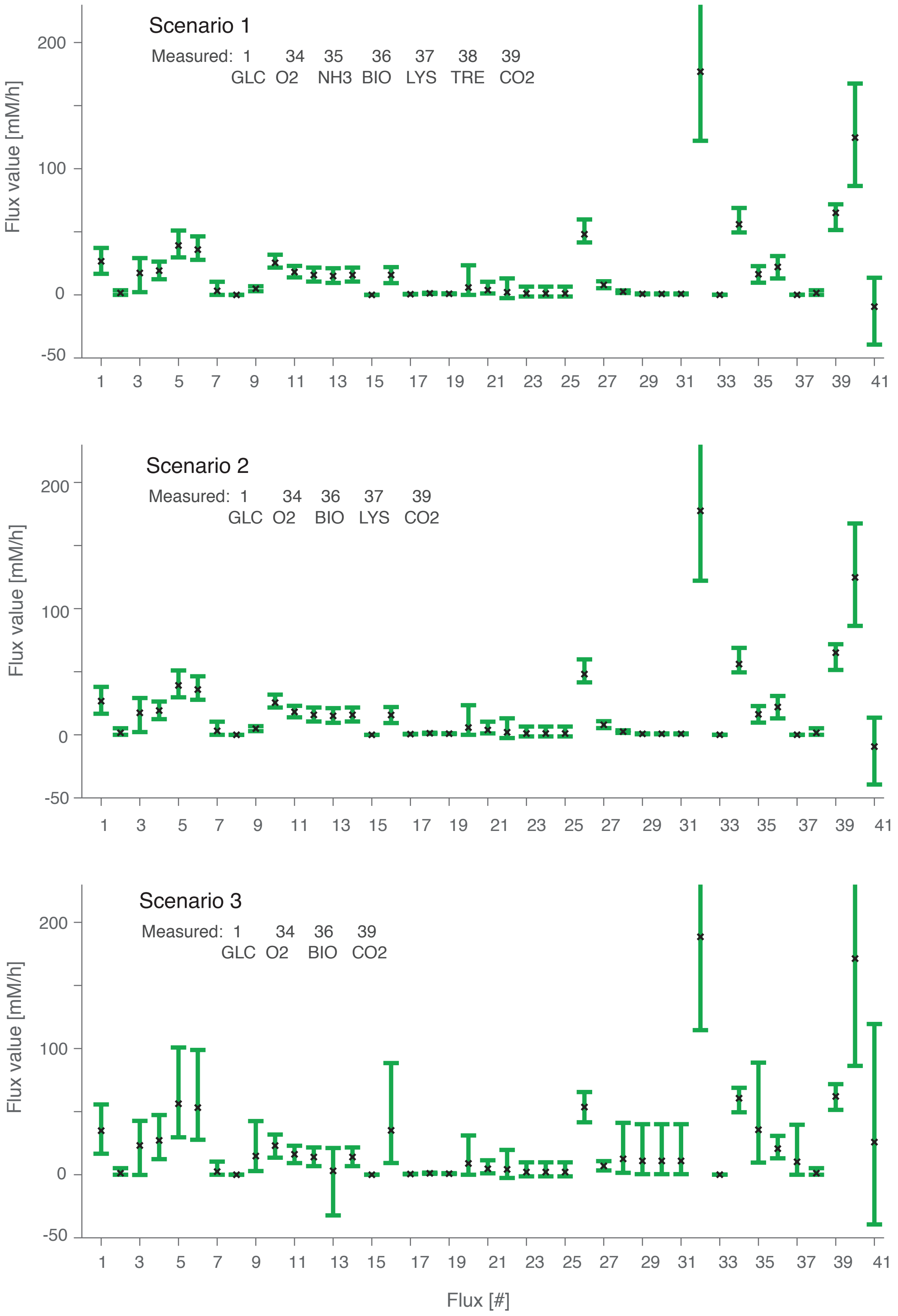


Figure 4.10. FS-MFA in three scenarios. The flux-spectrum intervals are indicated with light green intervals and the centroid with a green "x". NADPH is free (not-balanced) in this three scenarios.

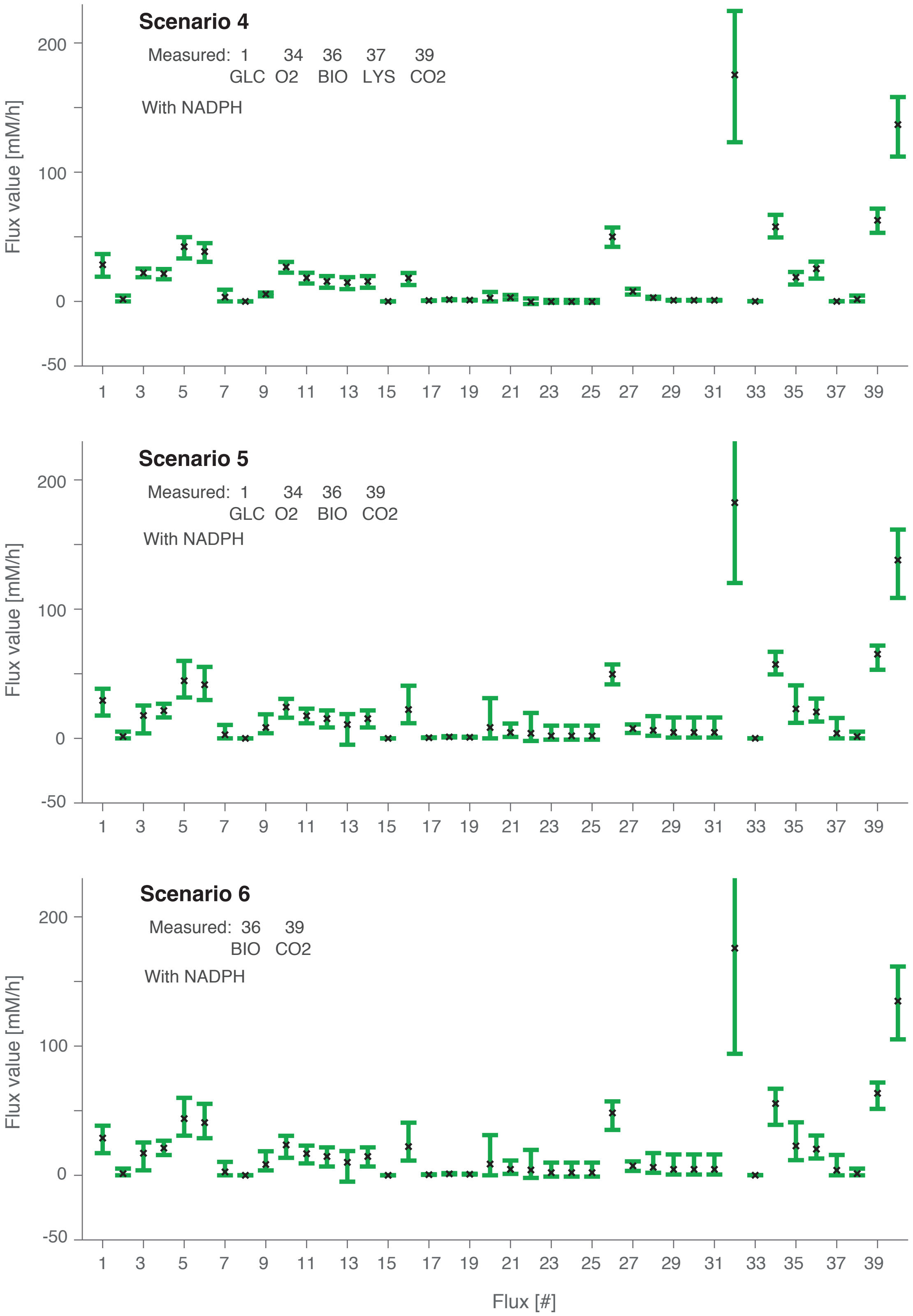


Figure 4.11. FS-MFA in three scenarios. The flux-spectrum intervals are indicated with light green intervals and the centroid with a green "x". NADPH is constraint to be balanced in this scenarios.

4.6 Conclusions

In this chapter we have presented an interval approach to estimate the metabolic fluxes that operate within cells. The method, called FS-MFA, is based on coupling a constraint-based model with a set of measurements. It is a variant of metabolic flux analysis particularly well suited to scenarios with data scarcity.

The main benefit of FS-MFA is that, instead of point-wise estimates, it provides interval estimates. These are richer and more reliable (uncertainty is explicit). The use of intervals also enables MFA in two scenarios: when there is a lack of measurable fluxes, and when the available measurements highly imprecise.

Two cases studies have been used with three objectives: (1) to pinpoint some limitations of traditional metabolic flux analysis (TMFA), (2) to validate FS-MFA against experimental data, and (3) to illustrate its main benefits. We have corroborated that, as expected, TMFA is unreliable if there are not redundant measurements. Moreover, we show that, even with redundant measurements, TMFA point-wise estimates can be highly deviated because of the uncertainty; FS-MFA is more reliable because its interval estimates are only as precise as allowed by the uncertainty. Finally, we have demonstrated that FS-MFA could provide good estimates even if only a few external fluxes are measurable.

Next chapters will further develop the work described here as follows:

- In chapter VI, the flux-spectrum will be used to estimate time-varying fluxes during a cultivation process. The presented procedure can be used as an off-line analysis of collected data, or for the on-line monitoring of a running process, mitigating the traditional absence of reliable on-line sensors in industry.
- In chapter VII, possibility theory will be used to extend the ideas underlying FS-MFA, resulting in a more complex methodology, but bringing several advantages. This methodology will be applied to the estimate metabolic fluxes (chapter VII an VIII), build dynamic FBA models (chapter VIII), and validate a constraint-based models (chapter IX).

In summary, the described FS-MFA is a powerful, yet simple, improvement of traditional metabolic flux analysis, which can be particularly valuable in scenarios where data are scarce, as it is common in industry. The key feature of the approach is that the method provides reliable estimates, since these are only as precise as allowed by the available data and knowledge.

Main references

- Llaneras F, Picó J (2007). An interval approach for dealing with flux distributions and elementary modes activity patterns. *Journal of Theoretical Biology*, 246:290-308.

- Stephanopoulos GN, Aristidou AA (1998). *Metabolic Engineering: Principles and Methodologies*. San Diego, USA: Academic Press.

- Klamt S, Schuster S, Gilles ED (2002). Calculability analysis in underdetermined metabolic networks illustrated by a model of the central metabolism in purple non-sulfur bacteria. *Biotechnology & Bioengineering*, 77:734-751.

- Schrijver A (1988). *Theory of linear and integer programming*. Amsterdam, Netherlands: Wiley.

- Provost A and Bastin G (2004). Dynamic metabolic modelling under the balanced growth condition. *Journal of Process Control* 14(7):717-728.

- Vallino JJ (1994). *Identification of branch-point restrictions in microbial metabolism through metabolic flux analysis and local network perturbation*. PhD thesis, Massachusetts Institute of Technology, Cambridge.

V

Translation of flux states into pathway activities under data scarcity

This chapter discusses how to translate a given metabolic flux state into a pattern of pathway activities. As in chapter IV, fluxes are represented by means of intervals to handle scenarios of data scarcity: scenarios where not all fluxes are known, and scenarios where the know fluxes are imprecise or uncertain. Experimental data from a cultivation of CHO cells will be used as case study.

Part of the contents of this chapter appeared in the following journal articles:

- Llaneras F, Picó J (2007). An interval approach for dealing with flux distributions and elementary modes activity patterns. *Journal of Theoretical Biology*, 246(2):290-308.

5.1 Introduction

In chapter II, it was explained that a metabolic flux state, the distribution of flux through a metabolic network, reflects the behaviour exhibit by cells at given conditions. Chapter III was devoted to network-based pathways and it was shown that every flux state can be seen as the aggregated action of these pathways. In other words, any flux state can be translated into a pattern of pathway activities. This enables the study of cellular states in a context of pathways instead of fluxes, which can be valuable to connect the intracellular state with regulation processes or with the exhibited phenotypes.

This chapter is devoted to study this translation. We will review methods to determine how much flux is being carried by each pathway at given conditions. It will be shown that in most cases there are multiple valid translations, that is, that a given a flux state can be represented with different patterns of pathway activities. Two approaches are usually followed: chose one pattern based on a reasonably assumption (Poolman, 2004; Schwartz, 2006), or deal with the whole space of possible patterns. The second approach relies on the so-called α-spectrum, the ranges of possible activities for each pathway (Wiback, 2003).

Here we will show how the α-spectrum can be computed when the fluxes are represented by means of intervals, what provides some benefits in scenarios of data scarcity: (i) the α-spectrum can be computed when the flux state is partially unknown, (ii) accounting for uncertainty, and (iii) handling high inconsistency. These advantages will be illustrated with a case study.

The chapter is structured as follows. In section 5.2 the translation problem is studied, and particular translation methods are discussed. In section 5.3 the α-spectrum is presented, and in section 5.4 its interval version is introduced. In section 5.4 a case study with CHO cells shows that the α-spectrum can be of use in scenarios of data scarcity. The chapter concludes with some conclusions.

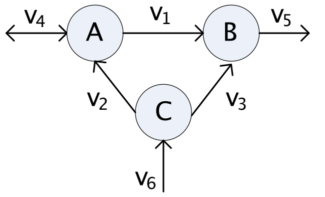
5.2 From fluxes to pathway activities

First, let us recall the formulation used in previous chapters. A simple *constraint-based model* is the flux space **P**, which could be defined as follows:

$$\mathbf{P} = \left\{ \mathbf{v} \in \mathbf{R}^n : \begin{array}{c} \mathbf{N}\cdot\mathbf{v} = \mathbf{0} \\ \mathbf{D}\cdot\mathbf{v} \geq \mathbf{0} \end{array} \right\} \qquad (1)$$

where **v** is the vector of fluxes that represent the mass flow through each of the *n* reactions in the network, **N** is the stoichiometric matrix, and **D** is a diagonal matrix with $\mathbf{D_{ii}}$ = 1 if the flux *i* is irreversible (and otherwise 0).

These constraints define the space of feasible flux (steady) states, which ideally comprises every possible phenotype: only those flux vectors **v** that fulfil (1) are valid cellular states.



$$\mathbf{N} = \begin{bmatrix} -1 & 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 & -1 & 0 \\ 0 & -1 & -1 & 0 & 0 & 1 \end{bmatrix}$$

Figure 5.1. Example of metabolic network and its stoichiometric matrix.

Network-based pathways generate the flux space

Now consider the network-based pathways discussed in chapters II and III. Network-based pathways are flux vectors[1] with certain properties that make them useful for the analysis of the modelled metabolism. For instance, *elementary modes* are all the simplest pathways in the network, those that cannot be decomposed in simpler ones, and a *minimal generating set* is a smallest set of pathways sufficient to span the flux space (Figure 5.2). Other network-based pathways are *extreme currents* and *extreme pathways*. A comparison among all these concepts was carried out in chapter III.

However, herein we are interested in one characteristic that these sets of pathways share: they all generate the flux space. That is, every feasible flux vector **v** in **P** can be translated into a pattern of pathway activities (Figure 5.3).

In particular, each flux vector **v** can be expressed as sum of pathway activities:

$$\mathbf{v} = \sum_{k}^{e} \mathbf{e_k} \cdot \alpha_k, \quad \alpha_k \geq 0 \tag{2a}$$

The same can be expressed in matrix form, as follows:

$$\mathbf{v} = \mathbf{E} \cdot \alpha, \quad \alpha_k \geq 0 \tag{2b}$$

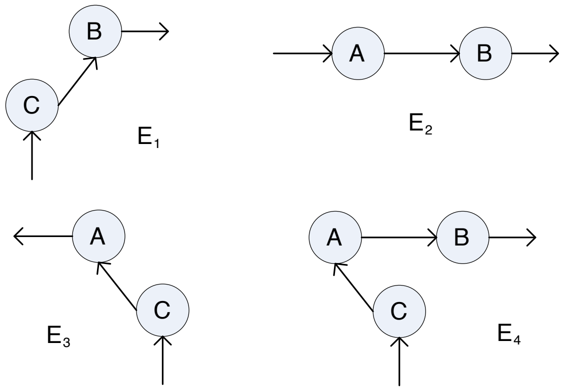
where each $\mathbf{e_k}$ denotes a generating pathway, and each α_k its non-negative activity. The matrix **E** is formed with pathways as a columns.

The patterns of pathway activities (α) express how much flux is being carried by each pathway, an information that can be simpler and more meaningful than reaction

[1] As each pathway is a flux vector, they can be represented as a vector $\mathbf{e} = (e_1,..., e_n)^T$ fulfilling (1).

fluxes. The translation connects the phenotype (the fluxes), with larger structures (the pathways), thus linking it with regulation and other high-level mechanisms.

Remark on nomenclature. Hereinafter we use the term *generating set* to refer to any of the network-based concepts: elementary modes, extreme pathways, or minimal generators. The results discussed hereinafter apply to all of them. We also use the term *pathway* to refer to each vector in a generating set (e.g., an elementary mode).



$$\mathbf{E} = \begin{array}{c} \begin{array}{cccc} E_1 & E_2 & E_3 & E_4 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

Figure 5.2. The elementary modes of the metabolic network depicted in Figure 5.1. There are 4 elementary modes; a minimal generating set is formed by E1, E2 and E3 (since E4 = E3 + E2).

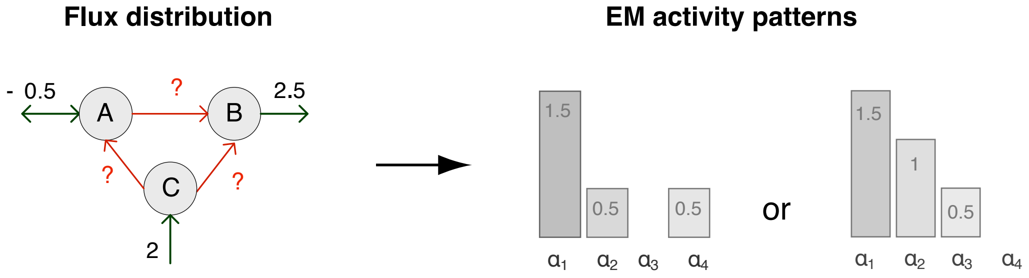


Figure 5.3. Translation of a flux state into a pattern of pathways activities. A flux state is translated into two different patterns. The pathways, elementary modes in this example, are given in Figure 5.2.

Analysis of the translation problem

The relationship between a given flux vector **v** and the corresponding pattern of pathway activities α is given by the system of linear equations (2b). One can study the determinacy and redundancy of the translation problem as follows.

Determinacy. The number of elementary modes n_e is always equal to or larger than *n-m,* the number of linear independent vectors needed to span the flux space (1).

Thus, the rank of **E** is *n-m*. In the particular case when the number of unknowns (n_e) is equal to the rank of **E** (*n-m*), the translation problem (2) is exactly determined and the unique pattern α can be calculated as follows:

$$\alpha = \mathbf{E}^{-1} \cdot \mathbf{v} \tag{3}$$

This is a rare case, however. In most cases, $n_e > (n\text{-}m)$, so the system (2) is underdetermined with n_e - (*n-m*) degrees of freedom, and there are infinite α's fulfilling (2). That means that, in general, a given flux vector **v** cannot be uniquely translated into a pattern of pathway activities.

Those α_k that are uniquely determined can be detected by considering the general solution α^G of the translation problem (2):

$$\alpha^G = \alpha^p + \mathbf{K} \cdot \lambda, \quad \alpha_k^G, \alpha_k^p \geq 0 \tag{4}$$

where α^p is a particular solution, **K** the null space of **E** and λ an arbitrary vector.

Those elements $\alpha^G{}_k$ of α^G whose corresponding row in **K** is a null row are classified as calculable. These elements do not depend on λ, so they are uniquely determined, and its value can be taken from any particular solution (e.g., the non-negative least square solution) because for these elements, $\alpha^G{}_k = \alpha^p{}_k$.

Redundancy. The rank of **E** is always less than *n*, and therefore the system is always redundant. That means that any given flux vector **v** will not be consistent with (1) in general due to measurement or modeling errors. A procedure to detect inconsistent fluxes and to adjust their values was described in the context of metabolic flux analysis in chapter II.

Particular translation methods

It has been shown that the translation of a flux vector into pathways activity patterns has multiple solutions: there can be infinite α's fulfilling (2). Two directions are possible to face with this problem: choose a particular solution based on a rational criterion or deal with the whole space of translations.

Several translation methods have been proposed following the first approach:

- Schwartz et al. (2006) select the translation that minimizes pathway activities α, because this decomposition makes maximum use of the closest pathways to the actual state of cells.
- Poolman et al. (2004) used the same assumption, and although the calculation procedure was different, very similar results were obtained.

- Schwarz et al. (2005) chose the shortest pathways assuming that they are those that most contribute to gene expression, as it has been experimentally observed in the metabolism of *E. coli* (Stelling, 2002). This is also supported by the fact that metabolic networks grow selectively around central metabolites to favor short metabolic paths (Wagner, 2001).
- Nookaew et al. (2007) proposed to maximize the number of used pathways, based on the assumption that cells are likely to use as many routes as possible to maintain robustness and redundancy, as required to survive under genetic and environmental stresses.

These methods are able to yield a unique translation among those that are possible, but the validity of these translations depends on the validity of the underlying assumptions. This methods should be only applied if there is reasonably evidence that the underlying assumptions are true. As an alternative, to investigate the translation without incorporating any of these assumptions, the α-spectrum concept can be used (Wiback, 2005).

5.3 The α-spectrum

The α-spectrum concept provides a simple way to represent the whole space of possible translations for a given flux vector **v**. Basically, the range of possible activities for each pathway are calculated and expressed with an interval, $\left[\alpha_k^m, \alpha_k^M\right]$.

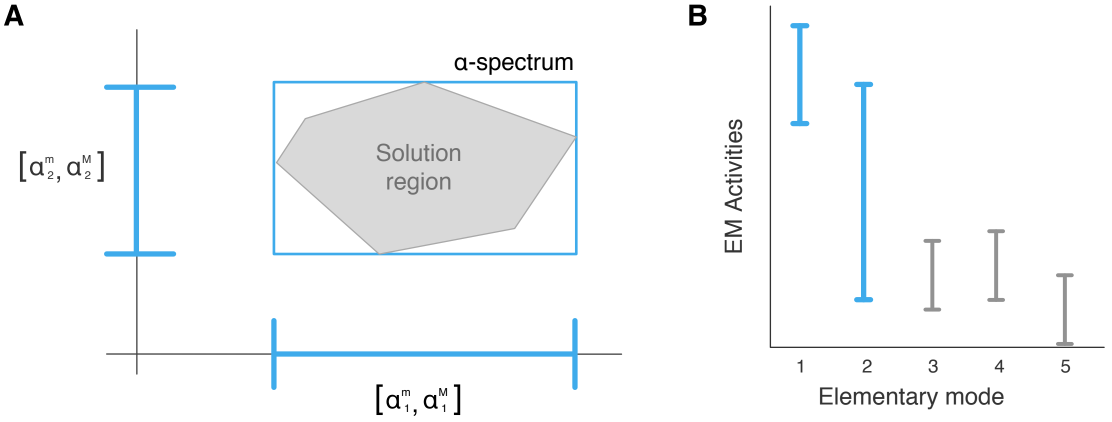


Figure 5.4. The α-spectrum. (A) A 2D projection of a high-dimensional α-spectrum. The polygon is the space of possible translations, and the rectangle is the α-spectrum. (B) The intervals of an α-spectrum represented with a bar-chart.

These intervals can be calculated solving two linear programming problems for each pathway (to get upper and lower bounds):

$$
\begin{aligned}
&\forall \alpha_j, \quad j = 1 \ldots n_e \\
&\alpha_j^m = \ \min\{\alpha_j\} \quad \text{s.t.} \begin{cases} \mathbf{v} = \mathbf{E} \cdot \alpha \\ \alpha_k \geq 0 \quad k = 1 \ldots n_e \end{cases} \\
&\alpha_j^M = \ \max\{\alpha_j\} \quad \text{s.t.} \begin{cases} \mathbf{v} = \mathbf{E} \cdot \alpha \\ \alpha_k \geq 0 \quad k = 1 \ldots n_e \end{cases}
\end{aligned}
\tag{5}
$$

The α-spectrum $\boldsymbol{\mathcal{A}}$ can be defined as the set of the obtained intervals:

$$
\boldsymbol{\mathcal{A}} = \left\{ \alpha \in \mathbf{R}^{\mathrm{n_e}} : \alpha_k^m \leq \alpha_k \leq \alpha_k^M \right\}
$$

In this way, the α-spectrum indicates which pathways can be responsible of the actual cell state. The intervals obtained can be plotted in a bar graph with the pathways represented on the x-axis and their activities on the y-axis (Figure 5.4B).

Let us now discuss some issues regarding the α-spectrum:

- *The α-spectrum contains all particular translations*. All the translations that can be yield based on different assumptions exist within the α-spectrum. This comes at the cost of indeterminacy: the α-spectrum cannot determine the *true* pathway activities.[1]

- *The α-spectrum is an overestimation*. The α-spectrum is a simple representation of the space of possible translations (2), but not an exact one; it is an overestimation (Figure 5.4A). The α-spectrum contains all the possible translations, but also combinations of pathway activities that do not fulfill (2). Notice, however, that this inexactitude is needed to give an independent activity for each pathway, and thus keep the representation simple and understandable (Figure 5.4B).

- *Pathway redundancy enlarges the α-spectrum*. It is well-known that the number of admissible paths through a network increases rapidly as the number of reactions increases (Schuster, 1999). This increment of pathway redundancy results in wider ranges for pathway activities.[2]

[1] Notice, indeed, that *true* pathway activities may not exist because the pathways are an idealisation that may not exactly correspond to a module with biological meaning.

[2] This problem can be explained with (2): when the number of pathways grows faster than the number of reactions, the degrees of freedom of the translation problem increase.

The α-spectrum: interval approach

A slight modification of the method proposed by Wiback et al. (2005) enables computing the α-spectrum when fluxes are represented by means of intervals:

$$
\begin{aligned}
&\forall \alpha_j, \quad j = 1 \ldots n_e \\
&\alpha_j^m = \min\{\alpha_k\} \quad \text{s.t.} \begin{cases} \mathbf{v}^{\mathbf{m}} \le \mathbf{E} \cdot \alpha \le \mathbf{v}^{\mathbf{M}} \\ \alpha_k \ge 0 \quad k = 1 \ldots n_e \end{cases} \\
&\alpha_j^M = \max\{\alpha_k\} \quad \text{s.t.} \begin{cases} \mathbf{v}^{\mathbf{m}} \le \mathbf{E} \cdot \alpha \le \mathbf{v}^{\mathbf{M}} \\ \alpha_k \ge 0 \quad k = 1 \ldots n_e \end{cases}
\end{aligned} \tag{6}
$$

where $\mathbf{v^M}$ and $\mathbf{v^m}$ are vectors with maximum and minimum values for each flux.[1]

In this way, the α-spectrum ($\mathcal{A}$) contains every pattern of pathway activities that fulfills (2) for any of the flux vectors within the interval representation ($[\mathbf{v^m}, \mathbf{v^M}]$).

This interval version of the α-spectrum bring some benefits:

- *The α-spectrum can be computed when the flux vector **v** is partially unknown.* Simply, the unknown fluxes are represented with intervals, e.g., [0, ∞], [-∞, ∞], [0, v^M]. The computed α-spectrum will contain all the α 's that correspond to flux vectors compatible with the available knowledge.

- *Uncertainty can be accounted for.* The uncertainty of the fluxes—consequence, for example, of measurement errors—can be represented with intervals, e.g., [0.9, 1.1]. The α-spectrum will be less precise, wider, but more reliable.

- *High inconsistency can be faced by means of uncertainty.* A given inconsistent flux vector **v** can be adjusted, but if its inconsistency is high, any point-wise adjusted flux vector will be unreliable, because the original values have proved uncertain. A more conservative approach would be define interval fluxes to enclose nearby consistent measurements, and get the α-spectrum from them.

- *Interval fluxes can be used to represent a range of cellular states.* This allows us to compute the ranges of pathway activities that are needed to represent this behavior. This may be useful, for instance, to build reduced kinetic models considering only those pathways that are active for a desired range of cellular states.

Some of these advantages will be illustrated in the case study that comes next.

[1] For instance, given two fluxes such as v_1 = [5, 6] and v_2 = [-1, 1], these vectors will be the following, $\mathbf{v^M} = [6, 1]^T$ and $\mathbf{v^m} = [5,-1]^T$.

5.4 Case study: CHO cells

The methods described above are now applied to the case of cultivation of CHO cells in batch mode. This problem was also addressed in chapter IV, where more details are available, including the metabolic network, the list of metabolites, and the stoichiometric matrix.

Preparation: compute the pathways

The elementary modes have been chosen as network-based pathways. Nevertheless, all the types of generating sets described in chapter III are equivalent in this example because all reactions are irreversible. The 7 elementary modes of CHO cells were computed with Metatool (Pfeiffer, 1999) and given in Table 5.1.

Table 5.1. Elementary modes of the model of CHO cells.

Reaction	E1	E2	E3	E4	E5	E6	E7
v_1	1	1	0	0	2	0	1
v_2	1	0	0	0	0	0	1
v_3	0	1	0	0	2	0	0
v_4	1	0	0	0	0	0	1
v_5	2	0	0	0	0	0	2
v_6	2	0	0	1	0	0	0
v_7	0	0	1	0	0	0	0
v_8	0	0	0	0	0	1	2
v_9	0	0	0	0	0	1	2
v_{10}	0	0	0	0	0	1	2
v_{11}	0	1	1	1	2	2	2
v_{12}	0	1	0	0	2	1	2
v_{13}	0	0	1	1	0	1	0
v_{14}	0	1	0	0	2	0	0
v_{15}	0	0	0	1	0	1	0
v_{16}	0	1	1	1	2	1	0
v_{17}	0	1	0	0	1	0	0
v_{18}	0	0	0	0	1	0	0
v_{19}	0	1	1	2	2	2	0
v_{20}	0	2	1	1	5	1	0
v_{21}	0	2	2	2	4	5	6
v_{22}	0	1	0	0	0	0	0

Analysis of the translation equation

Consider the flux vector given in Table 5.2, which was calculated in chapter IV applying metabolic flux analysis from a set of six measurements. We first analyse the translation problem for these data. The rank of **E** is 6 and there are 7 elementary modes, so the translation problem (2) is underdetermined and has multiple solutions. However, the inspection of the kernel of **E** shows that some activities are uniquely determined:

$$\mathbf{K} = \begin{pmatrix} -0.3 & 0 & 0 & 0.6 & 0 & 0.6 & 0.3 \end{pmatrix} \tag{7}$$

The activity of 3 elementary modes can be taken from a particular solution, such us the non-negative least square solution, resulting in: $\alpha_2 = 0$, $\alpha_3 = 0.268$ and $\alpha_5 = 0.143$. The activity of the other 4 elementary modes remains undetermined.

To estimate the possible activities of all the pathways, the α-spectrum can be computed with (5) or (6). The obtained intervals, used as reference hereinafter, are depicted in Figure 5.5. The results show that even if the activity of 4 pathways is not uniquely determined, its ranges of possible values can be narrow.

Table 5.2. A complete flux vector of CHO cells. Fluxes in mM/(d·10^9 cells).

Reaction	Flux	Reaction	Flux	Reaction	Flux
G (v_1)	4.05	L (v_6)	7.39	NH4 (v_{19})	0.96
Q (v_{20})	1.18	A (v_7)	0.26	CO2 (v_{21})	2.61
1: G→G6P	405	7: Pyr+Glu→A+aKG	0.26	13: Mal→Pyr+CO2	0.47
2: G6P→G3P+DAP	3.76	8: Pyr→ACA+CO2	0.34	14: Oxa+Glu→Asp+aKG	0.28
3: G6P→R5P+CO2	0.28	9: Oxa+ACA→Cit	0.34	15: Glu→aKG+NH4	0.20
4: DAP→G3P	3.76	10: Cit→aKG+CO2	0.34	16: Q→Glu+NH4	0.75
5: G3P→Pyr	7.53	11: aKG→Mal+CO2	1.10	17: R5P+Asp+Q→Pu	0.14
6: Pyr→L	7.39	12: Mal→Oxa	0.63	18: R5P+Asp+2Q→Py	0.14

The α-spectrum and partial knowledge

Let us consider that only v_1 (G), v_6 (L), v_{20} (Q) and v_{21} (CO_2) are measured. This is an underdetermined MFA problem, where the available measurements are insufficient to determine all the fluxes[1]. However, the α-spectrum can be computed.

First, the partially unknown flux vector has to be represented with intervals (Table 5.3, row B). Then, the intervals of the α-spectrum are computed using (6). The results are given in Table 5.4 and Figure 5.5B.

[1] The rank of $\mathbf{N_u}$ (16) is less than the number of unknown fluxes (22-3-1). See chapter II and IV for details about MFA problems.

This example shows that even from partial knowledge, the α-spectrum can be informative. In fact, the ranges obtained are very similar to those obtained from the complete flux vector, only the activities of elementary modes 2 and 3 are conservative.

Table 5.3. Different flux vectors of CHO cells. Fluxes in mM/(d·10^9 cells).

		v_1 (G)	v_2-v_5	v_6 (L)	v_7(A)	v_8-v_{18}	v_{19} (NH_4)	v_{20} (Q)	v_{21} (CO2)	v_{22}
Measure.		4.0546		7.3949	0.255		0.9617	1.186		0
Partial	B	4.0546	[0,∞]	7.3949	[0,∞]	[0,∞]	[0,∞]	1.186	2.557	[0,∞]
Uncertain	C	[3.5,4.5]	[0,∞]	[6,8]	[0.1,0.5]	[0,∞]	[0.6,1.4]	[1,1.5]	[0,∞]	0

A: Measured values (Provost, 2004). B: an uncertain flux vector defined around the measurements. C: partially unknown flux vector where only 4 fluxes are known.

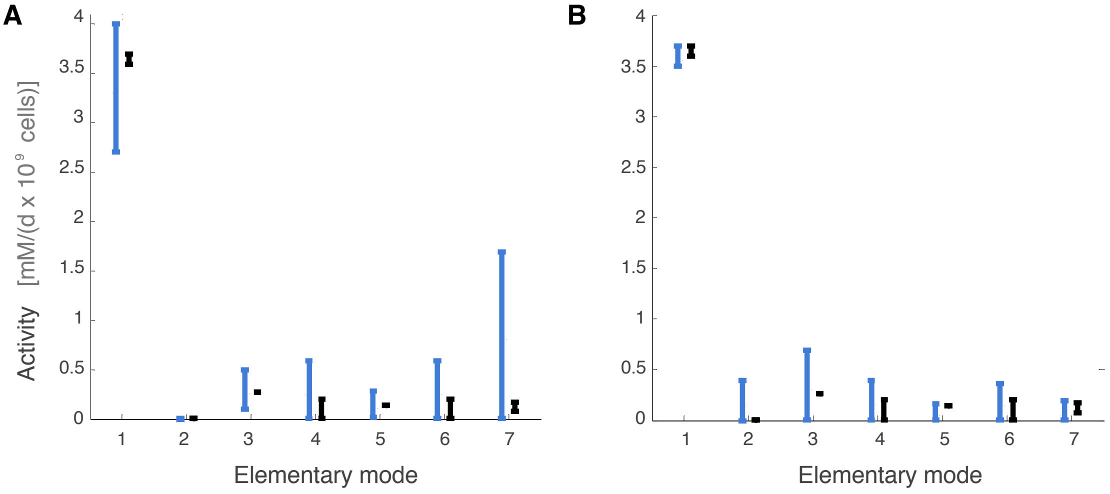


Figure 5.5. The α-spectrum in two scenarios of data scarcity. (A) The α-spectrum when measurements are uncertain (Table 5.3, row C). (B) The α-spectrum if the flux vector is partially unknown (Table 5.3, row B). The α-spectrum from the certain and complete flux vector is depicted in black.

The α-spectrum and uncertainty

The interval formulation in (6) makes it possible to compute the α-spectrum accounting for uncertainty. As an example, the α-spectrum has been computed using the uncertain measurements given in Table 5.3, row C. Results are given in Table 5.5 and Figure 5.5A. As expected, the α-spectrum intervals are wider, but more reliable if measurements are indeed uncertain.

The α-spectrum and consistency

Finally, let us consider an inconsistent flux vector generated adding random noise (±10%) to the flux vector given in Table 5.2. To approaches are possible to handle the inconsistency:

(a) Adjust the measurements to be consistent (as explained in chapter II), and then compute the α-spectrum from them using (5).

(b) Represent the measurements with intervals to consider their (obvious) uncertainty, thus enclosing the nearby consistent sets of measurements, and then compute the α-spectrum from these intervals using (6).

As shown in Table 5.6, the first approach (a) obtains a narrower α-spectrum, but deviated from the one that was obtained from the original flux vector (without the added noise). Following the second approach (b) we get an α-spectrum which is slightly wider, but which encloses the α-spectrum obtained from the original flux vector.

Table 5.4. The α-spectrum computed from a partially unknown flux vector (B).

	E1 activity	E2	E3	E4	E5	E6	E7
Complete	[3.59,3.69]	0	0.26	[0,203]	0.14	[0,0.203]	[0.07,0.17]
Partial	[3.50,3.69]	[0,0.39]	[0,0.69]	[0,0.39]	[0,0.16]	[0,0.36]	[0,0.19]

Table 5.5. The α-spectrum computed from an uncertain set of vector (C).

	E1 activity	E2	E3	E4	E5	E6	E7
Certain	[3.59,3.69]	0	0.26	[0,203]	0.14	[0,0.203]	[0.07,0.17]
Uncertain	[2.7,4]	0	[0.1,0.5]	[0,0.58]	[0.01,0.28]	[0,0.587]	[0,1.69]

Table 5.6. Computation of the α-spectrum from an inconsistent flux vector.

	E1 activity	E2	E3	E4	E5	E6	E7
Original	[3.59, 3.69]	0	26	[0, 203]	14	[0, 0.20]	[0.07, 0.17]
Ap. a	[3.60, 3.65]	[0, 0.02]	[0.36, 0.38]	[0, 0.07]	[0.13, 0.15]	[0, 0.07]	[017, 0.22]
Ap. b	[3.39, 3.76]	0	[0.26, 0.31]	[0, 0.24]	[0.13, 0.15]	[0, 0.24]	[0.06, 0.19]

5.5 Conclusions

Sometimes a pattern of pathways activities is a more meaningful (and simpler) representation than a vector of reaction fluxes, and therefore the translation between both representations is worth dealing with.

We have seen that there are proposals to choose one particular pattern of pathway activities among those that are possible. Yet, these methods rely on assumptions that are not easy to validate. As an alternative, one can calculate the α-spectrum, which represents the whole set of valid patterns. In particular, herein we have shown that the α-spectrum can be calculated even when the original fluxes are represented with intervals. This enhances the usage of experimental flux data, providing a way to handle common problems, such as sensor inaccuracy or lack of data.

The α-spectrum can be a useful tool in applications that connect the metabolic networks with experimental data. For instance, it may be of use for the on-line monitoring of the metabolic phases of a cells culture, if these phases are characterised by the active pathways. The α-spectrum could be also useful to build reduced dynamic models, which consider only those pathways active under the circumstances of interest.

The major limitation of computing patterns of pathways activities is that the number of pathways can be very large, resulting in several valid patterns. As explained in chapter III, the number of network-based pathways dramatically increases as the number of reactions in the network increases due to a combinatorial explosion. This effect is particularly intense with elementary modes, but occurs also with extreme pathways or minimal generators. This large number of pathways is necessary in many applications. For instance, we need all the elementary modes to predict the effect of knockouts, and all the minimal generators to exactly generate the flux space. Furthermore, redundancy is an inherent property of metabolism, so cells have multiple ways to produce similar behaviors. However, there are applications that may require a lower-dimensional set of pathways (Barrett, 2009), and computing pathway activities is probably one of them.

Main references

- Llaneras F, Picó J (2007). An interval approach for dealing with flux distributions and elementary modes activity patterns. *Journal of Theoretical Biology*, 246:290-308.

- Llaneras F, Picó J (2010). Which metabolic pathways generate and characterise the flux space? A comparison among elementary modes, extreme pathways and minimal generators. *J. Biomedicine and biotechnology*, 1:2010.

- Wiback SJ, Mahadevan R, Palsson BO (2003). Reconstructing metabolic flux vectors from extreme pathways: Defining the alpha-spectrum. *Journal of Theoretical Biology*, 224(3):313-324.

- Poolman MG, Venkatesh KV, Pidcock MK, Fell DA (2004). A method for the determination of flux in elementary modes, and its application to lactobacillus rhamnosus. *Biotechnology and Bioengineering*, 88(5):601-612.

- Klamt S, Schuster S, Gilles ED (2002). Calculability analysis in underdetermined metabolic networks illustrated by a model of the central metabolism in purple non-sulfur bacteria. *Biotechnology & Bioengineering*, 77:734-751.

- Pfeiffer T, Sanchez-Valdenebro I, Nuno JC, Montero F, Schuster S (1999). METATOOL: For studying metabolic networks. *Bioinformatics*, 15(3):251-257.

VI

Estimation of time-varying fluxes under data scarcity

This chapter describes a procedure to estimate time-varying metabolic fluxes during a cultivation process. The procedure is based on the results of chapter IV, so it handles measurements uncertainty and is particularly suitable in scenarios of data scarcity.

The procedure can be used as an off-line analysis of collected data to get insight on the dynamic behaviour of the organism, or to on-line monitoring a running process, mitigating the traditional absence of reliable on-line sensors in industry. The cultivation of CHO cells will be used as case study.

Part of the contents of this chapter appeared in the following journal article:

- Llaneras F, Picó J (2007). A procedure for the estimation over time of metabolic fluxes in scenarios where measurements are uncertain and/or insufficient. *BMC Bioinformatics*, 8:421.

6.1 Introduction

As seen in previous chapters, constraint-based models can be assembled for organisms of interest based on the mass balances around internal metabolites, which are assumed to be steady-state, and other constraints, such us transport capacities or thermodynamics. These constraints define a space containing every feasible metabolic state. The environmental conditions at particular circumstances would determine which of these corresponds are exhibited by the cells. One approach to determine the flux state of cells at a given moment, is to incorporate experimental measurements. This is the idea underlying metabolic flux analysis (MFA), as discussed in chapters II and IV.

MFA estimations are typically done under a static point of view. Therefore the obtained flux vector will be only valid during certain time, while the environmental conditions and the cells state remains steady (e.g., during growth phase). If these conditions change, as it happens in actual cultures, the flux vector may change. Clearly, following these changes over time will be useful to investigate the dynamic behaviour of cells and to monitor the progress of industrial fermentations (Mahadevan, 2005).

There are, in fact, several works in the field devoted to this problem. Mahadevan et al. (2002) extended classical FBA to predict the dynamic evolution of the metabolic fluxes. In (Gayen, 2006), elementary modes and the assumption of optimal behaviour are used to estimate the flux vector of *C. glutamicum* at different phases of fermentation. Elementary modes are also employed in (Provost, 2006b), where time-varying intracellular fluxes are obtained by switching the flux vectors calculated at different temporal phases. In (Herwig, 2002), on-line MFA is applied to quantify coupled intracellular fluxes. Takiguchi et al. (Takiguchi, 1997) use a similar approach to recognise the physiological state of cells culture, and show that this information improves Lysine production yield. Henry et al. (2007) presented an on-line estimation of intracellular fluxes applying MFA to an over-determined metabolic network.

Although these works consider that intracellular fluxes are in steady-state at each measurement step, the dynamic nature of the process is not disregarded: the intracellular fluxes will follow the changes of environmental conditions as mediated by the measured fluxes (e.g., substrate uptakes). Intracellular fluxes may undergo shifting from one state to another depending on the environmental conditions. The same idea can be found in several dynamic models (Provost, 2004; Provost, 2006; Lei, 2001; Teixeira, 2007; Sainz, 2003; Ren, 2003). In this way, we can study the (extracellular) dynamic behaviour of an organism, without considering the still not well-known intracellular kinetics.

Most of the works mentioned above use Traditional MFA to get the estimates during the cultivation. However, as explained in chapter II, traditional MFA has some limitations[1], and these are particularly critical under data scarcity, a situation common in industry, and worsen if measurements are needed on-line. To overcome these limitations, at least partially, in this chapter we will use the MFA variant described in chapter IV, the so-called flux-spectrum MFA (FS-MFA).

The objectives of this chapter are twofold: first, introduce a procedure to estimate time-varying metabolic fluxes that uses FS-MFA to be well-suited in scenarios of data scarcity. Second, illustrate the procedure applying it to a real case study: the cultivation of CHO cells in batch mode.

