

Universidade do Minho
Escola de Engenharia

Rafael de Castro Carreira

Algorithms and Computational Tools for Metabolic Flux Analysis

abril de 2015



Universidade do Minho
Escola de Engenharia

Rafael de Castro Carreira

Algorithms and Computational Tools for Metabolic Flux Analysis

PhD thesis in Biomedical Engineering

This work was realized under the supervision of:
Prof. Miguel Francisco de Almeida Pereira da Rocha
Prof. Isabel Cristina de Almeida Pereira da Rocha
Prof. Silas Granato Villas-Bôas

abril de 2015

I hereby declare having conducted my thesis with integrity. I confirm that I have not used plagiarism or any form of falsification of results in the process of the thesis elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

University of Minho, April 30, 2015

Rafael de Castro Carreira

ACKNOWLEDGEMENTS/AGRADECIMENTOS

Gostaria de manifestar o meu profundo agradecimento a todos aqueles que, directa ou indirectamente, me apoiaram na realização deste trabalho. Em particular, um agradecimento muito especial ao meus orientadores Doutor Miguel Francisco de Almeida Pereira da Rocha, Doutora Isabel Cristina de Almeida Pereira da Rocha e Doutor Silas Granato Villas-Bôas por todo o conhecimento que me transmitiram, e apoio, disponibilidade e paciência que revelaram para comigo durante a elaboração deste projecto.

À todos os elementos do BisBII pelos momentos de lazer e trabalho que me proporcionaram, onde o conhecimento era constantemente partilhado entre todos.

À Sónia Carneiro manifesto o meu agradecimento por toda a ajuda e discussões científicas que tivemos nas diferentes fases de desenvolvimento deste projecto, assim como os dados experimentais que disponibilizou para a minha investigação. Aqui agradeço também à Daniela Correia e ao grupo da Nova Zelândia por permitirem que pudesse trabalhar também sobre os seus dados.

Ao Paulo Vilaça agradeço a sua preciosa ajuda, nomeadamente na primeira parte to trabalho descrito neste documento, e ao Paulo Maia e Pedro Evangelhista pelas suas contribuições para essa mesma parte.

Agradeço ao financiamento proporcionado pela Fundação para a Ciência e Tecnologia na atribuição da bolsa de doutoramento que tornou este trabalho possível (SFRH/BD/66201/2009), assim como ao Centro de Ciências e Tecnologias de Computação, ao Centro de Engenharia Biológica, da Uni-

versidade do Minho e à Universidade de Auckland por me terem acolhido ao longo destes últimos anos.

Gostaria de reconhecer todo o apoio que a Silicolife disponibilizou para que conseguisse concluir este projecto enquanto trabalhador na empresa, onde o Simão Soares e Paulo Vilaça me motivaram constantemente a dar prioridade a este trabalho.

Como não podia deixar de ser, um obrigado muito especial para os meus pais, irmãos, amigos e familiares mais próximos por estarem sempre ao meu lado, dando-me suporte nas etapas mais complicadas destes anos, quando quase me deixava vencer pelo cansaço ou pelo desânimo. Em particular ao Hugo Costa por insistir sempre comigo e dizer que eu seria capaz.

Finalmente, à minha namorada que nestes últimos tempos aturou os efeitos secundários que esta recta final provocou, por ter paciência em ouvir a palavra "tese" inúmeras vezes nas nossas conversas, e por todo o esforço para manter-se acordada até horas impróprias, acompanhando-me o máximo que pôde nesta última fase.

ABSTRACT

Algorithms and Computational Tools for Metabolic Flux Analysis

Metabolic Engineering (ME) is increasingly used to design strains for industrial biotechnology, using *in silico* constraint-based methods defined over genome-scale models to determine flux distributions in different conditions. However, these methods present a number of limitations related to multiple optima, futile cycles and reliance on cofactor balances. These can be overcome adding constraints to the systems, coming from experimentally measured rates. Valuable information is obtained from ^{13}C tracer experiments, which allow to trace how labeling patterns propagate from substrates to products and how these relate to flux distributions through the application of ^{13}C metabolic flux analysis (^{13}C MFA).

Although many software tools exist, none provides a free application that makes available the full portfolio of flux analysis methods. This work presents *Constraint-Based Flux Analysis (CBFA)*, an open-source software application for flux analysis that implements several methods for phenotype prediction, allowing to define constraints related to measured fluxes and/or flux ratios, together with environmental conditions and gene knockouts. *CBFA* identifies the set of applicable methods based on the user inputs, encompassing algebraic and constraint-based methods. The integration within

the *OptFlux* framework for ME enables to use different model formats and standards and the integration with complementary methods for phenotype simulation and visualization.

To explore the information provided by ^{13}C tracer experiments, a framework is developed that enables the integration of fluxomics data, as ^{13}C labeling distributions, with models enriched with carbon atom transition maps. The algorithms allow to infer labeling distributions for fragments/ metabolites, building expressions for the relevant flux ratios. These can be used to generate flux ratio constraints to be added to flux analysis methods. This framework is generic not requiring any assumptions on the network or reaction reversibility. These algorithms were implemented in MetabolIc NEtwork Ratio AnaLysis (*MiNeRAl*), an user-friend generic tool to calculate metabolic flux ratios for any fragmentation used in GC-MS data. Both tools can thus be combined to perform tasks of ^{13}C MFA.

RESUMO

Algorithms and Computational Tools for Metabolic Flux Analysis

A Engenharia Metabólica (EM) é cada vez mais utilizada para projetar estirpes para a biotecnologia industrial, utilizando métodos baseados em restrições *in silico* definidos sobre modelos à escala genómica para determinar distribuições de fluxos em diferentes condições. No entanto, esses métodos apresentam várias limitações no que diz respeito a múltiplas soluções óptimas, ciclos fúteis e confiança no balanceamento de cofactores. Estas podem ser ultrapassadas através da adição de restrições aos sistemas, obtidas pela medição de fluxos experimentais. Informação relevante é obtida através de experiências de rastreamento de ^{13}C, o que permite perceber como os padrões de marcação são propagados desde os substratos para os produtos, e como estes estão relacionados com fluxos através da aplicação de análise de fluxos metabólicos baseados em ^{13}C.

Embora existam muitas ferramentas informáticas, nenhuma fornece uma aplicação livre que disponibilize um portfólio completo de métodos de análise de fluxos. Este trabalho apresenta a aplicação de código livre, *CBFA*, para análise de fluxos que implementa vários métodos de predição de fenótipos, permitindo a definição de restrições relacionadas com fluxos medidos e/ou rácios de fluxos, juntamente com condições ambientais e deleções de genes.

Esta aplicação identifica o conjunto de métodos aplicáveis com base nos dados introduzidos pelo utilizador, contendo métodos algébricos e baseados em restrições. A integração com o *OptFlux*, uma ferramenta para EM, permite o uso de diferentes formatos de modelos e a integração com métodos complementares de simulação de fenótipos e visualização.

Para explorar a informação obtida das experiências de carbono marcado, foi desenvolvida uma ferramenta que permite a integração de dados de fluxómica, tal como distribuição de marcação de ^{13}C, com modelos enriquecidos com mapeamentos de transição de átomos de carbono. Os algoritmos permitem inferir a distribuição de marcação para fragmentos/ metabolitos, construindo expressões para os rácios de fluxos relevantes. Estes podem ser utilizados para gerar restrições de rácios de fluxo para serem adicionadas à métodos de análise de fluxos. Esta ferramenta é genérica e não requer qualquer assunção sobre rede ou reversibilidades das reações. Estes algoritmos foram implementados no *MiNeRAl*, uma aplicação informática para calcular rácios de fluxos metabólicos para qualquer fragmentação utilizada em dados de GC-MS. Ambas as ferramentas podem ser então combinadas para a execução de análise de fluxos metabólicos baseada em ^{13}C.