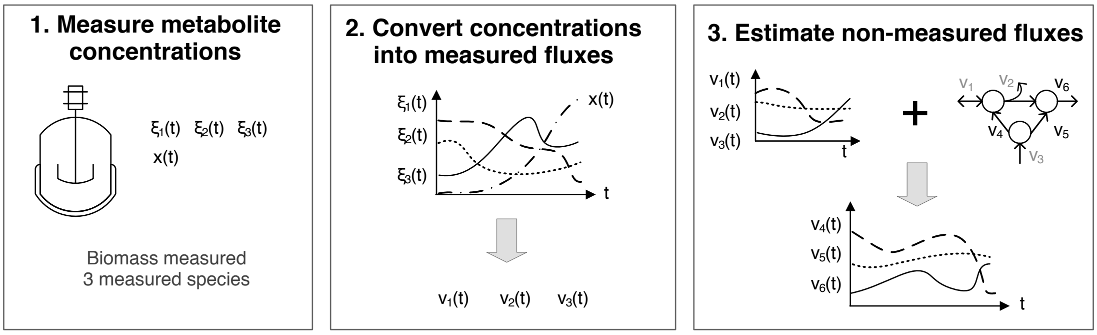


Figure 6.1. Three-step procedure to estimate the time-varying metabolic flux. e(t) denotes the concentration of an extracellular metabolite, *v*(t) its flux, and *x*(t) the biomass concentration. As an example, subindexes 1, 2 and 3 denote measured fluxes and 4, 5 and 6 non-measured ones.

6.2 Estimation procedure

In most cases, only a few extracellular metabolites are measurable during a fermentation. For this reason we follow an indirect approach to estimate those fluxes that cannot be measured: couple the available measurements with a constraint-based model. Under this philosophy, the proposed procedure is structured as follows (Figure 6.1): (1) measure the concentration of some extracellular metabolites and biomass, (2) convert these concentrations to “measured” fluxes and (3) estimate the non-measured fluxes with the flux-spectrum (FS-MFA).

It is often overlooked that extracellular fluxes are not directly measured. Instead, the concentrations of a set of metabolites are measured (step 1), and those data are converted to flux units or measured fluxes (step 2). The importance of a good conversion should not be disregarded: errors in the measured concentrations may be amplified,

[1] In brief, (i) traditional MFA cannot be used when measurements are scarce, (ii) it gives only point-wise estimates (insufficient if multiple flux values are reasonably possible due to the uncertainty), and (iii) it does not considers inequality constraints, such as reaction reversibility.

incorporated to the measured fluxes, and then propagated to the estimation of the non-measured ones. To minimise this hitch, the conversion should be careful.

Once the measured fluxes are available, the non-measured ones can be estimated by coupling them with the constraint-based model (step 3). This step has been done before using MFA (Herwig, 2002; Takiguchi, 1997; Henry, 2007; Ren, 2003), but herein the FS-MFA will be used instead.

The procedure can be useful in two ways: (a) as an off-line analysis of collected data, or (b) to on-line monitoring running process. The procedure scheme and its fundamental step (step 3) will be the same in both cases, but differences may arise in step 2.

Preliminaries: choose a constraint-based model

Recalling the formulation used in previous chapters, a simple *constraint-based model*, the flux space **P**, can be assembled assuming that internal metabolites are at steady-state and considering the irreversibility of some reactions, as follows:

$$\mathbf{P} = \begin{cases} \mathbf{N} \cdot \mathbf{v} = 0 \\ \mathbf{D} \cdot \mathbf{v} \geq 0 \end{cases} \tag{1}$$

where **v** is the vector of metabolic fluxes, representing the mass flow through each of the *n* reactions in the network, **N** is the stoichiometric matrix linking metabolites and fluxes, and **D** is a diagonal matrix with $\mathbf{D_{ii}} = 1$ if the flux *i* is irreversible (otherwise 0).

The constraints in (1) define a space of feasible steady-state flux vectors, or flux states, which ideally comprises every possible phenotype. Only flux vectors **v** that fulfil (1) are valid cellular states. That means that there are infinite **v** fulfilling (1).

As explained above, to determine which feasible flux vector is the actual one at given circumstances, measured fluxes can be incorporated as additional constraints to apply TMFA (see chapters II and IV). Unfortunately, these measurements tend to be scarce, which means that we need a reasonably small network to apply the estimation procedure—otherwise, the measurements cannot offset or reduce the under-determinacy of the model (1) to get valuable estimates. To keep reductions of the network at minimum, intracellular measurements from tracer experiments can be incorporated (Sauer, 2006; Wiechert, 2001), but those data are in most cases not available. FS-MFA will be also of help, because it gives estimates without completely offset network under-determinacy. However, we must kept in mind that the main fact remains: reasonably small networks are required.

Step 1: measuring metabolite concentrations

There are several alternatives to measure the concentration of metabolites—e.g., on-line sensors, isotopic tracer experiments or laboratory procedures—and describing them is out of the scope of this work. Just remember that the more measurements are available, the more non-measured fluxes may be accurately estimated. However, one should be prepared to deal with lack of measurements, especially when the procedure is done on-line (due to the lack of on-line sensors).

Step 2: converting measured concentrations into measured fluxes

A mass balance around an extracellular metabolite whose concentration is measurable can be stated as follows:

$$\frac{\mathrm{d}e}{\mathrm{dt}} = v_e \cdot x - \mathrm{D} \cdot e + \mathrm{F_e} \tag{2}$$

where e is the metabolite concentration, v_e its flux (substrate uptake or product formation), x the biomass concentration, D the dilution term and $\mathrm{F_e}$ the net exchange of the metabolite with the environment. This equation is only valid for extracellular metabolites, but biomass growth and mass balances of internal metabolites not at pseudo-steady state can be represented in a similar way (Bastin, 1990; Schüerl, 200).

One can calculate v_e as a function of e, x, D, $\mathrm{F_e}$ and de/dt, but this presents two main difficulties: (i) approximate a derivative (directly or indirectly) and (ii) deal with the presence of errors and noise in the measured e. The underlying problem is how precision can be combined with robustness against measurement errors.

We propose two alternatives two calculate the fluxes: (a) combine an Euler method with moving average filters, and (b) use a non-linear observer. The first one is suitable if the procedure is done off-line, the second one is better to work on-line (Figure 6.2).

Notice, however, that there is not a universal solution for the conversion. In real applications, the particularities of the measurements (accuracy, sample rate, importance and characteristics of the noise, etc.) and the operation mode (off-line, on-line with an acceptable delay or purely on-line) determines the most suitable approach.

Euler approximation and moving average filters

One approach is to approximate the derivative de/dt with a simple method, such as Euler or Runge-Kutta methods, and then solve (2) (Herwig, 2001). Euler methods provide the most straightforward approximations:

$$\text{Backward: } \frac{\mathrm{df}(k)}{\mathrm{dt}} \approx \frac{\mathrm{f}(k)-\mathrm{f}(k-1)}{\mathrm{t}(k)-\mathrm{t}(k-1)}$$

$$\text{Middle point: } \frac{\mathrm{df}(k)}{\mathrm{dt}} \approx \frac{\mathrm{f}(k+1)-\mathrm{f}(k-1)}{\mathrm{t}(k+1)-\mathrm{t}(k-1)}$$

The backward version does not introduce an intrinsic delay, but the middle point provides a less noisy approximation.

In most cases this straight approximations need to be combined with the use of filters to eliminate or reduce the presence of noise. Filters based on the moving average are good candidates because they are simple and versatile. Basically, moving average filters calculate a new signal by averaging the values of the original signal within a time window. Thus, the new signal becomes smoother. This kind of filters has already been applied to the calculation of metabolic fluxes (Herwig, 2001).

The centred moving average (CMA) provides best results because uses past and future information. The filtered value for instant *k* (*CMA$_k$*) is calculated by averaging the values of the original signal (*S*) between *k-n* and *k+n*:

$$\mathrm{CMA}_k = \frac{\sum_1^n \mathrm{S}_{k\text{-}i} + \mathrm{S}_k + \sum_1^n \mathrm{S}_{k+i}}{2 \cdot n + 1}$$

If only past values of the original signal are available, the standard moving average (SMA) can be used instead:

$$\mathrm{SMA}_k = \frac{\sum_0^n \mathrm{S}_{k\text{-}i}}{n+1}$$

The key parameter of moving average filters is the window size *n*, i.e., the number of averaged values.[1] The optimal size would be one observation, so as to be close to the original signal. However, to reject noise, the window size needs to be increased. There is a trade-off between sensitivity to noise and delay with respect to the original signal.

This simple approach to calculate the fluxes $v_e(k)$ provides particularly good results when centred methods can be used both to approximate the derivative and to filter the signals. That is, when past (*k-i*) and future information (*k+i*) is available. This is the case if the whole procedure is done off-line.

[1] A typical variant of these filters includes multiplying factors to give a different weight to each value within the time window (e.g., an exponential moving average), which must be also tuned.

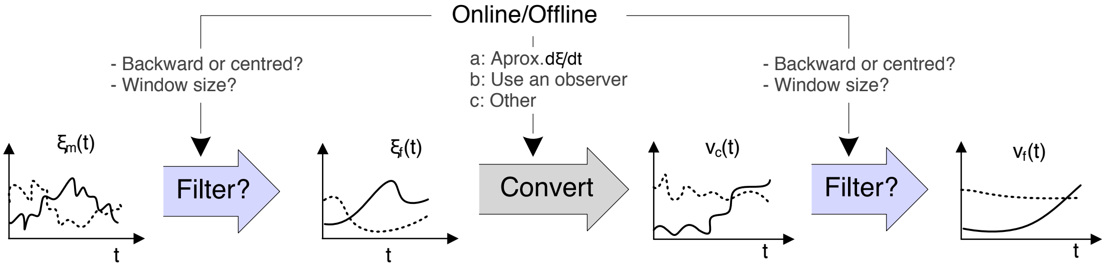


Figure 6.2. Conversion of measured concentrations into measured fluxes. First, the measured concentrations should be filtered. Then, fluxes are calculated from the concentration data (e.g., approximating the derivative or using a dynamic observer). Finally, the calculated fluxes may be further filtered to get a smoother signal. Each step is conditioned by the operation mode (on-line or off-line).

Non-linear observers and other alternatives

There are also methods especially aimed to the on-line approximation of derivatives. If the noise signal is well characterised (e.g., the frequency band or a stochastic feature is known) a linear differentiator (Pei, 1989) or even a Luenberger observer may be used (Luenberger, 1971). If nothing is known on the structure of the signal, then sliding mode techniques are profitable. For example, the method introduced in (Levant, 1998) combines exact differentiation for a large class of input signals with robustness against any small noises. An alternative based on Levant's super-twisting algorithm have been proposed to similar problems (Battista, 2010).

Finally, there are methods to calculate the extracellular fluxes that avoid the approximation of the derivative, for example, the use of extended Kalman filters (Henry, 2007; Dochain, 1988) or observers based on concepts from non-linear systems theory, such as the high gain estimators described in (Bastin, 1990). These methods do not use future information because they are aimed to on-line operation mode. For instance, a high-gain non-linear observer of the extracellular fluxes can be directly synthesised from (2) using the method proposed in (Farza, 1998):

$$\frac{\mathrm{d}e_o}{\mathrm{dt}} = v_e \cdot x - \mathrm{D} \cdot e_o - 2 \cdot \theta \cdot (e_o - e)$$

$$\frac{\mathrm{d}v_e}{\mathrm{dt}} = -\frac{\theta^2 \cdot (e_o - e)}{x}$$

where e_o denotes the observed concentration of the extracellular metabolite and v_e the observed flux. The unique adjustable parameter is θ. Not only these observers are proved to be stable, but also its asymptotic error can be made arbitrarily small by choosing sufficiently large values of θ. However, very large values need to be avoided in practice since the observer may become noise sensitive. Thereby, the choice of θ represents a trade-off between fast convergence (minor delay) and sensitivity to noise.

Remark. Filtering the fluxes calculated by the high-gain non-linear observer may be also advisable to get a smoother signal, although similar results may be achieved by tuning the parameter θ.

Step 3: estimating the metabolic fluxes with FS-MFA

Finally, at each time instant k, the measured fluxes obtained in step 2 are coupled with the constraint-based model (1) to estimate the non-measured fluxes (Figure 6.1). As explained in the introduction, previous works applied traditional MFA (TMFA) with this purpose, but herein we will apply a variant described in chapter IV, the so-called flux-spectrum (FS-MFA), which is particularly suitable in scenarios of data scarcity, where measurements are imprecise and most metabolites are unknown.

FS-MFA estimates of the non-measured fluxes at each time instant k can be computed with the following three-step procedure:[1]

As explained in chapter IV, the size of the intervals—the imprecision of the estimation—depends on the measurements and constraints: the more are available, the tighter intervals are obtained.

Step 3.1 Represent the measured fluxes in $\mathbf{v}(k)$ with an interval, $[v_{m,i}{}^{m}(k),\ v_{m,i}{}^{M}(k)]$ by means of inequalities:

$$\mathbf{v}_{\mathbf{m}}^{\mathbf{m}}(k) \le \mathbf{v}_{\mathbf{m}}(k) \le \mathbf{v}_{\mathbf{m}}^{\mathbf{M}}(k) \tag{3}$$

Step 3.2 Impose the constraints (1) to define the *current flux space* at k, $\mathbf{F}(k)$:

$$\mathbf{F}(k) = \left\{ \mathbf{v}(k) \in \mathbf{R}^{\mathrm{n}} : \begin{array}{c} \mathbf{N}\cdot\mathbf{v}(k) = \mathbf{0} \\ \mathbf{D}\cdot\mathbf{v}(k) \ge \mathbf{0} \\ \mathbf{v}_{\mathbf{m}}^{\mathbf{m}}(k) \le \mathbf{v}_{\mathbf{m}}(k) \le \mathbf{v}_{\mathbf{m}}^{\mathbf{M}}(k) \end{array} \right\} \tag{4}$$

The space $\mathbf{F}(k)$ contains all the flux vectors $\mathbf{v} \in \mathbf{P}$ compatible with the measurements at k, $\mathbf{v_m}(k)$.

[1] Notice that the same three-step procedure can be used to get an estimate of those fluxes that were measured, eventually reducing its uncertainty thanks to the coupling with the other measurements.

Step 3.3 Calculate the flux-spectrum, the interval of feasible values for each flux $v_i(k)$, solving a set of linear programming problems (LP):

$$\begin{aligned} &\forall v_i(k),\ i=1..n \\ &\quad v_i^m(k)= \quad \min\{v_i(k)\} \quad \text{s.t.} \quad \mathbf{F}(k) \\ &\quad v_i^M(k)= \quad \max\{v_i(k)\} \quad \text{s.t.} \quad \mathbf{F}(k) \end{aligned} \tag{5}$$

This gives an interval estimate per flux and time instant, $[v_i^m(k), v_i^M(k)]$.

Remember also that if uncertainty is not considered, all fluxes are reversible, and (4) is determined, the FS-MFA gives the same point-wise estimate that TMFA. However we saw in chapter IV that FS-MFA provides several advantages,[1] and new ones arise when it is used in a successive way (as here):

- FS-MFA may detect sensitivity problems. An interval estimate anomalously large at *k*, indicates that a sensitivity problem exists. With TMFA this sensitive problems may introduce peak values and misleading estimates.
- The inspection of past and future intervals, together with our qualitative knowledge on cells behaviour, may be useful to hypothesise which flux values are more likely among those that are feasible.

6.3 Case study: CHO cells

The three-step procedure described in the previous section is now applied to the estimation of the metabolic fluxes of CHO cells cultivated in batch mode. The available experimental data are the typical data measured off-line (accurate measurements of the concentration of a few metabolites, with a low sampling rate), and therefore this example will be approached assuming that the procedure is done off-line.

Hereinafter we pay special attention to the third step of the procedure, since is the more novel one. In particular, we compare the results given by FS-MFA with those provided by traditional MFA (TMFA), the well-established methodology that is the basis of similar procedures (Herwig, 2002; Takiguchi, 1997; Henry, 2007; Ren, 2003).

[1] Summarising, FS-MFA accounts for uncertainty, provides reliable and richer interval estimates (instead of point-wise ones), and can be used in scenarios of data scarcity.

The comparison discusses the benefits of the estimation procedure in three scenarios:

- S1. If measurements are *almost* sufficient. There are enough to determine all the non-measured fluxes, but there are not redundant measurements (the system (4) is determined and not redundant).
- S2. If measurements are sufficient. Measured fluxes are enough to determine the non-measured fluxes and there are also redundant measurements (the system (4) is determined and redundant).
- S3. If measurements are insufficient. There are not enough to determine all the non-measured fluxes (the system (4) is underdetermined and not redundant).

Preparation: metabolic network and constraint-based model

The metabolic network is the same that was used in chapter IV. However, in this case some reactions {2, 4, 5, 6 and 7} are considered reversible because the analysis is not restricted to the growth phase (e.g., when glucose is exhausted lactate and alanine are consumed instead of produced).

The complete list of metabolites and reactions, the stoichiometric matrix and a depiction of the network were given in chapter IV.

Step 1: measuring metabolite concentrations

Experimental data taken from (Provost, 2006a) is given in Figure 6.3. The cell density (X) and the concentration of 5 extracellular metabolites are measured: two substrates, glucose (G) and glutamine (Q), and three excreted products, lactate (L), alanine (A) and ammonia (NH4). This data was collected with a sample rate of 24 h. Notice that these measurements cannot be filtered because, due to the low sample rate, it is impossible to distinguish between noise and true changes of the signal.

Step 2: converting measured concentrations into measured fluxes

The second step of the procedure is convert the measured concentrations in measured fluxes. The measured fluxes calculated with three different approximations of the derivative are depicted in Figure 6.4. Since the procedure is being done off-line, a centred approximation is the most advisable choice, so fluxes calculated with the middle-point Euler method will be used hereinafter.[1]

[1] Similar results were obtained using a backward Euler approximation, which would be suitable in case the procedure were done on-line (not shown).

The results shown in Figure 6.4 already give the idea of uncertainty. The differences between different conversions are significant. Clearly, the reliability of the conversion, along with the precision of the measurements of metabolite concentrations, should be taken into account to define the uncertainty of the measured fluxes.

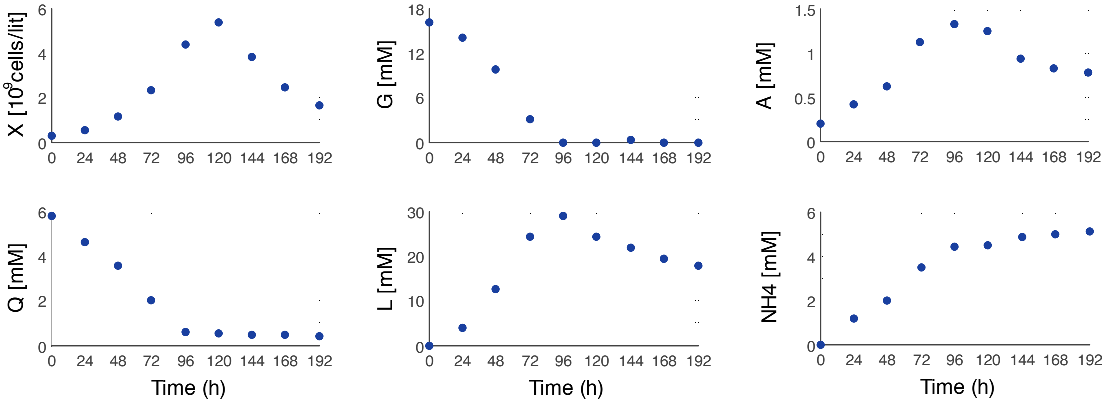


Figure 6.3. Concentration of measured extracellular metabolites and biomass during a cultivation of CHO cells. The measurements correspond to cell density (X), glucose (G), glutamine (Q), lactate (L), alanine (A) and ammonia (NH4).

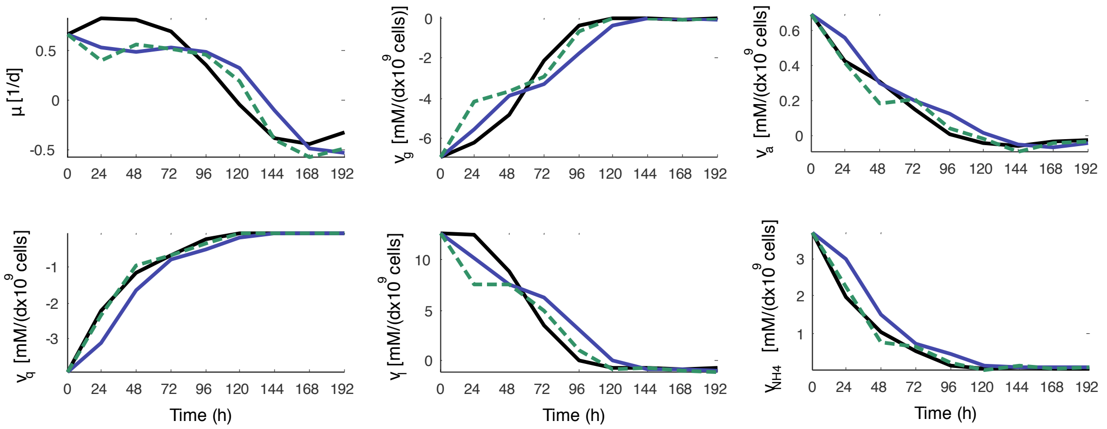


Figure 6.4. Extracellular fluxes (v_z) and biomass growth rate (μ) calculated from the measured concentrations. Fluxes have been calculated in three ways: using a middle-point Euler method (black, solid line), using a backward Euler method (green, dashed line), and using a backward Euler method coupled with a moving average filter of order 2 (blue solid line).

Step 3: flux estimation — measurements are *almost* sufficient (S1)

If the five measured fluxes are used $\{v_1\,(G), v_6\,(L), v_7\,(A), v_{19}\,(NH_4)$ and $v_{20}\,(Q)\}$ and it is assumed that the formation of purine and pyrimidine is the same $\{v_{22} = 0\}$, the MFA problem (4) is determined, but not redundant[1].

Using traditional MFA

We can use traditional MFA (TMFA) to determine the non-measured fluxes (see chapter IV for details). However, as it can be observed in Figure 6.5 (green solid line) the results obtained are not satisfactory:

- The estimated values at 24 h an 168 h for fluxes v_8, v_9, v_{10}, v_{11}, v_{12} and v_{21} seem unreasonable: the measured fluxes evolve smoothly, but these fluxes show peaks.
- The estimated fluxes v_8, v_9 and v_{10} do not fulfil the reversibility constraints (which, remember, are not considered in TMFA).
- To apply TMFA in an exactly determined case, we have to assume that there is no error in the measurements, which is unlikely, so the estimates are unreliable.

To show the last point, two new estimations have been done, one at 24 h with measured values for fluxes v_1 and v_6 slightly modified (+2% and -5% respectively), and another one at 168 h with a slight variation in the measurements for v_1 and v_6 (-0.05 and +0.05 mM/(d·10^9·cells), respectively). It can be observed in Figure 6.5 (red crosses) that the peak values in v_8, v_9, v_{10}, v_{11}, v_{12} and v_{21} are eliminated or reduced, while the values of the rest of fluxes remain almost unchanged. This indicates that the peaks at 24 h and 168 h were artefacts caused by slight errors in the measurements.

This illustrates the unreliability of TMFA in exactly determined cases: the impact of slight errors in the measured fluxes is not under control. These slight errors will exist in virtually all measured fluxes—they can even be consequence of the conversion step, as seen in Figure 6.4. This is why TMFA should not be used in scenarios without redundant measurements.

Using FS-MFA

The same scenario is now approached using FS-MFA instead of MFA.

If uncertainty is not considered and all reactions are considered reversible, FS-MFA provides the same solution that TMFA (results not shown). However, constraints can be incorporated for those reactions classified as irreversible (4). In this way we detect a high inconsistency at 24 h and a lower one at 144 h (i.e., the space $\mathbf{F}(k)$ is empty at these time instants). It must be pointed out that system (4) is not redundant, so TMFA

[1] The rank of $\mathbf{N_u}$ (16) is equal to the number of unknown fluxes (22-5-1). See chapter IV for details on this kind of analysis.

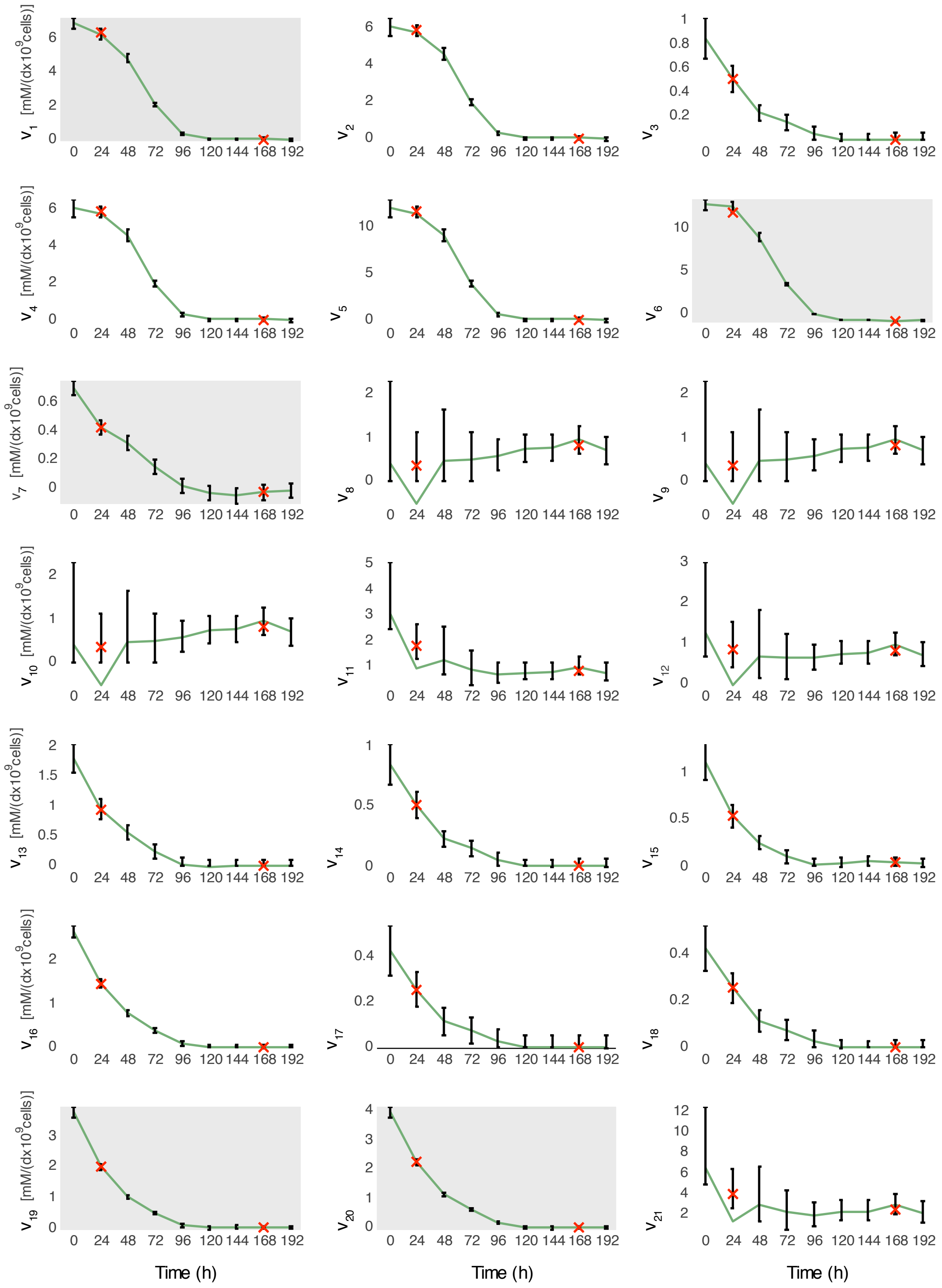


Figure 6.5. FS-MFA and TMFA in the determined and not redundant case (S1). Measured fluxes are, v_1 (G), v_6 (L), v_7 (A), v_{19} (NH_4), v_{20} (Q) and v_{22}. Measured fluxes have a grey background, and its uncertainty is represented with an interval. Fluxes estimated with FS-MFA are represented with an interval, and those estimated with TMFA with a green line. Two more TMFA estimations at 24 h and 168 h, from measurements of v_1 and v_6 slightly deviated from the original ones, are depicted with red crosses.

consistency analysis cannot be used; FS-MFA is detecting inconsistencies thanks to the reversibility constraints.

Now we consider the uncertainty in the measurements. We define a band of uncertainty around the measured values accounting for relative (5%) and absolute (0.1 mM/(d·10^9·cells)) errors around the values of the measured fluxes. Thus, the band around each measured flux $v_m(k)$ is defined as follows:

$$\text{If } v_m(k)\cdot \mathrm{E}_{\mathrm{rel}} \geq \mathrm{E}_{\mathrm{abs}} \quad \rightarrow \quad \text{band} = \left[v_m(k) + v_m(k)\cdot \mathrm{E}_{\mathrm{rel}}, v_m(k) - v_m(k)\cdot \mathrm{E}_{\mathrm{rel}}\right]$$

$$\text{Else} \quad \rightarrow \quad \text{band} = \left[v_m(k) + v_m(k)\cdot \mathrm{E}_{\mathrm{abs}}, v_m(k) - v_m(k)\cdot \mathrm{E}_{\mathrm{abs}}\right]$$

The relative error ($\mathrm{E}_{\mathrm{rel}}$) will be the dominant when the measured value is high, and the absolute one ($\mathrm{E}_{\mathrm{abs}}$) if it approaches zero. If more information about the measurements were available (e.g., sensors technical specifications), the range of uncertainty of each measured flux should be defined accordingly.

The results obtained with FS-MFA when uncertainty is accounted for are depicted in Figure 6.5. If compared with those given by TMFA, several conclusions arise:

- Reversibility constraints provide a method to detect inconsistencies. It can be easily checked that the solution provided by TMFA do not satisfy the reversibility constraints at 24 h (a negative value is given to three irreversible fluxes, v_8, v_9 and v_{10}). This inconsistency is detected and avoided with FS-MFA.
- Peaks at 24 h an 168 h for v_8, v_9, v_{10}, v_{11}, v_{12} and v_{21} are avoided with FS-MFA.[1]
- The uncertainty of experimental measurements is nontrivially propagated to the non-measured fluxes. For example, the estimates of v_8, v_9 and v_{10} are highly influenced by measurements uncertainty, while those of v_2, v_4, and v_5 are practically insensitive. Even if all fluxes could be estimated, FS-MFA says that the estimates for v_8, v_9 and v_{10} are less reliable, less precise, than those for v_2, v_4 and v_5. FS-MFA provides not only estimates for the fluxes, but also an indication of the reliability of these estimates.

In summary, this example shows that the described procedure gives richer estimates of time-varying fluxes in scenarios where there are not redundant measurements.

[1] We saw that the peaks were replaced by more sensible predictions if the measurements were slightly modified. As these modified measurements are enclosed by the band of uncertainty, the obtained intervals for v_8, v_9, v_{10}, v_{11}, v_{12} and v_{21} contain the sensible predictions. However, if a peak value violates the reversibility constraints, it will not be considered a valid solution, as happens at 24 h.

Step 3: flux estimation — measurements are sufficient (S2)

Consider now a scenario where the problem (4) is determined and redundant. Again, the use of TMFA is compared with that of FS-MFA.

- TMFA. When there are redundant measurements, TMFA can be applied with a two-step procedure: (1) exploit redundancies to detect gross errors and to adjust the measured fluxes, and (2) solve a weighted least squares problem to get a point-wise estimate for the non-measured fluxes. See chapter II for details.
- FS-MFA. It can be applied with the three step-procedure described in 6.2. Notice that it is still possible to exploit redundancies to detect gross errors, but instead of adjust the measured fluxes, FS-MFA defines a band of uncertainty around the measurements.

To get problem (4) determined and redundant, we need 7 measured fluxes.[1] There are only 6 available, so we assume that the flux of CO_2, v_{21}, was measured—assuming that it evolves smoothly and takes the values given by MFA in the previous section, except at 24 h and 168 h, which values are approximated by means of a spline curve.

First, we apply a χ^2-test to analyse the degree of inconsistency of the measurements at each time instant (see chapter II). The data fails the test at time 168 h (see Table 6.1), indicating that, if the model is correct, those measurements contains gross errors.[2]

Afterwards, we can estimate the non-measured fluxes during the cultivation of CHO cells using TMFA and FS-MFA. The results shown in Figure 6.6 indicate that FS-MFA can be also useful in this scenario:

- Even if there are no gross errors in the measurements, the point-wise estimate of TMFA can be unreliable due to uncertainty[3]. FS-MFA avoids this problem because its interval estimates are only as precise as allowed by the uncertainty, so they avoid this problem. To illustrate this problem, the fluxes that correspond to a set measurements near the original ones—within the band of uncertainty, and thus reasonably possible—have been highlighted in Figure 6.6 (dotted line). The evolution of some fluxes (e.g., v_8, v_9 and v_{10}) is clearly deviated from the estimates given by TMFA. Meanwhile, FS-MFA intervals indicate that two different interpretations of fluxes v_8, v_9 and v_{10} are possible: they can be stable around 0.6 or evolve from 0.2 to 0.7 mM/(d·10^9·cells). If there are evidences supporting one alternative over the other one, one could hypothesise which is more likely. In this way, accounting uncertainty in a richer way, FS-MFA reduces the number of wrong, or biased, predictions.

[1] The system will be redundant since the rank of $\mathbf{N_u}$ (15) is less than the number metabolites m (16).

[2] If we assume that the model is correct.

[3] Small changes in the measurements, which are expected, can have a large impact on the estimates.

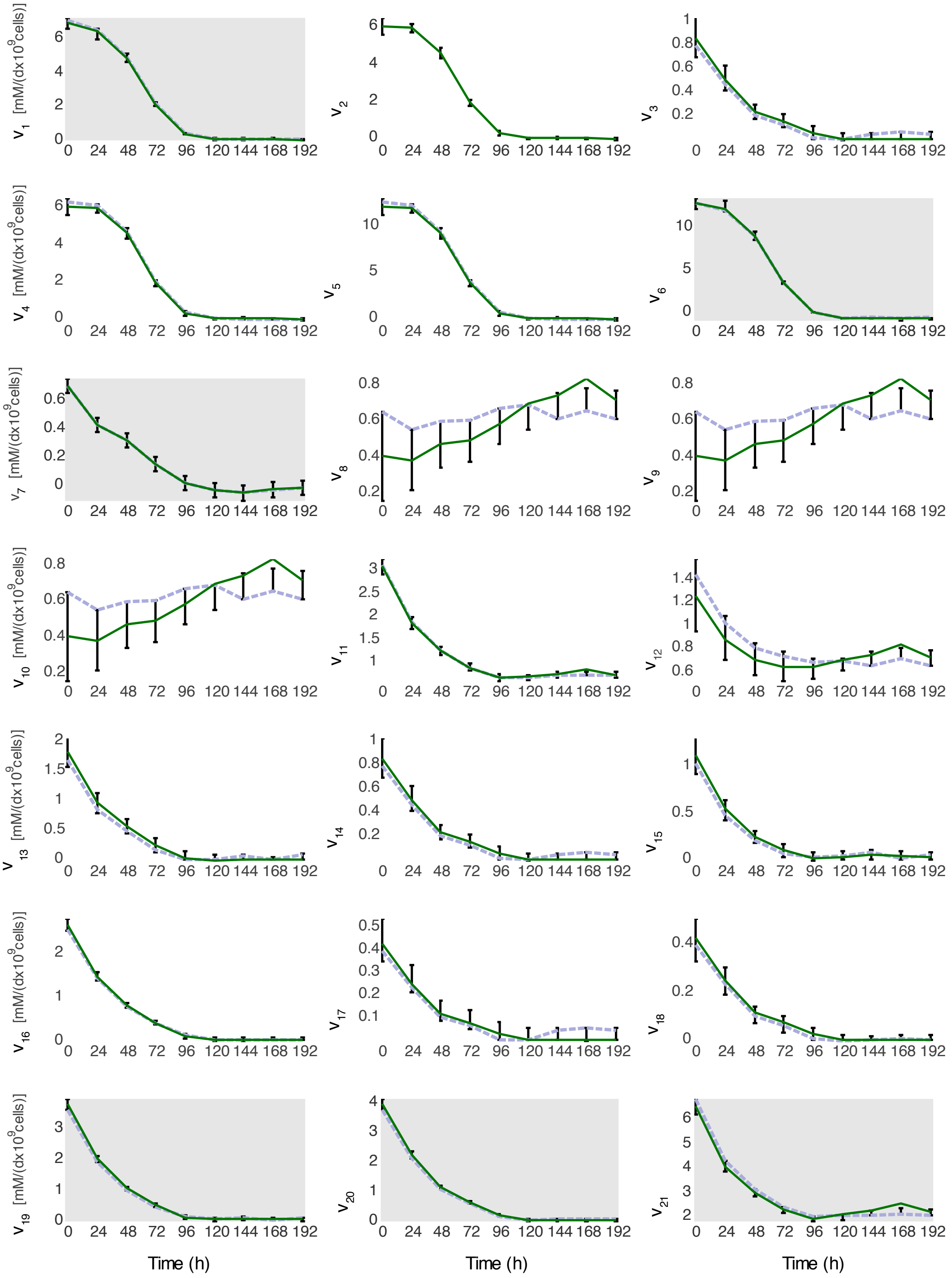


Figure 6.6. FS-MFA and TMFA in the determined and redundant case (S2). Measured fluxes are, *v_1 (G), v_6 (L), v_7(A), v_{19} (NH_4), v_{20} (Q), v_{21} (CO2)* and *v_{22}*. The measured fluxes have a grey background, and its uncertainty is represented with an interval. Fluxes estimated with FS-MFA are represented with an interval and those estimated with TMFA with a green line. Fluxes estimated with TMFA from measured values near the original ones—thus reasonably possible—are also depicted (blue dotted line) to show the undesired sensibility of TMFA results.

- Although there is a large error in measurements at 168 h, FS-MFA finds feasible flux vectors within the band of uncertainty (complementing the χ^2 test). Moreover, it gives an estimate accounting for the high uncertainty of measurements at 168 h. Conversely, TMFA estimates are sensitive to the large error—the value of v_{21} is significantly changed by the adjustment, resulting in a peak, and peaks appear also in v_8, v_9, v_{10}, v_{11} and v_{12}. In fact, TMFA estimates are usually discarded when measurements fail the χ^2 test because, being pointwise, they would be unreliable.
- Again, we see that measurements uncertainty is nontrivially transferred to the interval estimates. FS-MFA provides not only estimates of the fluxes, but also an indication of the reliability of these estimates.

This example shows that the described procedure provides richer estimates of time-varying fluxes also in scenarios where redundant measurements are available. This is the perfect scenario to apply TMFA—redundancies allow one to evaluate consistency and adjust the measurements—but FS-MFA still has some advantages: it gives more reliable estimates and handles larger inconsistency and uncertainty.

Table 6.1. χ^2 consistency test for a confidence level of 95%.

Time	0 h	24 h	48 h	72 h	96 h	120 h	144 h	168 h	192 h
Value h[a]	0	3.02	0.0001	0	1	2	294	37.94	0

[a] The test fails when h > 3.84 (i.e., h>χ^2).

Step 3: flux estimation — measurements are insufficient (S3)

In this section it is shown that the procedure can be used even when the available measurements are insufficient, i.e., when the problem (4) is underdetermined. Notice that in this situation TMFA cannot be applied.

The procedure is applied using different sets of 4 and 5 measured fluxes (remember that 6 was necessary to get a determined system). Uncertainty is also accounted for, using the band described above. All results are given in Table 6.2, and two illustrative cases are depicted in Figure 6.7.

With 4 sets of 5 measurements (G, F, E and C) the evolution of all the non-measured fluxes can be estimated. Case G, where v_{22} is not known, provides the best results. There is a mean interval increment of 39% over the determined case and the increment is minor than 25% for 12 fluxes out of 17. This case is depicted in Figure 6.7 (green). The interval estimates are practically the same as in the determined case for most fluxes (v_2, v_4, v_5, v_8, v_9, v_{10}, v_{11}, v_{12}, v_{13}, v_{15} and v_{21}). Estimates for v_3 and v_{14} are larger, but still accurate, and only the estimates for v_{16}, v_{17} and v_{18} are imprecise. Moreover, the temporal evolution—that can be roughly characterised by using the

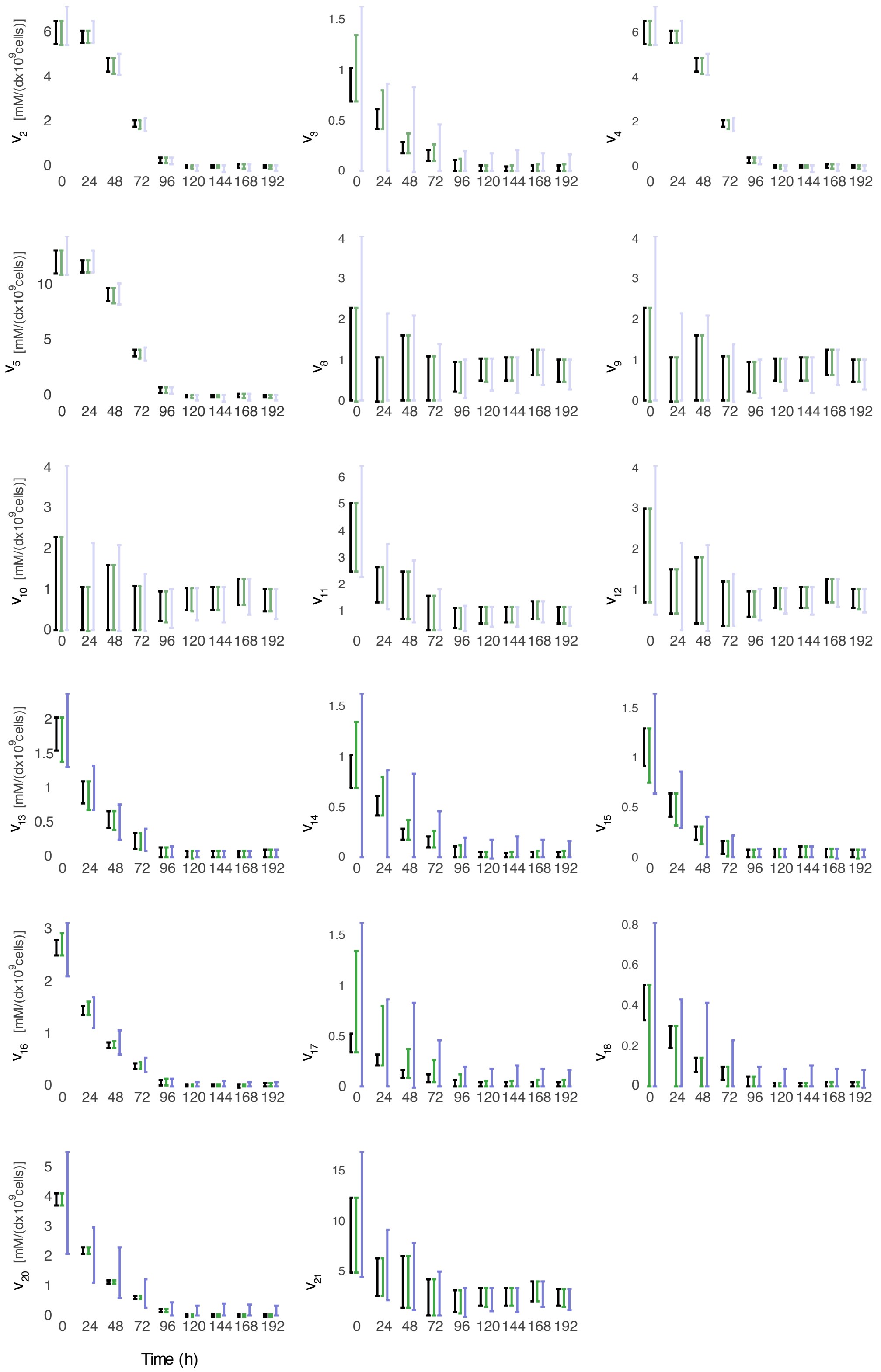


Figure 6.7. FS-MFA in two underdetermined cases (S3). Interval estimates obtained using 5 measurements {v_1, v_6, v_7, v_{19} and v_{20}} are depicted in green (second interval), and those obtained from {v_1, v_6, v_7 and v_{19}} in blue (third interval). To be used as reference, the estimates obtained in the determined case, when 6 fluxes were measured, are depicted in black (first interval).

Table 6.2. Comparison of different flux estimations.