TABLE OF CONTENTS

Agradecimentos/Acknowledgements **v**

Abstract **vii**

Resumo **ix**

Table of Contents **xi**

List of Acronyms **xv**

List of Figures **xvii**

List of Tables **xxi**

1 Introduction **1**
1.1 Context and motivation 1
1.2 Objectives 5
1.3 Thesis organization 6

2 State of the art **15**
2.1 Metabolic engineering: concepts and historical notes 15
2.2 Genome sequencing and omics data 18
2.3 Metabolic modeling 20
2.3.1 Systems biology 21
2.3.2 Genome-scale metabolic models 22
2.3.3 Constrained-based modeling 27
2.3.4 Phenotype constrained-based simulation 30
2.4 Stoichiometric metabolic flux analysis 33
2.4.1 Thermodynamics-based metabolic flux analysis . . . 33
2.4.2 Hybrid metabolic flux analysis 34
2.4.3 Software tools for stoichiometric metabolic flux analysis 35
2.4.4 Applications of stoichiometric metabolic flux analysis 36
2.5 Metabolomics 37
2.5.1 Analytical techniques 39
2.5.2 Data repositories and analysis 41

2.6 Fluxomics . . . 43
2.7 ^{13}C metabolite enrichment . . . 45
2.7.1 Labeling and metabolic stationary state . . . 45
2.7.2 Measuring the labeling pattern of amino acids and central intermediates . . . 47
2.7.3 Strategies for substrate labeling . . . 48
2.7.4 Parallel labeling experiments . . . 50
2.8 Stationary ^{13}C Metabolic flux analysis . . . 51
2.8.1 Traditional pipeline in ^{13}C metabolic flux analysis . . . 53
2.8.2 ^{13}C metabolic network model . . . 54
2.8.3 Isotopic labeling modeling and representation . . . 56
2.9 Methods of inference in ^{13}C tracer experiments . . . 65
2.9.1 Global methods of inference in ^{13}C-MFA . . . 66
2.9.2 Local paradigm for ^{13}C-MFA . . . 67
2.9.3 Fluxome profiling . . . 69
2.9.4 Software tools for ^{13}C metabolic flux analysis . . . 72

3 Methods for constraint-based flux analysis 109
3.1 Introduction . . . 109
3.2 Basics of constraint-based modeling . . . 110
3.2.1 Metabolite balance . . . 110
3.2.2 Other constraints . . . 111
3.3 System determination . . . 117
3.4 Methods and formulations . . . 120
3.4.1 Optimization methods . . . 121
3.4.2 Algebraic methods . . . 132

4 CBFA: a software tool for constraint-based flux analysis 147
4.1 Introdution . . . 147
4.2 Software functionalities . . . 150
4.2.1 System configuration . . . 151
4.2.2 Available methods for flux analysis . . . 152
4.2.3 User interaction . . . 154
4.3 Case studies . . . 157
4.3.1 *Saccharomyces cerevisiae* (over)determined configured system . . . 157
4.3.2 Large-scale constrained-based flux analysis in *Escherichia coli* . . . 158
4.4 Implementation . . . 161
4.4.1 *CBFA* software architecture . . . 161
4.4.2 Performing metabolic flux analysis simulations . . . 166

4.4.3 Availability and requirements 166

5 Algorithms to infer metabolic flux ratios from fluxomics data 171
5.1 Introduction 171
5.2 Definitions and problem statement 172
5.2.1 Mass isotopomer distribution vectors 172
5.2.2 Carbon atoms mapping 175
5.2.3 Inferring mass isotopomer distribution vectors 177
5.3 Linking source fragments to products 177
5.3.1 Labeling pattern of combined precursor fragments . . 178
5.3.2 Contribution of different pathways to the same product fragment 180
5.3.3 Labeling influence of carbon dioxide 181
5.4 Determining amino acids precursors 185
5.4.1 Possible scenarios when searching for precursors . . . 187
5.4.2 Meta-heuristics to compute precursors mass isotopomer distribution vectors 190
5.5 Calculating ratio expressions 191
5.5.1 Recurrent sub-expressions 193
5.5.2 Undetermined ratios 195
5.5.3 Carbon mapping cycles 200
5.5.4 Reversible reactions 202
5.6 Determination of metabolic flux ratios 204
5.6.1 Solving linear ratios expressions 205
5.6.2 Solving non linear ratios expressions 206
5.6.3 Estimating mass isotopomer distribution vectors for unknown paths 209

6 MiNeRAl - MetabolIc NEtwork Ratio AnaLysis 217
6.1 Introduction 217
6.2 Software functionalities 218
6.2.1 Metabolic flux ratio analysis workflow 218
6.2.2 Organization of the project components 220
6.2.3 Main operations 222
6.2.4 Carbon atom transitions model 227
6.2.5 Compounds fragmentation 227
6.2.6 Mass spectrometry data processing 231
6.2.7 Linking metabolites from different project components 235
6.2.8 Translating compounds labeling information 237
6.2.9 Predict calculable metabolic flux ratios 245

6.2.10 Computing metabolic flux ratios 247
6.3 Case studies 247
6.3.1 Calculable ratios for MCF derivatization in *Escherichia coli* 248
6.3.2 Computing metabolic flux ratios for *Pseudomonas putida* 250
6.3.3 Metabolic flux ratio constraints 254
6.4 Implementation 256
6.4.1 Workbench architecture 256
6.4.2 Integration with JECoLi 258
6.4.3 Core module organization 258
6.4.4 Implementing methods to compute ratios 261
6.4.5 Availability and requirements 265

7 Conclusions **271**
7.1 Summary of the work and main contributions 271
7.2 Future work 274

List of acronyms

AMDIS Automated Mass Spectral Deconvolution and Identification System

AMM Atom Mapping Matrices

API Application Programming Interface

BMM Bondomer Mapping Matrix

CSV Comma-Separated Values

CBFA Constraint-Based Flux Analysis

CBM Constraint-Based Modeling

DE Differential Evolution

EA Evolutionary Algorithm

EMU Elementary Metabolite Unit

FBA Flux Balance Analysis

FVA Flux Variability Analysis

GC-MS Gas Chromatography–Mass Spectrometry

GSMM Genome-Scale Metabolic Model

GPR Gene-Protein-Reaction

GUI Graphical User Interface

IDV Isotopomer Distribution Vector

IMM Isotopomer Mapping Matrices

JECoLi Java Evolutionary Computation Library

LP Linear Programming

MathML Mathematical Markup Language

MCF methyl chloroformate

MDV Mass isotopomer Distribution Vector

ME Metabolic Engineering

MFA Metabolic Flux Analysis

METAFoR Metabolic Flux Ratio

MILP Mixed-Integer Linear Programming

MOMA Minimization of Metabolic Adjustment

MS Mass Spectrometry

MVC Model-View-Controller

NMR Nuclear Magnetic Resonance

NLSIC Non-Linear Least Squares with Inequality Constraints

m/z mass-to-charge ratio

pFBA Parsimonious enzyme usage Flux Balance Analysis

QP Quadratic Programming

ROOM Regulatory On/Off Minimization

SA Simulated Annealing

SCC Strongly Connected Component

SBML Systems Biology Markup Language

TBDMS t-butyl-trimethylsilyl

UML Unified Modeling Language

LIST OF FIGURES

2.1 Illustration of a pseudo constraint-based metabolic genome-scale model. . . . 29
2.2 Illustration of two substrate labeling strategies. . . . 49
2.3 Traditional pipeline of ^{13}C metabolic flux analysis. . . . 55
2.4 Vector representation of the isotopic content of a molecule. . . . 58
2.5 Software tools for stationary ^{13}C-MFA. . . . 72

3.1 Allowable solution space of constraint-based modeling. . . . 112
3.2 Solution space reduced with additional constraints. . . . 112
3.3 Flux bounds obtained from different type constraints. . . . 114
3.4 Model degrees of freedom. . . . 119
3.5 Methods inputs and outputs 122
3.6 Extended stoichiometric matrix. . . . 133