	Ref (v_1, v_6 , v_7, v_{19}, v_{20}, v_{22})	G (no v_{22})		F (no v_{20})		E (no v_{19})		B (no v_6)		A (no v_1)		C (no v_7)		I (no v_{20} v_{22})		H (no v_{19} v_{22})	
Reactions [d]	MI[a] [b]	MI [a]	[%[c]]	MI [b]	[%]	MI [b]	[%]	MI [b]	[%]	MI [b]	[%]	MI [b]	[%]	MI [b]	[%]	MI [b]	[%]
1: G→G6P	0.267[e]	0.267	-	0.267	-	0.267	-	0.267	-	x	-	0.267	-	0.267	-	0.267	-
2: G6P→G3P+DAP	0.367	0.387	5%	0.628	71%	0.541	47%	0.398	8%	x	-	0.627	71%	0.628	71%	0.572	56%
3: G6P→R5P+CO2	0.131	0.199	53%	0.526	303%	0.340	160%	0.131	0%	0.131	0%	0.401	207%	0.526	303%	0.383	193%
4: DAP→G3P	0.367	0.387	5%	0.628	71%	0.541	47%	0.398	8%	x	-	0.627	71%	0.628	71%	0.572	56%
5: G3P→Pyr	0.735	0.774	5%	1.256	71%	1.082	47%	0.795	8%	x	-	1.253	71%	1.256	71%	1.144	56%
6: Pyr→L	0.475	0.475	-	0.475	-	0.475	-	x	-	0.475	-	0.475	-	0.475	-	0.475	-
7: Pyr+Glu→A+aKG	0.100	0.100	-	0.100	-	0.100	-	0.100	-	0.100	-	1.488	inf	0.100	-	0.100	-
8: Pyr→ACA+CO2	1.031	1.031	0%	1.562	51%	1.901	84%	x	-	x	-	0.957	-7%	1.562	51%	1.906	85%
9: Oxa+ACA→Cit	1.031	1.031	0%	1.562	51%	1.901	84%	x	-	x	-	0.957	-7%	1.562	51%	1.906	85%
10: Cit→aKG+CO2	1.031	1.031	0%	1.562	51%	1.901	84%	x	-	x	-	0.957	-7%	1.562	51%	1.906	85%
11: aKG→Mal+CO2	1.156	1.156	0%	1.604	39%	2.532	119%	x	-	x	-	1.443	25%	1.604	39%	2.530	119%
12: Mal→Oxa	0.994	0.994	0%	1.398	41%	1.769	78%	x	-	x	-	1.093	10%	1.398	41%	1.769	78%
13: Mal→Pyr+CO2	0.209	0.240	15%	0.352	68%	0.920	341%	0.209	0%	0.209	0%	0.903	332%	0.352	68%	0.918	340%
14: Oxa+Glu→Asp+aKG	0.131	0.199	53%	0.526	303%	0.340	160%	0.131	0%	0.131	0%	0.401	207%	0.526	303%	0.383	193%
15: Glu→aKG+NH4	0.150	0.182	21%	0.298	98%	0.870	479%	0.150	0%	0.150	0%	0.586	289%	0.298	98%	0.870	479%
16: Q→Glu+NH4	0.117	0.145	23%	0.325	177%	0.553	372%	0.117	0%	0.117	0%	0.569	386%	0.325	177%	0.548	367%
17: R5P+Asp+Q→Pu	0.104	0.277	165%	0.293	181%	0.200	91%	0.104	0%	0.104	0%	0.225	116%	0.526	404%	0.383	267%
18: R5P+Asp+2Q→Py	0.078	0.132	69%	0.283	262%	0.177	126%	0.078	0%	0.078	0%	0.209	168%	0.263	237%	0.163	108%
19:→ NH4	0.141	0.141	-	0.141	-	1.419	904%	0.141	-	0.141	-	0.141	-	0.141	-	1.412	899%
20:→ Q	0.132	0.132	-	1.127	752%	0.132	-	0.132	-	0.132	-	0.132	-	1.107	737%	0.132	-
21:→ CO2	3.338	3.338	0%	4.770	43%	6.966	109%	x	-	x	-	3.843	15%	4.770	43%	6.966	109%
22: Pu-Py (constraint)	0.100	0.354	254%	0.100	-	0.100	-	0.100	-	0.100	-	0.100	-	0.526	426%	0.383	283%
Mean	0.554	0.587	39%	0.899	155%	1.138	196%	0.217	2%	0.156	0%	0.802	122%	0.927	180%	1.168	214%
Measured fluxes [number]	6	5		5		5		5		5		5		4		4	
<25% (≈Ref)			12		0		0		10		7		5		0		0
25-100% (<2·Ref)			3		11		8		0		0		4		11		7
100-300% (2 -4·Ref)			2		3		5		0		0		5		2		7

Column Ref: FS-MFA is applied using the six available measurements (determined case). Columns F, G, E B, A and C: FS-MFA is applied using sets of 5 measurements in each case (underdetermined, 1 degree of freedom). Columns I and H: FS-MFA is applied using two different sets of 4 measurements (underdetermined, 2 degrees of freedom). In all cases the band of uncertainty described in the text has been used. [a] Mean interval size along time evolution; [b] in [mM/(d·10^9·cells)]; [c] Intervals enlargement w.r.t. case Ref. (in percentage); [d] The nomenclature was given in the chapter IV; [e] Measured values are in bold.

middle point of the intervals—is always similar to the determined case. Case C, where v_7 is not measured, provides also very good results; all fluxes are predicted with a mean interval increment of 122%. The interval increment is minor than 25% for 5 fluxes, and minor than 100% for 9 fluxes. Case F, where v_{20} is not measured, provides good results too. Case E, where v_{19} is not measured, provides slightly worse results than F. With the other two sets of 5 measurements (B and A), some non-measured fluxes cannot be estimated, but the estimated ones (10 and 7, respectively) are exactly the same that in the determined case.

Two sets of 4 measurements have been also considered (I and H). Case I, where v_{20} and v_{22} are not measured, provides remarkable results. There is a mean interval increment of 180% over the determined case, and the increment is minor than 100% for 11 fluxes. This case is depicted in Figure 6.7 (blue). The interval estimates are similar to the determined case for most fluxes. Those for v_{16} and v_{20} are wider, but still useful, and only v_3, v_{14}, v_{17} and v_{18} are highly imprecise.

This section illustrates an important feature of the procedure: it is able to estimate the metabolic fluxes during a cultivation process in scenarios of data scarcity, when measurements are uncertain and scarce.

6.4 Case study: CHO cells under uncertainty

In this section CHO cells case is used to analyse two issues regarding measurements uncertainty. We first discuss how the uncertainty is propagated to the estimations. Afterwards, we describe a simple approach to investigate which measurements should be more accurate to improve the precision of particular estimates.

The propagation of the uncertainty is unbalanced

As shown in previous sections, the uncertainty of the experimentally measured fluxes is not equally propagated to all the estimated fluxes. The structure of the constraint-based model (stoichiometry and reactions reversibility) determines how the uncertainty is propagated. A convenient way to investigate this effect is to calculate the interval sizes of each estimated flux and time instant (both in absolute and relative terms).

First, consider the aggregated average interval size (AIS) of each estimated flux (Table 6.3). It can be observed (determined case) that certain fluxes, such as v_{10}, v_{12} and v_{21}, are highly affected by the uncertainty of the measurements—they have an average interval size larger than 1 mM/(d·10^9·cells). Other fluxes, such as v_{14} and v_{17}, are less sensitive (values around 0.1 mM/(d·10^9·cells)). Obviously, smaller fluxes tend to be more affected by the uncertainty, in relative terms, but this phenomenon is not the only responsible for the unbalanced propagation of the uncertainty. For example, the estimated v_8 and v_{14} have similar magnitudes, but the effect of the uncertainty over them is dramatically different: v_8 is the more influenced flux (AIS of 90% in relative terms), while v_{14} is quite insensitive (AIS of 15%). Another example is given by v_{21}, one of the fluxes with larger magnitude, but highly affected by uncertainty (AIS interval size of 3.4 mM/(d·10^9·cells) or 39%).

The data given in Table 6.3 also provides a quantitative indication of the benefits of redundant measurements. When seven fluxes are measured instead of six, the estimates are more precise: the intervals are reduced a 71% on average. This is improvement is particularly significant for those fluxes poorly estimated in the determined case (reduction of 78% for v_8, v_9 and v_{10} and 76% for v_{12}).

Similar data, but aggregated with respect to time instants instead of fluxes, are given in Table 6.4. The same analysis could be done to evaluate the imprecision of each estimate, per flux and time instant, if this is considered necessary.

Table 6.3. Imprecision of the estimated fluxes caused by measurements uncertainty.

	Determined case			Determined / redundant case			Comparison	
	Max. [a]	AIS [a]	AIS [%b]	Max. [a]	AIS [a]	AIS [%b]	Diff. [a]	Diff. [%]
v_2	6.041	0.377	6.25%	6.032	0.321	5.32%	0.057	14.97%
v_3	0.853	0.129	15.12%	0.859	0.123	14.35%	0.006	4.41%
v_4	6.041	0.377	6.25%	6.032	0.321	5.32%	0.057	14.97%
v_5	12.081	0.755	6.25%	12.065	0.642	5.32%	0.113	14.98%
v_8	1.166	1.053	90.37%	0.715	0.231	32.32%	0.822	78.07%
v_9	1.166	1.053	90.37%	0.715	0.231	32.32%	0.822	78.07%
v_{10}	1.166	1.053	90.37%	0.715	0.231	32.32%	0.822	78.07%
v_{11}	3.769	1.180	31.30%	3.073	0.165	5.37%	1.015	86.02%
v_{12}	1.854	1.017	54.89%	1.263	0.241	19.05%	0.777	76.34%
v_{13}	1.813	0.209	11.52%	1.809	0.195	10.78%	0.014	6.58%
v_{14}	0.853	0.129	15.12%	0.859	0.123	14.35%	0.006	4.41%
v_{15}	1.113	0.150	13.52%	1.109	0.147	13.27%	0.003	2.11%
v_{16}	2.665	0.117	4.39%	2.668	0.114	4.26%	0.003	2.91%
v_{17}	0.426	0.101	23.64%	0.442	0.087	19.60%	0.014	14.10%
v_{18}	0.426	0.079	18.42%	0.417	0.063	15.17%	0.015	19.48%
v_{21}	8.698	3.407	39.17%		-	-	-	-
Mean		699	3.231%		202	1.432%	497	71,09%

Max: maximum value of the estimated flux along time. AIS: average interval size for each estimated fluxes (over time). Diff: difference between determined and over-determined cases; [a] in mM/(d·10^9·cells). [b] average interval size for each estimated flux expressed w.r.t. its maximum value.

Table 6.4. Summary of results for each time instant (determined and overdetermined cases).

	Determined case		Determined / redundant case		Comparative	
	AIS [a]	AIS [%b]	AIS [a]	AIS [%b]	Diff. IS [a]	Diff. [%]
0h	1.617	73.99%	0.518	34.14%	1.099	67.96%
24 h	0.835	38.66%	0.295	21.68%	0.540	64.69%
48h	1.083	48.18%	0.283	16.30%	0.799	73.85%
72h	0.737	34.37%	0.202	14.19%	0.536	72.66%
96h	0.468	22.64%	0.151	11.66%	0.317	67.75%
120h	0.382	17.93%	0.089	7.42%	0.293	76.71%
144h	0.382	17.90%	0.102	8.08%	0.280	73.23%
168 h	0.392	18.43%	0.077	6.94%	0.315	80.31%
192h	0.397	18.68%	0.103	8.46%	0.295	74.18%
mean	0.699	32.31%	0.202	14.32%	0.497	71.09%

AIS: average interval size of the estimated fluxes at each time instant. Diff: difference between determined and overdetermined cases. [a] in mM/(d·109·cells). [b] average at each time instant of the interval sizes of the calculated fluxes expressed w.r.t. the maximum value.

The propagation of the uncertainty is nonlinear

In the previous section it was show that the propagation of the uncertainty from the measured fluxes to the estimated ones is not balanced. Herein the non-linearity of this propagation is analysed.

We have performed 15x15 instances of the estimation procedure for different degrees of uncertainty in two measured fluxes, v_1 and v_6 (between ±2% and ±30%). Then, we calculate the averaged interval size for one of the estimated fluxes, v_2. In this way, we can analyse the effect over the estimate of both sources of uncertainty.

Figure 6.8 shows the averaged interval size (AIS) of the estimated v_2 for each instance. As expected, the intervals tend to increase as uncertainty increases. It is also clear that the uncertainty of the two measurements has not the same effect. The effect of uncertainty in v_6 over v_2 is larger than the effect of uncertainty in v_1.

Figure 6.8 also shows the non-linearity of the propagation of the uncertainty from the measurements to the estimates. Let $\mathrm{f}(u_i)$ be the interval size of an estimated flux, such as v_2, when measurements uncertainty is u_i, then:

c. The propagation does not satisfy the principle of superposition,

$$\mathrm{f}(u_1) + \mathrm{f}(u_2) \neq \mathrm{f}(u_1 + u_2)$$

d. The propagation of uncertainty does not satisfy the principle of homogeneity,

$$\mathrm{f}(k \cdot u_1) \neq k \cdot \mathrm{f}(u_1)$$

To highlight (a), the result of summing up the independent effect of the uncertainty of v_6 and v_1 has been depicted with black dots in Figure 6.8. When the uncertainty is low, f(u_1) + f(u_1) > f(u_1 + u_2), but this is inverted when uncertainty increases, and, f(u_1) + f(u_1) < f(u_1 + u_2). It can be observed that the effect of the uncertainty of v_1 is not important by itself, but it is boosted in combination with the uncertainty of v_6. Regarding (b), Figure 6.8 clearly show that, f($k \cdot u_1$) ≠ $k \cdot$f(u_1). For example, assume that the uncertainty of v_6 is fixed in 10% (fourth row in the right top figure). The effect of adding the first 4% of uncertainty to v_1 is higher than the effect of adding a second one, and after 16%, more uncertainty has practically no effect (there is a saturation).

Therefore, the relationship between the uncertainty of the measurements and the precision of the estimates is a complex one: different for each estimate and clearly non-linear. Interestingly, this means that the estimation procedure described earlier in the chapter provides non-trivial information in this respect.

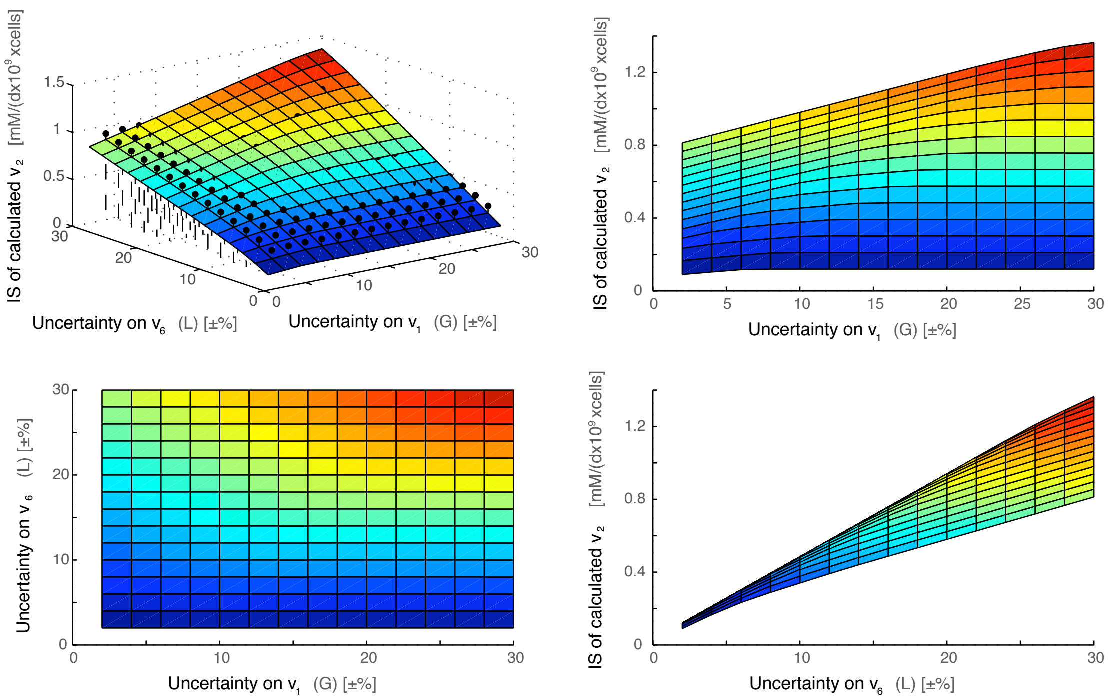


Figure 6.8. Effect over the estimated v_2 of the uncertainty of the measured v_1 and v_6. The surface (and its projections) represents the averaged interval size (AIS) of the estimated v_2 when different degrees of uncertainty are considered for the measured v_1 and v_6. In the top left figure, the result of summing up the independent effect of v_6 uncertainty and v_1 uncertainty is shown with black dots.

Analysing the effect of the uncertainty of each measurement

In this section two methods are proposed to investigate how the precision of the estimated fluxes can be improved acting on the measurements.

- Direct approach. Calculate the *increase* of the imprecision of the estimates when the uncertainty of one measured flux is *increased*.
- Indirect approach. Calculate the *reduction* of the imprecision of the estimates when the uncertainty of one measured flux is *decreased*.[1]

The direct approach, similar to a classical analysis of sensitivity, will be useful during the setting-up of a process plant to choose the equipment, sensors, and the measuring protocols. On the other hand, given a current setting (equipment, protocols, etc.), the indirect approach indicates which fluxes should be more accurately measured (e.g., using an accurate sensor or taking redundant measurements), if we want to improve the precision of a particular estimate, for a flux of interest, and even at a critical phase of the cultivation process.

The procedure to perform the indirect analysis can be outlined as follows:

For each measured flux $v_{m,x}(k)$ in $\mathbf{v_m}(k)$

Step 1 Apply FS-MFA to get interval estimates for each flux $v(k)$, considering:

±5% of uncertainty $\forall v_{m,i}(k),\ i \neq x$

±2% of uncertainty $v_{m,x}(k)$

**Particular values of 5% and 2% are just an example.*

Step 2 Calculate the interval size for each estimated flux $v(k)$.

Step 3 Quantify the reduction of imprecision for each estimated flux $v(k)$:

$$Red = 100 - \frac{(\mathrm{BC} - \mathrm{D_X})}{(\mathrm{BC} - \mathrm{WC})} \cdot 100$$

where $\mathrm{D_x}$ is the interval size of $v(k)$, WC its interval size in a worst-case [±5% of uncertainty $\forall v_m(k)$], and BC its interval size in a best-case [±2% of uncertainty $\forall v_m(k)$].

Note: the direct analysis can be formulated in an analogous way.

[1] Notice that increase and reduction are not inverse, i.e., $f(u+x) + f(u-x) \neq f(0)$.

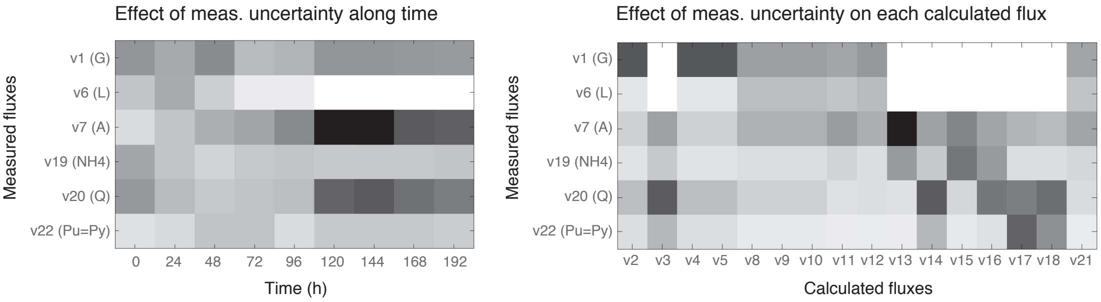


Figure 6.9. Improving the accuracy of the measurements. (Left) Average reduction of the imprecision at each time instant when the uncertainty of measured fluxes decreases a 3%. (Right) Average reduction of the imprecision of each estimated flux when the uncertainty of measured fluxes decreases a 3%. Reductions quantified between 0% (white) and 100% (black).

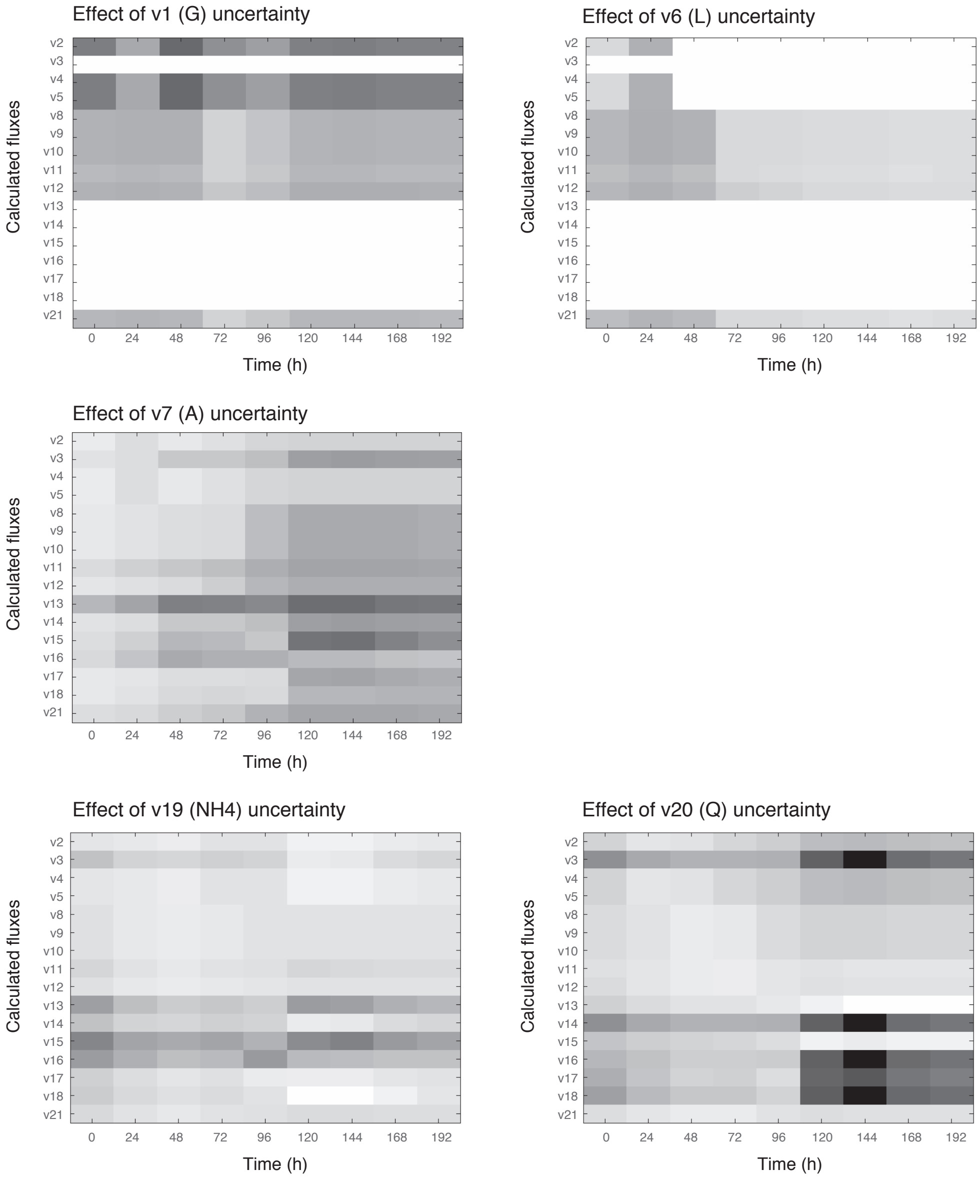


Figure 6.10. Effect over the estimates of improving the accuracy of the measurements. Each figure depicts the reduction of the imprecision of the estimated fluxes when the uncertainty of a measured one decreases a 3%. Reductions quantified between 0% (white) and 100% (black).

The indirect analysis has been applied to the cultivation of CHO cells. The results, given in Figure 6.10, show how the imprecision of the estimated fluxes is reduced when the uncertainty of measured fluxes decreases a 3%. For example, the results indicate that the larger improvement of the estimates will occur if the uncertainty of v_{20} is reduced at 144 h: the imprecision of the estimates for v_3, v_{14}, v_{16} and v_{18} is reduced by more than an 85%. It can be also observed that during the first 96 h, reducing the uncertainty in v_{20} reduces only slightly the imprecision of the estimated v_{16}, but this reduction is very important between 120 h and 192 h. It is also clear that reducing the uncertainty of v_1 or v_6 has no effect over the estimates v_3, v_{14}, v_{15}, v_{16} v_{17} and v_{18}. These data will be valuable to improve the estimations (or an on-line monitoring system).

A summary of the direct analysis can be given in a more compact way, as in in Figure 6.9. These figures can be used to improve our estimations in a rational manner. Some examples are given below:

- If one is interested in increasing the precision of the estimated v_3, the best intervention will be to reduce the uncertainty of the measured v_{20}.
- If we want to improve the estimations during the transition phase (between 72 h and 120 h) we should reduce the uncertainty of v_7.
- If we prefer to improve the overall precision of all the estimations, we should reduce the uncertainty in the measured v_7, although reducing the uncertainty of v_1 or v_{20} brings similar benefits.

6.5 Conclusions

In this chapter we have presented a procedure to estimate time-varying metabolic fluxes during a cultivation process, which handles data scarcity and measurements uncertainty. The procedure has been illustrated with a real case study: the estimation of the intra- and extracellular fluxes of CHO cells cultivated in batch mode.

Previous approaches to this problem used traditional MFA to perform the flux estimation (Herwig, 2002; Takiguchi, 1997; Henry, 2007; Ren, 2003). However, it has been shown that the flux-spectrum, introduced in chapter IV, has advantages. The flux-spectrum gives interval estimates instead of point-wise ones, thus allowing to apply the estimation procedure even if measurements are insufficient, or imprecise.

The flux estimation procedure can be applied off-line (with collected data), providing insight into the time-varying behaviour of the organism. This can help to understand its dynamic regulation, and its adaptation to environmental conditions. It can be also useful for physiological studies, to characterise strains, or to guide improvements of strains and processes. The procedure could serve as basis for on-line monitoring processes in industrial environments, where reliable on-line sensors are lacking.

In summary, it has been shown how a constraint-based model and set of measurements of metabolites concentrations can be used to estimate time-varying metabolic fluxes during a cultivation process, even in scenarios of data scarcity and uncertainty.

Main references

- Llaneras F, Picó J (2007). A procedure for the estimation over time of metabolic fluxes in scenarios where measurements are uncertain and/or insufficient. *BMC Bioinformatics*, 8:42.

- Llaneras F, Picó J (2007). An interval approach for dealing with flux distributions and elementary modes activity patterns. *Journal of Theoretical Biology*, 246:290-308.

- Henry O, Kamen A, Perrier M (2007). Monitoring the physiological state of mammalian cell perfusion processes by on-line estimation of intracellular fluxes. *Journal of Process Control*, 17:241-251.

- Takiguchi N, Shimizu H, Shioya S (1997). An on-line physiological state recognition system for the lysine fermentation process based on a metabolic reaction model. *Biotechnology & Bioengineering*, 55:170-181.

- Herwig C, Marison I, von Stockar U (2001). On-line stoichiometry and identification of metabolic state under dynamic process conditions. *Biotechnology & Bioengineering*, 75:345-354.

- Provost A (2006b). *Metabolic design of dynamic bioreaction models*. PhD Thesis, Université catholique de Louvain, Louvain-la-Neuve.

Part III: Possibilistic methods

VII

Possibilistic framework to analyse consistency and estimate the metabolic fluxes

This chapter discusses the use of possibility theory in the context of constraint-based models. We introduce a unifying possibilistic framework to (a) evaluate consistency between model and measurements, and (b) provide rich estimates of the metabolic fluxes. The framework is shown to be flexible, reliable, usable under data scarcity, and computationally efficient.

Part of the contents of this chapter appeared in the following journal article:

- Llaneras F, Sala A, Picó J (2009). Possibilistic framework for constraint-based metabolic flux analysis. *BMC Systems Biology*, 3:79.

7.1 Introduction

Constraint-based models define the possible metabolic states or behaviours that can be exhibited by the cell; however, they do not predict which of these are likely under given circumstances. One approach to perform these predictions is flux balance analysis (FBA), which assumes that cells behaviour has evolved to be optimal in a certain sense (Price et al., 2003). It has been shown that FBA is able to predict the actual fluxes (Schuetz, 2007; Edwards, 2001; Schilling, 2002), but this requires to identify which are the relevant objectives for different conditions (Schuster, 2008; Schuetz, 2007). As an alternative, one could perform a metabolic flux analysis (MFA) which, generally speaking, is the exercise of estimating the fluxes shown by cells by combination of a constraint-based model and the available experimental measurements.

One difficulty to be tackled by MFA is that the available measurements are often insufficient to estimate the intracellular fluxes, particularly in large-scale networks, because there may be different flux states compatible with the measurements. To face this situation, one could choose one flux vector among those that are compatible with the measurements. For instance, Nookaew et al. have proposed to estimate the intracellular fluxes based on the assumption that cells are likely to use as many pathways as possible to maintain robustness and redundancy (Nookaew, 2007). Related hypotheses have been formulated using the concept of elementary modes (Poolman, 2004; Schwartz, 2006). The assumption of optimal cell behavior typically used in FBA could be also used (Schuetz, 2007). Another option to face a lack of measurements is to incorporate intracellular information obtained from stable isotope tracer experiments (Sauer, 2006; Szyperski, 1998; Wiechert, 2001). Yet, data from isotope tracer experiments are still rarely available, and will not be considered in this work. Instead, we follow a constraint-based modelling approach, in the sense that we do not attempt necessarily to predict the actual fluxes with precision, but rather to distinguish "most possible" from "impossible" flux states, based on a suitable definition of "possibility".

With this purpose in mind, this chapter presents a possibilistic framework for MFA. Uncertainty, lack of measurements and model imprecision will be handled introducing the notion of "degree of possibility". Then, an efficient optimisation-based approach will be employed to query the most possible fluxes and their possibility distributions.[1] The methodology is based on a reinterpretation of the consistent causal reasoning paradigm (Dubois, 1995) as an equivalent problem of feasibility subject to equality and inequality constraints. Preferences under uncertain knowledge are incorporated by transforming the feasibility problem into a linear optimisation one, which may be interpreted in possibilistic terms. The optimisation approach to logic reasoning has been previously explored in (Sala, 1998; Sala, 2001; Sala, 2008).

[1] Our proposal is a new example of how profusely mathematical optimization is used to research in systems biology. Other examples can be found in (Banga, 2008).

The main features of the framework introduced herein, that will be called Possibilistic MFA (Poss-MFA), can be summarised as follows:

- Poss-MFA exploits a constraint-based model, not only stoichiometric balances.
- It considers measurements uncertainty and model imprecision in a flexible way (e.g., non-symmetric error or a band of uncertainty due to systemic error).
- It provides possibility distributions (and intervals) which are more informative than point-wise estimations if multiple flux values are be reasonably possible.
- It is reliable even if only a few fluxes are measurable.
- It can detect, and handle, inconsistencies between measurements and model.
- Furthermore, it has high computational efficiency.

The chapter is organised as follows. Preliminaries on possibility, optimisation and metabolic flux analysis are first addressed. Afterwards, the basics of Possibilistic MFA and some refinements are discussed, and the framework is illustrated with examples and with a case study using a well-know model of *C. glutamicum*. The main conclusions are outlined to close the chapter.

7.2 Preliminaries: possibility and optimisation

In an abstract ideal situation, many estimation problems in science and engineering can be cast as estimating some decision variables δ given the known values of a set of other ones m (e.g., measurements) and a model expressed as a set of equality and inequality constraints (involving decision variables, measurements and some model parameters). Then, the valid estimations will be the feasible solutions of a constraint satisfaction problem (Kumar, 1992; Russell, 2003).

However, in many practical cases, the measurements are imprecise and the model parameters and constraints are also not accurate, so real data violates them. This is the reason why most real-life models should include uncertainty. The most basic representation of uncertainty would be giving interval values to measurements and model parameters. Refinements of the uncertainty representation lead to probabilistic (Russell, 2003; Jensen, 1996; Hand, 1993) and possibilistic (Yager, 1983; Dubois, 1988; Zadeh, 1981) frameworks.

Probabilistic frameworks have an underlying interpretation in terms of the frequency in which some conditions appear; on the other hand, possibilistic frameworks measure the degree of compliance (consistency) of some decision variables with some (soft) modeling constraints. In this sense, the basic assumptions of both paradigms of inference under uncertainty are different.

In the following subsections the possibilistic framework will be described. Afterwards, the relationship between probability and possibility will be discussed to justify the use the possibilistic framework.

Soft constraint satisfaction problems: a possibilistic approach

As explained above, the possibilistic framework is the chosen representation for the problem under study, following the ideas in (Dubois, 1996), where possibilistic constraint satisfaction problems (CSP) are presented. There, the authors introduce constraints which are satisfied to a degree, transforming the feasibility/unfeasibility of a potential solution into a *gradual* notion: given a CSP with multiple solutions $\delta \in \Delta$ (where Δ denotes the search space over which feasible values for the decision variables will be searched), a function $\pi : \Delta \rightarrow [0,1]$ was suggested in order to represent preference or priority as a "consistency degree". The meaning of $\pi(\delta)=1$ would indicate that δ is in full agreement with the model and measurement constraints; the meaning of $\pi(\delta)=0$ indicates that δ is in "absolute, total contradiction" with the problem constraints, and never should be considered a feasible value. Intermediate values would denote values of decision variables which "somehow mildly" violate the problem constraints but could be considered "partially possible" from the "practical" knowledge of the "expert" modeller who defined π. The higher the value of $\pi(\delta)$, the higher the accordance with the problem constraints should be (subjectively interpreted as a higher "possibility" of the decision variable choice δ). Given this subjective meaning of π, it is denoted in literature as *possibility distribution*. The possibilistic calculus (Dubois, 1988; Dubois, 1996) refers then to computations with possibility distributions from a series of axioms. Basic ideas on it will be outlined below in this section. A simple example illustrates the basic idea.

Example Consider a flux balance $\{f_1 = f_2\}$, stating equality between two flows, f_1 and f_2, supposedly measured in a biological or chemical reaction. The measurements $m_a = (5, 7)$ and $m_b = (5, 5.1)$ are unfeasible, whereas $m_c = (5, 5)$ is feasible. However, it seems clear that the subjective "possibility" of m_b is higher than that of m_a —m_b can be thought to be quite reasonable in practice due to measurement errors. The idea can be easily formalised for further computations by defining a possibility distribution, for instance: $\pi(f_1, f_2) = e^{-(f_1 - f_2)^2}$.

In this way, potential solutions can be ranked: $\pi(m_a) = 0.018$, $\pi(m_b) = 0.99$ and $\pi(m_c) = 1$. The search space in which to define the possibility, Δ, could be defined as, say, $\Delta = \{(\delta_1, \delta_2) \,|\, 0 \le \delta_i \le 10\}$.

Usually, the function $\pi(\delta)$ is built by conjunction of possibility functions of individual relations $\pi_i(\delta_i)$ (expressing user-defined preference or priority on each individual constraint, in many cases in a problem-dependent way). Such conjunction will be latter discussed in this section. The best CSP solutions are defined to be those which satisfy the global problem to the maximal degree.

In this way, once the user has defined such function expressing how a particular combination of system variables is "consistent" with its model, the basic idea on possibilistic calculus is, given a subset of the system variables (assumed as known or measured), estimate the "most possible" values of all the remaining variables via an optimisation problem. The close relationship between possibilistic calculus and optimisation is discussed below.

Possibility theory

The basic building block of possibility theory is a user-defined possibility distribution $\pi:\Delta\rightarrow[0,1]$. This defines the possibility of each "point" δ in Δ. A consistent problem formulation is defined to be the one in which there exists at least one point with possibility equal to one.

The second building block are events, formally defined as subsets of Δ, in order to address problems such as, in the above example, determining the possibility of event $A = \{(f_1,f_2) \in \Delta \mid 0 \le f_1 \le 3, 4 \le f_2 \le 10\}$.

Possibility calculus as optimisation. By definition, the possibility of an event A (subset of Δ) is computed via:

$$\pi(\mathrm{A}) = \sup_{\delta\in \mathrm{A}} \pi(\delta) \tag{1}$$

and, obviously, given two events A and B, $A \subset B$ entails $\pi(A) \le \pi(B)$. Hence, possibility computations are optimisation problems[1].

For a multidimensional $\Delta = \Delta_1\times\Delta_2$, $\delta = (\delta_1, \delta_2) \in \Delta$, the marginal possibility distribution of δ_1 is defined as:

$$\pi(\delta_1^*) = \sup_{\delta_2\in\Delta_2} \pi(\delta_1^*, \delta_2) \tag{2}$$

i.e., the possibility of the event $\{\delta_1 = \delta_1^*\}$.

Optimisation as possibility calculus. Conversely, consider a cost function $J:\Delta\rightarrow\mathbf{R}^+$ (i.e., verifying $J(\delta) \ge 0$ for all $\delta \in \Delta$), so that there exists $\delta_0 \in \Delta$ such that $J(\delta_0) = 0$. Then, a consistent possibility distribution may be defined on Δ via:

[1] Cf. with probability computations, which are integration problems.

$$\pi(\delta) = \mathrm{e}^{-\mathrm{J}(\delta)} \quad \delta \in \Delta \tag{3}$$

and the possibility of an event A is given by replacing the possibility definition (3) in (1), resulting in:

$$\pi(\mathrm{A}) = \mathrm{e}^{-\inf\limits_{\delta \in \mathrm{A}} \mathrm{J}(\delta)} \tag{4}$$

In the next sections, abusing notation, an event A will be usually described by a set of constraints on the decision variables δ. In this way, numeric constrained optimisation problems may be subjectively interpreted in possibilistic terms: the cost J(δ) will be interpreted as the log-possibility of δ and, by definition, unfeasible values of decision variables will be assigned zero possibility.

Let us now review some other relevant definitions and issues in possibilistic calculus.

Necessity

To assert that an event A is *necessarily* true (in our context, that all problem solutions belong to A), saying that A is "possible" may be not enough: it must also be true that the complementary event "not A" is not possible. This motivates the introduction of a necessity measure:

$$\mathrm{N}(\mathrm{A}) = 1 - \pi(\neg \mathrm{A}) \tag{5}$$

In a binary setting, all solutions belong to a subset A if and only if π(A) = N(A) = 1; there exist solutions in A (and solutions outside A) if π(A) = 1 but N(A) = 0, and there are no solutions in A if π(A) = 0.

Extending the measures π(A), N(A) to [0,1] provides a natural gradation of such concepts: π(A) = 0.95, N(A) = 0.1 would indicate that there are very possible solutions in A, but not all of them are in there (there are solutions with possibility 1 – 0.1 = 0.9 outside A).

Interactivity and possibilistic conjunction

The possibilistic analogue to statistical independence is the non-interactivity. If the joint possibility of two variables $\Delta = \Delta_1 \times \Delta_2$, $\delta = (\delta_1, \delta_2) \in \Delta$ can be expressed as the product of two univariate ones:

$$\pi(\delta_1, \delta_2) = \pi_1(\delta_1)\pi_2(\delta_2) \tag{6}$$

then variables δ_1 and δ_2 are said to be non-interactive. Thus, given two events $A_1 \subset \Delta_1$ and $A_2 \subset \Delta_2$, it is straightforward to prove that:

$$\pi(A_1 \cap A_2) = \pi_1(A_1)\pi_2(A_2) \tag{7}$$

which can be read as “the possibility of event A_1 *and* event A_2 is the product of the individual possibilities when the events relate non-interactive variables”, interpreting, as usual in literature, set intersection as a linguistic conjunction.

Under the non-interactivity assumption, if the possibility is defined as the logarithm of a cost index (3), the product (6) gets transformed into a sum:

$$J(\delta_1,\delta_2) := J_1(\delta_1) + J_2(\delta_2) \tag{8}$$

On the following, given individual cost indices $J_1(\delta_1)$, $J_2(\delta_2)$, etc. relating to different constraints, the expression above (8) will be the one used in most cases to define a possibility distribution in the product space. In this way, we are interpreting the possibilistic conjunction operator in (Dubois, 1996) as an algebraic product of possibilities, i.e., stating an underlying non-interactivity assumption between different constraints.

Note, however, that the interactivity assumption is not always intuitively needed. In the other extreme (total interactivity: variables δ_1 and δ_2 fully “correlated”, for instance equal), we would have: $\pi(A_1 \cap A_2) \leq \max(\pi(A_1), \pi(A_2))$, which would suggest the maximum possibility as the conjunction operator when two events affect exactly the same decision variables. In between those two extremes, other choices may be also possible, e.g., T-norm operators (Benferhat, 1997).

Conditional possibility

The possibilistic analogue to conditional probability is conditional possibility. Consider an event B with nonzero possibility. A quotient definition for conditional possibility of an event A given event B will be used in this work:

$$\pi(A \mid B) := \pi(A \cap B) / \pi(B) \tag{9}$$

In this way, given a (multivariate) possibility distribution $\pi(\delta)$, the conditional possibility can be computed as:

$$\pi(A \mid B) := \frac{\sup_{\delta \in A \cap B} \pi(\delta)}{\sup_{\delta \in B} \pi(\delta)} \tag{10}$$

so, if the possibility distribution is actually the exponential of a cost index, we get:

$$\pi(A \mid B) = e^{-\left(\min_{\delta \in A \cap B} J(\delta) - \min_{\delta \in B} J(\delta)\right)} \tag{11}$$

that is, computing the possibility by subtracting the cost associated to event B from the cost of any of its subsets.

To get a conditional possibility distribution of a variable δ, we assume event A being an individual point δ^*, getting:

$$\pi(\delta^* \mid \mathrm{B}) = \begin{cases} \dfrac{e^{-J(\delta^*)}}{e^{-\min_{\delta \in B} J(\delta)}} & \delta \in \mathrm{B} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

That is, the conditional distribution can be obtained by dividing the possibility distribution function for all points in a set by the maximum possibility of them, i.e., normalising the possibility distribution on a restricted conditioning domain B to a maximum equal to one.

The conditional definitions allow for an analogy to Bayesian inference: if we assume that B is actually certain (whatever the *a priori* possibility π(B) was), then conditional possibility may be understood as an *a posteriori* possibility.

Possibility versus probability

Both possibility theory and probability theory are frameworks for handling uncertainty in constraint satisfaction problems. Basically, a subjective interpretation would assign high possibility to events with high probability. Hence, in a first approximation, user-defined probabilities and possibilities should be related by an implicit monotonically increasing function. Possibility-necessity measures have also been linked to imprecise probabilities (Dubois, 2005). However, once aggregation takes place (via sums and integration in probability, via maximisation in possibility), although the subjective interpretation might be considered similar, there is no longer an implicit function relating probability and possibility. For further discussion of possibility, probability, and other uncertain reasoning frameworks, and their interrelations, the reader is referred to (Klir, 1992, Dubois, 2001, Dubois, 2005).

Ideally, probabilistic results would be preferable (to confidently assert that, e.g., 95% of cases a flux estimate will lie in a particular interval). However, there are some drawbacks: (i) exact probabilistic inference under equality and inequality modeling constraints is computationally hard (multivariate integration on irregular sets) (ii) some of the *a priori* Bayesian probabilities are in practice rough user-given estimates, (iii) some of the assumptions (linearity of transformation, Gaussian distributions) may not hold in practice, and (iv) there may be some uncertainty in the model parameters or in the model probabilities. Thus, as practical use of probability does not fully adhere to the theoretical assumptions, its results should be interpreted with some flexibility. As this work will discuss, the proposed possibilistic framework is much less demanding computationally (using optimisation instead of integrals, so large-scale cases become tractable) and gives similar results to the probabilistic approach in realistic cases.

The objective of the next sections is to set up a possibilistic framework for efficient computations in metabolic flux analysis.

7.3 Preliminaries: metabolic flux analysis

As explained in previous chapters, the metabolic networks encoding the elementary biochemical reactions taking place within a cell can be translated to a matrix **N**, where rows are the *m* internal metabolites and columns the *n* reactions. If these metabolites are at steady state, mass balances can be formulated as follows (Stephanopoulos, 1998):

$$\mathbf{N}\cdot\mathbf{v} = \mathbf{0} \tag{13}$$

where $\mathbf{v} = (v_1, v_2,..., v_3)^{\mathrm{T}}$ is the *n*-dimensional vector of metabolic fluxes.

Hence, a (steady-state) flux vector **v** represents the metabolic state of the cells at a given time, without any information on the kinetics of the reactions; it shows the contribution of each reaction to the overall metabolic processes of substrate utilization and product formation. Notice that as typically *n* is larger than *m*, the system (13) is underdetermined, i.e., there is a wide range of stoichiometrically-feasible flux vectors.

Assuming now that some fluxes in **v** have been measured (denoted as $\mathbf{v_m}$), while the rest remain unknown (denoted as $\mathbf{v_u}$), equation (13) can be rearranged as follows:

$$\mathbf{N_u}\cdot\mathbf{v_u} = -\mathbf{N_m}\cdot\mathbf{v_m} \tag{14}$$

As measurements are imprecise in practice, such measurement imprecision can be incorporated as constraints:

$$\mathbf{v_m} = \mathbf{w_m} + \mathbf{e_m} \tag{15}$$

where $\mathbf{e_m}$ represents measurements errors and $\mathbf{w_m}$ represents the actual measured flux value. In our approach, the measurement uncertainty is translated into an *a priori* possibility distribution for $\mathbf{e_m}$ from sensor characteristics. Other approaches consider different choices, as discussed below.

As seen in previous chapter, traditional metabolic flux analysis (TMFA) can be defined as the estimation of the flux vector satisfying (14) and compatible with the measurements (15). In particular, TMFA is often formulated as a two step procedure (Heijden, 1994a and 1994b): (1) analyse measurements consistency (and detect gross errors) using chi-square tests, and (2) solve a least squares problem to estimate the actual flux vector **v**:

$$\begin{aligned} &\min \quad \mathbf{e_m^T} \cdot \mathbf{F}^{-1} \cdot \mathbf{e_m} \\ &\text{s.t.} \quad \begin{cases} \mathbf{N_u} \cdot \mathbf{v_u} = -\mathbf{N_m} \cdot \mathbf{v_m} \\ \mathbf{v_m} = \mathbf{w_m} + \mathbf{e_m} \end{cases} \end{aligned} \tag{16}$$

where it is assumed that $\mathbf{e_m}$ are distributed normally with a mean value of zero and a variance-covariance matrix **F**.

Since all the constraints are linear equalities, the analytic solution of this minimisation problem can be obtained, resulting in the expressions to estimate $\mathbf{v_u}$ and $\mathbf{v_m}$ that are typically seen in literature, (Stephanopoulos, 1998, Gambhir, 2003). Details about TMFA calculations can be found in chapter II, section 2.8.

Unfortunately, with this formulation TMFA has some important limitations: (i) irreversibility constraints, or any other inequality constraints, cannot be considered, (ii) measurement errors have to be assumed to be normally distributed, (iii) it only provides point-wise flux estimates, and (iv) it requires a high number of measurable fluxes to be of use—system (14) has to be determined and redundant (Klamt, 2002).

Several alternatives have been suggested to face those limitations (Table 7.1). Quadratic programming solves the least squares problem (16) allowing to include irreversibility constraints, but inherits the rest of drawbacks (and introduces a drawback: χ^2-tests loose validity). The flux-spectrum, described in chapter IV, follows an interval approach to overcome the limitations mentioned before, but its estimations tend to be conservative because represents measurements uncertainty only with lower and upper bounds. Monte Carlo has been also used in the context of 13C-MFA (Wiechert, 2001; Kadirkamanathan, 2006; Schmidt, 1999), but rarely in absence of isotopic data. Moreover, sometimes it has been used incorrectly: Monte Carlo cannot be performed just solving a quadratic programming problem for each simulated set of measurements, because this introduce a bias on the results. Anyway, the major drawback of Monte Carlo is its high computational cost, which restricts its use for medium metabolic networks as an impractical number of samples is required to assess probabilities within a reasonable accuracy.

In the following we introduce a possibilistic framework for MFA that brings several interesting features: (i) it overcomes all the mentioned limitations of TMFA, (ii) is able to detect, and handle, inconsistencies between measurements and model, and furthermore (iii) with high computational efficiency.

7.4 Possibilistic MFA

In this section the possibilistic framework for MFA flux estimations is discussed. First, we define a set of time-invariant constraints derived from the metabolism being modelled. Then we incorporate the constraints imposed by the measured fluxes, representing its uncertainty by means of auxiliary slack decision variables and a cost index. In this way the notion of "degree of possibility" is incorporated. Finally, we show how (linear) optimisation problems are able to settle queries about the most possible fluxes, the possibility distributions, etc.

Table 7.1. Possibilistic MFA (Poss-MFA) is compared with four approaches for metabolic flux analysis, Traditional MFA (TMFA), constraint least-squares MFA (LS-MFA) and the flux-spectrum (FS-MFA). Legend: (x) provided feature, (-) partially provided feature and (○) potentially provided feature.

Feature	TMFA	LS-MFA	FS-MFA	M. Carlo	Poss-MFA
Considers irreversible reaction		x	x	x	x
Usable in scenarios lacking measurements			x	o	x
Includes a check of consistency	x	-	-	o	x
Flexible description of meas. errors			-	x	x
Richer estimations (not only point-wise)			-	x	x
Computational efficiency	x	x	x		x

Problem statement

Let us define a set of invariant constraints that every steady-state flux vector must satisfy; they do not depend on environmental conditions, do not change through evolution, etc. (Palsson, 2006). In this work these *model constraints*, denoted as $\mathcal{MOC}$, will be the stoichiometric relationships (13) and irreversibility constraints, described by means of inequalities:

$$\mathcal{MOC} = \begin{cases} \mathbf{N} \cdot \mathbf{v} = \mathbf{0} \\ \mathbf{D} \cdot \mathbf{v} \geq \mathbf{0} \end{cases} \tag{17}$$

where $\mathbf{D}$ is a diagonal nxn-matrix with $\mathbf{D_{i,i}} = 1$ if the flux i is irreversible (otherwise 0).

Other model-based constraints can be defined in an analogous way. For instance, elementary balances or degree of reduction balances might be incorporated into (17) as additional constraints (Stephanopoulos, 1998). It may be also possible to add constraints based on standard Gibbs free energy changes (Henry, 2007; Feist, 2007) or extracellular metabolites concentrations (Mo, 2009).

Incorporating the measurements

Estimating the non-measured fluxes would amount for solving the above equations (17), where some of the elements in vector **v** are measured ($\mathbf{v_m}$). However, this simple approach will be impractical in two very common situations:

- Measurements are very few, so the system has many—infinite—solutions.
- Real measurements do not exactly satisfy the constraints due to measurements (and modelling) errors. Therefore, no solution will be found[1].

Hence, the approach needs refinements to deal with a lack of measurements and to introduce the "possibility" of sensor errors and imperfect models. As shown below, such difficulties can be overcome by the introduction of slack variables and a cost index, enabling a grading of different candidate flux vectors as more or less "possible".

Possibilistic description of measurements. Each experimental measurement w_m can be described by a constraint as follows:

$$v_m = w_m + e_m \tag{18}$$

where e_m is a decision variable that represents the intrinsic uncertainty of the experimental measurements, i.e., the discrepancy between the actual flux v_m, and the measured value w_m. for convenience (see remark below), e_m is substituted by two non-negative decision variables, ε_1 and μ_1:

$$v_m = w_m + \varepsilon_1 - \mu_1 \quad \text{with:} \quad \varepsilon_1, \mu_1 \geq 0 \tag{19}$$

These decision variables $\delta = \{\varepsilon_1, \mu_1\}$ relax the basic assertion $w_m = v_m$, conforming a possibility distribution in (w_m, v_m) associated to some cost index $J_m(\delta)$. Among different possible choices, a simple—yet sensible—one is the linear cost index:

$$J(\delta) = \alpha \cdot \varepsilon_1 + \beta \cdot \mu_1 \tag{20}$$

with $\alpha \geq 0$ and $\beta \geq 0$. As explained in a section below, the weights α and β should be defined related to each measurement's "*a priori* accuracy" (usually, if sensor error is "symmetric", α and β should be defined to be equal).

[1] Notice, for instance, that an unfeasible set results with the constraint $v_1 = v_2$ and the two measurements $\{v_1 = 0.5, v_2 = 0.5001\}$.

Recalling the ideas introduced the preliminaries section, the interpretation of (19) and (20) may be the following: "$v_m = w_m$ is fully possible; the more v_m differs from w_m, the less possible such situation is."

Indeed, the event $A = \{v_m = w_m\} \equiv \{\varepsilon_1 - \mu_1 = 0\}$ will be fully possible, because:

$$\inf_{\delta=(\varepsilon_1,\mu_1)\in A} J(\delta) = 0$$

achieved at $\varepsilon_1 = \mu_1 = 0$, and then $\pi(A) = e^{-0} = 1$.

On the other hand, the possibility of the event A corresponding to v_m being different from w_m—to say, $A = \{v_m = w_m + \partial\} \equiv \{\varepsilon_1 - \mu_1 = \partial\}$—will be given by:

$$\pi(A) = e^{-\inf\limits_{\delta\in A} J(\delta)}$$

For example, with a cost index $J(\delta) = 5\varepsilon_1 + 5\mu_1$, and a measurement $w_m = 0.1$, the possibility of the actual flux v_m being $v_m = 0.2$ is $e^{-5 \cdot 0.1} = 0.6065$ ("quite" possible), and the possibility of $v_m = 1.1$ is $e^{-5 \cdot 1} = 0.0063$ ("almost" impossible).

A global cost index. Consider now a set of measurements $w_m = (w_1, \ldots, w_m)$ with its associated slack variables $\delta_1 = (\varepsilon_1, \mu_1), \ldots, \delta_m = (\varepsilon_m, \mu_m)$, and individual cost indices $J_1(\delta_1), \ldots, J_m(\delta_m)$. These results in what we call *measurement constraints*, $\mathcal{MEC}$:

$$\mathcal{MEC} = \begin{cases} \mathbf{v_m} = \mathbf{w_m} + \varepsilon_1 - \mu_1 \\ \varepsilon_1, \mu_1 \geq \mathbf{0} \end{cases} \tag{21}$$

In order to have a possibility distribution, under the non-interactivity assumption (6), the cost index is defined as a linear function, as follows:[1]

$$J(.) = \alpha \cdot \varepsilon_1 + \beta \cdot \mu_1 \tag{22}$$

where α and β are the row vectors of sensor accuracy coefficients and ε_1 and μ_1 correspond to stacking in vectors the slack variables from individual constraints.

[1] The Poss-MFA will be cast as a linear programming problem, and this is why the decision variables ε_1 and μ_1 were introduced instead of $\mathbf{e_m}$. However, it can be formulated using any other optimisation framework, such as quadratic programming. Throughout the thesis, linear programming will be assumed due to its great computational performance (solvable in polynomial time), which is a great advantage when dealing with large networks. Nevertheless, an example using quadratic programming will be described in a next section to point out the flexibility of the Poss-MFA.

The possibilistic MFA problem

At this point, we can define the Poss-MFA problem by means of the cost index J (22) and the set of constraints $\mathcal{CB}$:

$$\mathcal{CB} = \mathcal{MOC} \cap \mathcal{MEC} \tag{23}$$

where the decision variables δ are the actual fluxes $\mathbf{v} = (\mathbf{v_u}, \mathbf{v_m})$, and the slack variables ε_1 and μ_1.

The cost index J(δ) reflects the log-possibility of a particular combination of the decision variables, that is, the log-possibility of a particular flux vector $\mathbf{v}$.

Example 1 *Problem statement.* Consider the toy metabolic network depicted at the top of Figure 7.1, and the corresponding constraints, $\mathcal{MOC}$ and $\mathcal{MEC}$. Let us consider that the measurement of v_2 is "very accurate", that of v_5 is moderately accurate and those of v_3 and v_4 are quite unreliable. The weights α and β associated to the slack variables (ε_1 and μ_1) can be defined in accordance with this information: if we take $\alpha_2 = \beta_2 = 2$, $\alpha_5 = \beta_5 = 0.5$, and $\alpha_3 = \beta_3 = \alpha_4 = \beta_4 = 0.15$, for supposed measurements $w_2 = 9$, $w_5 = 31$, $w_3 = 30$, $w_4 = 10$, the measurements will be represented as depicted at the bottom of Figure 7.1.

Flux estimations: point-wise

The simplest outcome of a Poss-MFA problem is a point-wise flux estimation: the minimum-cost (maximum possibility) flux vector. This problem can be conveniently cast as the optimisation of a linear functional subject to linear constraints.

According to (4), the maximum possibility flux vector $\mathbf{v_{mp}}$ corresponding to a given set of measurements is obtained as the solution to the linear programming (LP) optimisation problem:

$$\begin{aligned} \min_{\mathbf{v},\varepsilon_1,\mu_1} \quad & \mathrm{J} = \alpha\cdot\varepsilon_1 + \beta\cdot\mu_1 \\ & \text{s.t. } \mathcal{CB} \end{aligned} \tag{24}$$

being its degree of possibility $\pi(\mathbf{v_{mp}}) = \exp(\mathrm{J}_{\min})$.

The obtained $\mathbf{v_{mp}}$ contains the most possible flux values consistent with the model and the measurements. A possibility equal to one must be interpreted as the flux vector being in complete agreement with model and original measurements. Lower values of possibility imply that $\mathbf{v_{mp}}$ corresponds to fluxes $\mathbf{v_m}$ deviated from the measurements $\mathbf{w_m}$.

Notice that as $\pi(\mathbf{v_{mp}}) = \pi(\mathcal{CB})$, it can be interpreted as the "*a priori*" possibility of encountering the measurements $\mathbf{w_m}$. If $\pi(\mathbf{v_{mp}})$ is low, this implies that either (a) there is a gross error in the measurements, (b) there is an error in the model, or (c) both. Therefore, the maximum possibility can be used to evaluate consistency and detect errors. We will come back to this point in a subsequent section.

Example 1 *Continued*. Consider again the model and the measurements given in Figure 7.1. The maximum possibility flux vector resulting from (24) is $\mathbf{v_{mp}} = (0.75, 9, 30.25, 8.25, 31, 39.3)^T$, with a possibility of $e^{-0.3} = 0.74$. The most possible flux vector being not fully possible (peak value not equal to 1) indicates that the measurements and the model are not in complete agreement. Indeed, the model says that $v_2 - v_4 = v_5 - v_3$, but $w_2 - w_4 = -1$ and $w_5 - w_3 = 1$. If the measurements had been fully compatible with the constraints imposed by the model—i.e., $w_2 = 10$, $w_5 = 30$, $w_3 = 30$ and $w_4 = 10$—the maximum possibility flux vector would have been $\mathbf{v_{mp}} = (0, 10, 30, 10, 30, 40)^T$, with a possibility of $\pi(\mathbf{v_{mp}}) = 1$.

Notice also that the possibility depends on the reliability associated to each measurement. If all the measurements were supposed to be more reliable, say $\alpha^* = 10\alpha$ and $\beta^* = 10\beta$—the possibility distribution functions would be narrower. The interpretation of the new coefficients would, therefore, be that the same deviation from the fluxes of maximum possibility will be now be considered as a less possible fact.

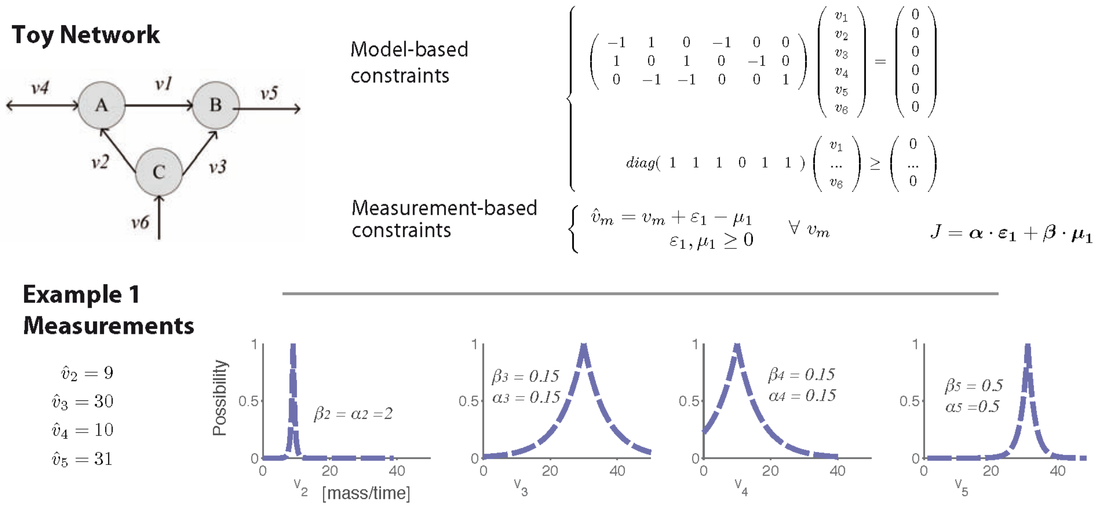

Toy Network
v4
v1
v5
A
B
v2
C
v3
v6
Model-based constraints
$\left\{ \begin{pmatrix} -1 & 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 & -1 & 0 \\ 0 & -1 & -1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \right.$
$diag(\ 1\ \ 1\ \ 1\ \ 0\ \ 1\ \ 1\ ) \begin{pmatrix} v_1 \\ \dots \\ v_6 \end{pmatrix} \geq \begin{pmatrix} 0 \\ \dots \\ 0 \end{pmatrix}$
Measurement-based constraints
$\left\{ \begin{array}{r} \hat{v}_m = v_m + \varepsilon_1 - \mu_1 \\ \varepsilon_1, \mu_1 \geq 0 \end{array} \right. \quad \forall\ v_m$
$J = \boldsymbol{\alpha} \cdot \boldsymbol{\varepsilon_1} + \boldsymbol{\beta} \cdot \boldsymbol{\mu_1}$
Example 1 Measurements
$\hat{v}_2 = 9$
$\hat{v}_3 = 30$
$\hat{v}_4 = 10$
$\hat{v}_5 = 31$
Possibility
1
0.5
0
0
20
40
$\beta_2 = \alpha_2 = 2$
$\beta_3 = 0.15$
$\alpha_3 = 0.15$
$\beta_4 = 0.15$
$\alpha_4 = 0.15$
$\beta_5 = 0.5$
$\alpha_5 = 0.5$
V2 [mass/time]
V3
V4
V5

Figure 7.1. Example 1: problem statement. A toy network and the corresponding constraints are given at the top. A possibilistic distribution representing a set of measurements is at the bottom.

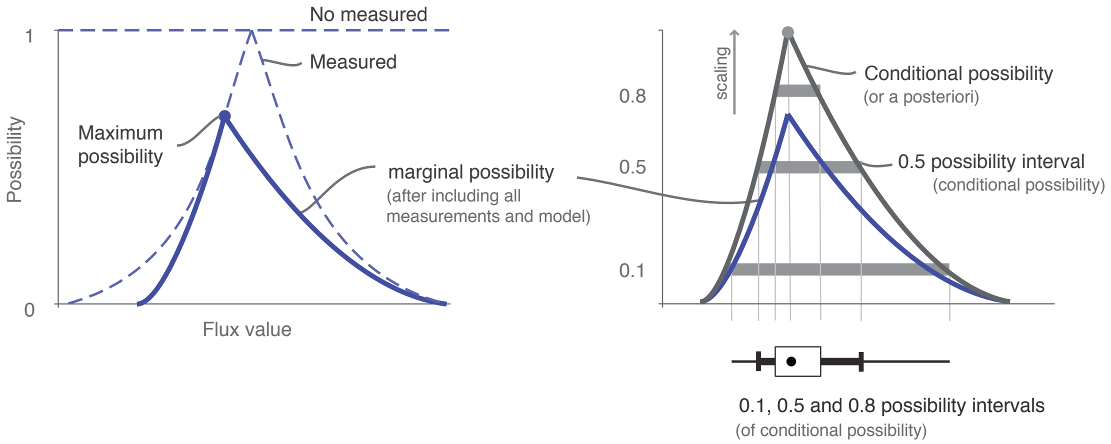


Figure 7.2. Possibilistic flux estimations. (Left) the figure shows possibilistic distributions representing the original measurement, the point-wise maximum possibility flux estimation, and the distribution of marginal possibility given by Poss-MFA. (Right) the figure shows the distributions of marginal and conditional (*a posteriori*) possibility. The flux intervals for conditional possibilities of π=0.8, 0.5 and 0.1, and the maximum possibility estimation, are depicted in a box-plot chart.

Flux estimations: distributions and intervals

Clearly, the validity of a point-wise flux estimation is limited in a situation where multiple flux values might be reasonably possible. To face these situation, marginal and conditional possibility distributions (and intervals) can be obtained, again, by solving linear optimisation problems. These flux estimations, illustrated in Figure 7.2, will be described next.

Marginal possibility distributions

Marginal possibility distributions (2) can be easily plotted and give a valuable information for the end user: they show, and rank, the possible values for each flux in the network.

The possibility of v_i being equal to a given value f, $\pi(v_i = f \cap \mathcal{CB})$, is computed simply adding a constraint to (24):

$$\min_{\mathbf{v},\varepsilon_1,\mu_1} \quad \mathrm{J} = \alpha \cdot \varepsilon_1 + \beta \cdot \mu_1 \qquad \text{s.t.} \begin{cases} \mathcal{CB} \\ v_i = f \end{cases} \tag{25}$$

Hence, plotting the marginal possibility for a range of fixed given values f (taken within a pre-specified range) will provide the marginal possibility distributions that be interpreted as the "distribution of the possible values for each flux in the network, given the measurements" (see Figure 7.2, left).

Notice that "cuts" $[v_{i,g}^{m}, v_{i,g}^{M}]$ of a possibility distribution, containing those values of v_i with a marginal possibility higher than γ, can be obtained solving two LP problems:

$$\begin{aligned} v_{i,g}^{m} &= \min v_i \quad \text{s.t.} \begin{cases} \mathcal{CB} \\ J < -\log\gamma \end{cases} \\ v_{i,g}^{M} &= \max v_i \quad \text{s.t.} \begin{cases} \mathcal{CB} \\ J < -\log\gamma \end{cases} \end{aligned} \tag{26}$$

This provides an efficient procedure to get a possibility distribution: compute "cuts" of possibilities between 0 and 1, say, $\pi = 0.1, 0.2$, etc.[1] This approach is better (computationally) than defining a range of values f and computing its possibility with (25), because avoids the problem of determining the most convenient step size and bounds for the flux (which, usually, are not known beforehand).

Conditional possibility distributions

Using the definition given in the preliminaries (12), the conditional possibility distribution of a flux v_i can be computed as follows:

$$\pi(v_i = f \mid \mathcal{CB}) = \begin{cases} \dfrac{\pi(v_i = f \cap \mathcal{CB})}{\pi(\mathcal{CB})} & f \in \mathcal{CB} \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

This is equivalent to normalise the marginal possibility distribution to a maximum equal to one (see Figure 7.2).

Conditional possibility may be understood as an *a posteriori* possibility: the possibility of v_i having the value f, if we assume that $\mathcal{CB}$ is actually true, i.e., that the model and the measurements are correct.

(A posteriori) Possibilistic intervals

In analogy to (26), the interval of flux values $[v_{i,g}^{m}\ v_{i,g}^{M}]$ with a degree of conditional (or *a posteriori*) possibility higher than γ can be obtained solving two LP problems:

[1] Notice that, remarkably, computing the marginal possibility of all the fluxes in the network by means of a grid of points is linear in the number of grid points and polynomial in the number of fluxes.

$$v_{i,g}^{m} = \begin{array}{c} \min\limits_{\mathbf{v},\varepsilon_1,\mu_1} \; v_i \\ \text{s.t.}\begin{cases} \mathcal{CB} \\ \mathrm{J} - \log \pi(\mathcal{CB}) < -\log \gamma \end{cases} \end{array} \qquad (28)$$

The upper bound $v_{i,g}^{M}$ would be obtained by replacing minimum by maximum.

These possibilistic intervals have a similar interpretation to *confidence intervals* (*credible intervals*) in Bayesian statistics, providing a concise flux estimation that can be represented by means of a box-plot chart (see Figure 7.2, right).

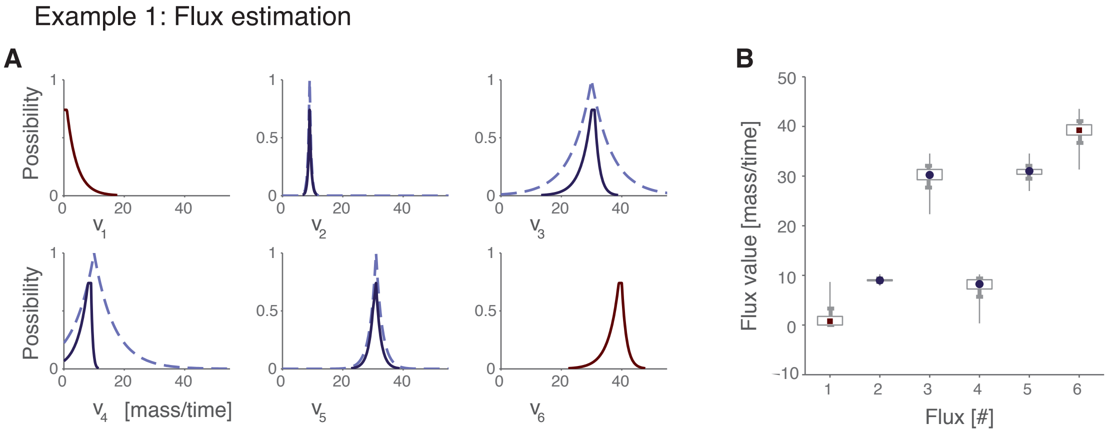


Figure 7.3. Example 1: flux estimation. Poss-MFA estimations were obtained for the example described in Figure 7.1. (A) The measured values are depicted with dashed lines, and the computed possibility distributions with solid lines. (B) The figure shows the flux intervals of conditional possibility 0.8 (box), 0.5 (thick line) and 0.1 (narrow line), and the maximum possibility flux estimation (squares and circles for non-measured and measured fluxes, respectively).

Example 1 *Continued*. Given the measurements in Figure 7.1, the obtained marginal possibility distributions for each flux are plotted in Figure 7.3A. They show that, for instance, the most possible value of v_1 is 0.75 ($\pi = 0.74$), that v_1 being 2.25 is quite possible, but that v_1 bigger than 10 is almost impossible ($\pi < 0.05$). The possibility distributions also reflect the reliability of the estimation of each flux: the estimation of v_6 is less reliable than the one of v_1 or v_2.

Notice too that the uncertainty on the measurements is often strikingly reduced through the flux estimation. For instance, the estimation of v_4—whose measurement was quite unreliable *a priori*—has been significantly improved, once the model constraints and the rest of measurements are incorporated. This reflects the already noticed fact that the network structure greatly constrains the possible values of fluxes for a given, typically small, set of measured flux values. The plots of marginal possibility can also detect multiple flux vectors with maximum possibility (possibility distribution functions with flat top). Figure 7.3B depicts the maximum possibility flux estimation and three possibilistic intervals by means of a box-plot chart. The intervals point out that, for instance, the highly possible *a posteriori* values of v_5 are those in [30.75, 31] ($\pi > 0.9$) and that those in [29.5, 32] are also quite possible ($\pi > 0.5$), while those outside [27, 34.5] are almost impossible ($\pi < 0.1$).

7.5 Possibilistic MFA: refinements

Now that the basics of the Poss-MFA framework have been introduced, some refinements will be discussed.

A better description of measurement's uncertainty

The formulation used above to describe the uncertainty of the experimental measurements might be considered somehow limited in some applications. Fortunately, it is very easy to add new slack variables, and modify the $\mathcal{CB}$ (23) and the cost index (22), allowing to work with possibility distribution functions of different characteristics.

As an example, the constraints (29) and cost (30) below describe an interval measurement plus some possibility of having outlier measurements:

$$v_m = w_m + \varepsilon_1 - \mu_1 + \varepsilon_2 - \mu_2 \qquad \text{with:} \begin{cases} \varepsilon_1, \mu_1 \geq 0 \\ 0 \leq \varepsilon_2 \leq \varepsilon_2^M \\ 0 \leq \mu_2 \leq \mu_2^M \end{cases} \tag{29}$$

and

$$\mathrm{J}(.) = \alpha \cdot \varepsilon_1 + \beta \cdot \mu_1 \tag{30}$$

The possibility of $w_m \in [v_m - \varepsilon_2^M, v_m + \mu_2^M]$ is one and the possibility of the actual flux v_m being out of the referred interval depends on the cost index weights (α and β).

For instance, a band with possibility equal to one can be used to account for systemic errors in measuring a particular flux, and a couple of additional slack variables may be defined to account for the decreasing possibility of random errors. These kind of representation of measurement uncertainty will be illustrated in subsequent examples.

Notice that more slack variables can be added to achieve a more complex representations of the measurements uncertainty. In fact, any convex representation of the log-possibility uncertainty can be approximated if sufficient slack variables are incorporated. Details are omitted for the sake of brevity.

Considering uncertainty in the model structure

Until now, the model-based constraints (23) have been considered as hard constraints; only those flux vectors **v** that exactly satisfy them could be feasible solutions. However, these constraints can be "softened" via suitable slack variables to consider uncertain knowledge. Then, these additional slack variables may be used in a cost index to generate a possibility distribution.

Consider, as an example, an equality restriction $a = b$. A relaxed ("softened") version of such restriction may be written as:

$$a = b + \zeta - \nu, \quad \zeta, \nu \geq 0 \tag{31}$$

with ζ and ν being slack variables penalised in an optimisation index J = f(ζ, ν), typically with linear cost index terms, $\gamma \cdot \zeta + \tau \cdot \upsilon$, in an analogous way to the discussion of uncertain measurements.

Notice also that a "softened" inequality restriction is nothing but an equality one with no penalisation on one of the slack variables above. For instance $a \leq b + \varepsilon$ can be expressed as $a = b + \varepsilon - \mu$ with free μ.

Such softened model constraints may be used to roughly incorporate imprecision in the model arising, for instance, from non-compliance with the pseudo-steady-state assumption, partial unbalance of some metabolites or uncertain yields. Although these issues require further research, let us outline some ideas below.

Relaxing the pseudo-steady state assumption. Equation (13) derives from the dynamic mass balance around the internal metabolites **c**, where it is assumed that d**c**/dt ≈ **0**. Adding slack decision variables to (13) makes it possible to relax this assumption.

Partial unbalance of metabolites. Sometimes, a metabolite cannot be assumed to be balanced, for example if there are reactions producing or consuming this metabolite that have not been taken into account in the network, as it is often the case for the cofactors, ATP, NADP, etc. This unknown consumption or production can be represented by means of slack variables if some interval bounds are known.

Uncertainty in stoichiometric yields. Sometimes the value of a yield coefficient is not exactly known, as it happens with the yield coefficients of lump reactions used to represent biomass synthesis. Let v_r be the flux through a reaction with an uncertain yield $Y_{i,r}$ for the metabolite i. The row corresponding to this metabolite in (13) can be rewritten as:

$$[n_{i,1} \ldots Y_{i,r} \ldots n_{i,n}] \cdot \mathbf{v} = [n_{i,1} \ldots n_{i,n}] \cdot \mathbf{v} + Y_{i,r} \cdot v_r = 0 \tag{32}$$

If $Y_{i,r} \in [Y_{i,r}^{min}, Y_{i,r}^{max}]$ and v_r is irreversible, equation (32) can be substituted by:

$$[n_{i,1} \ldots n_{i,n}] \cdot \mathbf{v} + u_r = 0 \tag{33}$$

$$Y_{i,r}^{min} v_r \leq u_r \leq Y_{i,r}^{max} v_r \tag{34}$$

However, if the flux v_r is reversible, inequalities in (33) cannot be set up, and the approach is no longer applicable. Integrating modal interval arithmetic (Gardeñes, 2001) could be an option to face this problem.

7.6 Possibilistic MFA: illustrative examples

Other features of Possibilistic MFA (Poss-MFA) will be briefly illustrated using the same example used above, which metabolic network is depicted in Figure 7.1.

Example 2: detecting errors in measurements and model

As mentioned earlier, the value of the peak possibility in the resulting flux distribution provides an indication of the agreement between the model ($\mathcal{MOC}$) and the measurements ($\mathcal{MEC}$). A low degree of possibility means that the model and the measurements are inconsistent. That is, that there is no flux vector “near” the measured values satisfying the model-based constraints. If the maximum possibility flux vector has a low value, one must assume that either (a) there is an error in one or more measurements, (b) there is an error in the model (e.g., a mass balance is not closed, or a metabolite is not at steady state), or (c) both.

If a high inconsistency (low possibility) is detected, it is possible to investigate what is causing it, and thus correct the measurements or improve the model. For instance, we can remove one measured flux at a time and perform the flux estimation to determine if the removed measurement was causing the low possibility[1]. If this is the case, we may consider the following alternatives: (a) consider that w_m is a totally unreliable measurement and accept the flux estimation inferred from the others measurements, (b) measure either w_m again, or a different flux that could provide new information, or (c) consider w_m a reliable piece of data and, hence, conclude that there is an error in the model. In case (c), a similar approach can be used to investigate which particular model-based constraint is causing the low possibility.

A simple example of the procedure just described is shown in Figure 7.4. Initially, a Poss-MFA estimation using all the measured fluxes was performed, obtaining a maximum possibility flux vector with low possibility, $\pi(\mathbf{v}) = \pi(\mathbf{w_m}) = 0.15$. We then repeated the estimation removing the flux w_4, but the maximum possibility does not increase. However, when the estimation was performed removing w_6, the maximum possibility was significantly higher (0.7). This suggests that there is a large error in w_6, or an error in the model around metabolite C which involves fluxes v_2, v_3 and v_6.

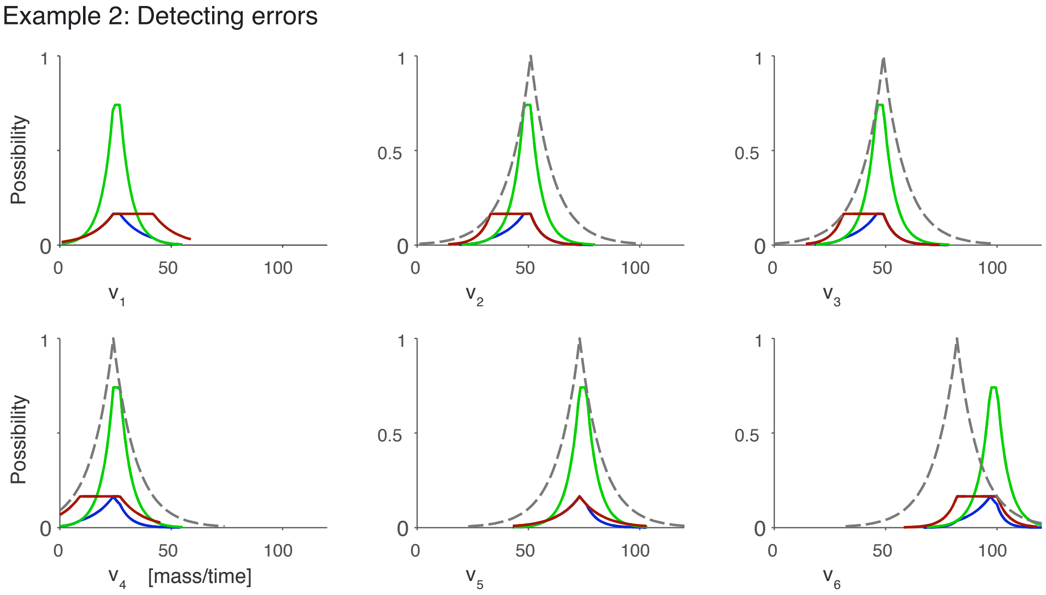


Figure 7.4. Example 2: Poss-MFA to detect errors in measurements and model. The metabolic network depicted in Figure 7.3 is used, assuming that five fluxes have been measured: w_2, w_3, w_4, w_5 and w_6 (dotted line). The possibility distributions for each flux are depicted in three cases: using all the measurements (deep blue), removing the flux w_4 (red) and removing the flux w_6 (light green).

[1] Another approach to analyse consistency with possibilistic MFA, based on the inspection of the slack variables, will be presented in chapter VIII.

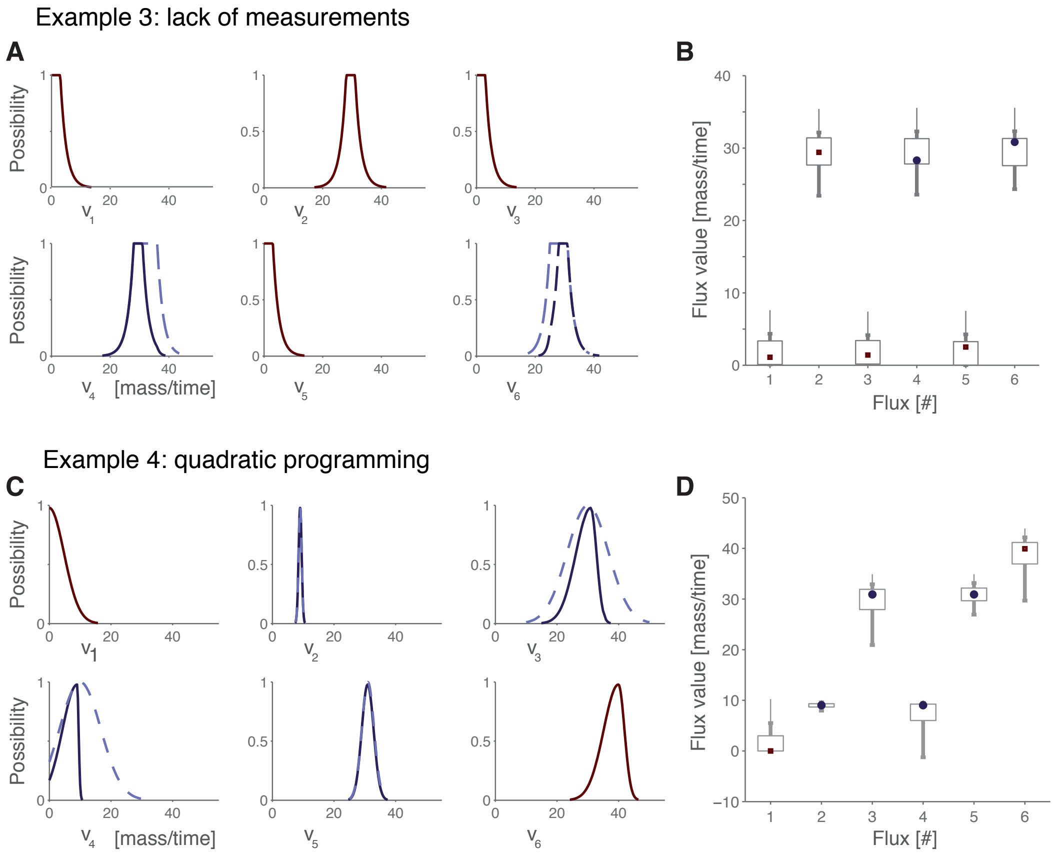


Figure 7.5. Examples 3 and 4. Both examples use the simple model described in Figure 7.1, assuming that some fluxes are measured (dashed lines). (A) (C) Possibility distributions of measured and non-measured fluxes (solid line). (B) (D) flux intervals for conditional possibilities of 0.8 (box), 0.5 (thick line) and 0.1 (narrow line) and the maximum possibility flux estimation (squares and circles for non-measured and measured fluxes, respectively).

Example 3: scenario of data scarcity

One of the features of Poss-MFA is that it can be used even if there is a lack of measurements; i.e., even if (14) is underdetermined or not redundant (Klamt, 2002). Let us continue with our example assuming now that only two fluxes are measured. Poss-MFA flux estimates are shown in Figure 7.5. Notice that crisp estimates will be only obtained if the irreversibility constraints, or other inequalities, are able to “bound” the under-determinacy of (14). Interestingly, our experience shows that this is often the case for medium size networks. Moreover, if this is not the case, the possibilistic flux estimation will be less precise—large intervals and flat distributions—but still reliable. The estimates will always be only as precise as allowed by the available data.

Example 4: using quadratic programming

To show how Poss-MFA can be cast within other optimisation frameworks, an example using quadratic programming will be discussed. We define $\mathcal{MEC}$ as $\mathbf{w_m} = \mathbf{v_m} +$

$\mathbf{e_m}$ and J = $\mathbf{e_m}^{\mathrm{T}} \cdot \mathbf{Y} \cdot \mathbf{e_m}$, where **W** is a diagonal matrix of weights. Hence, the possibility for each measurement is given by:

$$\pi(v_m) = \mathrm{e}^{-\mathrm{y_i}(w_m - v_m)^2}$$

In this way, measurements are represented with a quadratic possibility distribution.

We continue with our example using the measurements of Figure 7.1, but representing them with the quadratic formulation just introduced. The original possibility distribution of single measurements (dashed lines) and the possibility distributions computed with Poss-MFA (solid lines) are depicted in the Figure 7.5. Notice that results are similar to those obtained in the previous example (Figure 7.1), where the standard linear programming framework was used. However, the qualitative similarity between the results makes the author think that, in most cases, the linear programming setup is expressive enough and much more efficient than quadratic or other more complex optimisation frameworks.

Example 5: comparison with other methods

This example compares Poss-MFA with traditional MFA and some of its extensions. We perform estimations with Poss-MFA, but also with traditional MFA (TMFA), MFA as a constraint least-squares problem (LS-MFA) and the flux-spectrum (FS-MFA).

To show that Poss-MFA is able to represent measurements in a flexible way, we consider that errors in v_2 and v_3 are non-symmetric, and we add a band of uncertainty to account for systemic errors (Figure 7.6). Conversely, errors have to be approximated with a normal distribution so that TMFA and LS-MFA can be used (see preliminaries). To apply FS-MFA we represent the measurements with interval of 95%, or 2σ (see chapter IV). All the results are depicted in Figure 7.6.

Notice that TMFA assigns a negative value to an irreversible flux, v_1, because it is not taking reversibility constraints into account. This was clearly predictable, but it must be highlighted because TMFA is still widely used in the literature. The results also point out that the possibilistic estimates, distribution and intervals, are much more informative than the point-wise estimations of TMFA and LS-MFA, or the intervals of FS-MFA. Basically, point-wise estimations fail when several flux values reasonably possible, whereas the flux-spectrum interval tend to be conservative. Remember also that TMFA and LS-MFA cannot be used in scenarios lacking data, such as example 3, where Poss-MFA and was shown to be valuable.

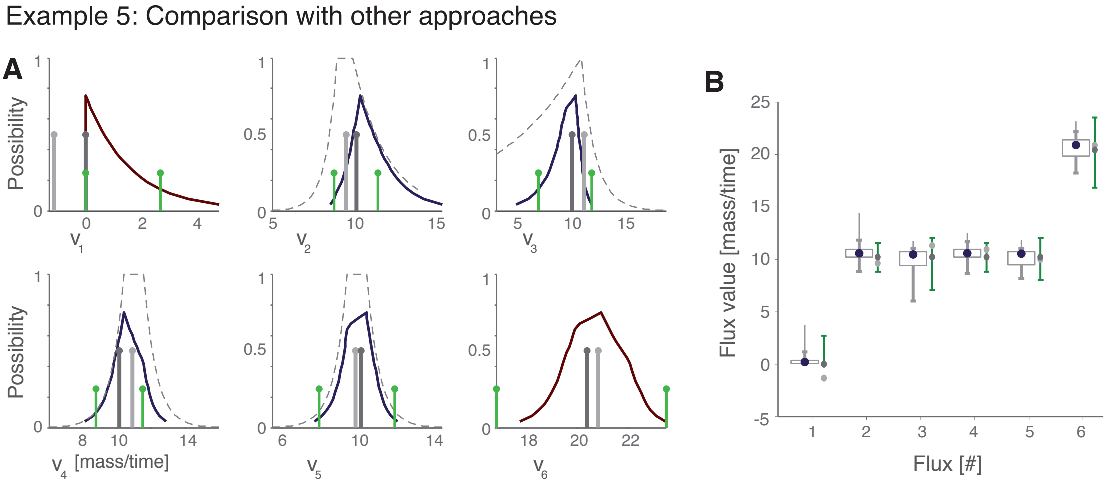


Figure 7.6. Example 5: comparison of Poss-MFA and alternative methods. We use the model described in Figure 7.1 considering v_2, v_3, v_4 and v_5 have been measured (depicted in grey). (A) The marginal distribution computed with Poss-MFA are depicted in blue, the point-wise estimations of TMFA and LS-MFA in light and dark grey, respectively, and the intervals of FS-MFA in green. (B) The maximum possibility flux estimate and the flux intervals for conditional possibilities 0.8 (box), 0.5 (thick line) and 0.1 (narrow line) are compared with the estimates given by TMFA, LS-MFA and FS-MFA.

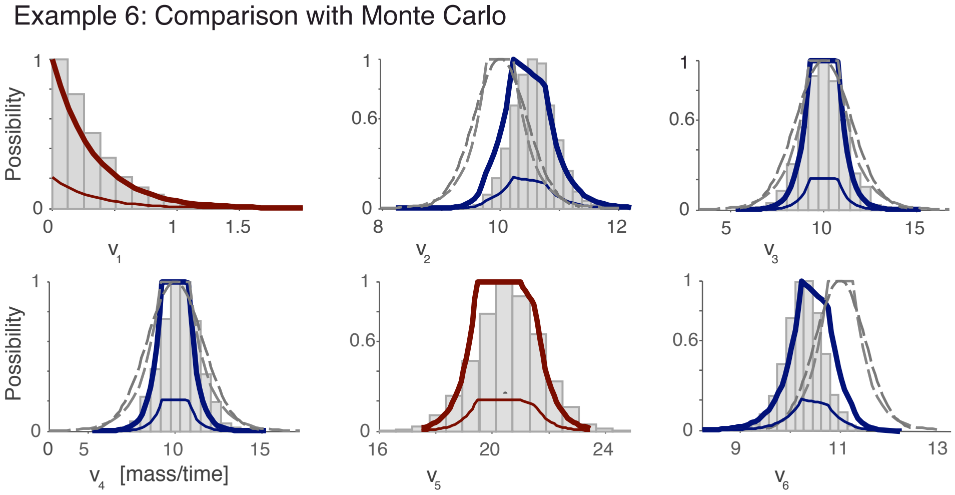


Figure 7.7. Example 6: comparison of Poss-MFA and Monte Carlo. We use the simple model described in Figure 7.1 considering that v_2, v_3, v_4 and v_5 have been measured. Poss-MFA: the measurements represented in possibilistic terms are depicted in grey, and the possibility distributions calculated from them in blue (thin lines for marginal distributions and thick lines for conditional ones). (2) Monte Carlo approach: the measurements represented assuming that errors are normally distributed are depicted in grey, and the histograms are those resulting from the Monte Carlo simulations.