4.1 Overview of the application. . . . 150
4.2 Screenshots of *CBFA*. . . . 156
4.3 Visualization of flux analysis results for *Saccharomyces cerevisiae*. 159
4.4 Visualization of flux analysis results for *Escherichia coli*. . . . 162
4.5 *CBFA* layers. . . . 163
4.6 Application layer MVC architecture. . . . 164
4.7 Functional modules of the *CBFA* application, including the developed datatypes, views and operations. . . . 165
4.8 UML sequence diagram for the configuration of an execution of the pFBA method. . . . 167
4.9 Continuation of the UML sequence diagram for executing a pFBA method. . . . 168

5.1 Carbon atom transitions between the metabolites in the transaldolase reaction. . . . 176
5.2 Possible scenarios of carbon labeling contribution of the precursors to their products. . . . 178
5.3 Schematization of how the amino acid carbon skeletons derive from their precursors in *Escherichia coli*. . . . 186
5.4 Examples of how carbon atoms can be transferred between intermediate metabolites to amino acids. . . . 189

5.5 Small excerpt of a network and the tree-shaped representation of the expression that enables to calculate the Mass isotopomer Distribution Vector (MDV) of a product fragment 192
5.6 Pseudo network originating recurrent ratios sub-expressions. . . 194
5.7 Radius of the search procedure. 196
5.8 Illustration of a pseudo network with undetermined ratios. . . . 197
5.9 Substitution of expression sub-trees by unknown nodes. 199
5.10 Illustration of a network with carbon mapping cycles. 200
5.11 Reversible reactions might originate several expressions for different assumptions. 203
5.12 Illustration of a meta-heuristic method to estimate metabolic flux ratios. 208
5.13 Illustration of a meta-heuristic method to estimate both flux ratios and unknown fragment MDV. 210

6.1 Workflow with the main tasks in *MiNeRAl*. 219
6.2 *MiNeRAl* project elements organization. 221
6.3 Main *MiNeRAl* operations. 222
6.4 A project is defined in the context of a carbon transitions model. 228
6.5 After loading a derivatization file, a datatype representing the fragmentation is created. 230
6.6 Activity diagram for filtering compounds spectra. 234
6.7 Visualization of the scan spectra obtained for a loaded sample replicate. 236
6.8 Correction of mass distribution vectors for natural isotope abundance. 240
6.9 Calculable ratios operation dialog. 246
6.10 Calculable ratios datatype view. 249
6.11 Computed metabolic flux ratios for *Pseudomonas putida*. 253
6.12 Illustration of the process to build metabolic flux ratio constraints. 255
6.13 *MiNeRAl* layers. 257
6.14 UML class diagram for the main *MiNeRAl* components data structuring. 259
6.15 UML class diagram for the organization of the estimation processes implemented in *MiNeRAl*. 260
6.16 Sequence diagram of the ratios estimation method based on the Evolutionary Algorithm (EA) approach. 262
6.17 Sequence diagram of the EA algorithm configuration for the estimation of flux ratios to solve ratios to metabolite fragment expressions. 263

6.18 Continuation of the sequence diagram of the EA algorithm configuration for the estimation of flux ratios to solve ratios to metabolite fragment expressions. 264

7.1 Overview of the developed software applications during this work. 273

LIST OF TABLES

2.1 Summary of the datasets generated by the main "omic" technologies. 20
2.2 Summary of some published genome-scale metabolic models. . . 24
2.3 Summary of applications with ratios constrained metabolic flux analysis. 70

3.1 Symbols used in the equations and formulations of the methods. 142

4.1 *CBFA* technical information. 167

5.1 Symbols used in the equations and formulations of ^{13}C-MFA methods. 212

6.1 Amino acids methyl chloroformate derivatives 229
6.2 Methyl chloroformate derivatives fragments atoms composition . 230
6.3 Isotopic compositions of the most biologically relevant atoms . . 241
6.4 Intermediate precursors of methyl chloroformate derivatives fragments. 243
6.5 Summary of some flux metabolic ratios expressions generated by *MiNeRAl*. 249
6.6 Central carbon atoms transitions model of *Pseudomonas putida* 251
6.7 *MiNeRAl* technical information. 265

CHAPTER 1

INTRODUCTION

1.1 Context and motivation

Nowadays, many useful materials are produced through the use of cell factories. The growing adoption of industrial biotechnology for producing pharmaceuticals, food ingredients and fine chemicals, among others [Griffiths, 2001], has originated an increasing attention to the field of Metabolic Engineering (ME) [Nielsen, 1998]. ME is a multidisciplinary field that deals with the analysis and design of metabolic systems towards particular goals, such as increasing the production of useful compounds.

One of the major aims of ME is the rational development of strains, enabling them to optimize their productivity in biotechnological processes. The emergence of high-throughput tools for genome and metagenome sequencing [Warnecke et al., 2007, Tringe and Rubin, 2005, Alm et al., 2005], gene expression, protein and metabolite profiling [Sauer, 2004, Heux et al., 2012, Wiechert, 2001] have enabled to better understand cell metabolism [Tang et al., 2009] and, consequently, empowered rational strain development approaches. However, more accurate knowledge on the physiological state of cells might be obtained from metabolic flux profiles [Sauer, 2004, Sauer, 2006].

Current biochemical knowledge and the information collected from the annotation of numerous genome sequencing projects allowed the development of metabolic networks for numerous organisms, within the emergent

field of Systems Biology [Kitano, 2002, Ideker et al., 2001]. Metabolic networks provide information about possible biochemical pathways for the conversion of nutrients into products, energy and reducing equivalents, but do not provide much information about the actual fluxes through different pathways under certain conditions [Stephanopoulos et al., 1998].

Genome-Scale Metabolic Model (GSMM)s are related to metabolic networks, but attempt to go one step further, supporting the simulation of the metabolic phenotype, i.e. determining fluxes for the reactions. Although the process is still only semi-automatic, a number of GSMMs have been reconstructed for different organisms [Henry et al., 2010].

Based on these models, Metabolic Flux Analysis (MFA) methods can be used to calculate intracellular fluxes, thus revealing the degree of pathway commitment in different metabolic processes [Jørgensen et al., 1995, Nissen et al., 1997, Vallino and Stephanopoulos, 1993], and to elucidate the metabolic state of cells *in vivo*. Through the configuration of steady-state mass balance equations for each intracellular metabolite in a given metabolic model and considering reaction reversibilities, a system is constructed and reaction rates (fluxes) can be calculated.

However, these systems are often underdetermined [Varma and Palsson, 1994], specially for GSMMs, and the flux values cannot be calculated algebraically. Therefore, linear/ quadratic programming optimization methods are commonly used to solve the systems, such as the Flux Balance Analysis (FBA) method [Edwards et al., 2001], where an objective function is devised, e.g., the maximization of growth rate.

However, different flux distributions can be obtained for objective function optimal value [Mahadevan et al., 2003]. Moreover, the used metabolic networks comprise reversible reactions and futile cycles [Stephanopoulos

et al., 1998, Wiechert and de Graaf, 1997], which might hamper the flux estimations, as does the reliance on balances for cofactors [Christensen and Nielsen, 1999].

To overcome these limitations, additional information can be provided to the system in the form of problem constraints. These constraints can translate experimental data, such as measured consumption and production rates. However, usually, this information is not enough to obtain flux distributions that reflect the actual cell's physiological state. Valuable information can be obtained from experiments where cells are grown in a media with labeled substrates. Through the application of ^{13}C tracer experiments, it is possible to obtain more reliable values for the cell's internal fluxes [Blum and Stein, 1982, Malloy et al., 1988, Chance et al., 1983, Walsh and Koshland, 1984, Walker et al., 1982].