Example 6: comparison with Monte Carlo

Continuing with our example, now measurements are represented (a) in possibilistic terms (linear case) and (b) with a "similar" probabilistic formulation assuming that errors are normally distributed. Both representations are depicted in Figure 7.7 (dashed lines). Then, we perform two flux estimations using (a) Poss-MFA and (b) Monte Carlo simulations (1.7 millions of combinations of values of measured fluxes were generated, taken into account their normal distribution). The conditional possibility distributions and the histograms resulting from Poss-MFA and Monte Carlo, respectively, are depicted in Figure 7.7. Even if probability and possibility are not truly equivalent, a reasonable similarity between the results from both approaches exists.

Notice that this is a simple case where Monte Carlo can be applied. Nonetheless, its worst performance is clear: the cost of computing the possibility distributions is polynomial in the number of fluxes (as shown above), whereas the cost of a Monte Carlo approach grows exponentially with the number of independent decision variables.

7.7 Case study: C. glutamicum

In this section we apply Possibilistic MFA (Poss-MFA) to a medium-size example. For illustrative purposes, we have chosen a very well-know metabolic model of *Corynebacterium glutamicum*.

Preparation: metabolic network and constraint-based model

The metabolic network of *C. glutamicum* has been taken from (Gayen, 2006) and is a slight variation of the one originally described in (Vallino, 1994; Vallino, 2000). The network describes the biochemistry of the primary metabolism of *C. glutamicum* necessary to support lysine and biomass synthesis from glucose. A reaction of ATP dissipation is included in the network, so that the ATP balance could be maintained, without actually constraining the flux space. On the contrary, the co-factors NADP, NAD and FAD are supposed to be balanced. The reaction for biomass formation is an approximation using as reactants those amino acids that explicitly appear in the network and the precursors of the other amino acids synthesized by *C. glutamicum*. This same example was used in chapter IV (section 4.5), where more details can be found, including the lists of reactions and metabolites, and the stoichiometric matrix.

Poss-MFA setting. The stoichiometric relationships, embedded in a 36×40 stoichiometric matrix $\mathbf{N}$, and the irreversibility of certain reactions, embedded in a 40×40 diagonal matrix $\mathbf{D}$, define our model-based constraints ($\mathcal{MOC}$) according to (17). Both matrices are given in chapter IV (section 4.5).

Preparation: experimental measurements

Experimental data of a batch fermentation of *C. glutamicum* cultured on minimal glucose medium was taken from (Vallino, 1994). There, the growth rate and the fluxes (production or consumption rates) of the external metabolites—lactate, acetate, glucose, O2, CO2, NH3, lysine and trehalose—were experimentally measured. Since the accumulation of lactate and acetate was negligible, their flux is zero in this case study. The measured fluxes v_{GLC} (1), v_{O2} (34), v_{NH3} (35), v_{LY} (37), v_{Thre} (38) and v_{CO2} (39) and the growth rate v_{Bio} (36), and their standard deviations, are given in Figure 7.8.

Poss-MFA setting. Using the data in Figure 7.8, we have built a possibilistic representation of single measurements defining convenient auxiliary variables and weights (Figure 7.8). The criteria to choose the weights was the following:

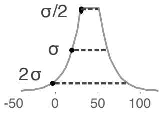
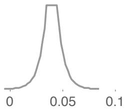
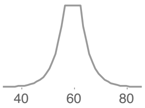
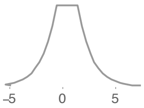
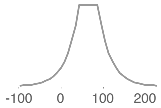
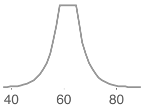
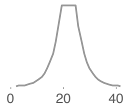
Metabolite	Flux (mM/h)			Metabolite	Flux (mM/h)		
	Measured	*Possibilistic rep.*			*Measured*	*Possibilistic rep.*	
GLC (1) Consump.	40.6 ± 22		α_1= 0.069 β_1= 0.069 μ_2^{max}= 11 ξ_2^{max}= 11	TREHAL (38) Production	0.4 ± 2		α_1= 150 β_1= 150 μ_2^{max}= 0.005 ξ_2^{max}= 0.005
O2 (34) Consump.	59.2 ± 5.9		α_1= 0.25 β_1= 0.25 μ_2^{max}= 2.95 ξ_2^{max}= 2.95	Biomass (36) Production	21.9 ± 5.4		α_1= 0.75 β_1= 0.75 μ_2^{max}= 1 ξ_2^{max}= 1
NH3 (35) Consump.	64.8 ± 44		α_1= 0.034 β_1= 0.034 μ_2^{max}= 22 ξ_2^{max}= 22	CO2 (39) Production	61.9 ± 6.2		α_1= 0.24 β_1= 0.24 μ_2^{max}= 3.1 ξ_2^{max}= 3.1
LYSE (37) Production	0.04 ± .01		α_1= 0.28 β_1= 0.28 μ_2^{max}= 2.7 ξ_2^{max}= 2.7				

Figure 7.8. Experimentally measured fluxes during a batch fermentation of *C. glutamicum*. The second column contains the experimental measurements and their standard deviation, taken from (Vallino, 1994). The possibility distribution representing each single measurement is depicted in the third column, when the used weights are given.

$$\pi = 1, \text{ for } v_m \in w_m \pm \sigma/2$$

$$\pi = 0.5, \text{ for } v_m \in w_m \pm 1\sigma$$

$$\pi = 0.1, \text{ for } v_m \in w_m \pm 2\sigma$$

where σ denotes the standard deviation of the measurement. If errors were assumed to be normally distributed, these levels would correspond to the probabilistic confidence intervals of 38%, 68% and 95%, respectively.

Possibilistic flux estimation

First, we obtained the maximum possibility flux vector considering all the available measurements, v_{GLC}, v_{O2}, v_{NH3}, v_{LY}, v_{Thre} and v_{CO2} and v_{Bio}. Its possibility was $\pi = 0.38$, which could be considered relatively low if one considers that a significant uncertainty was already being taken into account (Table 7.1). We then obtained the marginal possibility distributions for each flux, which inspection indicates that the low possibility is almost completely caused by only one measured flux, v_{NH3} (35). This suggests that this

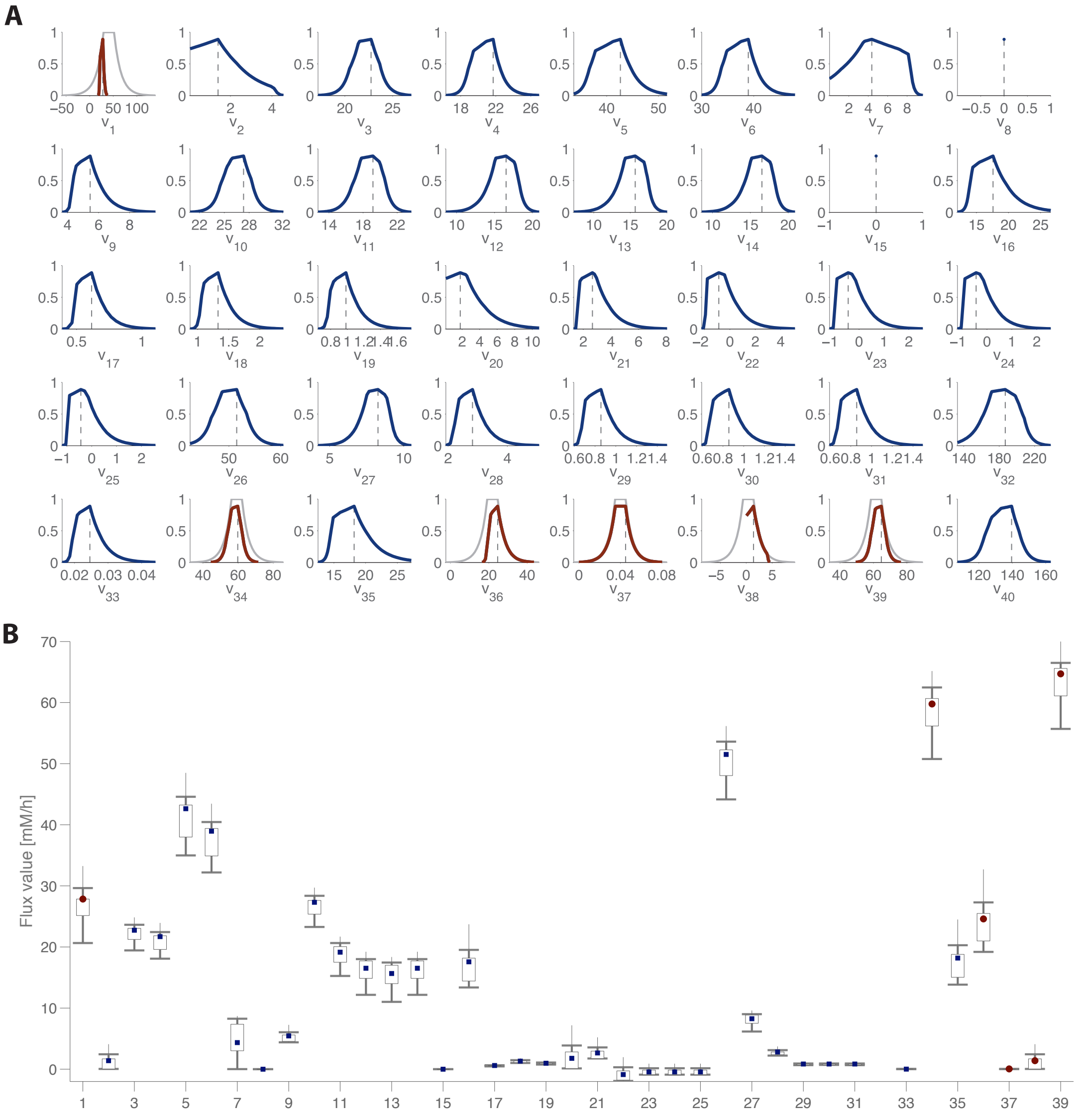


Figure 7.9. Possibilistic flux estimation for *C. glutamicum*. The measured fluxes are v_{GLC} (1), v_{O2} (34), v_{NH3} (35), v_{LY} (37), v_{Thre} (38) and v_{CO2} (39) and v_{Bio} (36). (A) Marginal possibility distributions for each flux. The original distribution of single measurements appear in grey (thick line). (B) The maximum possibility flux estimation (circles and squares for measured and non-measured fluxes, respectively) and the flux intervals for conditional possibilities of 0.8 (box), 0.5 (thick line) and 0.1 (narrow line) are depicted. All fluxes are in mM/h.

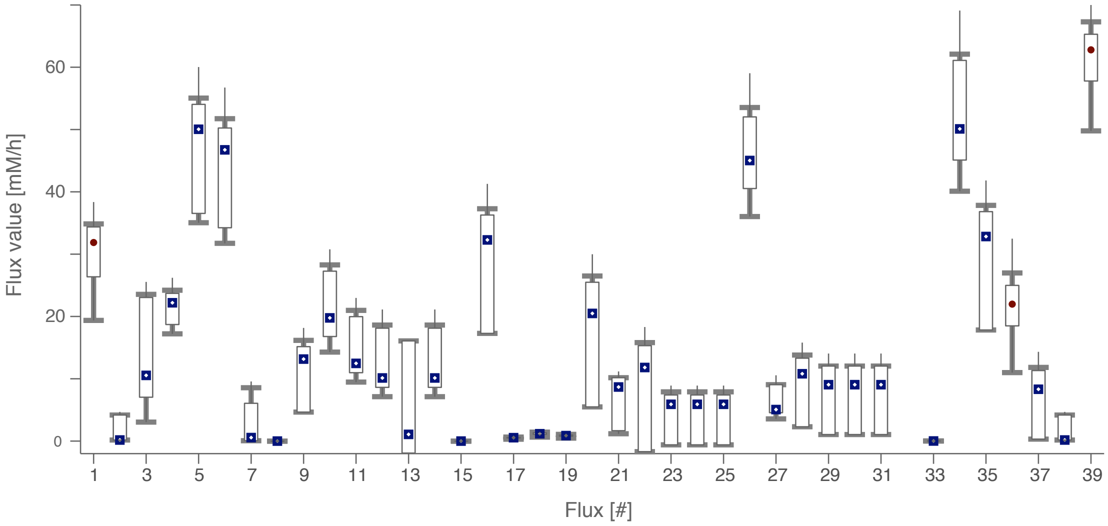


Figure 7.10. Possibilistic flux estimation for *C. glutamicum* under data scarcity. Only three measured fluxes are available v_{GLC} (1), v_{CO2} (39) and v_{Bio} (36). The maximum possibility flux estimation (circles and squares for measured and non-measured fluxes, respectively) and the flux intervals for conditional possibilities of 0.8 (box), 0.5 (thick line) and 0.1 (narrow line) are depicted. All fluxes are in mM/h.

measurement was inaccurate, or that its standard deviation was underestimated. Interestingly, this flux was indeed the most uncertain one in the original dataset (its standard deviation was a huge 44mM/h for a nominal value of 64.8mM/h).

As a result of this analysis—which is a rough example of the procedure mentioned in a previous section—we decided to remove the measurement and repeat the calculations. As expected, this time we obtained a maximum possibility flux vector with a similar shape, but higher possibility (π = 0.88). The possibility distributions for this case are depicted in Figure 7.9A, and the flux intervals are depicted in Figure 7.9B.

Possibilistic flux estimation under data scarcity

We have also performed a flux estimation using only three measured fluxes that can be measured with standard equipment: v_{GLC}, v_{CO2} and v_{Bio}. In this case the obtained maximum possibility flux vector is fully possible. This flux vector and the flux intervals are depicted in Figure 7.10. Remarkably, even if few measurements are available, the possibilistic estimates are quite precise (narrow).

Possibilistic flux estimation with an uncertain model

As explained above, the model-based constraints can be soften to relax the pseudo-steady state assumption. As an example, we have assumed a degree of uncertainty around all the mass balances in (17) introducing decision variables ζ_I and υ_I and the weights $\gamma_I = \tau_I = 2$ (see Figure 7.11). Thus, flux vectors that imply small accumulations of a metabolite will be accepted, yet considered less possible.

It can be also stated that the metabolic network used above, the one introduced by Vallino et al., relies on an unrealistic assumption: that cofactors NADP, NAD and FAD are balanced (Yang, 2006; Marx, 1996). To avoid this, we can remove these metabolites from the stoichiometric matrix or, as an alternative, use the expressivity of

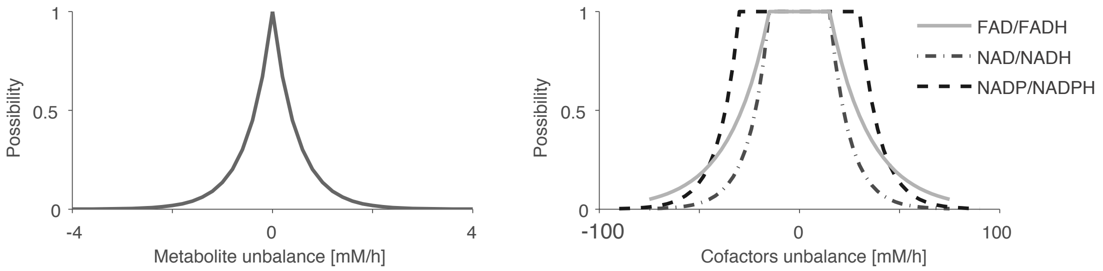


Figure 7.11. Possibilistic representation of two different kinds of model uncertainty.

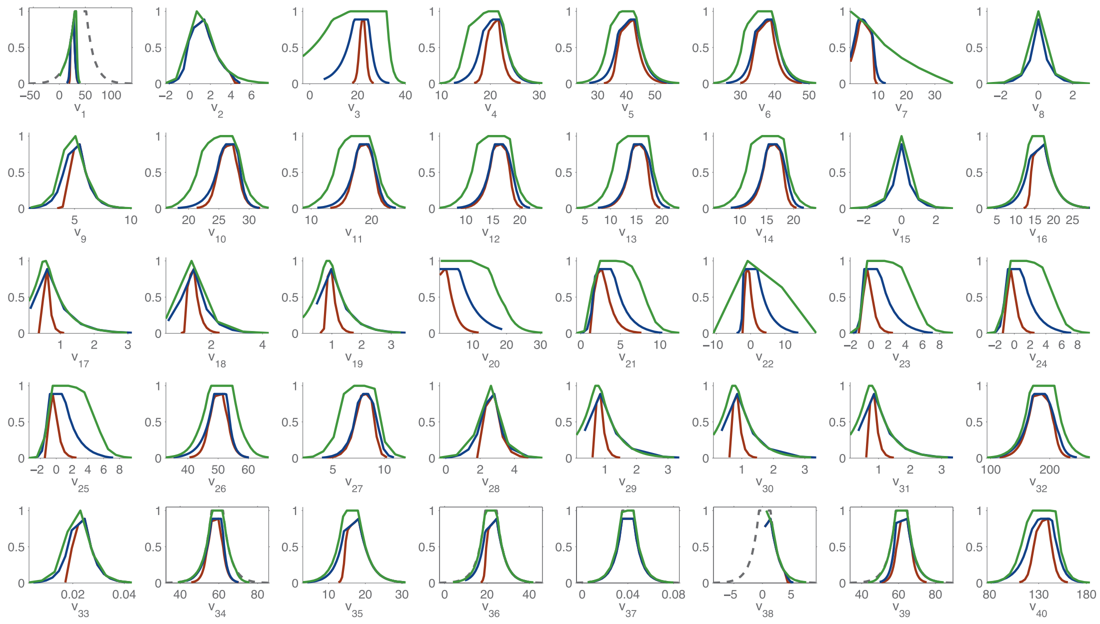


Figure 7.12. Possibilistic flux estimation for *C. glutamicum* when uncertainty is incorporated into the model. Marginal possibility distributions for each flux are depicted in three cases: (a) the model-based constraints are not relaxed (red) (b) the pseudo-steady state assumption is relaxed and NADP/NADPH is allowed to be unbalanced (deep blue), and (c) the pseudo-steady state assumption is relaxed and the three cofactors are allowed to be unbalanced (light green). The original distribution of single measurements are depicted with dashed lines. All fluxes are in mM/h.

the possibilistic framework to allow a certain degree of unbalance for these metabolites. Just as an example, we have assumed that cofactors can be unbalanced with some limits: 30 mM/h for NADP/NADPH, and 15 mM/h for FAD/FADH and NAD/NADH. This “knowledge” can be easily incorporated into the model defining the convenient auxiliary variables and weights as explained above (see Figure 7.11).

At this point, Poss-MFA was performed in three scenarios: (a) the model-based constraints are not relaxed (reference case); (b) the pseudo-steady state assumption is relaxed and NADP/NADPH is allowed to be unbalanced; (c) the pseudo-steady state assumption is relaxed and the three cofactors, NADP/NADPH, FAD/FADH and NAD/NADH, are allowed to be unbalanced.

The possibility distributions obtained in each case are compared in Figure 7.12. It can be observed how model uncertainty is translated into the flux estimates; consider uncertainty results in less precise estimates, given the less reliable model equations.

7.8 Conclusions

In this chapter we have discussed a unifying, possibilistic framework to evaluate consistency and estimate metabolic fluxes, which is shown to be flexible, reliable, usable under data scarcity and computationally efficient.

Considering ordinary constraint-satisfaction problems, the metabolic fluxes fulfilling a set of model-based constraints and compatible some experimental measurements are “possible”, otherwise “impossible”. Herein, this idea is refined to handle uncertain knowledge by introducing the notion of “degree of possibility”, which enables grading the candidate flux values.

Possibilistic MFA overcomes several limitations of traditional MFA and some of its extensions. It considers measurements uncertainty and model imprecision in a flexible way (e.g., non-symmetric error), and is reliable even if few fluxes are measurable (a common scenario). Possibilistic MFA provides distributions (and intervals) that are more informative than point-wise estimates when multiple flux values are reasonably possible. These are also better than the intervals of the flux-spectrum. In addition, Possibilistic MFA detects and handles inconsistencies between the measurements and the model. Finally, Possibilistic MFA has been cast as linear optimisation problems, for which widely known and efficient tools exist. This great computational performance makes the methodology suitable for large-scale metabolic networks.

There is, however, a challenge when estimating fluxes in large networks because there may be many flux vectors compatible with the (few) available measurements (Bonarius, 1997). Interestingly, Possibilistic MFA is still of use in this situation: it will detect all these equally possible flux vectors (or those similarly possible) by means of possibilistic distributions or intervals (e.g., example 3). Unfortunately, if there is a wide range of candidates, the estimation may be little informative (but reliable, since all reasonably possible flux vectors are captured). To face this difficulty one can promote particular flux vectors among those that are equally possible. For instance, it can be assumed that fluxes are optimally regulated depending on the given environmental conditions, and invoke this principle to choose particular flux vectors (Schuetz, 2007; Palsson, 2006; Schilling, 2002). There might be still alternate optima, but the approach will reduce the range of candidate flux vectors. The use of this optimality principle in a possibilistic framework will be discussed in chapter VIII.

In summary, the combination of computational efficiency and flexibility of the assumptions is a distinctive advantage of Possibilistic MFA over other approaches which either may rely on stronger assumptions (chi-squared distributions, interval-only descriptions, absence of irreversibility), or be only data-based (so they do not incorporate, say, stoichiometric model balances), or provide only point-wise estimates (for fluxes or consistency), or be computationally intensive (multi-variate integration in a general Bayesian estimation problem).

Main references

- Llaneras F, Sala A, Picó J (2009). A possibilistic framework for metabolic flux analysis. *BMC Systems Biology*, 3:73.

- Sala A (2008). Encoding fuzzy possibilistic diagnostics as a constrained optimisation problem. Information Sciences, 178:4246–4263.

- Dubois D, Fargier H, Prade H (1996). Possibility theory in constraint satisfaction problems: handling priority, preference and uncertainty. *Applied Inteligence*, 6(4):287–309.

- Stephanopoulos GN, Aristidou AA (1998). *Metabolic Engineering: Principles and Methodologies*. San Diego, USA: Academic Press.

- Klamt S, Schuster S, Gilles ED (2002). Calculability analysis in underdetermined metabolic networks illustrated by a model of the central metabolism in purple nonsulfur bacteria. *Biotechnology & Bioengineering*, 77:734-751.

- Vallino JJ (1994). *Identification of branch-point restrictions in microbial metabolism through metabolic flux analysis and local network perturbation*. PhD thesis, Massachusetts Institute of Technology, Cambridge.

VIII

Possibilistic, dynamic prediction of fluxes and metabolites

In this chapter the possibilistic framework is used to get predictions from a constraint-based model accounting for extracellular dynamics. We consider both predictions given by metabolic flux analysis (MFA), and by flux balance analysis (FBA).

The methods described provide rich estimates for time-varying fluxes and metabolite concentrations, taking into account uncertainty, alternate optima and sub-optimality. The approach can also be used to monitor consistency and detect faults.

Part of the contents of this chapter appeared in the following publications:

- F. Llaneras, A. Sala and J. Picó. Dynamic flux estimations from constrain-based models: a possibilistic approach (In preparation)
- Llaneras F, Sala A, Picó J (2010). Possibilistic estimation of metabolic fluxes during a batch process accounting for extracellular dynamics, *Computer applications in biotechnology 2010.*
- Llaneras F, Sala A, Picó J (2010). Dynamic flux balance analysis: a possibilistic approach, *Systems biology of microorganisms 2010.*

8.1 Introduction

There are two main approaches to get predictions from a given constraint-based model—which, recall, defines a space of feasible cellular states based on the operating constraints:

(a) Use experimental measurements to perform a metabolic flux analysis (MFA), following a traditional approach (Heijden, 1994) or one of the proposals described earlier in this thesis. (See chapters IV and VII.)

(b) Assume that cells have evolved to be optimal in some sense and apply flux balance analysis (FBA). (See chapter II.)

These predictions are typically static, aimed to study cells at a given state, but extracellular dynamics can be easily taken into account. As seen in chapter II, mass balances around the extra-cellular species can be established as follows:

$$\frac{\mathrm{d}\mathbf{e}}{\mathrm{dt}} = \mathbf{v}_\mathbf{e} \cdot x - \mathrm{D} \cdot \mathbf{e} + \mathbf{F}_\mathbf{e} \tag{1}$$

where **e** denotes the concentration of extracellular metabolites (substrates and products), $\mathbf{v}_\mathbf{e}$ the vector of extracellular reaction rates (uptake or production), D the dilution rate (inflow per volume) and $\mathbf{F}_\mathbf{e}$ the inflow of extracellular metabolites.

Given a metabolic network of the modelled cells, and extracting its stoichiometric matrix, mass balances around the intracellular metabolites can also considered:

$$\frac{\mathrm{d}\mathbf{c}}{\mathrm{dt}} = \mathbf{N} \cdot \mathbf{v} - \mu \cdot \mathbf{c} \tag{1b}$$

where **c** is the *m*-dimensional vector of intracellular metabolite concentration, **v** the *n*-dimensional vector of flux through each reaction, μ is the growth rate of biomass cells, and **N** is the stoichiometric matrix linking fluxes and internal metabolites.

However, since reaction kinetics are still rarely known, internal metabolites are often assumed to be at steady-state. In this way a model of cells considering extracellular dynamics can be as follows:

$$\mathbf{0} = \mathbf{N} \cdot \mathbf{v} \tag{2a}$$

$$\frac{\mathrm{d}\mathbf{e}}{\mathrm{dt}} = \mathbf{N}_\mathbf{e} \cdot \mathbf{v} - \mathrm{D} \cdot \mathbf{e} + \mathbf{F}_\mathbf{e} \tag{2b}$$

where $\mathbf{N}_\mathbf{e}$ is a selection matrix linking each external metabolite with its flux. Without loss of generality, each extracellular metabolite can be represented by two nodes, one intra- and one extra- cellular, so that there is only one reaction in **v** accounting for its total uptake or consumption. Biomass can be considered as another external metabolite and its synthesis represented with a flux in **v**. The formulation in (2) has been used, for instance, to seek extracellular or macroscopic models compatible with the underlying metabolic network (Provost, 2004; Haag, 2005; Provost, 2006, Bastin, 2007).

Along with the mass balances, other constraints can then be imposed, such as the irreversibility of certain reactions:

$$\mathbf{D}\cdot\mathbf{v} \geq \mathbf{0} \tag{3}$$

where **D** is a diagonal matrix with $\mathbf{D_{ii}} = 1$ if the flux *i* is irreversible (otherwise 0).

The resultant constraint-based models are typically used under a static point of view to analyse the metabolic fluxes at a given state. Therefore extracellular dynamics are not considered and derivatives are replaced by constant uptake or production rates in (2b). However, several works accounting for extracellular dynamics can be found in the literature, both in the context of MFA (Herwig, 2002; Takiguchi, 1997; Henry, 2007) and FBA (Mahadevan, 2002; Hjersted, 2009).

In this chapter the extracellular dynamics are considered in a similar way, to explore the benefits that the possibilistic framework introduced in chapter VII can bring in this context.

(i) Possibilistic metabolic flux analysis (Poss-MFA) is extended to get dynamic (time-varying) estimations of fluxes and metabolite concentrations.

(ii) Is it also discussed how Poss-MFA can be used to monitor consistency, as a fault detection procedure.

(iii) A possibilistic version of flux balance analysis (Poss-FBA) is presented herein. It gives dynamic predictions for fluxes and metabolites invoking an optimality assumption, and accounting for alternate optima and sub-optimality.

The chapter is organised as follows. Dynamic, possibilistic MFA is presented in section 8.2, and illustrated with a case study in section 8.3. The dynamic, possibilistic version of FBA is discussed in section 8.4, and illustrated with a case study in 8.5. The chapter is closed with a summary and the discussion of future work.

8.2 Dynamic Possibilistic MFA

In chapter VII we described Poss-MFA, a framework to formulate metabolic flux estimations as possibilistic constraint satisfaction problems (thus following a *constraint-*

based modelling approach). Herein we extend this idea to take extracellular dynamics into account.

Consider a batch process[1] during a period of time [0, T] divided in *t* intervals given by the sampling rate of the measurements. First, we consider the constraints conforming the model at successive time instants *k*, hereinafter referred as $\mathcal{MOC}(k)$[2]:

$$\mathcal{MOC}(k) = \begin{cases} \mathbf{0} = \mathbf{N} \cdot \mathbf{v}(k) & \text{(4a)} \\ \dfrac{\mathbf{e}(k) - \mathbf{e}(k-1)}{\Delta\mathrm{T}} = \mathbf{N_e} \cdot \mathbf{v}(k) & \text{(4b)} \\ \mathbf{D} \cdot \mathbf{v}(k) \geq \mathbf{0} & \text{(4c)} \\ \mathbf{e}(k) \geq \mathbf{0} & \text{(4d)} \end{cases}$$

Initial conditions should be given, at least, for each metabolite **e**(0). For convenience, hereinafter the set of system variables will be denoted as **var**(*k*)={**v**(*k*), **e**(*k*)}.

Then, measured concentrations of extracellular species are also incorporated as constraints, $\mathcal{MEC}(k)$:

$$\mathcal{MEC}(k) = \begin{cases} \mathbf{e_m}(k) = \mathbf{f_m}(k) + \varepsilon_1(k) - \mu_1(k) + \varepsilon_2(k) - \mu_2(k) & \text{(5a)} \\ \varepsilon_1(k), \mu_1(k) \geq 0 & \text{(5b)} \\ 0 \leq \varepsilon_2(k) \leq \varepsilon_2^{\max}(k) & \text{(5c)} \\ 0 \leq \mu_2(k) \leq \mu_2^{\max}(k) & \text{(5d)} \end{cases}$$

where $\mathbf{e_m}(k)$ represent the actual concentrations of each metabolite and $\mathbf{f_m}(k)$ the measured values. Slack variables ε and μ are introduced to consider uncertainty and relax the assertions $\mathbf{f_m}(k) = \mathbf{e_m}(k)$, conforming a possibility distribution associated to a cost index J(*k*):

$$\mathrm{J}(k) = \alpha(k) \cdot \varepsilon_1(k) + \beta(k) \cdot \mu_1(k) \qquad (6)$$

where α(*k*) and ß(*k*) are row vectors of user-defined, sensor accuracy coefficients.

[1] With no inflow or outflows, and thus with Fe and D equal to zero

[2] We use a backward approximation of derivatives for simplicity, but alternatives might be considered.

The index J(k) reflects the log-possibility of each $\mathbf{e_m}(k)$. The interpretation of (5-6) may be: "$\mathbf{f_m}(k) = \mathbf{e_m}(k)$ is fully possible; the more $\mathbf{f_m}(k)$ differs from $\mathbf{e_m}(k)$, the less possible such situation is".

Two pairs of slack variables are used to represent each measurement, the bounds for ε_2 and μ_2 define an interval of values with possibility equal to one (fully possible), and the possibility of the actual concentration being out of this interval depends on the chosen $\alpha(k)$ and $\beta(k)$. This allows to account for systemic and random errors. Slack variables can be added to achieve more complex representations of the measurements. See chapter VII for more details on this issue.

Dynamic Possibilistic MFA: simultaneous approach

Now that the constraint-based model has been formulated, two main problems can be addressed: (1) the estimation of fluxes and metabolite concentrations along the process duration, and (2) the monitoring of measurements consistency.

The most straightforward way to approach both problems is to consider the operating constraints at each time instant simultaneously. In this way all the available knowledge and information is taken into account to get each estimate. Clearly, this approach can be computationally expensive, and even non solvable if the sampling rate is high (the number of constraints will be extremely large). However, this difficulty will rarely arise because extracellular dynamics are typically slow and measurements are taken with relatively long sampling periods.

Monitoring consistency of measurements and model

To detect errors in measurements (or in the model) is it possible to monitor the consistency between measurements and model along the process evolution. The maximum possibility (minimum-cost) solution of the constraint satisfaction problem (4-5) can be obtained solving a linear programming problem (LP):

$$\begin{aligned} &\min \quad \mathrm{J_T} = \sum_{k=0}^{t} \mathrm{J}(k) \\ &\text{s.t.} \begin{cases} \mathcal{MOC}(k) & \forall k \\ \mathcal{MEC}(k) & \forall k \end{cases} \end{aligned} \tag{7}$$

The possibility π^{mp} of the most possible solution $\mathbf{var^{mp}}$ is given by the minimised cost:

$$\pi^{mp} = \pi(\mathbf{var^{mp}}) = \exp(-\mathrm{J_T^{min}}) \tag{8}$$

The value of π^{mp} provides a measure of consistency: possibility equal to one must be interpreted as complete agreement between the model and measurements, lower values imply that there is some error in one of them.

To analyse which measurements might be causing the inconsistency, the values of the slack variables can be inspected, just noticing that:

$$\pi^{mp}=\prod_{k}\pi_{k}^{mp}=\prod_{k}\exp\left(-\mathrm{J}_{\mathrm{T}}^{\min}(k)\right) \tag{9}$$

$$\pi^{mp}=\prod_{k}\prod_{i}\pi_{k,i}^{mp}=\prod_{k}^{t}\exp\left(-\sum_{i}\left(\alpha_{i}(k)\cdot\varepsilon_{1,i}^{min}(k)+\beta_{i}(k)\cdot\mu_{1,i}^{min}(k)\right)\right) \tag{10}$$

where index k denotes the time instants, index i the measurements elements of $\mathbf{e_m}(k)$, and $\varepsilon_{1,i}^{min}(k)$ and $\mu_{1,i}^{min}(k)$ are the values of the slack variables in $\mathrm{J}_{\mathrm{T}}^{\min}(k)$.

Thus, it can be investigated which measurements are those that most likely are causing the inconsistency by plotting the values of π_{k}^{mp} and $\pi_{k,i}^{mp}$ can (see example in 8.3).

Monitoring consistency using QP

It can be argued that it is better to formulate the consistency analysis as a quadratic programming problem (QP), instead of using LP. If two variables (measurements) are inconsistent, LP solution concentrates the error in the less penalised one (the less reliable), whereas QP solution distributes the error between both variables. This second alternative is more convenient because guaranties that all the possible sources of inconsistency are detected, even if they seem less likely.

> *Example.* Consider the constraints {A=B, A_m=6, B_m=10}, where the measured B_m is more reliable. The LP solution will be A=B=10, while the QP solution will be something in between, say A=B=9, suggesting that the problem can be due to A_m or B_m but that the first is more likely according to the considered uncertainty.

The consistency analysis can be formulated as a QP problem replacing (5) with:

$$\mathbf{e_m}(k)=\mathbf{f_m}(k)+\theta(k) \tag{11}$$

and defining the cost index as:

$$\mathrm{J}(k)=\theta(k)'\cdot\mathbf{W}(k)\cdot\theta(k) \tag{12}$$

where $\mathbf{W}(k)$ is a matrix of user-defined, sensor accuracy coefficients, analogous to $\alpha(k)$ and ß(k) in the LP case.

The benefits of formulating the consistency analysis as a QP problem, instead of an LP one, will be illustrated with an example in section 8.3.

Dynamic estimation of fluxes and metabolites

The simplest estimate is given by the solution of (7), which contains the most possible value for each flux and metabolite, **var**(*k*). However, these point-wise estimates are insufficient if multiple solutions are reasonably possible, as discussed in chapter VII. As an alternative, possibilistic intervals can be obtained.

The interval of values with conditional possibility higher than γ for a given flux or metabolite, $[var_{i,\gamma}^{m}(k), var_{i,\gamma}^{M}(k)]$, can be computed solving two LP problems:

$$\begin{aligned} & var_{i,\gamma}^{m}(k) = \min\, var_{\mathrm{i}}(k) \\ \text{s.t.} & \begin{cases} \mathcal{MOC}(k) & \forall k \\ \mathcal{MEC}(k) & \forall k \\ \sum \mathrm{J}(k) - \log \pi(var_{mp}) < -\log \gamma \end{cases} \end{aligned} \tag{13}$$

The upper bound can be obtained by replacing minimum by maximum.

These possibilistic intervals provide a rich and concise estimation. Remember also that the possibility distribution of a particular variable can be reconstructed obtaining the intervals for a grid of possibilities, to say π=1, 0.9, 0.8, ... 0.1. These and other details about possibilistic calculus and optimisation can be consulted in chapter VII.

Notice that (13) can be used both to estimate the metabolic fluxes **v**(*k*) and the metabolite concentrations **e**(*k*). Regarding the last ones, it is remarkably that even the evolution of non-measured metabolites could be estimated, such as the concentration of a product of interest. An example of this feature is given below.

Dynamic Possibilistic MFA: isolated approach

The approach described above can be computationally intractable if the sampling rate of measurements highly increases. This contingency will be rare, as mentioned above. However, if one needs to reduce the computational cost, the problem can be divided in *t* small problems, considering only the constraints operating at each time instant *k*.

This kind of "isolated" approaches, even imperfect, were indeed followed by the majority of works that can be found in the literature accounting for extracellular dynamics in the context of constraint-based modelling (Herwig, 2002; Takiguchi, 1997; Henry, 2007; Mahadevan, 2002; Hjersted, 2009).

Monitoring consistency of measurements and model

With the isolated approach, the consistency analysis now requires solving *t* smaller LP problems, one at each time instant *k*:

$$\forall k \qquad \begin{array}{l} \min \quad \mathrm{J_T} = \mathrm{J}(k) + \mathrm{J}(k\text{-}1) \\ \text{s.t.}\begin{cases} \mathcal{MOC}(k) \\ \mathcal{MEC}(k) \\ \mathcal{MEC}(k\text{-}1) \end{cases} \end{array} \tag{14}$$

With possibilities, $\pi_k^{mp} = \pi(\mathbf{var_{mp}}(k)) = \exp(-\mathrm{J_T^{min}}(k))$.

At this point, the same approach described above to investigate which measurements could be causing inconsistency, can be used in an analogous way, by the inspection of the values of π_k^{mp} and $\pi_{k,i}^{mp}$.

Dynamic estimation of fluxes and metabolites

The interval estimate for a given flux or metabolite, at a given time *k* and with conditional possibility higher than γ, can now be obtained solving two smaller LP problems:

$$\begin{array}{l} var_{i,\gamma}^{m}(k) = \min var_i(k) \\ \text{s.t.}\begin{cases} \mathcal{MOC}(k) \\ \mathcal{MEC}(k) \\ \mathcal{MEC}(k\text{-}1) \\ \sum \mathrm{J}(k) - \log \pi(var_{mp}) < -\log \gamma \end{cases} \end{array} \tag{15}$$

The upper bound obtained by replacing minimum by maximum.

Using this isolated approach, the possibilistic intervals are obtained solving LP problems that do not grow with the sampling rate. There is, however, a price: as less constraints are considered at a time, the solution space will be larger and the computed intervals will eventually become wider than those given by (13). Being the intervals conservative, the isolated approach may lead to less insight, but in no case will lead to wrong results with respect to the simultaneous case.

Remark: a mixed version. In the previous sections we have described two different procedures to get the dynamic, possibilistic MFA estimates. The first one considers simultaneously the operating constraints at every time instant, each *k* in [1, *t*], while the second one divides the problem in a succession of smaller sub-problems, one per each time instant *k* in [1, *t*], which consider only constraints at *k* and *k-1.* Clearly, a mixed approach considering a time window of user-defined size can be easily implemented.

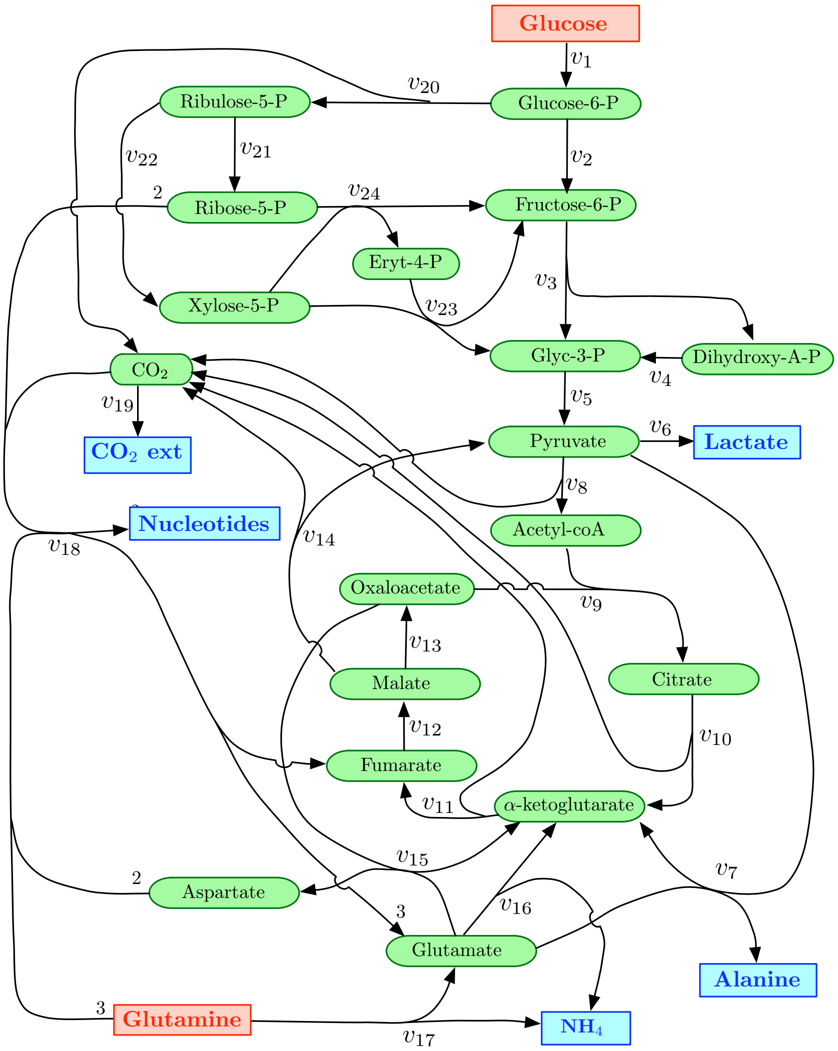


Figure 8.1. Intracellular metabolic network of CHO cells (Bastin, 2007).

8.3 Case study: CHO cells

To illustrate the described methods we consider the example of Chinese Hamster Ovary cells (CHO cells) cultivated in batch mode.

Preparation: metabolic network and constraint-based model

A metabolic network that describes the metabolism concerned with the two main energetic nutrients, glucose and glutamine has been taken from (Bastin, 2007).[1] The network is depicted in Figure 8.1.

The network includes 31 reactions (24 internal, 6 exchanges and the biomass growth) and 25 metabolites (these are listed in tables 2 and 3). There are no redundant mass balances, therefore the network has 6 degrees of freedom. The corresponding 25×31 stoichiometric matrix **N** is given in Table 8.1. The vector of reactions irreversibility, which defines the diagonal of the matrix **D**, is also given in Table 8.1. The 7 fluxes that represent uptakes or productions of extracellular metabolites are the last ones in **v**. In this way, the constraint-based model (4) is completely defined.

Preparation: measurements

Measurements of concentration for glucose (G), alanine (A), lactate (L), glutamine (Q) and ammonia (NH4) and the growth rate (μ) were taken from (Provost, 2006). Those data were collected with a sample rate of 24 h. The uncertainty of the measurements is represented in possibilistic terms as follows:

- Values near the measured ones, within ±2% deviation, are considered fully possible, π=1 (to account for systemic errors).
- A decreasing possibility is assigned to larger deviations: values with a deviation of ±5% have a possibility of π=0.5 and those with a deviation of ±10% a possibility of π=0.15 (to account for random errors).

Notice that possibility has been defined by conjunction; thus, if two measurements are deviated with possibilities 0.8 and 0.5 respectively, their joint possibility will be 0.4.

Dynamic Poss-MFA: estimating fluxes and metabolites

First, we show how the dynamic Poss-MFA can be used to estimate all the metabolic fluxes (measured or not) and the metabolite concentrations along the cultivation process (0-196 h). We used (13) to compute three possibilistic interval (π=1, π=0.5,

[1] The network is an extension of the one given in (Provost, 2004), which was used in chapter IV and V.

π=0.15) for each variable at each time instant. This implies solving 2·3·9 LP problems (2·*p*·*t*) for each variable.

The evolution of the metabolite concentrations is depicted in Figure 8.2. Remarkably, it is possible to estimate the evolution of the concentration of non-measured metabolites, such as CO2 (dynamic Poss MFA is thus being used as an observer). The estimations of measured concentrations are also valuable since they can reduce the uncertainty of the measurements, and correct them if they are inconsistent—however, this effect was not significant in the considered example.

The estimated fluxes are depicted in Figure 8.3. It can be observed that some of them are estimated with precision (v_5 or v_7), whereas other estimates are wider (v_8 or v_{12}). However, even the wider ones can be valuable: for instance, v_{12} indicates that this reaction is always active during exponential growth (0-120 h). Uptake or production rates for the extracellular metabolites can also be estimated (v_{25} or v_{26}).

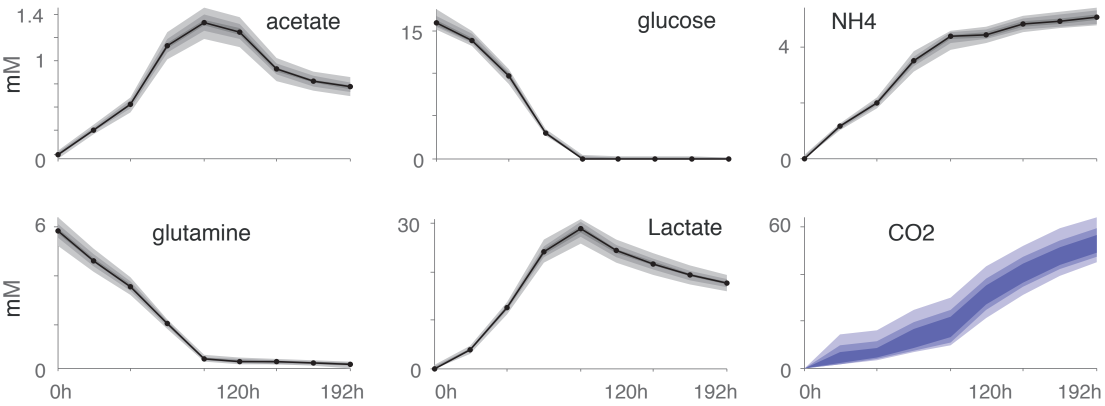


Figure 8.2. Measured and estimated metabolite concentrations during a cultivation of CHO cells. Measurements are denoted with black dots. The concentrations estimated with Poss-MFA for three degrees of possibility (π=1, π=0.5 and π=0.15) are denoted with grey and blue areas.

Dynamic Poss-MFA: estimating fluxes and metabolites (isolated)

The estimation is now performed with the isolated formulation described in section 8.3, instead of the simultaneous one used above. The results are depicted in Figure 8.4. As expected, the obtained estimates are similar, but wider. The increase of the estimated areas (one per variable and degree of possibility) with respect to those obtained with the simultaneous approach have been calculated: on average, the estimation of fluxes is 3.2% larger (between 0% and 12.4%) and the estimation of metabolites is 4.3% larger (values between 0% and 7.8%). The oversize is depicted in Figure 8.5. It can be checked that it is reasonably small, at least for this particular example.

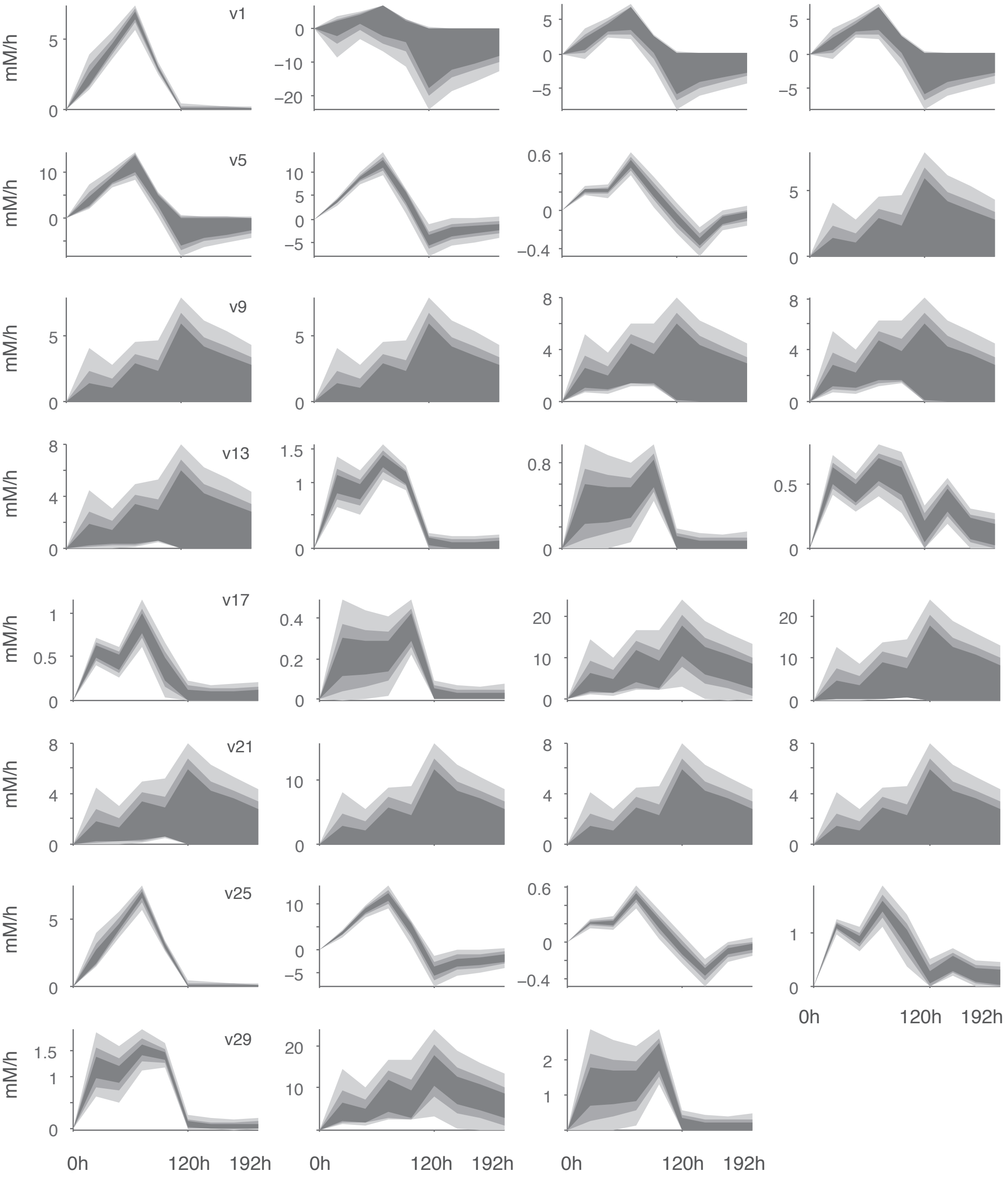


Figure 8.3. Estimated fluxes during a cultivation of CHO cells. The fluxes estimated with Poss-MFA for three degrees of possibility (π=1, π=0.5 and π=0.15) are denoted with grey areas.

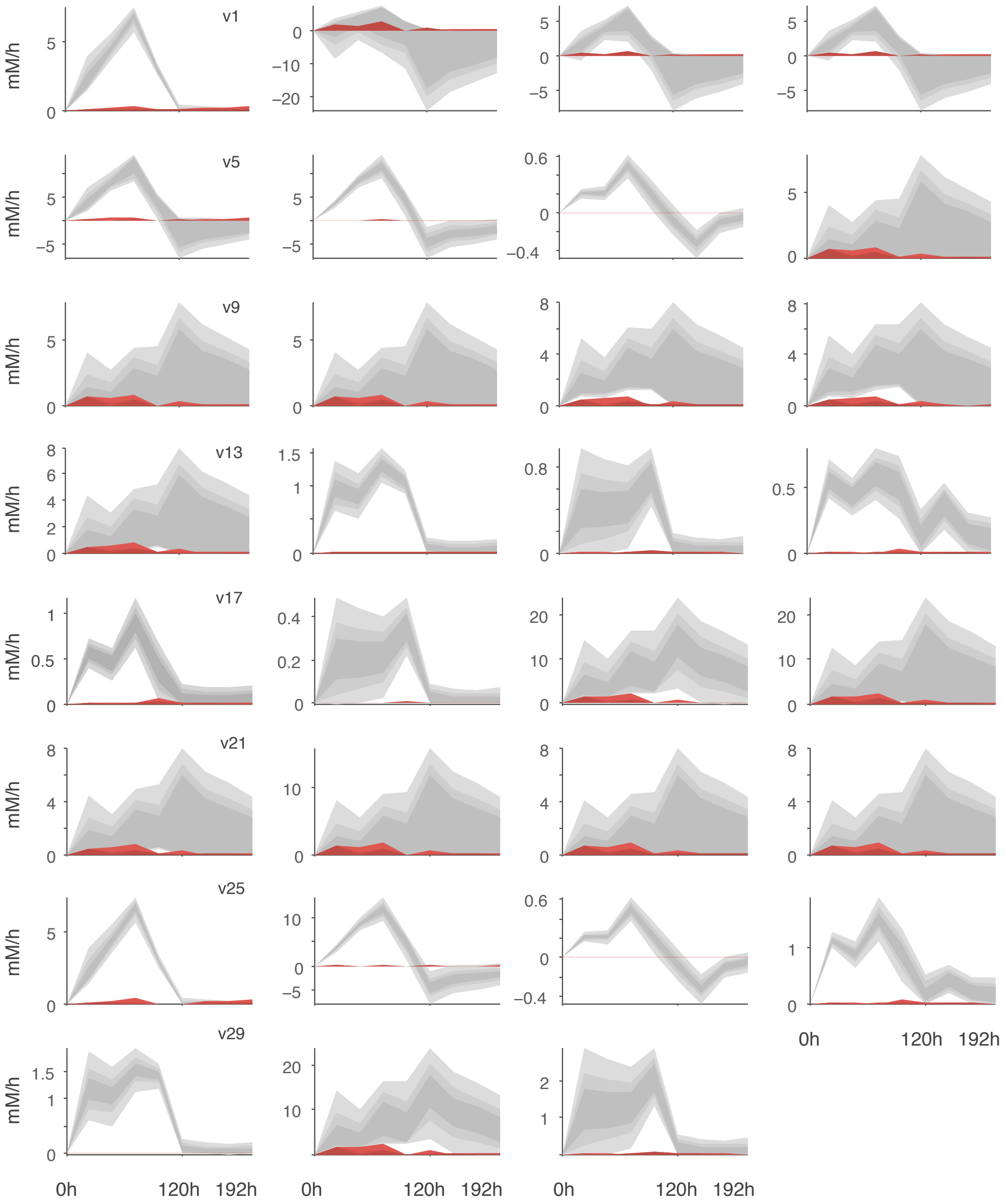


Figure 8.4. Estimated fluxes during a cultivation of CHO cells. The fluxes estimated with Poss-MFA (isolated approach) for three degrees of possibility (π=1, π=0.5 and π=0.15) are denoted with grey areas. The oversize respect to the results obtained with the simultaneous approach (Figure 8.3) is represented with red areas.

Dynamic Poss-MFA: monitoring consistency

Herein we apply the ideas described in section 8.2 to detect errors by monitoring the degree of consistency between measurements and model along the cultivation. To perform this analysis we solved one LP problem (7) to obtain the maximum possibility solution of the constraint satisfaction problem (4-5). The solution obtained is fully possible ($\pi^{mp} = 1$), indicating that the original measurements were consistent. That is, the measurements show full agreement with the model during the whole batch process (for the considered degree of uncertainty).

For the shake of illustration, we repeated the analysis after introducing two errors in the measurements:

(a) A deviation of 65% in the measurement of glucose at 48 h,

(b) A deviation of 0.5 mM in the measured NH4 at 120 h.

Now the solution of (7) showed a very low possibility ($\pi^{mp} = 0.04$), meaning that errors are detected. We then performed the same analysis using QP, after choosing an appropriate matrix **W** so that measurements uncertainty is represented in a similar way. The QP analysis also detected the inconsistency ($\pi^{mp} = 0.06$).

To investigate the candidate sources of inconsistency, he values of the slack variables were calculated as explained in section 8.2, which can be inspected with the monitoring charts given in Figure 8.5. The upper charts represent the contribution to the total

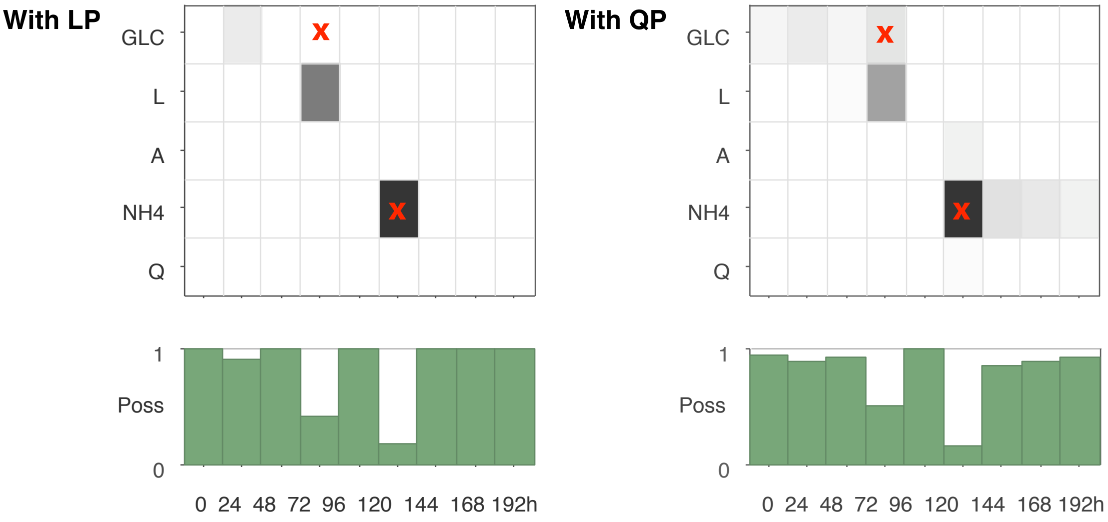


Figure 8.5. Possibilistic monitoring of consistency with Poss-MFA. On the left, the analysis with LP, on the right, the one with QP. The upper charts represent the contribution to the inconsistency of each measurements (white denotes no inconsistency, π=1, black total contradiction, π=0). The charts at the bottom monitors the aggregated consistency per time instant. Red crosses represent the measurements were the errors were introduced.

Table 8.1. Stoichiometric matrix for CHO cells.

	Irrevers.	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1
	Reaction	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	vG	vL	vA	vNH_4	vQ	vCO_2	vBio
1	G6P	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0
2	F6P	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
3	G3P	0	0	1	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
4	DAP	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Pyr	0	0	0	0	1	-1	-1	-1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	ACO	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	Cit	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	aKG	0	0	0	0	0	0	1	0	0	1	-1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Fum	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
10	Mal	0	0	0	0	0	0	0	0	0	0	0	1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	Oxa	0	0	0	0	0	0	0	0	-1	0	0	0	1	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	Glu	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	-1	-1	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0
13	Asp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	-2	0	0	0	0	0	0	0	0	0	0	0	0	0
14	RU5P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	-1	0	0	0	0	0	0	0	0	0
15	RI5P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-2	0	0	1	0	0	-1	0	0	0	0	0	0	0
16	X5P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	-1	0	0	0	0	0	0	0
17	E4P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	1	0	0	0	0	0	0	0
18	CO2i	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	0	0	-1	-1	1	0	0	0	0	0	0	0	0	0	0	0
19	NUC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	-0,17
20	G	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
21	L	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0
22	A	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0
23	NH4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	-1	0	0	0
24	Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-3	0	0	0	0	0	0	0	0	0	0	1	0	0
25	CO2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	-1	0

Table 8.2. List of initial substrates, extracellular and intracellular products.

G	Glucose	Substrates	Q	Glutamine	initial substrates
L	Lactate	extracell. product	A	Alanine	extracell. product
NH4	Ammonia	extracell. product	CO2	Carbon dioxide	extracell. product
Nuc	Nucleotides	intracell. product			

Table 8.3. List of internal metabolites.

G6P	Glucose-6-phosphate	Mal	Malate
F6P	Fructosa-6-phosphate	Oxa	Oxaloacetate
G3P	Glyceraldehyde-3-phosphate	Glu	Glutamate
DAP	Dihydroxy-acetone Phosphate	Asp	Aspartate
Pyr	Pyruvate	Ri5P	Ribose-5-Phosphate
ACO	Acetyl-coenzyme A	Ru5P	Ribulose-5-Phosphate
Cit	Citrate	X5P	Xylose--Phosphate
aKG	α-ketoglutarate	E4P	Eryt-4-Phosphate
Fum	Fumarate	CO2i	Carbon dioxide (intracellular node)

inconsistency of each measurement: a white background means no inconsistency (π=1) and a black one total contradiction (π=0). The lower charts monitor the aggregated degree of consistency at each time instant.

Clearly, both LP and QP consistency analysis are able to identify the error in the measured NH4 at time 120 h. However, the LP analysis fails to detect the error in the measured glucose: it detects that the error is at 120 h, but it erroneously suggests that the source is the measured lactate. Conversely, the QP analysis is able to detect that an error in the measured glucose can also be causing the problem (even if it still suggests that an error in lactate is a more likely cause given the declared measurements uncertainty). This example illustrates why the use of QP is a better choice to perform the consistency analysis.

8.4 Dynamic Possibilistic FBA

As explained in chapter II, flux balance analysis (FBA) is a methodology that uses optimisation to get predictions from a constraint-based model invoking an assumption of optimal cell behaviou. Basically, one particular state among those that cells can show, accordingly to a constraint-based model, is promoted based on the assumption that cells have evolved to be optimal (and that their "objective" is known and can be expressed, at least approximately, in convenient mathematical terms).

FBA is typically used to analyse cells at a particular state, but extracellular dynamics have been taken into account to predict fluxes and external metabolites during a cultivation (Mahadevan, 2002; Hjersted, 2009). The novelty of the approach described hereinafter is that optimality is defined in a gradual way using possibility theory: the optimal state is considered fully possible, and the more a state differs from it, the less possible such situation is considered. This enables getting dynamic FBA predictions accounting for alternate optima and a desired degree of sub-optimality.

Problem setting

Dynamic FBA considers the model constraints at each time instant k, as in (4), including dynamic mass balances for the extracellular metabolites, assuming that the internal ones are at steady-state, and imposing constraints on reactions reversibility. Notice, however, that measurements are not incorporated. Instead, constraints on a few uptake fluxes are imposed based on known capacities, on a kinetic expression or on the availability of substrates. These constraints are denoted as $\mathcal{CAP}(k)$:

$$\mathcal{CAP}(k) = \mathbf{v}_{\mathbf{u}}^{\mathrm{m}}(k) \geq \mathbf{v}_{\mathbf{u}}(k) \geq \mathbf{v}_{\mathbf{u}}^{\mathrm{M}} \tag{15}$$

Then, to get FBA predictions one has to invoke an optimal use of resources (e.g. maximum growth), expressed by means of a linear cost index Z(*k*):

$$\mathrm{Z}(k) = \mathbf{d} \cdot \mathbf{v}(k) \tag{16}$$

FBA predictions at each time instant *k* could be now obtained maximising Z(*k*) subject to the operating constraints, $\mathcal{MOC}(k)$ and $\mathcal{CAP}(k)$. However, some refinements can be easily incorporated.

Considering sub-optimality

To account for optimality in a gradual way, the following constraints are defined:

$$\mathrm{Z}(k) = \mathrm{Z}_{\max}(k) \cdot \left(1 - \phi(k)\right) \tag{17a}$$

$$0 \le \phi(k) \le 1 \tag{17b}$$

where $\phi(k)$ is a slack variable that represents sub-optimality.

Now, possibility can be redefined in terms of optimality using a new cost index $\mathrm{J}_{\mathrm{opt}}$:

$$\mathrm{J}_{\mathrm{opt}}(k) = \alpha_{\mathrm{s}} \cdot \phi(k) \tag{18}$$

where α_s is user-defined weight linking possibility and optimality. For instance, if one chooses α_s = -log(0.5), then $\pi = 0.5$ is assigned to $\mathrm{Z}=0.5\cdot\mathrm{Z}_{\max}$.

In this way, only an optimal cells state, **var**(*k*), maximising Z(*k*) is considered fully possible, and the more a state differs from this optimal one, the less possible such situation is considered.

Dynamic Poss-FBA: predicting fluxes and metabolites

Predictions with sub-optimality (possibility) γ are obtained successively, for each *k*, following a two-step procedure:

Step 1

$$\max\ Z(k)\qquad \text{s.t.}\begin{cases} \mathcal{MOC}(k) & 1\ldots k \\ \mathcal{CAP}(k) & 1\ldots k \\ Z(k)=Z^{\max}(k)\cdot\left(1-\phi(k)\right) & 1\ldots(k\text{-}1) \\ 0\leq\phi(k)\leq 1 & 1\ldots(k\text{-}1) \\ \alpha_s\cdot\phi(k)<\log\gamma & 1\ldots(k\text{-}1) \end{cases} \tag{19}$$

The last three constraints guarantee that the optimal solution at *k*, $Z^{\max}(k)$, does not violate the optimality γ at previous time instants $\{1\ \ldots\ k\text{-}1\}$.

In the second step, $Z^{\max}(k)$ is used as reference to get the sub-optimal predictions as possibilistic intervals, $[var^{m}_{i,\gamma}(k)\ ,var^{M}_{i,\gamma}(k)\,]$:

Step 2

$$var^{m}_{i,\gamma}(k)=\min\ var_i(k)\qquad \text{s.t.}\begin{cases} \mathcal{MOC}(k) & 1\ldots k \\ \mathcal{CAP}(k) & 1\ldots k \\ Z(k)=Z^{\max}(k)\cdot\left(1-\phi(k)\right) & 1\ldots k \\ 0\leq\phi(k)\leq 1 & 1\ldots k \\ \alpha_s\cdot\phi(k)<\log\gamma & 1\ldots k \end{cases} \tag{20}$$

Bound $var^{\mathrm{M}}_{\mathrm{i},\gamma}(k)$ is obtained by replacing minimum by maximum.

This two-step procedure can be repeated for different degrees of possibility—to say, π=1, π=0.8 and π=0.5—thus getting a rich prediction that considers sub-optimality. and accounts for alternate optima (those in the interval estimate of π=1).

8.5 Case study: *E. coli*

To illustrate the kind of results that can be obtained with the dynamic Poss-FBA, we use an example of diauxic growth of *E. coli* on glucose and acetate. The example has been taken from Mahadevan et al. (2002), where dynamic FBA was presented.

Preparation: metabolic network and constraint-based model

Mahadevan et al. (2002) chose 4 pathways from a genome-scale reconstruction of *E. coli*, and used them to define a simplified network with 3 extracellular metabolites, glucose (G), acetate (A) and oxygen (O), and biomass (*x*):

$$\begin{array}{lrl}
v_1: & 39.43\ \mathrm{A} + 35\ \mathrm{O2} & \rightarrow x \\
v_2: & 9.46\ \mathrm{G} + 12.92\ \mathrm{O2} & \rightarrow x \\
v_3: & 9.84\ \mathrm{G} + 12.73\ \mathrm{O2} & \rightarrow 1.24\ \mathrm{A} + x \\
v_4: & 19.23\ \mathrm{G} & \rightarrow 12.12\ \mathrm{A} + x
\end{array} \tag{21}$$

A constraint-based model accounting for these metabolites and biomass can be defined with the constraints $\mathcal{MOC}(k)$ and $\mathcal{CAP}(k)$. We consider a duration of the batch of 10 h, divided in 21 intervals, so that $k = [1, 2,..., 21]$.

The first constraints in $\mathcal{MOC}(k)$ are the mass balances around the extracellular metabolites and biomass, as in (4), are the following:

$$\begin{gathered}
\frac{\mathrm{G}(k)-\mathrm{G}(k-1)}{\Delta\mathrm{T}} = \begin{pmatrix} 0 & -9.46 & -9.84 & -19.23 \end{pmatrix} \cdot \mathbf{v}(k) \\
\frac{\mathrm{A}(k)-\mathrm{A}(k-1)}{\Delta\mathrm{T}} = \begin{pmatrix} -39.43 & 0 & 1.24 & 12.12 \end{pmatrix} \cdot \mathbf{v}(k) \\
\frac{\mathrm{O2}(k)-\mathrm{O2}(k-1)}{\Delta\mathrm{T}} = \begin{pmatrix} -35 & -12.92 & -12.73 & 0 \end{pmatrix} \cdot \mathbf{v}(k) + k_L\left(0.21-\mathrm{O2}(k-1)\right) \\
\frac{x(k)-x(k-1)}{\Delta\mathrm{T}} = \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \cdot \mathbf{v}(k)
\end{gathered} \tag{22}$$

where G, A and O denote the metabolite concentrations (in mM), and x the biomass concentration (in g/L). The mass transfer coefficient for oxygen, k_L, is 7.5 h^{-1} accordingly to (Edwards et al., 2001), and the oxygen concentration in the gas phase is assumed to be a constant and equal to 0.21 mM.

Constraints are also incorporated to define fluxes as irreversible, and to impose a positiveness condition to the concentrations (which, obviously, cannot be negative):

$$\mathbf{D}\cdot\mathbf{v}(k) \geq 0 \tag{23a}$$

$$\mathbf{e}(k) \geq 0 \text{ and } \mathrm{x}(k) \geq 0 \tag{23b}$$

The constraints $\mathcal{CAP}(k)$ only bound the glucose uptake by the measured concentrations[1] of glucose $\mathrm{G_m}(k)$:

[1] Similar results were obtained when the uptake was modelled with Michaelis-Menten kinetics instead of using the measured values of glucose concentration.

$$\frac{G(k)-G(k-1)}{\Delta T}=\frac{G_m(k)-G_m(k-1)}{\Delta T}\Rightarrow G(k)=G_m(k) \tag{24}$$

In this way, the constraint operating at each time instant have been defined, $\mathcal{MOC}(k)$ constraints are defined with (22) and (23), and $\mathcal{CAP}(k)$ constraints with (24).

The last step is define cells optimality. In this example, following (Mahadevan, 2002) it is considered that the cells objective is to maximise growth. This assumption can be expressed with the following objective function:

$$Z(k)=\begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix}\cdot \mathbf{v}(k) \tag{25}$$

To account for sub-optimality, the parameter α_s (18) is defined as α_s = -log(0.5), so that possibility is 0.5 when the biomass growth is 50% of maximum.

Dynamic Poss-FBA: predictions of fluxes and metabolites

The two-step procedure described above (19-20) is applied to get dynamic estimates for all the variables, fluxes and metabolites, for three degrees of optimality, π=0.95, π=0.8 and π=0.5. The results are depicted in figures 6 and 7.

It can be observed that Poss-FBA detects alternative optima. It provides alternative predictions for v_2 and v_3 (see Figure 8.7) even if only a slight sub-optimality is allowed (π=0.95), what seems sensible because both pathways have similar yields (i.e., both are nearly exchangeable in terms of biomass growth). This could indicate that both pathways can be efficiently used by the organism, or more likely, that the selection of one of them depends on phenomena not captured by the model (e.g., the choice depends on a secondary objective or its regulated by an environmental condition different from the substrates availability).

The results in Figure 8.6 also show that considering sub-optimality gives a richer prediction, and better agreement with the actual concentrations. Cells behaviour can be reasonably captured with the simple model considered here, even if it seems clear that the assumption of "maximisation of growth" is not perfect. During the phase of growth on glucose, the growth rate is between 80% and 50% of the maximum. In the second phase, when acetate is consumed, actual behaviour seems nearer to the optimal one.

Considering sub-optimality and alternate optima also provides an indication of the uncertainty of each prediction. For instance, as expected, the assumption of "maximisation of growth" provides a narrower prediction for biomass, than for oxygen or acetate, for which wider ranges of values are reasonably possible.

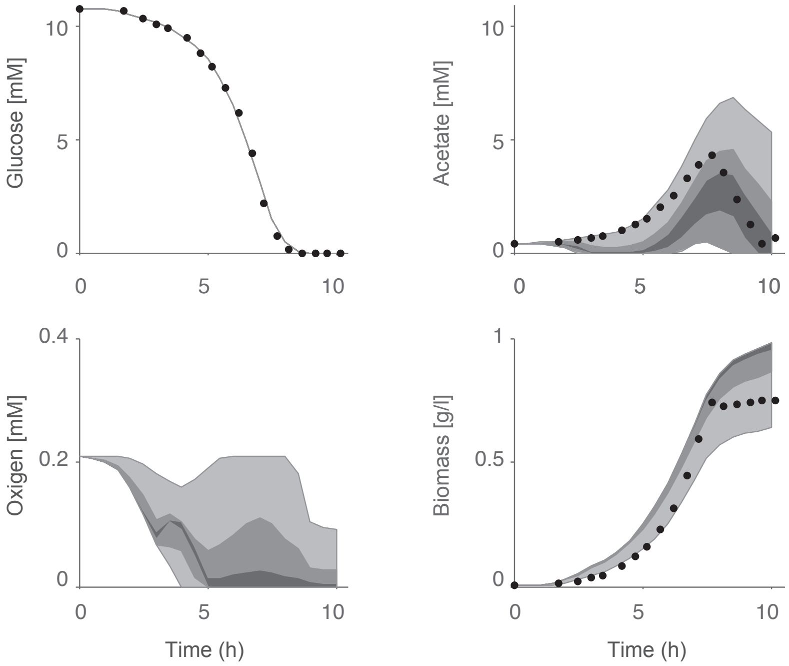


Figure 8.6. Measured and predicted metabolite concentrations during a cultivation of *E. coli*. Measurements are denoted with black dots. The concentrations estimated with Poss-FBA for three degrees of optimality (π=0.95, π=0.8 and π=0.5) are denoted with grey areas. Recall that Poss-FBA only uses glucose measurements to perform the estimation.

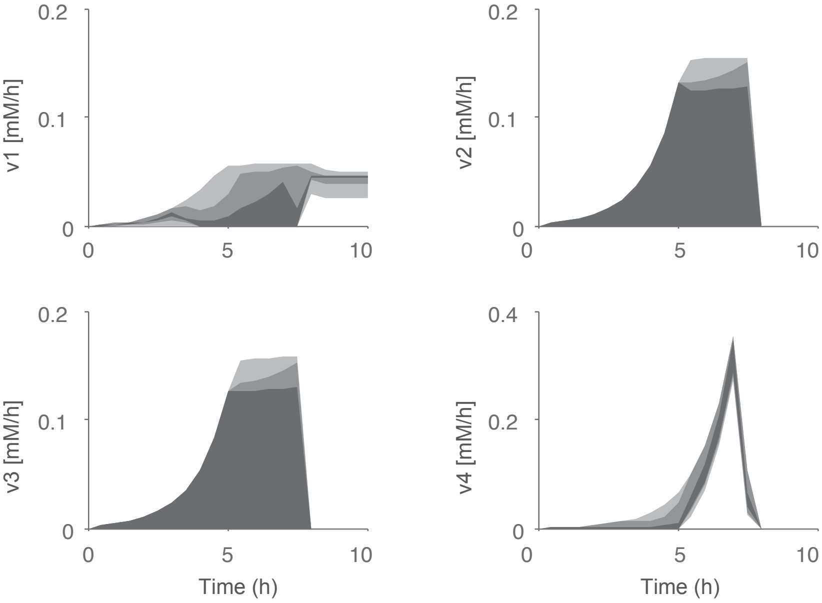


Figure 8.7. Estimated fluxes during a cultivation of *E. coli*. The fluxes estimated with Poss-FBA for three degrees of optimality (π=0.95, π=0.8 and π=0.5), are denoted with grey areas.

8.6 Conclusions

In this chapter we have discussed the benefits that the possibilistic framework introduced in chapter VII brings when getting predictions from a constraint-based model accounting for extracellular dynamics.

In the context of MFA, it has been shown how to estimate time-varying fluxes and extracellular metabolite concentrations considering uncertainty and dealing with data scarcity. We have also outlined a procedure for monitoring the consistency between measurements and mode during a cultivation, which can be a useful tool for on-line fault detection in industrial processes. Notice also that dynamic Poss-MFA inherits other benefits of the possibilistic framework that were discussed in chapter VII. For instance, Poss-MFA handles data scarcity, and represents knowledge in a flexible way to account for measurements uncertainty or model imprecision.

As stated above, the first method described to perform dynamic Poss-MFA computations, which considers all the constraints simultaneously, is computationally expensive when the sampling rate is high. Fortunately, this problem will be rare because extracellular dynamics are typically slow and measurements are taken with low sampling rates. To deal with high sampling rates, the so-called "isolated" approach, or a mixed one, can be used, but this comes at the cost of wider estimations. Future work should address this issue, because a better approach to deal with faster sampling rates could make the methodology suitable to other problems.

In the context of FBA we have shown that the possibilistic approach is able to provide rich predictions, for fluxes and external metabolites, which (a) consider sub-optimality, thus improving the agreement with measured data, and (b) include alternate and quasi-alternate optima solutions.

In summary, we have shown that the possibilistic framework enables getting richer dynamic predictions from a constraint-based model, using measurements, as in Poss-MFA, or invoking optimal cell behaviour, as in Poss-FBA.

Main references

- Llaneras F, Sala A, Picó J (2009). A possibilistic framework for metabolic flux analysis. *BMC Systems Biology*, 3:73.

- Henry O, Kamen A, Perrier M (2007). Monitoring the physiological state of mammalian cell perfusion processes by on-line estimation of intracellular fluxes. *Journal of Process Control*, 17:241-251.

- Bastin G (2007). Quantitative analysis of metabolic networks and design of minimal bioreaction models. *International Conference in Honor of Claude Lobry*.

- Edwards JS, Covert M, Palsson B (2002). Metabolic modelling of microbes: the flux-balance approach. *Environmental Microbiology*, 4:133-140.

- Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Molecular Systems Biology* 3:119.

- Hjersted JL, Henson MA (2009). Steady-state and dynamic flux balance analysis of ethanol production by *Saccharomyces cerevisiae*. *IET Systems Biology*, 3(3):167–79.

- Mahadevan R, Edwards JS, Doyle FJ (2002). Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophysics Journal*, 83:1331–1340.

IX

Possibilistic validation of a constraint-based model of *P. pastoris*

In this chapter elementary modes analysis and Possibilistic MFA are used to validate against experimental data a model of *P. pastoris*, a yeast used in industry for the expression of recombinant proteins.

This work follows a systematic, yet simple, procedure to validate small-sized constraint-based models in a common scenario of data scarcity.

Part of the contents of this chapter have been published in the journal paper:

- Tortajada M, Llaneras F, Picó J. Validation of a constraint-based model of *Pichia pastoris* growth under data scarcity. *BMC Systems Biology*, 4:115.

9.1 Introduction

The biochemical reactions involved in the metabolism of cells are assembled in networks, which can then be used to build constraint-based models, assuming that internal metabolites not accumulate (thus avoiding reaction kinetics) and incorporating other constraints, such as enzyme kinetics, thermodynamics, or the irreversibility of certain reactions (see chapter II for details). These constraint-based models are often build upon large, or genome-scale, networks of well-characterised organisms such as *E. coli*, *S. cerevisiae,* or *P. putida,* (Feist, 2007; Nogales, 2008) but also in simpler networks that consider only a few key metabolites (Schuetz, 2007; Teixeira, 2007; Nookaew, 2007).

As seen in previous chapters, a constraint-based model can be combined with extracellular measurements to perform metabolic flux analysis (MFA) and estimate the non-measured fluxes in the network. This provides information about the state of cells at given circumstances.

The main difficulty to be faced to apply MFA is the lack of measurements[1]. If one considers all the complexity of the metabolic network of a cell, the available measurements (and known constraints) cannot offset its under-determinacy nor reduce it enough to get valuable estimates. This is why MFA can only be performed using reasonably small networks. To keep reductions of the network at minimum, intracellular measurements from tracer experiments can be incorporated (Sauer, 2006; Wiechert, 2001), but those data are in most cases not available. Interval and possibilistic methods (chapters IV and VII) are also helpful because do not require to completely offset the network under-determinacy to get the estimates. However, the main fact remains: reasonably small networks are required.

Unfortunately, these small-sized networks are sometimes not properly validated, even if they are simplifications of the whole (known) metabolism of a cell, and rely necessarily on reductionist hypothesis. They are often not evaluated against datasets different from the one of interest, which is thus inconveniently used both to validate the model and to perform the MFA analysis. Herein we discuss a procedure seeking for a more exhaustive validation of these networks.

We will follow a systematic, yet simple, procedure to validate a small-sized model of *P. Pastoris* using only data from extracellular measurements. The same procedure could be used with other organisms of industrial interest.

We work with a model of *Pichia pastoris*, a methylotrophic yeast recognised world-wide as a reference platform for the expression of recombinant proteins in eukaryotes, due to the possibility to grow cultures to very high cell densities, its ability to produce post-

[1] Notice that, indeed, this lack is the reason for the existence of MFA. If one could measure all the fluxes in the network—with accuracy and on-line—MFA will be barely useful.

translational modifications, and the good protein yield per cost ratio. Heterologous genes are cloned under *P. pastoris*' strong and tightly regulated alcohol oxidase promoter, and thus expressed when the cells grow on methanol as sole or combined carbon source. The optimisation of recombinant protein expression in *P. pastoris* has been usually addressed heuristically. Only a few publications describe rational, model-based optimisation of *Pichia* growth and protein production. Among these, structured or metabolism-based models representing intracellular behaviour are particularly rare (Ren, 2003; Solà, 2007).

The chapter is organised as follows. First, a constraint-based model of *P. pastoris* will be described and validated against the available experimental data. Then, its ability to predict non-measured fluxes will be illustrated by estimating the biomass growth rate. The potential use of the model to predict intracellular fluxes will be discussed to close the chapter.

9.2 Methods

Recalling the formulation used in previous chapters, a *constraint-based model*—assuming steady-state for internal metabolites and considering the irreversibility of some reactions—can be described with a set of model constraints ($\mathcal{MOC}$) as follows:

$$\mathcal{MOC} = \begin{cases} \mathbf{N} \cdot \mathbf{v} = \mathbf{0} \\ \mathbf{D} \cdot \mathbf{v} \geq \mathbf{0} \end{cases} \tag{1}$$

Where $\mathbf{v}$ is the flux vector representing the mass flow through each of the n reactions in the network, $\mathbf{N}$ is the stoichiometric matrix, and $\mathbf{D}$ is a diagonal matrix with $\mathbf{D_{ii}}$ = 1 if the flux i is irreversible (otherwise 0).

The constraints in (1) define a space of feasible steady-state flux vectors, or flux states, which ideally comprises every theoretically possible phenotype: only flux vectors $\mathbf{v}$ that fulfil (1) are considered valid cellular states.

Consistency analysis

The simplest consistency analysis could be performed checking that the flux states shown by cells fulfils the constraints imposed by the model. However, this simple approach would be impractical because, as measurements are imprecise, they do not *exactly* satisfy the constraints. Such difficulty is overcome by taking into account uncertainty, as follows:

$$\mathbf{v_m} = \mathbf{w_m} + \mathbf{e_m} \tag{2}$$

where $\mathbf{e_m}$ represents the deviation between the fluxes $\mathbf{v_m}$ in $\mathbf{v}$ and the measurements $\mathbf{w_m}$.

Model and measurements will be consistent if there is a flux vector **v** fulfilling (1) and (2) for "reasonably small" deviations $\mathbf{e_m}$. Otherwise, we will conclude that model and measurements are inconsistent. An easy way to evaluate the consistency is finding the flux vector **v** fulfilling (1-2) that minimises the (variance-weighted) sum of measurements errors:

$$\min \quad \Phi = \mathbf{e}_\mathbf{m}^\mathrm{T} \cdot \mathbf{F}^{-1} \cdot \mathbf{e}_\mathbf{m} \quad \text{s.t.} \quad \mathcal{MOC} \tag{3}$$

where it is assumed that $\mathbf{e_m}$ are distributed normally with a mean value of zero and a variance-covariance matrix **F**.

If only linear equality constraints are considered in $\mathcal{MOC}$, the residual ϕ is a stochastic variable following a χ^2-distribution, and therefore a χ^2-test can be used to detect and evaluate the inconsistency. The χ^2-test is based upon statistical hypothesis testing to determine if the deviation is within expected experimental error (See chapter II). However, we want to consider inequality constraints in (2), so the χ^2-test cannot be used because its assumptions are not fulfilled (ϕ does not follows a χ^2-distribution anymore). Yet, the residual ϕ provides at least a rough indication of consistency.

Consistency analysis with Possibilistic MFA

The consistency analysis can also be formulated as a possibilistic constraint satisfaction problem, following the ideas presented in chapter VII. The basic idea is that a flux vector fulfilling the model constraints (1) and compatible with the measurements will be considered "possible", otherwise "impossible". This idea can be refined to handle measurements errors by using the notion of "degree of possibility".

As explained in chapter VII, we can introduce a set of measurement constraints ($\mathcal{MEC}$) considering measurement imprecision, as in (2), but where $\mathbf{e_m}$ is substituted by two pairs of nonnegative decision variables:

$$\mathcal{MEC} = \begin{cases} \mathbf{v_m} = \mathbf{w_m} + \varepsilon_1 - \mu_1 + \varepsilon_2 - \mu_2 \\ \varepsilon_1, \mu_1 \geq 0 \\ 0 \leq \varepsilon_2 \leq \varepsilon_2^{max} \\ 0 \leq \mu_2 \leq \mu_2^{max} \end{cases} \tag{4}$$

These decision variables $\{\varepsilon_1, \mu_1, \varepsilon_2, \mu_2\}$ relax the basic assertion $\mathbf{w_m} = \mathbf{v_m}$, conforming a possibility distribution in $(\mathbf{w_m}, \mathbf{v_m})$ associated to some cost index J.

Among different possible choices, a simple –yet sensible– one is the linear cost index:

$$\mathrm{J} = \alpha \cdot \varepsilon_1 + \beta \cdot \mu_1 \tag{5}$$

with $\alpha \geq 0$ and $\beta \geq 0$ being row vectors of user-defined, sensor reliability coefficients.

The cost index J reflects the log-possibility of a particular combination of the decision variables $\delta = \{\mathbf{v}, \varepsilon_1, \mu_1, \varepsilon_2, \mu_2\}$, that is, the log-possibility of a particular flux vector **v**. The possibility of each solution is given by:

$$\pi(\delta) = \mathrm{e}^{-\mathrm{J}(\delta)} \quad \delta \in \mathcal{MOC} \cap \mathcal{MEC} \tag{6}$$

The interpretation of (4) and (5) may be: "$\mathbf{w_m} = \mathbf{v_m}$ is fully possible; the more $\mathbf{w_m}$ differs from $\mathbf{v_m}$, the less possible such situation is".

The maximum possibility (minimum-cost) flux vector $\mathbf{v_{mp}}$ corresponding to a given set of measurements can be obtained solving a linear programming (LP) problem:

$$\min_{\varepsilon,\mu,\mathrm{v}} \mathrm{J} \quad \text{s.t.} \begin{cases} \mathcal{MOC} \\ \mathcal{MEC} \end{cases} \tag{7}$$

The possibility of the most possible flux vector $\mathbf{v_{mp}}$ being, $\pi_{\mathrm{mp}} = \mathrm{e}^{-\mathrm{J}^{\min}}$.

This degree of possibility provides an indication of the consistency between model ($\mathcal{MOC}$) and measurements ($\mathcal{MEC}$): a possibility equal to one must be interpreted as complete agreement between the model and the original measurements; lower values of possibility imply that certain error in the measurements is necessary to find a flux vector fulfilling the model constraints.

See chapter VII for further technical details on the possibilistic framework.

Estimating the non-measured fluxes with Possibilistic MFA

Possibilistic MFA can also estimate the non-measured fluxes, based on the model and the available measurements (as discussed in chapter VII). The simplest point-wise estimate is the minimum-cost flux vector resulting from (7), which contains most possible value for each flux. However, a point-wise estimate is limited when multiple combinations might be reasonably possible; in this situation, a possibilistic interval estimate is a better choice.

Remember that the interval of values with conditional possibility higher than γ for a given variable, $\left[v_{i,g}^{m}, v_{i,g}^{M}\right]$, can be computed solving two LP problems:

$$v_{i,g}^{m} = \min_{\varepsilon,\mu,v} v_i \quad \text{s.t.} \begin{cases} \mathcal{MOC} \cap \mathcal{MEC} \\ \mathrm{J} - \log \pi(\mathbf{v_m}) < -\log \gamma \end{cases} \tag{8}$$

The upper bound $v_{i,g}^{M}$ would be obtained by replacing minimum by maximum.

9.3 Constraint-based model of *P. pastoris*

The metabolic network presented in Figure 9.1 is based on the stoichiometric model defined in (Dragosits, 2009) for *P. pastoris* growth on glucose, which has been extended with reactions representing methanol and glycerol metabolism.

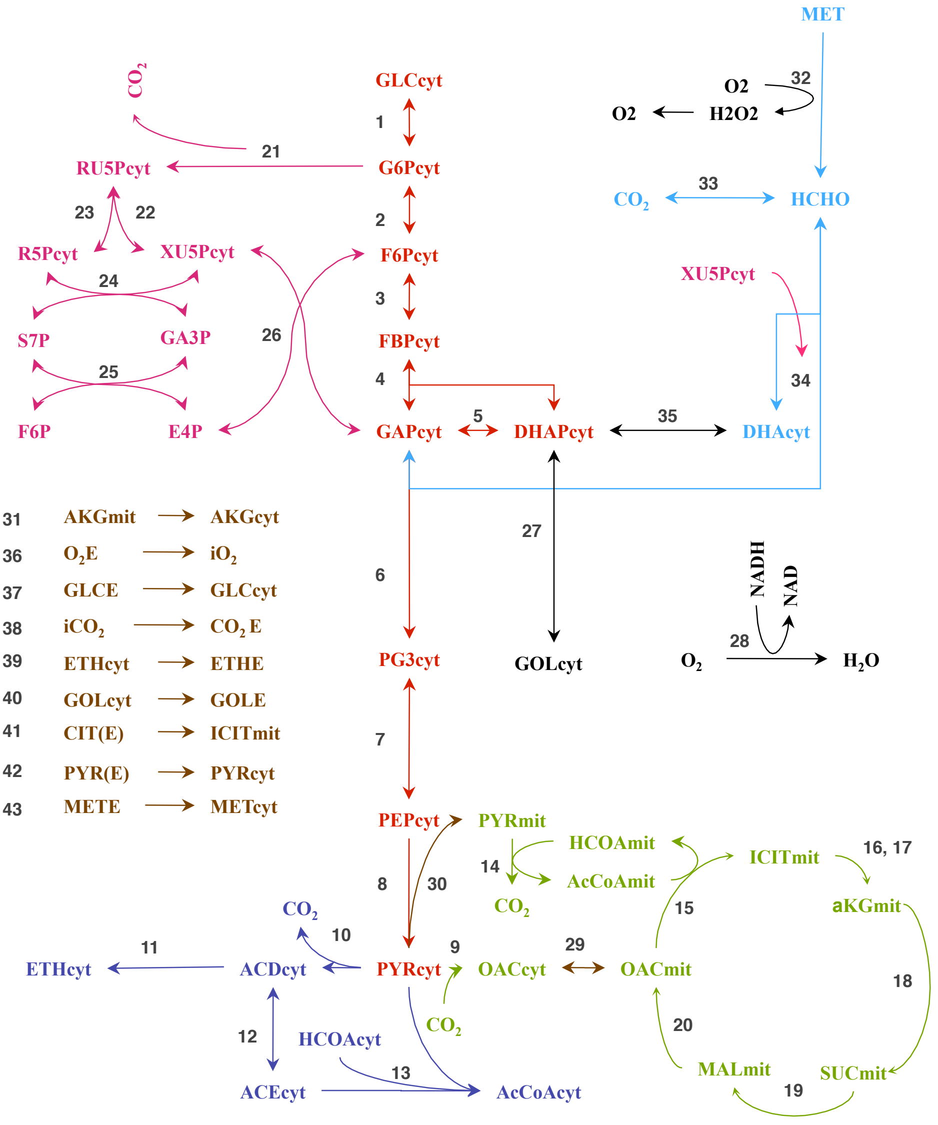


Figure 9.1. Metabolic network of *P. pastoris*. The reaction representing the biomass formation is not depicted, but given in Table 9.3.

The main catabolic pathways—Embden-Meyerhof-Parnas pathway, citric acid cycle, pentose phosphate and fermentative pathways—of the yeast *P. pastoris* are represented for growth on the substrates mainly used for its culture: glucose, glycerol and methanol. A mean biomass equation derived from the macromolecular composition of the yeast is used to summarise the anabolic pathways (Dragosits, 2009). Key metabolites such as NAD, NADP, AcCoA, oxalacetate and pyruvate are considered in distinct cytosolic (*cyt*) and mitochondrial pools (*mit*).

The model considers 45 compounds and 44 metabolic reactions (tables 1-3). The steady-state assumption can by applied to 36 metabolites, resulting in 8 degrees of freedom. The corresponding 36×44 stoichiometric matrix **N** and the vector of reactions reversibility—the diagonal of matrix **D**—is given in Table 9.4. Matrices **N** and **D** define the constraint-based model used in this rest of the chapter.

Table 9.1. Extracellular metabolites.

O2 (E)	Oxygen	Cit (E)	Citric Acid
GLU (E)	Glucose	Pyr (E)	Pyruvic acid
CO2 (E)	Carbon dioxide	Met (E)	Methanol
EtH (E)	Ethanol	Biom	Biomass
GOL (E)	Glycerol		

Table 9.2. List of internal metabolites.

GLCcyt	Glucose	ACCOAmit	Acetyl coenzyme A (mitochondrial)
G6Pcyt	Glucose-6-phosphate	OAAmit	Oxalate (mitochondrial)
F6Pcyt	Fructose-6-phosphate	ICITmit	Isocitric acid (mitochondrial)
FBPcyt	Fructose-6-biphosphate	AKGmit	2-Amino-6-ketopimelate (mitochondrial)
DHAPcyt	Dihydroxyacetone phosphate	PYRmit	Pyruvate (mitochondrial)
GAPcyt	D-glyceraldehyde 3-phosphate	SUCmit	Sucinate (mitochondrial)
PG3cyt	Glyceraldehydes-3-phosphate	MALmit	Malate (mitochondrial)
PEPcyt	Phosphoenolpyruvate	NADPHmit	NADPH (mitochondria)
PYRcyt	Pyruvate	ACDcyt	Acetaldehyde
GOLcyt	Glycerol	ACEcyt	Acetate
RU5Pcyt	Ribulose-5-phosphate	iCO2	Carbon dioxide
R5Pcyt	Ribose-5-phosphate	iO2	Oxygen
XU5Pcyt	Xylulose-5-phosphate	NADH	NADH
S7Pcyt	Sedoheptulose-7-phosphate	EtOH cyt	Ethanol
E4Pcyt	Erythrose--4-phosphate	MeOHcyt	Methanol
OAAcyt	Oxalate	HCHOcyt	formaldehyde
AKGcyt	2-Amino-6-ketopimelate	DHAcyt	dihydroxyacetone
ACCOAcyt	Acetyl coenzyme A	NADPHcyt	NAD

Cytosolic (cyt) and mitochondrial pools (mit) are considered.

Table 9.3. List of considered reactions in the model of *P. pastoris*.

System	Reaction
Embden Meyerhoff Parnas (Glycolysis)	GLCcyt > G6Pcyt
	G6Pcyt <> F6Pcyt
	F6Pcyt <> FBPcyt
	FBPcyt <> DHAPcyt + GAPcyt
	DHAPcyt <> GAPcyt
	GAPcyt + NADcyt <> PG3cyt + NADHcyt
	PG3cyt <> PEPcyt + H2O
	PEPcyt <> PYRcyt
Pyruvate branch point	PYRcyt + iCO2 > OAAcyt
	PYRcyt <> ACDcyt + iCO2
Fermentative patways	ACDcyt + NADHcyt > ETHcyt + NADcyt
	ACDcyt + NADPcyt > ACEcyt + NADPHcyt
	ACEcyt + HCOAcyt > ACCOAcyt
TCA cycle	PYRmit + HCOAmit + NADmit > ACCOAmit + iCO2 + NADHmit
	ACCOAmit + OAAmit <>ICITmit + HCOAmit
	ICITmit + NADmit >AKGmit + iCO2 + NADHmit
	ICITmit + NADPmit > AKGmit + iCO2 + NADPHmit
	AKGmit + NADmit > SUCmit + iCO2 + NADHmit
	SUCmit + NADmit > MALmit + NADHmit
	MALmit + NADmit > OAAmit+ NADHmit
Pentose phosphate pathway	G6Pcyt + 2 NADPcyt > RU5Pcyt + iCO2 + 2 NADPHcyt
	RU5Pcyt >XU5Pcyt
	RU5Pcyt > R5Pcyt
	R5Pcyt + XU5Pcyt > S7Pcyt + GAPcyt
	S7Pcyt + GAPcyt > E4Pcyt + F6Pcyt
	E4Pcyt + XU5Pcyt>F6Pcyt + GAPcyt
Glycerol formation	DHAPcyt + NADHcyt > GOLcyt + NADcyt
Oxidative phosphorylation	NADH + 0.5 iO2 > NAD
Transport reactions	OAAcyt <> OAAmit
	PYRcyt >PYRmit
	AKGmit >AKGcyt
	O2(E) > iO2
	GLC(E) > GLCcyt
	iCO2 >CO2(E)
	ETHcyt > ETH(E)
	GOL(E)> GOLcyt
	CIT(E) <> ICITmit
	PYR(E) >PYR cyt
	MET(E) > METcyt
Methanol metabolism	METcyt + 1/2 O2 > HCHOcyt + H2O
	HCHOcyt + 2 NADcyt > 2 NADHcyt + iCO2
	HCHOcyt + XU5Pcyt <> DHAcyt + GAPcyt
	DHAcyt > DHAPcyt
Biomass Synthesis	0,0033 ACCOAcyt + 0,008 ACCOAmit + 0,0266 AKGcyt + 0,0146 E4Pcyt + 0,0363 F6Pcyt + 0,0165 PG3cyt + 0,0363 G6Pcyt + 0,0000003 GOLcyt + 0,000002 iO2 + 0,0242 OAAcyt + 0,00079 OAAmit + 0,0252 PEPcyt + 0,0294 PYRmit + 0,011 R5Pcyt + 0,199 NADPHcyt + 0,056 NADPHmit + 0,0626 NAD > 1 BIOM + 0,0127 iCO2 + 0,0626 NADH + 0,0033 HCCOAcyt + 0,008 HCCOAmit + 0,199 NADPcyt + 0,056 NADPmit

Table 9.4. Stoichiometric matrix of *P. pastoris*.

	Irreversible	1	0	0	0	0	0	0	0	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	0	0	0	0	1	0	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1
	Reaction	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	O2	GLC	CO2	ET	GOL	Cit	Pyr	MET	BIO
1	GLCcyt	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
2	G6Pcyt	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0,036
3	F6Pcyt	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0,036
4	FBPcyt	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	DHAPcyt	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
6	GAPcyt	0	0	0	1	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
7	PG3cyt	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0,017
8	PEPcyt	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0,025
9	PYRcyt	0	0	0	0	0	0	0	1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	-1	0	0
10	GOLcyt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
11	NADPHcyt	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0,199
12	iCO2	0	0	0	0	0	0	0	0	-1	1	0	0	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	-1	0	0	0	0	0	0,0177
13	RU5Pcyt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	R5Pcyt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0,011
15	XU5Pcyt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	-1	0	-1	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0
16	S7Pcyt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	E4Pcyt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0,015
18	OAAcyt	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0,024
19	PYRmit	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	-0,029
20	ACCOAmit	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0,008
21	OAAmit	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0,001
22	ICITmit	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0
23	NADH	0	0	0	0	0	1	0	0	0	0	-1	0	0	1	0	1	0	1	1	1	0	0	0	0	0	0	-1	-1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0,0627
24	AKGmit	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0
25	NADPHmit	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0,056
26	AKGcyt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	-0,027
27	SUCmit	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	MALmit	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	ACDcyt	0	0	0	0	0	0	0	0	0	1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	ACEcyt	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	ACCOAcyt	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0,003
32	iO2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	-1	0	0	0	1	0	0	0	0	0	0	0	-2E-05
33	EtOH cyt	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0
34	MeOHcyt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	1	0
35	HCHOcyt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	-1	0	0	0	0	0	0	0	0	0	0
36	DHAcyt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0

9.4 Analysis of the elementary modes

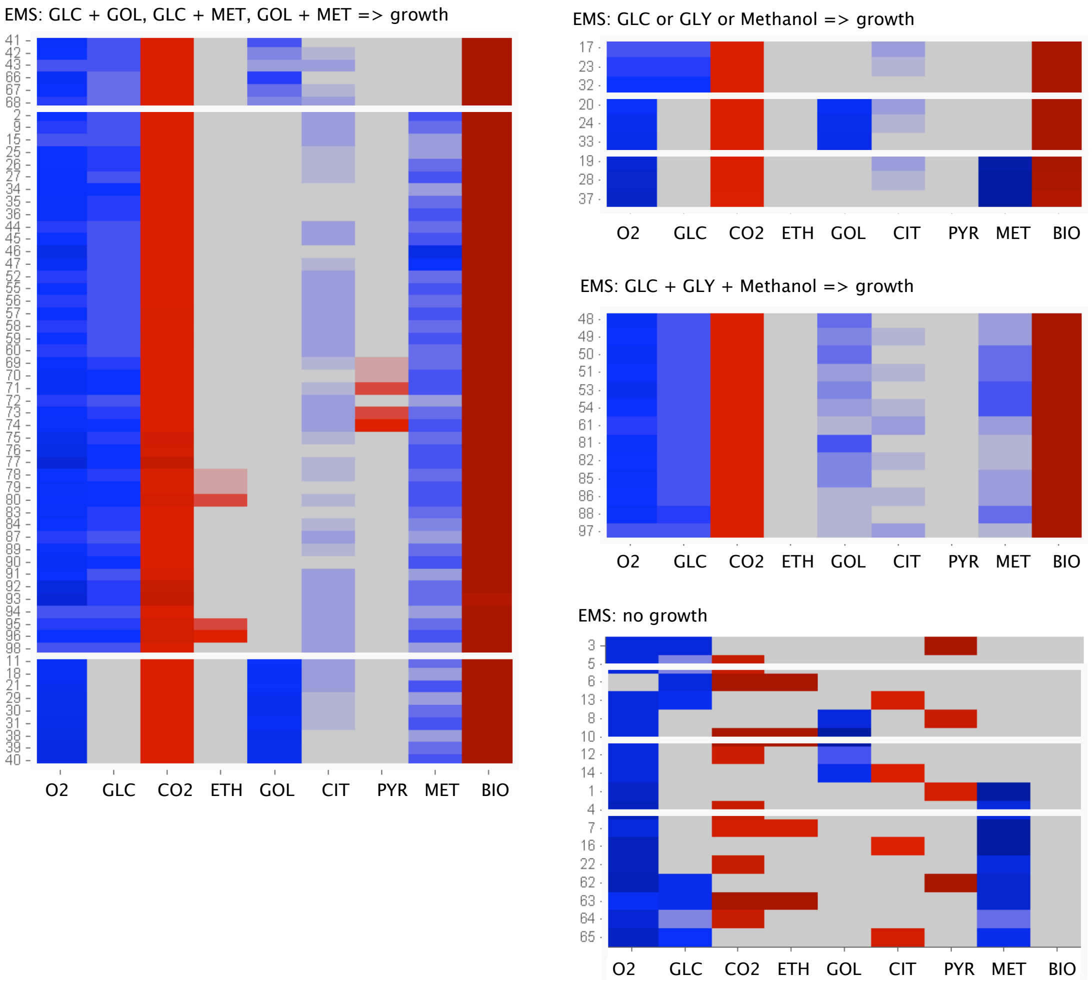
As explained in chapters II and III, elementary modes analysis provides a way to systematically identify a set of relevant pathways of a metabolic network (Schuster, 1999). The elementary modes (EMs) are the simplest (steady-state) flux vector that cells can show; whereas the remaining feasible states can be seen as its aggregated action (without cancelations of reversible fluxes). Moreover, the fact that they comprise all the simple pathways in the network—the functional states or non-decomposable vectors—makes it possible to investigate the infinite behaviours that cells can show by simply inspecting them. They have been used, for instance, to identify pathways with optimal yields (Schuster, 2002), determine minimal medium requirements (Schilling, 2000), and infer viability of mutants (Stelling, 2002).

The 98 elementary modes for the model described in the previous section were obtained using Metatool (Pfeiffer, 1999). The set of EMs can be classified as shown in Figure 9.2 depending first on its ability to produce biomass, and second on the carbon source used: glucose, methanol or glycerol. There are 17 EMs that do not result in biomass production, whereas 9 generate ethanol. No ethanol is produced in single substrate EMs when growing.

The carbon yields for biomass obtained for each EM are shown in Table 9.5. The maximum yield is 4.93 Cmol·dcw/Cmol, and is achieved with glucose as solely substrate. Glucose is the most efficient substrate for growth also in combination with glycerol or methanol. Methanol is the worst biomass yielding substrate. The distribution of the EMs according to their biomass yield is illustrated in Figure 9.3.

Table 9.5. Maximal biomass yields (Cmol·DW/mol)

Glu	Glyc	Met	YTotal	EM
x			4.93	32
	x		2.46	33
		x	0.82	37
x	x		3.68	41
	x	x	2.25	38
x		x	3.98	34
x	x	x	3.47	85



EMS: GLC + GOL, GLC + MET, GOL + MET => growth

EMS: GLC or GLY or Methanol => growth

EMS: GLC + GLY + Methanol => growth

EMS: no growth

Figure 9.2. Macroscopic equivalents of the elementary modes. Blue denotes substances being consumed by the EM, and red those being produced (the darker, the higher stoichiometric coefficient).

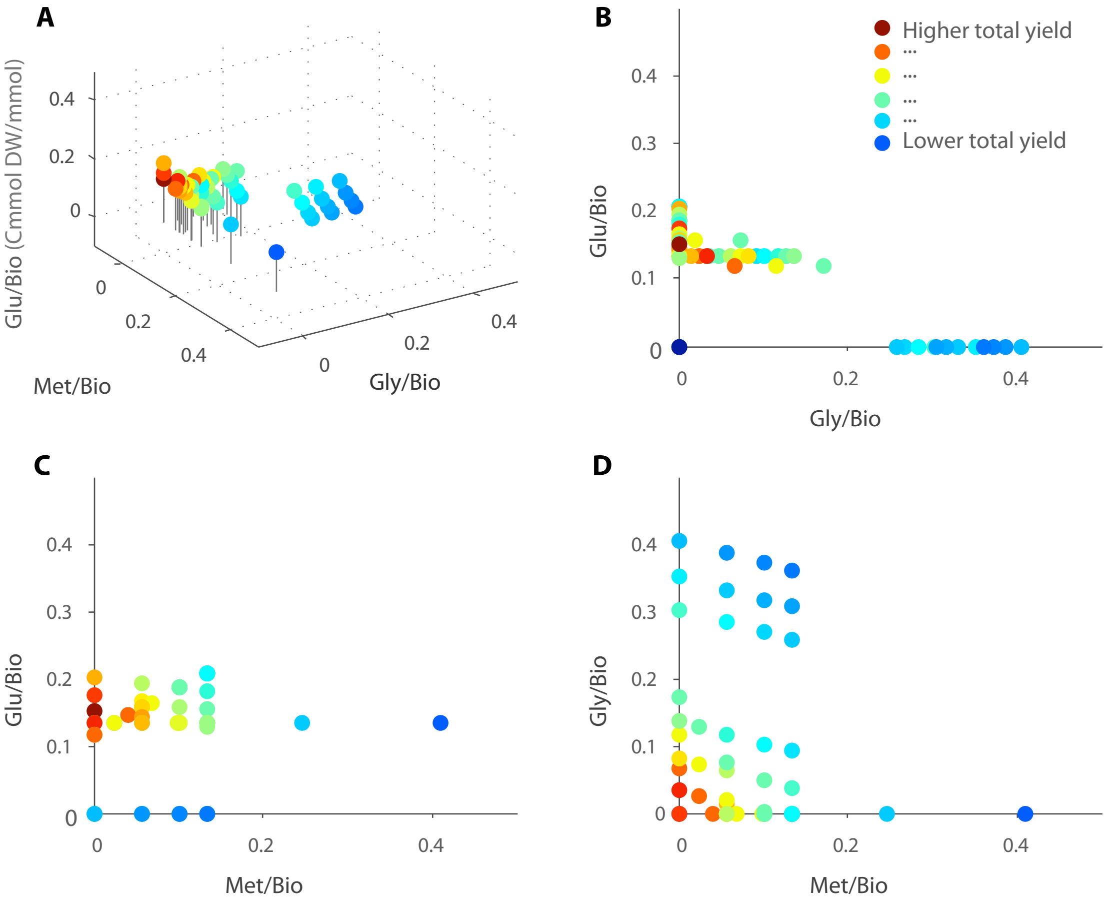


Figure 9.3. Biomass yields for each elementary mode of the network of *P. pastoris*.

9.5 Validating the model against experimental data

In this section a total of 11 different datasets compiled from the literature (tables 6 and 7) are used to determine whether the simplified model described above is coherent with the available experimental data.

Validation: experimental versus theoretical yields

As a first validation, we checked that the experimental growth yields do not exceed the maximum theoretical ones given by the model (which have been obtained by inspection of the elementary modes). For instance, the theoretical yield for growth on glucose is 4.93, whereas the experimental one is 3.98 (Cmmol·DW/mmol). The maximum yield on glycerol and methanol is 2.25, and the experimental ones—at different ratios of glycerol and methanol—range between 1.31 and 0.63. It also seems that the experimental yields decrease for combinations of substrates with lower theoretical yields.

Thus, no experimental yield violates the maximum theoretical ones (the contrary would indicate errors in the model because theoretical yields were obtained from it). However, the experimental yields tend to be lower than theoretical ones. There are multiple reasons for this deviation: (a) the model does not consider restrictions on energy cofactors, such as ATP, nor the resources devoted to recombinant protein production, (b) the EM analysis do not takes into account the ratio between the different substrates in mixed cases, and (c) even if they are feasible, cells does not necessarily make use of the pathways optimal for growth (Schuetz, 2007).

Table 9.6. Validation of the model against experimental data (yields).

Ref*	μ	Q_{Glu}	Q_{Gly}	Q_{Met}	Q_{et}	**OUR**	**CPR**	Q_{P}	**Yields** Exp. / Theo.	
	Cmol/ (Kg·h)	Cmol/ (Kg·h)	"	"	"	"	"	mg/ (g·h)	Cmol·DW/ (mol)	"
D1	3.86	0.97	0.00	0.00	0.00	2.02	2.07	0.020	3.98	< 6.62
A1	1.88	0.00	1.09	0.00	0.00	2.16	1.56	0.000	1.73	< 2.46
A2	2.07	0.00	0.95	0.63	0.00	2.70	1.70	0.001	1.31	< 2.25
A3	1.72	0.00	0.74	1.48	0.00	3.90	2.10	0.014	0.77	< 2.25
A4	2.02	0.00	0.57	2.33	0.00	4.85	2.21	0.024	0.70	< 2.25
B1	6.17	0.00	2.75	0.00	0.00	3.62	2.35	0.000	2.24	< 2.46
B2	6.18	0.00	2.22	1.87	0.00	7.19	4.18	0.001	1.51	< 2.25
B3	6.24	0.00	2.23	2.73	0.00	7.20	3.60	0.012	1.26	< 2.25
C1	2.32	0.00	0.74	2.22	0.00	3.58	2.05	0.012	0.78	< 2.25
C2	2.32	0.00	0.37	3.33	0.00	4.44	2.55	0.021	0.63	< 2.25
C3	2.32	0.00	0.00	4.44	0.00	5.29	2.82	0.022	0.52	< 0.82

*All the datasets correspond to continuous fermentation in defined chemical media. Further detail can be found in D, (Dragosits, 2009); A, (Solà, 2007); B, (Solà, 2007); C, (Jungo, 2007). Citrate and Pyruvate are assumed not to be produced nor consumed, except for dataset D1 in which citrate is consumed at 0.007 Cmol/(Kg·h).

Validation: consistency between model and experimental data

The same datasets are now used to check that the experimental measurements, which reflect the metabolic state of cells, are feasible states according to the model. Two different analysis of consistency were performed: one based on minimized, variance-weighted sum of squared residuals (ϕ) and another one based on the possibility of the most possible flux state or vector (π). Both were described in the methods section. The possibilistic approach is preferred in this case because the analysis of least squares residuals has limitations due to the presence of inequalities in the model.

In all weighted least squares problems, a standard deviation of 10% is assigned to each measurement of the set trying to capture their uncertainty. The variance-covariance matrix **F** in (4) is defined accordingly.

In Possibilistic MFA problems, the uncertainty of the measurements was represented as follows:

- Full possibility (π=1) is assigned to values near the measured ones, less than ±5% deviation, to account for random errors.
- A decreasing possibility is assigned to larger deviations so that values with a deviation equal to ±20% have a possibility of π=0.1 (those values with a deviation of ±9.5% will have possibility of π=0.5).[1]

This representation is achieved choosing the necessary bounds (ε_2^{max}, μ_2^{max}) and weights (α, β) for each measurement $\mathbf{w_m}$. Due to (a), the bounds are simply defined as $\varepsilon_2^{max}=\mu_2^{max}=0.05\cdot\mathbf{w_m}$. Then we operate with equations (5-7) to achieve (b). From (5) we have that, $0.2\cdot\mathbf{w_m}=\varepsilon_1^{20\%}+\varepsilon_2^{max}$, and from (6) and (7), $\log(0.1)=-\alpha\cdot\varepsilon_1^{20\%}$. As a result we get that, $\alpha=-\log(0.1)/(0.2\text{-}0.05)/\mathbf{w_m}$. Since uncertainty is symmetric, $\beta=\alpha$.

The results for each dataset are shown in Table 9.7, where the minimised, variance-weighted sum of squared residuals (ϕ) and the possibility of the most possible flux state or vector $\pi(\mathbf{v_{mp}})$ are given. The last column contains another useful indicator of consistency: the degree of measurements uncertainty needed to find a flux vector in full agreement with the model constraints (i.e., with π=1). All the computations were performed with MATLAB (MathWorks Inc., 2003), and YALMIP toolbox (Lofberg, 2004) was used to perform Possibilistic MFA.

The consistency between model and experimental measurements is very high, except for a pair of datasets. In these cases, the inconsistency pinpoints especial characteristics of these sets of data, as explained below.

The dataset D1, which corresponds to *Pichia* growing on glucose, shows very good agreement. The measured data has full possibility (π=1), meaning that there is a flux vector compatible with model and measurements; a band of 1% around the measured values is encloses this flux vector. The residual is also very low.

Datasets A1 and A2, which correspond to cultures growing totally or mainly on glycerol and producing a small amount of protein, also show a good agreement. The discrepancy between measurements and model is bigger for A3 (π=0.25), but still a band of 10% of deviation around measurements encloses a flux vector compatible with the model. Dataset A3 corresponds to a culture growing mainly on methanol, but supplemented on glycerol, and producing larger amounts of protein. The discrepancy is larger for A4, which corresponds to a scenario with high protein productivity.

[1] Notice that possibility has been defined by conjunction (see methods), so that if two measurements are deviated, for instance with possibilities 0.8 and 0.5 respectively, their joint possibility will be 0.4. Hence, a maximum possibility of 0.36 implies that there is an error between 10% and 20% in one measurement, or maybe an error between 5% and 10% in two measurements.

Similar results are obtained with cultures at a higher growth rate: B1 is highly consistent, while protein producing B2 and B3 show similar behaviour to A3-A4. This reveals the existence of non-modelled phenomena, probably related with protein production. The agreement is quite good for the three datasets C1-C3, but the increase of the discrepancy along with higher protein expression is also noticeable.

Finally, we used two batteries of random datasets to assess whether the model is indeed able to reject flux vectors that do not correspond to actual states of *P. pastoris* cultures. These datasets were defined taking random combinations of values for each flux within predefined bounds (see Table 9.7). Most of these random scenarios were highly inconsistent with the model (possibilities lower than 0.1 in 99% and 95% of the datasets, for each battery).

In summary, the constraint-based model shows acceptable agreement with the experimental data reported by different groups for *P. pastoris* cultures, and at the same time, rejects artificially generated invalid datasets. The scenarios with lower agreement pinpoint non-modelled phenomena, possibly related to protein expression.

Table 9.7. Validation of the model against experimental data (consistency).

Ref*	μ	Q_{Glu}	Q_{Gly}	Q_{Met}	Q_{et}	OUR	CPR	Q_P	Consistency**		
	Cmmol/ (g·h)	mmol/ (g·h)	"	"	"	"	"	mg/ (g·h)	ϕ	π	To π=1
D1	3.86	0.97	0.00	0.00	0.00	2.02	2.07	0.020	0.03	1.00	2%
A1	1.88	0.00	1.09	0.00	0.00	2.16	1.56	0.000	0.28	1.00	7%
A2	2.07	0.00	0.95	0.63	0.00	2.70	1.70	0.001	1.20	0.73	12%
A3	1.72	0.00	0.74	1.48	0.00	3.90	2.10	0.014	2.81	0.25	20%
A4	2.02	0.00	0.57	2.33	0.00	4.85	2.21	0.024	5.36	0.09	29%
B1	6.17	0.00	2.75	0.00	0.00	3.62	2.35	0.000	0.07	1.00	4%
B2	6.18	0.00	277	1.87	0.00	7.19	4.18	0.001	0.88	0.82	12%
B3	6.24	0.00	2.23	2.73	0.00	7.20	3.60	0.012	2.34	0.32	19%
C1	2.32	0.00	0.74	2.22	0.00	3.58	2.05	0.012	0.06	1.00	3%
C2	2.32	0.00	0.37	3.33	0.00	4.44	2.55	0.021	0.79	1.00	10%
C3	2.32	0.00	0.00	4.44	0.00	5.29	2.82	0.022	1.63	0.49	15%
Random	0-10	0-10	0-10	0-10	0-10	0-10	0-10	-	>10 99%	<0.1 99%	-
Random	1.5-6	0-2	0-2.7	0-2.7	0-0.1	2.1-7.2	1.5-4	-	>10 86%	<0.1 95%	-

*All the datasets correspond to continuous fermentation in defined chemical media. Further detail can be found in D, (Dragosits, 2009); A, (Solà, 2007); B, (Solà, 2007); C, (Jungo, 2007).

Citrate and Pyruvate are assumed not to be produced nor consumed, except for dataset D1 in which citrate is consumed at 0.007 Cmol/(Kg·h).

**Abbreviations refer to: minimized sum of squared residuals (ϕ), possibility of the most possible flux vector (π) and degree of measurements uncertainty to π=1.

9.6 Using the model to predict growth

Possibilistic MFA can now be applied to estimate the biomass growth rate for each of the previous datasets. Details of this estimation can be found in the methods section. Basically, Possibilistic MFA is applied to the datasets shown above excluding the measured value for the growth rate (which will be used to validate the estimates).

The estimated growth rate is found to be in very good agreement with the measured one for the vast majority of the analysed scenarios (D1, A1, A3, A4, B1, B2, B3, C1 and C2), which correspond to cultures at different growth rates, using different substrates, and coming from three independent literature references. For two other scenarios (A2 and C3), the most possible estimate is still accurate.

The fact that, although limited, the model has predictive capacity provides further validation for the constraint-based representation. This conclusion is strengthened if we consider that the growth rate is highly interconnected along the whole network, since the biomass equation takes into account several metabolic precursors (Table 9.3), and thus accurate correspondence between substrate uptake, respiratory fluxes and growth cannot be inferred from the network in a straightforward way.

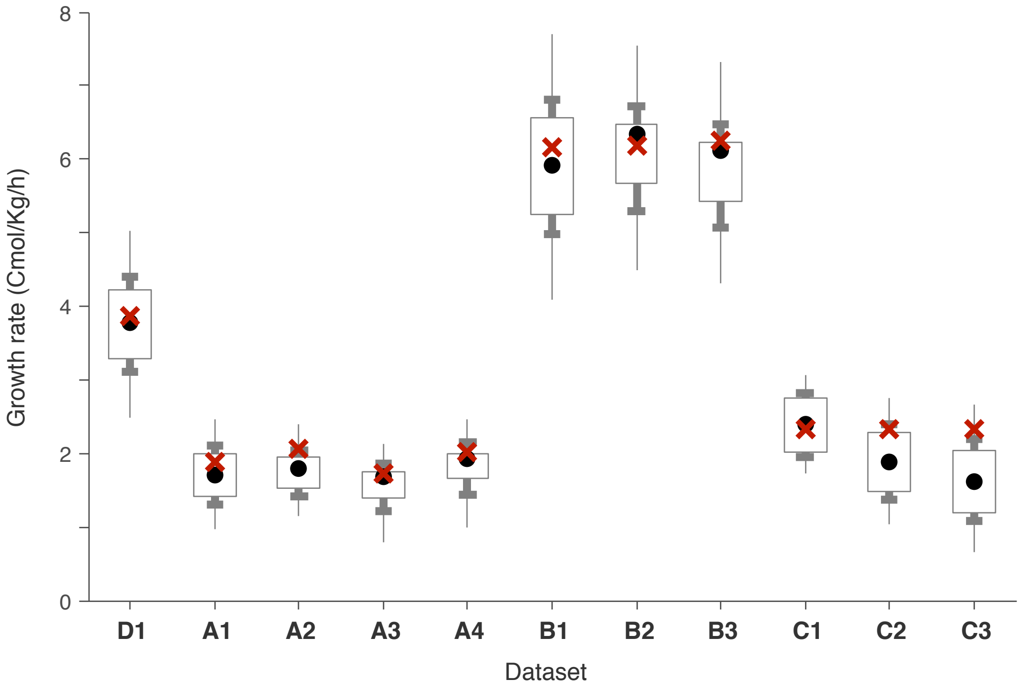


Figure 9.4. Prediction of growth rate for *P. pastoris* cultures using Possibilistic MFA. Crosses denote the measured values and circles most possible estimates. The intervals of possibilities of 0.8 (box), 0.5 (bar) and 0.1 (lines) are also depicted.

9.7 Using the model to estimate every flux

Once a validated model is available, possibilistic MFA could be used to estimate all the fluxes, intracellular or extracellular, as it has been done with the growth rate in the previous section (and as it was deeply discussed in chapter VII). For illustration purpose, the whole distribution of fluxes for the scenario A2 is depicted in Figure 9.5.

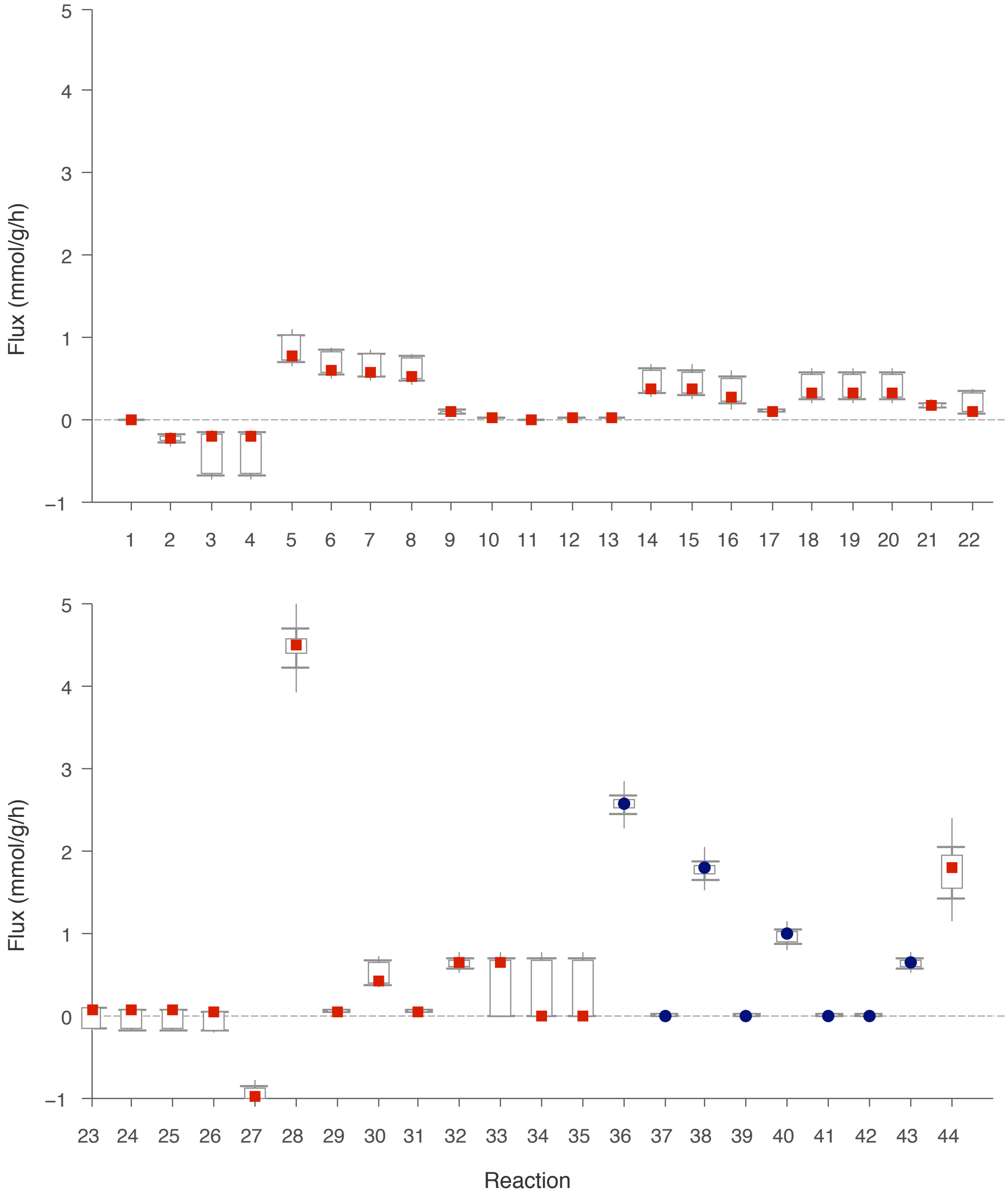


Figure 9.5. Possibilistic MFA estimates for every flux in the scenario A2. Most possible values (circles and squares for measured and non measured fluxes, respectively) and intervals of conditional possibilities 0.8, 0.5 and 0.1 are depicted for each flux.

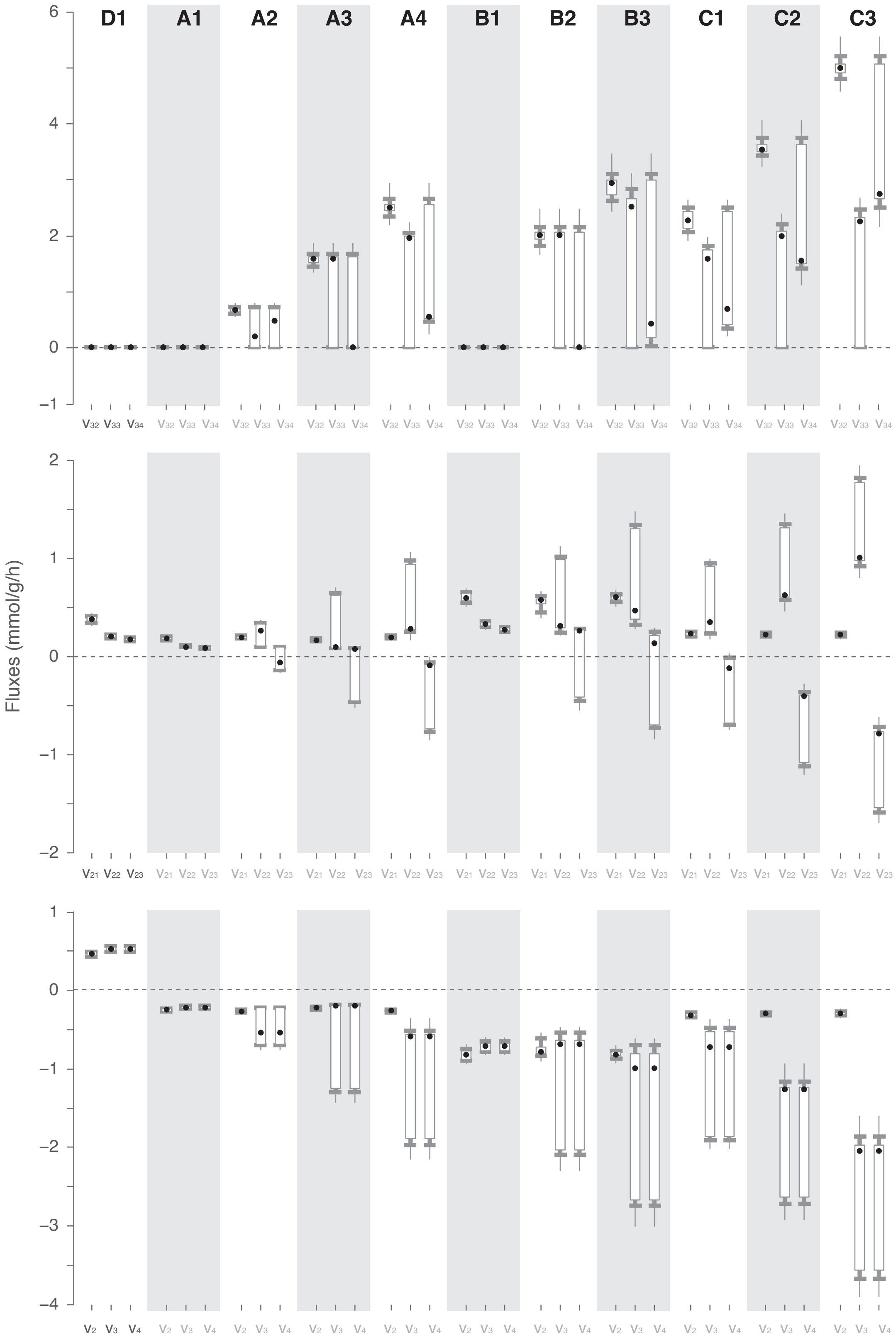


Figure 9.6. Estimations for a set of relevant fluxes in each scenario. Most possible values (circles and squares for measured and non measured fluxes, respectively) and intervals of conditional possibilities 0.8, 0.5 and 0.1 are depicted for each flux.

Notice that these estimations could not be done with traditional MFA because the measurements would be insufficient to get a determined system. The network has 8 degrees of freedom (44 fluxes and 36 linear equations) and there are 9 measured fluxes. However, these measurements introduce only 7 independent linear constraints, so the system remains underdetermined with 1 degrees of freedom. Possibilistic MFA can be used because considers reactions irreversibility and gives interval estimates (or even distributions) if there are multiple reasonably possible flux values.

It is also possible to estimate fluxes of particular interest to compare the different scenarios. For instance, the estimates for three relevant groups of fluxes, which represent splitting nodes within the network, are depicted in Figure 9.6.

- Fluxes v_2, v_3 and v_4 belong to glycolysis pathway, are positive as expected in cultures grown in glucose, and appear inverted in glycerol and/or methanol fed cultures.
- Fluxes v_{21}, v_{22} and v_{23} represent the isomerization of R5P into Ru5P and Xu5P. Note how v_{23} inverts its direction at growing methanol fluxes, as increased methanol consumption demands higher amounts of Xu5P thus requiring more R5P precursor.
- Fluxes v_{32}, v_{33} and v_{34} represent the branch-point related to methanol usage, that is, how this flux is split between direct oxidation and catabolic pathways. High methanol fluxes are necessarily conducted through CO_2 generation: see how flux v_{34} becomes distinct from zero in A4, B4, C2 and C3 scenarios.

These results, even if may require to be tested experimentally, can lead intervention within cells or optimization through manipulation of extracellular variables.

9.8 Conclusions

This chapter has described the application of the possibilistic framework (introduced in chapter VII) to validate a constraint-based model of *Pichia pastoris* in a real scenario of data scarcity where only a few extracellular measurements are available.

The model of *Pichia pastoris* has shown a reasonably agreement with the measurements in several scenarios, and at the same time, is able to rejects artificial, invalid datasets. Besides, it has been verified that the model has predictive capacity for cell growth rate, an attractive target for industrial fermentation monitoring and control. Interestingly, the accuracy of predictions worsens for higher protein producing scenarios, showing how the model, derived for a wild-type strain, is increasingly less precise as wider resources are devoted to recombinant protein generation.

It must be highlighted that the model has been strictly constructed upon first-principles and sensible hypothesis. At this point, the model can be curated, extended, and its parameters tuned to improve the consistency with the investigated scenarios. Particularly, energy requirements, strongly related to protein expression, are not yet considered by the model. Possibilistic MFA becomes a useful tool to systematise this procedure of model improvement.

Under a general perspective, the work described in this chapter shows how a small-sized network can be assessed following a rational, quantitative procedure even when measurements are scarce. This approach enables validation considering the stoichiometric balances and also reactions reversibilities, and accounting for measurements imprecision. The use of Possibilistic MFA also makes it possible to predict non-measured fluxes without removing the network underdeterminancy. There is, however, a challenge when validating networks with higher number of degrees of freedom because there may be many flux vectors compatible with the (few) available measurements. It is expected that the datasets will be highly consistent, so the approach in this case would be to check if the model rejects the artificially generated invalid datasets.

This chapter also illustrates the potential of the possibilistic estimates in scenarios lacking data. For instance, when a validated model is available—ideally incorporating measurements for some intracellular fluxes—the kind of comparative analysis described in the last section can provide insight on how the internal state of the cells determines its external behavior. This knowledge can potentially lead intervention within cells, suggesting target metabolites or biochemical branch-points, and optimize through manipulation of extracellular variables, such as feeding strategies and substrate selection.

Main references

- Tortajada M, Llaneras F, Picó J (2010). Validation of a constraint-based model of *Pichia pastoris* metabolism under data scarcity. *BMC Systems Biology*, 4:115.

- Llaneras F, Sala A, Picó J (2009). A possibilistic framework for metabolic flux analysis. *BMC Systems Biology*, 3:73.

- Llaneras F, Picó J (2008). Stoichiometric modelling of the cell metabolism. *Journal of Bioscience and Bioengineering*, 1, 1-12.

- Palsson BO (2006). *Systems biology: properties of reconstructed networks*. New York, USA: Cambridge University Press New York.

- Stephanopoulos GN, Aristidou AA (1998). *Metabolic Engineering: Principles and Methodologies*. San Diego, USA: Academic Press.

- Dragosits M, Stadlmann J, Albiol J, Baumann K, Maurer M, Gasser B, Sauer M, Altmann F, Ferrer P and Mattanovich D (2009). The effect of temperature on the proteome of recombinant *Pichia pastoris*. *Journal of Proteome Research*, 8(3):1380–92.

- Solà A, Jouhten P, Maaheimo H, Sánchez-Ferrando F, Szyperski T, Ferrer P (2007). Metabolic flux profiling of *Pichia pastoris* grown on glycerol/methanol mixtures in chemostat cultures at low and high dilution rates. *Microbiology*, 153(1):281–90.

- Jungo C, Marison I, Stockar U (2007). Mixed feeds of glycerol and methanol can improve the performance of *Pichia pastoris* cultures: A quantitative study based on concentration gradients in transient continuous cultures. *Journal of Biotechnology*, 128(4):824–37.

"Maturity of mind is the capacity to endure uncertainty"

John Finley

Conclusions

This thesis addressed problems related to constraint-based metabolic models. The objective was to find simple ways to handle the difficulties that arise in practice due to uncertainty: models of organisms of interest are incomplete, there is a lack of measurable variables, those available are imprecise, etc. With this purpose in mind, we have developed tools to analyse, estimate and predict the metabolic behaviour of cells.

The contributions of this work were listed in the introduction, and particular conclusions can be found in each chapter. Here, some general conclusions are discussed together with lines for future work.

- **The application of constraint-based models show that much valuable information can be extracted from them even if intracellular kinetics are unknown.** Constraint-based models are being employed to analyse the modelled organisms (e.g., identify optimal pathways), to simulate genetic modifications (e.g., gene deletions), to estimate which reactions are active at certain conditions, and to predict cells behaviour. Moreover, new and better knowledge will improve the models in an iterative way since they are easily extensible. Indeed, we expect that the increasing availability of biological data will fuel the use of mathematical models in biology.
- **Interval and possibilistic MFA-wise methods provide better estimates of the metabolic state of cells (chapters IV and VII).** The estimation of the metabolic fluxes provides insight on the internal state of cells,

which determine the behaviour exhibit at given environmental conditions. This knowledge can potentially lead to intervention within cells, suggesting target metabolites or biochemical bifurcation, and process optimisation through manipulating external variables, such as feeding strategies or substrate selection. The interval approach (FS-MFA) is a simple extension of traditional MFA that considers inequality constraints and measurements uncertainty, and can be applied even if measurements are scarce or imprecise. The possibilistic methodology (Poss-MFA) is slightly more complex, but also more powerful. It has a distinctive advantage over other approaches which either rely on stronger assumptions (chi-squared distributions, absence of irreversibility), or are only data-based (so they do not incorporate a model), or provide only point-wise estimates (instead of the richer possibility distributions and intervals), or are computationally intensive (e.g., multi-variate integration in a general Bayesian estimation problem). For these reasons, FS-MFA and Poss-MFA are a better alternative than traditional MFA in many current applications. An interesting extension of Poss-MFA would be to incorporate other constraints or measurements from stable isotope tracer experiments. This would be straightforward if the constraints are linear equalities or inequalities, but this is often not the case (e.g., thermodynamic constraints include integer variables). Although the possibilistic framework could be still of use, most likely computational efficiency will be lost.

- **The combination of a constraint-based model with measurements enables monitoring the intracellular state of cells during a running process (chapter VI and VIII).** This information is of great use for fault-detection and manual or automatic control of industrial processes. Although similar approaches have been described in the literature before, real applications remain difficult due to the scarcity of reliable online sensors. Interestingly, the methods proposed in this thesis mitigate this problem (FS-MFA and Poss-MFA). Yet, more variables should be measurable online to boost these model-based monitoring systems. Meanwhile monitoring could be applied quasi-online using (fast) measurements even if those require manual intervention. Current work is also being done to generalise the possibilistic monitoring as model-based observers suitable in other fields.

- **The major challenge regarding MFA-wise methods in large networks is the lack of information;** many metabolic flux states are often compatible with the (known) constraints and the (few) available measurements. Conversely to traditional methods, those proposed here are still of use in this situation. Poss-MFA detects all the equally possible flux states (or “similarly” possible) capturing them by means of possibilistic distributions or intervals. If there is a wide range of candidates, however, the estimation may be little informative. If this is the case, one could decide to incorporate a rational assumption, as it is done by FBA.

- **A possibilistic approach to FBA allows to account for alternate optima and sub-optimality (Chapter VIII).** FBA predicts the state of cells at given conditions based on the assumption that cells evolved to be optimal in some sense. Defining possibility for optimality, Poss-FBA gives predictions that capture alternate optima (cell states with equal "performance") and grades sub-optimality, somehow relaxing the original assumption. So far, this approach has been used to predict the fluxes and metabolite concentrations during a cultivation process. However, the same ideas should be used to analyse flux spaces as it has been done with Monte Carlo sampling methods.[1] Other relevant issues for FBA could be investigated under the possibilistic perspective, such as non-linear or multi-objective functions (both to better represent the strategies that cells acquired through evolution). Some suggestive questions arose in this respect when Poss-FBA was applied considering extracellular dynamics: can be assumed that cells behave optimally at each instant or should a temporal horizon be considered? Are cells optimal in rare environments (e.g., lack of competitors) or they anticipate that the environment is likely to change? Although these are speculative questions, constraint-based models and FBA-wise methods may be of help for those interested in answering them.

- **A constraint-based model can be validated even if experimental data is scarce (chapter IX).** Many medium-sized metabolic models are not properly validated, ignoring that they are simplifications of the whole metabolism and rely on reductionist hypothesis. For instance, some models are only evaluated against one set of data, which is thus inconveniently used both to validate the model and perform the analysis. Trying to face this problem, this thesis proposes a simple procedure to validate models against data from different cultures that can be of use if data is scarce. First, elementary modes are used to check that the experimental growth yields do not exceed the maximum theoretical ones given by the model. Then, Poss-MFA is used to check if the model shows acceptable agreement with the experimental data, and at the same time rejects artificially generated invalid data. This way, the data available is exploited to build more reliable reduced models. The procedure may be extended to detect limitations of a model and guide its improvement. This procedure has been applied to validate a model of *P. pastoris*, a yeast used in industry for the expression of recombinant proteins.

- **Possibility theory and a constraint-based model can be used to detect errors in a set of experimental measurements (chapter VII).** The approach is similar to the χ^2-tests used in traditional MFA, but more flexible: it is not necessary to assume that errors are normally distributed and inequality constraints can be considered besides equalities (e.g., irreversibility). Notice that

[1] The approach would have a slightly different interpretation (the frequency of a flux value within the space is not relevant to rank its possibility) and it could be more efficient computationally.

this approach can be seen as the inverse of the validation procedure mentioned above ("check a model against reliable measurements" versus "check measurements against a reliable model"). Future work may apply these ideas to find errors in other measurements, such as metabolite concentrations.

- **Elementary modes have advantages over other similar network-based pathways (Chapter III).** Although the minimal generating set will be preferred in some applications due to its reduced size and because their computation is more efficient, the elementary modes allow to answer several questions by simply inspecting them (such as which reactions are essential to produce a compound, or which would be the effect of a reaction knockout). There is, however, a major limitation of all these approaches regarding large models: the number of pathways dramatically increases, reducing understandability and becoming not computable. Recent works in literature face this problem looking for better ways to compute the elementary modes and proposing other pathways, smaller in number, but holding some of their properties.

The work described in this thesis shows the importance of accounting for uncertainty when modelling living cells. We have seen that constraint-based models provide a way to handle uncertainty: maybe we cannot exactly model how cells operate,[1] but the available knowledge allow us to distinguish what is possible (*as far as we know*) from what is not. Following this idea, we have developed interval and possibilistic methods to analyse, estimate and predict the metabolic behaviour of cells. These methods start by representing our knowledge accounting for its uncertainty, and then exploit this knowledge to generate reliable new information.

Uncertainty is still present in biological systems, it cannot be neglected, and it really makes things more difficult. But it can be handled. This way imperfect mathematical models of living cells can be used with success.

[1] Some people would say that this just reflects a lack of understanding: if you cannot model a phenomena, you do not understand it completely. Richard Feynman stated that, "What I cannot create I do not understand." and we would rephrase him to say: "What I cannot *recreate* I do not understand."

References

16. Bailey JE (2001). Complex biology with no parameters. *Nature Biotechnology*, 19:503–504.

17. Bailey JE (1998). Mathematical modeling and analysis in biochemical engineering: past accomplishments and future opportunities. *Biotechnology Progress*, 14:8–20.

18. Banga JR, Balsa-Canto E, Moles CG, Alonso AA (2005). Dynamic optimisation of bioprocesses: effcient and robust numerical strategies. *Journal of Biotechnology*, 117:407–419.

19. Banga JR (2008). Optimization in computational systems biology. *BMC systems biology*, 2(1):47.

20. Banga JR, Alonso AA, Singh RP (2008b). Stochastic dynamic optimization of batch and semicontinuous bioprocesses. *Biotechnology Progress*, 13(3):326–335.

21. Barrett CL, Herrgard MJ, Palsson BO (2009). Decomposing complex reaction networks using random sampling, principal component analysis, and basis rotation. *BMC Systems Biology*, 3(1):30.

22. Bastin G, Dochain D (1990). *On-line Estimation and Adaptative Control of Bioreactors*. Amsterdam, Netherlands: Elsevier.

23. Bastin G (2007). Quantitative analysis of metabolic networks and design of minimal bioreaction models. *International Conference in Honor of Claude Lobry*.

24. Battista H, Picó J, Garelli F, Vignoni A (2010). Specific Growth Rate Estimation in Bioreactors Using Second-Order Sliding Observers. *Computer Applications in Biotechnology*.

25. Beard DA, Liang S, Qian H (2002). Energy balance for analysis of complex metabolic networks. *Biophysics Journal*, 83(1):79–86.

26. Bell SL, Palsson B (2005). Expa: a program for calculating extreme pathways in biochemical reaction networks. *Bioinformatics*, 21(8)1739–40.

27. Benferhat S, Dubois D, Prade H (1997). Syntactic Combination of Uncertain Information: A Possibilistic Approach. *Lecture notes in computer science*, 30–42.

28. Bonarius H, Schmid G, Tramper J (1997). Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends in Biotechnology*, 15(8):308–314.

29. Bonarius H, Hatzimanikatis V, Meesters K, de Gooijer CD, Schmid G, Tramper J (1996). Metabolic flux analysis of hybridoma cells in different culture media using mass balances. *Biotechnology & Bioengineering*, 50:299–318.

30. Braunstein A, Mulet R, Pagnani A (2008). The space of feasible solutions in metabolic networks. *Physics Journal*, 95:012–017.

31. Camacho J and Picó J (2007). Self-tuning run to run optimisation of fed-batch processes using unfold-PLS. *AIChE Journal*, 53(7):1789–1804.

32. Cakir T, Kirdar B, Ulgen KO (2004). Metabolic pathway analysis of yeast strengthens the bridge between transcriptomics and metabolic networks. *Biotechnology & Bioengineering*, 86:251–260.

33. Calik P, Ozdamar TH (2002). Metabolic flux analysis for human therapeutic protein productions and hypothesis for new therapeutical strategies in medicine. *Biochemical Engineering Journal*, 11:49–68.

34. Carlson R, Fell D, Srienc F (2002). Metabolic pathway analysis of a recombinant yeast for rational strain development. *Biotechnology & Bioengineering*, 79(2):121–34.

35. Casti JL (1992). *Reality Rules: Picturing the World in Mathematics*. New York, Wiley.

36. Chassagnole C, Noisommit-Rizzi N, Schmid JW, Mauch K, Reuss M (2002). Dynamic Modelling of the central carbon metabolism of *Escherichia coli*. *Biotechnology & Bioengineering*, 79:53–73.

37. Chernikova NV (1965). Algorithm for finding a general formula for the non-negative solutions of a system of linear inequalities. *USSR Computational Mathematics and Mathematical Physics*, 5(2):228–233.

38. Clarke BL (1988). Stoichiometric network analysis. *Cell Biophysics*, 12:237–253.

39. Cornish-Bowden A, Cardenas ML (2000). From genome to cellular phenotype-a role for metabolic flux analysis? *Nature biotechnology*, 18:267–268.

40. Covert MW, Palsson BO (2003). Constraints-based models: regulation of gene expression reduces the steady-state solution space. *Journal of Theoretical Biology*, 221(3):309–325.

41. Covert MW, Schilling CH, Palsson B (2001). Regulation of gene expression in flux balance models of metabolism. *Journal of Theoretical Biology*, 213:73–88.

42. Dochain D, Pauss A (1988). On-line Estimation of Microbial Specific Growth-Rates: An Illustrative Case Study. *The Canadian Journal of Chemical Engineering*, 66:626.

43. Dragosits M, Stadlmann J, Albiol J, Baumann K, Maurer M, Gasser B, Sauer M, Altmann F, Ferrer P and Mattanovich D (2009). The effect of temperature on the proteome of recombinant *Pichia pastoris*. *Journal of Proteome Research*, 8(3):1380–92.

44. Dubois D and Prade H (1995). Fuzzy relation equations and causal reasoning. *Fuzzy Sets and Systems*, 45(2):119–134.

45. Dubois D and Prade H (2005). Interval-valued fuzzy sets, possibility theory and imprecise probability. *Proceedings of International Conference in fuzzy Logic and Technology*.

46. Dubois D and Prade H (1988). *Possibility theory: an approach to computerized processing of uncertainty*. New York, USA: Wiley.

47. Dubois D and Prade H (2001). Possibility theory, probability theory and multiple-valued logics: a clarification. *Annals of Mathematics and Artificial Intel ligence*, 32(1):35–66.

48. Dubois D, Fargier H, Prade H (1996). Possibility theory in constraint satisfaction problems: handling priority, preference and uncertainty. *Applied Inteligence*, 6(4):287–309.

49. Dunn IJ, Heinzle E, Ingham J, Prenosil E (2000). *Biological Reaction Engineering: Dynamic Modelling Fundamentals with Simulation Examples*. Wiley, Zürich.

50. Edwards JS, Ibarra RU, Palsson BO (2001). In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature biotechnology*, 19(2):125–30.

51. Edwards JS, Palsson, BO (1999). Systems properties of the *Haemophilus influenzae* Rd Metabolic genotype. *Journal of Biological Chemistry*, 274:17410–6.

52. Edwards JS, Palsson, BO (2000). The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences*, 97:5528–33.

53. Edwards JS, Covert M, Palsson B (2002). Metabolic modelling of microbes: the flux-balance approach. *Environmental Microbiology*, 4:133–140.

54. Edwards JS, Ibarra RU, Palsson BO (2001) In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature biotechnology*, 19:125–130.

55. Elbassioni K, Tiwary H (2009). Complexity of Approximating the Vertex Centroid of a Polyhedron. *Lecture Notes In Computer Science*, 5878:413–422.

56. Farza M, Busawon K, Hammouri H (1998). Simple nonlinear observers for online estimation of kinetic rates in bioreactors. *Automatica*, 34:301–318.

57. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO (2007). A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology*, 3:121.

58. Feist AM, Scholten JCM, Palsson BO, Brockman FJ, Ideker T (2006). Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. *Molecular Systems Biology*, 2:2006.004.

59. Figueiredo LF, Podhorski A, Rubio A, Kaleta C, Beasley JE, Schuster E, Planes FJ (2009). Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics*, 25(23):3158–65.

60. Follstad BD, Balcarcel RR, Stephanopoulos G, Wang DI (1999). Metabolic flux analysis of hybridoma continuous culture steady state multiplicity. *Biotechnology & Bioengineering*, 63:675–683.

61. Forster J, Gombert AK, Nielsen J (2002). A functional genomics approach using metabolomics and in silico pathway analysis. *Biotechnology & Bioengineering*, 79(7):703–712.

62. Forster J, Famili I, Fu P, Palsson BO, Nielsen J (2003). Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Researchearch*, 13:244–253.

63. Fukuda K, Prodon A (1996). Double description method revisited. *Combinatorics and computer science*, 1120:91–111.

64. Gagneur J, Klamt S (2004). Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics*, 5:175.

65. Galvanauskas V, Simutis R, Volk N, Lübbert A (1998). Model based design of a biochemical cultivation process. *Bioprocess and Biosystems Engineering*, 18:227–234.

66. Gambhir A, Korke R, Lee J, Fu PC, Europa A, Hu WS (2003). Analysis of cellular metabolism of hybridoma cells at distinct physiological states. *Journal of Bioscience and Bioengineering*, 95(4):317–327.

67. Gayen K, Venkatesh KV (2006). Analysis of optimal phenotypic space using elementary modes as applied to *Corynebacterium glutamicum*, *BMC Bioinformatics*, 7:445.

68. Gerdtzen ZP, Daoutidis P, Hu WS (2004). Non-linear reduction for kinetic models of metabolic reaction networks. *Metabolic Engineering*, 6:140–154.

69. Gombert AK, Nielsen J (2000). Mathematical modelling of metabolism. *Current Opinion in Biotechnology*, 11:180–186.

70. Guardia MJ, Gambhir A, Europa AF, Ramkrishna D, Hu WS (2000). Cybernetic Modelling and regulation of metabolic pathways in multiple steady states of hybridoma cells. *Biotechnology Progress*, 16:847–853.

71. Haag J, Wouwer A, Bogaerts P (2005). Systematic procedure for the reduction of complex biological reaction pathways and the generation of macroscopic equivalents. *Chemical Engineering Science*, 60:459–465.

72. Hand DJ. Statistical reasoning with imprecise probabilities. *Applied Statistics*, 42(1):237–238.

73. Heijden RT, Romein B, Heijnen JJ, Hellinga C, Luyben KC (1994). Linear Constraint Relations in Biochemical Reaction Systems: I. *Biotechnology & Bioengineering*, 43(1):3–10.

74. Heijden RT, Romein B, Heijnen JJ, Hellinga C, Luyben KC (1994). Linear Constraint Relations in Biochemical Reaction Systems: II. *Biotechnology & Bioengineering*, 43(1):11–20.

75. Heinrich R, Schuster S (1996). *The regulation of cellular systems*. New York, USA: Chapman & Hall.

76. Henry CS, Broadbelt LJ, Hatzimanikatis V (2006). Thermodynamics-based metabolic flux analysis. *Biophysics Journal*, 92(5):1792–805.

77. Henry O, Kamen A, Perrier M (2007). Monitoring the physiological state of mammalian cell perfusion processes by on-line estimation of intracellular fluxes. *Journal of Process Control*, 17:241–251.

78. Henson MA (2003). Dynamic Modelling of microbial cell populations. *Current Opinion in Biotechnology*, 14:460–467.

79. Herwig C, Marison I, von Stockar U (2001). On-line stoichiometry and identification of metabolic state under dynamic process conditions. *Biotechnology & Bioengineering*, 75:345–354.

80. Herwig C, von Stockar U (2002). A small metabolic flux model to identify transient metabolic regulations in *Saccharomyces cerevisiae*. *Bioprocess and Biosystems Engineering*, 24:395–403.

81. Hjersted JL, Henson MA (2009). Steady-state and dynamic flux balance analysis of ethanol production by *Saccharomyces cerevisiae*. *IET Systems Biology*, 3(3):167–79.

82. Hoppe A, Hoffmann S, Holzhütter HG (2007). Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. *BMC Systems Biology*, 1:23.

83. Ideker T, Galitski T, Hood L (2001). A new approach to decoding life: Systems Biology. *Annual Review of Genomics and Human Genetics*, 2:343–372.

84. Ishii N, Robert M, Nakayama Y, Kanai A, Tomita M (2004). Toward large-scale Modelling of the microbial cell for computer simulation. *Journal of Biotechnology*, 113:281–294.

85. Jensen FV (1996). Introduction to Bayesian networks. Secaucus, USA: Springer-Verlag New York.

86. Jungo C, Marison I, Stockar U (2007). Mixed feeds of glycerol and methanol can improve the performance of *Pichia pastoris* cultures: A quantitative study based on concentration gradients in transient continuous cultures. *Journal of Biotechnology*, 128(4):824–37.

87. Kadirkamanathan V, Yang J, Billings SA, Wright PC (2006). Markov chain monte carlo algorithm based metabolic flux distribution analysis on *Corynebacterium glutamicum*. *Bioinformatics*, 22(21):2681–2687.

88. Kaleta C, Figueiredo LF, Schuster E (2009). Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Research*, 19(10):1872–83.

89. Kannan R, Lovász L, Simonovits M. (1998). Random walks and an o(n5) volume algorithm for convex bodies. *Random Structures and Algorithms*, 11(1):1–50.

90. Kauffman, KJ, Prakash P, and Edwards JS (2003). Advances in flux balance analysis. *Current Opinion in Biotechnology*, 14:491–496.

91. Soh K, Hatzimanikatis V (2010). Network thermodynamics in the post-genomic era. *Current Opinion in Microbiology*, 13(3):350–357.

92. Kitano H (2002). Computational systems biology. *Nature*, 420:206–210.

93. Klamt S, Schuster S, Gilles ED (2002). Calculability analysis in underdetermined metabolic networks illustrated by a model of the central metabolism in purple nonsulfur bacteria. *Biotechnology & Bioengineering*, 77:734–751.

94. Klamt S, Stelling J, Ginkel M, Gilles ED (2003). FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps. *Bioinformatics*, 19(2):261–269.

95. Klamt S, Stelling J (2003). Two approaches for metabolic pathway analysis? *Trends in Biotechnology*, 21(2):64–69.

96. Klamt S, Gilles ED (2004). Minimal cut sets in biochemical reaction networks. *Bioinformatics*, 20(2):226–234.

97. Klamt S, Gagneur J, Kamp A (2005). Algorithmic approaches for computing elementary modes in large biochemical reaction networks. *BMC Systems Biology*, 152(4):249–55.

98. Klamt S, Saez-Rodriguez J, Gilles E (2007). Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Systems Biology*, 1(2).

99. Klipp E, Herwig R, Kowald A, Wierling C, Lehrach H. (2005). *Systems biology in practice: concepts, implementation and application*. Weinheim, Germany: Wiley-VCH.

100. Klir GJ, Parviz B (1992). Probability-Possibility Transformations: a Comparison. *International Journal of General Systems*, 21(3):291–310.

101. Komives C, Parker RS (2003). Bioreactor state estimation and control. *Current Opinion in Biotechnology*, 14:468–474.

102. Kompala DS, Ramkrishna D, Jansen NB, Tsao GT (1986). Investigation of bacterial growth on mixed substrates: experimental evaluation of cybernetic models. *Biotechnology & Bioengineering*, 28:1044–1055.

103. Kumar V (1992). Algorithms for constraint-satisfaction problems: A survey. *AI magazine*, 13(1):32–44.

104. Kümmel A, Panke S, Heinemann M (2006). Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics*, 7 :512.

105. Lange BM (2006). Integrative analysis of metabolic networks: from peaks to flux models? *Current Opinion in Plant Biology*, 9:220–226.

106. Larhlimi A, Bockmayr A (2009). A new constraint-based description of the steady-state flux cone of metabolic networks. *Discrete Applied Mathematics*, 157 (10):2257–2266.

107. Le Verge H (1992). A note on Chernikovas algorithm. Research Report 635.

108. Lee J, Lee SY, Park S, Middelberg APJ (1999). Control of fed-batch fermentations. *Biotechnology Advances*, 17:29–48.

109. Lei F, Jorgensen SB (2001). Estimation of kinetic parameters in a structured yeast model using regularisation. *Journal of Biotechnology*, 88:223-237.

110. Lei F, Rotbøll M, Jørgensen SB (2001). A biochemically structured model for Saccharomyces cere6isiae. *Journal of Biotechnology*, 88:205-221.

111. Levant A (1998). Robust exact differentiation via sliding mode technique. *Automatica*, 34:379-384.

112. Liao J, Hou S, Chao Y (1996). Pathway analysis, engineering, and physiological considerations for redirecting central metabolism. *Biotechnology & Bioengineering*, 52(1):129-140.

113. Llaneras F, Bastin G, Picó J (2007). On metabolic flux analysis when measurements are insufficient and/or uncertain. *IAP Dysco workshop.*

114. Llaneras F, Picó J (2006). The linkage between flux distributions and elementary modes activity patterns: An interval Approach. *International Symposium on Systems Biology*.

115. Llaneras F, Picó J (2007). A procedure for the estimation over time of metabolic fluxes in scenarios where measurements are uncertain and/or insufficient. *BMC Bioinformatics*, 8:42.

116. Llaneras F, Picó J (2007). An interval approach for dealing with flux distributions and elementary modes activity patterns. *Journal of Theoretical Biology*, 246(2):290-308.

117. Llaneras F, Picó J (2008). Stoichiometric modelling of the cell metabolism. *Journal of Bioscience and Bioengineering*, 1, 1-12.

118. Llaneras F, Picó J (2010). Which metabolic pathways generate and characterise the flux space? A comparison among elementary modes, extreme pathways and minimal generators. *J. Biomedicine and biotechnology*, 1:2010.

119. Llaneras F, Sala A, Picó J (2008). A possibilistic framework for metabolic flux analysis. *Reunión de la red Española de Biología de Sistemas.*

120. Llaneras F, Sala A, Picó J (2009). A possibilistic framework for metabolic flux analysis. *BMC Systems Biology*, 3:73.

121. Llaneras F, Sala A, Picó J (2009). Applications of possibilistic reasoning to intelligent system monitoring: a case study. *IEEE Multi-conference on Systems and Control.*

122. Llaneras F, Sala A, Picó J (2010). Dynamic flux balance analysis: a possibilistic approach. *Systems Biology of Microorganisms Conference*.

123. Llaneras F, Sala A, Picó J (2010). Possibilistic estimation of metabolic fluxes during a batch process accounting for extracellular dynamics. *Computer Applications in Biotechnology*.

124. Llaneras F, Tortajada M, Picó J (2007). Structural analysis of metabolic pathways applied to heterologous protein production in *P. pastoris*. *European Congress on Biotechnology*, *Journal of Biotechnology*, 131(2):S209.

125. Llaneras F. and Picó J. (2008) Stoichiometric modelling of cell metabolism J *Journal of Bioscience and Bioengineering*, 105 (1), 1–11.

126. Lofberg J (2004). YALMIP: A toolbox for modeling and optimization in MATLAB. *IEEE International Symposium on Computer Aided Control Systems Design*, 284-289.

127. Luenberger D (1971). An introduction to observers. *IEEE Transactions on Automatic Control*, 16:596–602.

128. Mahadevan R, Burgard A, Famili I, Van Dien S, Schilling C (2005). Applications of metabolic modeling to drive bioprocess development for the production of value-added chemicals. *Biotechnology and Bioprocess Engineeringineering*, 10:408.

129. Mahadevan R, Edwards JS, Doyle FJ (2002). Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophysics Journal*, 83:1331–1340.

130. Marx A, Graaf A, Wiechert W, Eggeling L, Sahm H (1996). Determination of the fluxes in the central metabolism of *Corynebacterium glutamicum* by nuclear magnetic resonance spectroscopy combined with metabolite balancing. *Biotechnology & Bioengineering* 49:111–129.

131. Mashego MR, Rumbold K, De Mey M, Vandamme E, Soetaert W, Heijnen JJ (2007). Microbial metabolomics: past, present and future methodologies. *Biotechnology Letters*, 29:1–16.

132. Mo ML, Palsson BO, Herrgard MJ (2009). Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Systems Biology*, 3(1):37.

133. Montagud A, Navarro E, de Córdoba PF, Urchueguía JF, Patil KR (2010). Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. *BMC Systems Biology*, 4:156.

134. Nielsen J, Villadsen J (1992). Modelling of microbial kinetics. *Chemical Engineering Science*, 47:4225–4270.

135. Nogales J, Palsson BO, Thiele I (2008). A genome-scale metabolic reconstruction of *Pseudomonas putida* KT2440: iJN746 as a cell factory. *BMC Systems Biology*, 2:79.

136. Nolan RP, Fenley AP, Lee K (2006). Identification of distributed metabolic objectives in the hypermetabolic liver by flux and energy balance analysis. *Metabolic Engineering*, 8:30–45.

137. Nomikos P, MacGregor JF (1995). Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics*, 37:41–59.

138. Nookaew I, Meechai A, Thammarongtham C, Laoteng K, Ruanglek V, et al. (2007). Identification of flux regulation coefficients from elementary flux modes: A systems biology tool for analysis of metabolic networks. *Biotechnology & Bioengineering*, 97(6):1535–49.

139. Nyberg GB, Balcarcel RR, Follstad BD, Stephanopoulos G, Wang DI (1999). Metabolism of peptide amino acids by Chinese hamster ovary cells grown in a complex medium. *Biotechnology & Bioengineering*, 62:324–335.

140. Palsson BO (2000). The challenges of in silico biology. *Nature biotechnology*, 18(11):1147–50.

141. Palsson BO (2006). *Systems biology: properties of reconstructed networks*. New York, USA: Cambridge University Press New York.

142. Papin JA, Price ND, Palsson BO (2002). Extreme pathway lengths and reaction participation in genome-scale metabolic networks. *Genome Research*, 12(12):1889–1900.

143. Papin JA, Price ND, Edwards JS, Palsson BO (2002). The genome-scale metabolic extreme pathway structure in *Haemophilus influenzae* shows significant network redundancy. *Journal of Theoretical Biology*, 215, 67–82.

144. Papin JA, Stelling J, Price ND, Klamt S, Schuster S, Palsson BO (2004). Comparison of network-based pathway analysis methods. *Trends in Biotechnology*, 22(8):400–405.

145. Pei SC, Shyu JJ (1989). Eigenfilter design of higher-order digital differentiators. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37:505–511.

146. Pfeiffer T, Sanchez-Valdenebro I, Nuno JC, Montero F, Schuster S (1999). METATOOL: For studying metabolic networks. *Bioinformatics*, 15(3):251–257.

147. Picó-Marco E, Navarro JL, Bruno-Barcena JM (2006). A closed loop exponential feeding law: Invariance and global stability analysis. *Journal of Process Control*, 16(4):395–402.

148. Picó-Marco E (2004). *Nonlinear Robust Control of Biotechnological Processes*. PhD Thesis, Universidad Politécnica de Valencia, Valencia.

149. Poolman MG, Venkatesh KV, Pidcock MK, Fell DA (2004). A method for the determination of flux in elementary modes, and its application to lactobacillus rhamnosus. *Biotechnology and Bioengineering*, 88(5):601–612.

150. Poolman MG, Fell DA, Raines CA (2003). Elementary modes analysis of photosynthate metabolism in the chloroplast stroma. *FEBS Journal*, 270:430–439.

151. Price ND, Papin JA, Palsson BO (2002). Determination of redundancy and systems properties of the metabolic network of *Helicobacter pylori* using genome-scale extreme pathway analysis. *Genome Research*, 12(5):760–769.

152. Price ND, Papin JA, Schilling CH, Palsson BO (2003). Genome-scale microbial in silico models: the constraints-based approach. *Trends in Biotechnology*, 21(4):162–9.

153. Provost A and Bastin G (2004). Dynamic metabolic modelling under the balanced growth condition. *Journal of Process Control* 14(7):717–728.

154. Provost A, Bastin G, Agathos SN, Schneider YJ (2006a). Metabolic design of macroscopic bioreaction models: application to Chinese hamster ovary cells. *Bioprocess and Biosystems Engineering*, 29 (5-6):349–66.

155. Provost A (2006b). *Metabolic design of dynamic bioreaction models*. PhD Thesis, Université catholique de Louvain, Louvain-la-Neuve.

156. Rademacher LA (2007). Approximating the centroid is hard. *Proceedings of the twenty-third annual symposium on Computational geometry*.

157. Mahadevan R, Edwards JS, Doyle FJ (2002). Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophysics Journal*, 83(3):1331–40.

158. Ramakrishna R, Ramkrishna D, Konopka AE (1996). Cybernetic modelling of growth in mixed, substitutable substrate environments: preferential and simultaneous utilization. *Biotechnology & Bioengineering*, 52:141–151.

159. Rani KY, Rao VSR (1999). Control of fermenters: a review. *Bioprocess Engineering*, 21:77–88.

160. Ratcliffe RG, Shachar-Hill Y (2006). Measuring multiple fluxes through plant metabolic networks. *Plant Journal*, 45:490–511.

161. Reder C (1988). Metabolic control theory: a structural approach. *Journal of Theoretical Biology*, 135:175–201.

162. Reed JL, Vo TD, Schilling CH, Palsson BO (2003). An expanded genome-scale model of *Escherichia coli* K-12. *Genome Biology*, 4:R54.

163. Ren HT, Yuan JQ, Bellgardt KH (2003). Macrokinetic model for methylotrophic *Pichia pastoris* based on stoichiometric balance. *Journal of Biotechnology*, 5–106 (1):53–68.

164. Rizzi M, Baltes M, Theobald U, Reuss M (1997). In vivo analysis of metabolic dynamics in *Saccharomyces cerevisiae*: II. Mathematical model. *Biotechnology & Bioengineering*, 55:592–608.

165. Rockafellar RT (1996). *Convex analysis*. Princeton, USA: Princeton University Press.

166. Rocha I, Maia P, Evangelista P, Vilaça P, Soares S, Pinto JP, Nielsen J, Patil KR, Ferreira EC (2010). OptFlux: an open-source software platform for in silico metabolic engineering. *BMC Systems Biology*, 4–45.

167. Russell S, Norvig P. Artificial Intelligence: a modern approach (3rd edition). New Jersey, USA: Prentice-Hall.

168. Sainz J, Pizarro F, Perez-Correa JR, Agosin E (2003). Modeling of yeast metabolism and process dynamics in batch fermentation. *Biotechnology & Bioengineering*, 81:818–828.

169. Sala A, Albertos P (1998). Fuzzy systems evaluation: The inference error approach. *IEEE Transactions on Systems, Man and Cybernetics*, 28(2):268–275.

170. Sala A, Albertos P (2001). Inference error minimisation: fuzzy modelling of ambiguous functions. *Fuzzy Sets and Systems*, 121(1):95–111.

171. Sala A (2008). Encoding fuzzy possibilistic diagnostics as a constrained optimisation problem. *Information Sciences*, 178:4246–4263.

172. Sauer U (2006). Metabolic networks in motion: 13C-based flux analysis. *Molecular Systems Biology*, 2:62.

173. Savinell JM, Palsson BO (1992). Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. *Journal of theoretical biology*, 154(4):421–454.

174. Savinell JM, Palsson BO (1992b). Network analysis of intermediary metabolism using linear optimization.II. Interpretation of hybridoma cell metabolism. *Journal of theoretical biology*, 154(4):455–473.

175. Schilling CH, Palsson BO (2000). Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. *Journal of Theoretical Biology*, 203(3):249–283.

176. Schilling CH, Letscher D, Palsson BO (2000). Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology*, 203 (3) :229–248.

177. Schilling CH, Covert MW, Famili I, Church GM, Edwards JS, Palsson BO (2002). Genome-scale metabolic model of *Helicobacter pylori* 26695. *Journal of Bacteriology*, 184(16):4582–93.

178. Schilling CH, Schuster S, Palsson BO, Heinrich R (1999). Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnology Progress*, 15:296–303.

179. Schmidt K, Nørregaard LC, Pedersen B, Meissner A, Duus JO, Nielsen JO, Villadsen J (1999), Quantification of intracellular metabolic fluxes from fractional

enrichment and 13C-13C coupling constraints on the isotopomer distribution in labeled biomass components. *Metabolic Engineering*, 1(2):166–79.

180. Schmidt K, Nielsen J, Villadsen J (1999). Quantitative analysis of metabolic fluxes in *Escherichia coli*, using two-dimensional NMR spectroscopy and complete isotopomer models. *Journal of Biotechnology*, 71:175–189.

181. Schrijver A (1988). *Theory of linear and integer programming.* Amsterdam, Netherlands: Wiley.

182. Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Molecular Systems Biology* 3:119.

183. Schügerl K, Bellgardt KH (2000). *Bioreaction Engineering: Modelling and Control.* Heidelberg, Germany: Springer-Verlag.

184. Schuster S, Dandekar T, Fell DA (1999). Detection of elementary flux modes in biochemical networks: A promising tool for pathway analysis and metabolic engineering. *Trends in Biotechnology*, 17(2):53–60.

185. Schuster S, Fell DA, Dandekar T (2000). A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature biotechnology*, 18(3):326–332.

186. Schuster S, Hilgetag C, Woods JH, Fell DA (2002). Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism. *Journal of Mathematical Biology*, 45(2):153–181.

187. Schuster S, Pfeiffer T, Moldenhauer F, Koch I, Dandekar T (2002b). Exploring the pathway structure of metabolism: decomposition into subnetworks and application to *Mycoplasma pneumoniae*. *Bioinformatics*, 18:351–361.

188. Schuster S, Pfeiffer T, Fell DA (2008). Is maximization of molar yield in metabolic networks favoured by evolution? *Journal of Theoretical Biology*, 252 (3):497–504

189. Schwartz JM, Kanehisa M (2006). Quantitative elementary mode analysis of metabolic pathways: the example of yeast glycolysis. *BMC Bioinformatics*, 7:186.

190. Schwarz R, Musch P, von Kamp A, Engels B, Schirmer H, Schuster S, Dandekar T (2005). YANA: a software tool for analyzing flux modes, gene-expression and enzyme activities. *BMC Bioinformatics*, 6(1):135.

191. Schwender J, Ohlrogge J, Shachar-Hill, Y (2004). Understanding flux in plant metabolic networks. *Current Opinion in Plant Biology*, 7:309–317.

192. Segre D, Vitkup D, Church GM (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, 99:15112–117.

193. Sharma NS, Ierapetritou MG, Yarmush ML (2005). Novel quantitative tools for engineering analysis of hepatocyte cultures in bioartificial liver systems. *Biotechnology & Bioengineering*, 92:321–335.

194. Shirai T, Matsuzaki K, Kuzumoto M, Nagahisa K, Furusawa C, Shioya S, Shimizu H (2006). Precise metabolic flux analysis of coryneform bacteria by gas chromatography–mass spectrometry and verification by nuclear magnetic resonance. *Journal of Bioscience and Bioengineering*, 102:413–424.

195. Solà A, Jouhten P, Maaheimo H, Sánchez-Ferrando F, Szyperski T, Ferrer P (2007). Metabolic flux profiling of *Pichia pastoris* grown on glycerol/methanol mixtures in chemostat cultures at low and high dilution rates. *Microbiology*, 153(1):281–90.

196. Sonnleitner B, kappeli O (1986). Growth of *Saccharomyces cerevisiae* is controlled by its limited respiratory capacity: formulation and verification of a hypothesis. *Biotechnology & Bioengineering*, 28:927–937.

197. Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED (2002). Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420:190–193.

198. Stelling J (2004). Mathematical models in microbial Systems Biology. *Current Opinion in Microbiology*, 7:513–518.

199. Stephanopoulos GN, Aristidou AA (1998). *Metabolic Engineering: Principles and Methodologies*. San Diego, USA: Academic Press.

200. Steuer R, Nesi AN, Fernie AR, Gross T, Blasius B, Selbig J (2007). From structure to dynamics of metabolic pathways: application to the plant mitochondrial TCA cycle. *Bioinformatics*, 23:1378–85.

201. Szyperski T (1998). 13C-nmr, ms and metabolic flux balancing in biotechnology research. *Quarterly Reviews of Biophysics*, 31(1):41–106.

202. Takiguchi N, Shimizu H, Shioya S (1997). An on-line physiological state recognition system for the lysine fermentation process based on a metabolic reaction model. *Biotechnology & Bioengineering*, 55:170–181.

203. Teixeira AP, Alves C, Alves PM, Carrondo MJ, Oliveira R (2007). Hybrid elementary flux analysis/nonparametric modeling: application for bioprocess control. *BMC Bioinformatics*, 8:30.

204. Terzer M and Stelling J (2008). Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics*, 24 (19):2229–35.

205. Thiele I, Price ND, Vo TD, Palsson BO (2005). Candidate metabolic network states in human mitochondria. Impact of diabetes, ischemia, and diet. *Journal of Biological Chemistry*, 280, 11683–95.

206. Tomita M, Hashimoto K, Takahashi K, Shimizu TS, Matsuzaki Y, et al. (1999). E-CELL: software environment for whole-cell simulation. *Bioinformatics*, 15:72–84.

207. Tomita M (2001). Whole-cell simulation: a grand challenge of the 21st century. *Trends in Biotechnology*, 19:205–210.

208. Tortajada M, Llaneras F, Picó J (2008). Constraint-based modelling applied to heterologous protein production with *P. pastoris*. *Reunión de la red Española de Biología de Sistemas.*

209. Tortajada M, Llaneras F, Picó J (2010). Possibilistic validation of a constraint-based model for *P. pastoris* under data scarcity. *Computer Applications in Biotechnology*.

210. Tortajada M, Llaneras F, Picó J (2010). Validation of a constraint-based model of *Pichia pastoris* growth under data scarcity. *BMC Systems Biology*, 4:115.

211. Urbanczik R (2006). SNA: a toolbox for the stoichiometric analysis of metabolic networks. *BMC Bioinformatics*, 7:129.

212. Vallino JJ and Stephanopoulos G (1993). Metabolic flux distributions in *Corynebacterium glutamicum* during growth and lysine overproduction. *Biotechnology & Bioengineering*, 67(6):872–85 (Reprinted).

213. Vallino JJ (1994). *Identification of branch-point restrictions in microbial metabolism through metabolic flux analysis and local network perturbation*. PhD thesis, Massachusetts Institute of Technology, Cambridge.

214. Varma A, Palsson BO (1994). Metabolic flux balancing: basic concepts, scientific and practical use. *Nature Biotechnology*, 12(10):994–998.

215. Varma A, Palsson BO (1994). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *applied and Environmental Microbiology*, 60(1):3724–31.

216. Varner J, Ramkrishna D (1999). Metabolic engineering from a cybernetic perspective. I. Theoretical preliminaries. *Biotechnology Progress*, 15:407–425.

217. Visser E, Srinivasan B, Palanki S, Bonvin D (2000). A feedback-based implementation scheme for batch process optimisation. *Journal Process Control*, 10:399–410.

218. Veloso ACA, Rocha I, Ferreira EC (2009). Monitoring of fed-batch *E. coli* fermentations with software sensors. *Bioprocess and biosystems engineering*, 32(3):381–8.

219. Wagner C, Urbanczik R (2005). The geometry of the flux cone of a metabolic network. *Biophysics Journal*, 89(6):3837–3845.

220. Wiback SJ, Mahadevan R, Palsson BO (2003). Reconstructing metabolic flux vectors from extreme pathways: Defining the alpha-spectrum. *Journal of Theoretical Biology*, 224(3):313–324.

221. Wiback SJ, Famili I, Greenberg HJ, Palsson BO (2004). Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space. *Journal of Theoretical Biology*, 228:437–447.

222. Wiechert W, Möllney M, Petersen S, Graaf AA (2001). A universal framework for 13C metabolic flux analysis. *Metabolic Engineering*, 3(3):265–83.

223. Wiechert W (2001). 13C metabolic flux analysis. *Metabolic Engineering*, 3(3):195–206.

224. Wittmann C, Heinzle E (2002). Genealogy profiling through strain improvement by using metabolic network analysis: metabolic flux genealogy of several generations of lysine-producing corynebacteria, *Applied Environmental Microbiology* 68:5843–5859.

225. Wold S, Geladi P, Esbensen K, Ohman J (1987). Multi-way principal components-and PLS-analysis. *Journal of Chemometrics*, 1:41–56.

226. Wold S, Kettaneh N, Friden H, Holmberg A (1998). Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemometrics Intelligent Laboratory Systems*, 44:2.

227. Yager RR (1983). An introduction to applications of possibility theory. *Human Systems Management*, 3:246–269.

228. Yang TH, Wittmann C, Heinzle E (2006). Respirometric 13C flux analysis Part II: in vivo flux estimation of lysine-producing *Corynebacterium glutamicum*, *Metabolic Engineering*, 8,:432–446.

229. Zadeh LA (1981). Possibility theory and soft data analysis. *Mathematical frontiers of Social and Policy Sciences*, 69–129. Boulder, USA: Westview Press.

230. Zhong JJ (2002). Plant cell culture for production of paclitaxel and other taxanes. *Journal of Bioscience and Bioengineering*, 94:591–599.