A number of analytical tools can be used to perform the comprehensive determination of the ^{13}C labeling distribution of intracellular metabolites, including Nuclear Magnetic Resonance (NMR) and Mass Spectrometry (MS) [Wiechert and de Graaf, 1996]. On metabolic networks, the carbon backbones of metabolites can be derived from alternative pathways. These experiments enable to track how the ^{13}C labeling patterns of the substrates are propagated throughout the metabolites in the network, since the labeling distribution is very sensitive to the reaction rates. MFA methods can be significantly enhanced with this knowledge [Sauer et al., 1997, Wiechert et al., 1997, Zelle et al., 2008, McKinlay and Harwood, 2010].

This more accurate way to perform flux estimation is the so called ^{13}C-MFA, and existing approaches can be classified as implementing global or local methods of inference. The former consists in procedures formulated as nonlinear optimization problems that iteratively search for the optimum

flux distribution that fits data observed from labeling substrate experiments [Zupke and Stephanopoulos, 1994, Schmidt et al., 1997, Wiechert et al., 1999, Wiechert et al., 2001, van Winden et al., 2002, Antoniewicz et al., 2007, Nöh et al., 2007, Young et al., 2008].

Due to the complexity in the global formulations, their results might be hard to interpret. Moreover, no direct evidence is given for specific fluxes in the metabolic network [Emmerling et al., 2002]. Metabolic Flux Ratio (METAFoR) analysis [Sauer et al., 1999, Szyperski, 1995b] is a local approach that allows to access direct evidence about particular fluxes. This is accomplished by building probabilistic equations that express the fractions of labeling content that are propagated between certain metabolites. The computation of these flux ratios is independent of the flux values, so they can be computed separately of the flux estimations. These ratios can be posteriorly used to reduce the degrees of freedom of the systems in stationary MFA.

However, usually, the flux ratios are calculated from manually derived equations [Sauer et al., 1997, Szyperski et al., 1999, Maaheimo et al., 2001, Fischer and Sauer, 2003, Blank et al., 2005]. In previous work, a computational framework has been developed [Rantanen et al., 2008] that aims to infer, from experimental data, an equation system to constrain the flux distribution to a low-dimensional convex set. However, this framework still shows some limitations since it depends on simplifying assumptions about the network topology of the used model, as it resorts only to linear optimization techniques, or flux directionality, and it does not provide a general-purpose way of creating bounds over flux ratios.

1.2 Objectives

In this context, the main aim of this work is to provide methods and computational tools for MFA, which can be used by the community as important tools for ME and biological discovery. To achieve this aim, two complementary goals are sought, related to stationary and to ^{13}C based MFA.

The first objective is, therefore, to provide a tool for performing stationary MFA with different types of measurements and/or knowledge. The tool should be able to, given a genome-scale model together with different measurements or knowledge (including flux measurements, knowledge on genetic modifications or on non-active reactions, and/ or metabolic flux ratios), first recalculate the corresponding problem and, afterwards, allow the user to calculate flux distributions with or without further assumptions. The key objectives of such tool are:

- to be applicable for any model and type of data/knowledge;
- encompass different state of the art methods for flux analysis;
- exploit the functionalities of *OptFlux*, an in-house software platform [Rocha et al., 2010] to build an user-friendly application that integrates well with other ME tools.

A second objective is, as mentioned, to develop methods and tools to handle MFA based on ^{13}C data. This encompasses, firstly, the development of a novel automated approach to determine metabolic flux ratios from ^{13}C data. This framework generates all possible equations to calculate metabolic flux ratios for a given metabolic model, taking as input a set of intact carbon fragments of internal metabolites for which relative abundances have been measured (or determined from the abundance of

the amino acids for which they are precursors [Szyperski, 1995a]). More specifically, this tool:

- is generic for any carbon atoms transition model and compound fragmentation;
- estimates flux ratio bounds, when not enough labeling measurements are given to the system;
- is independent of any assumptions on the metabolic network and on the reactions reversibility.

Finally, a second tool must be developed to facilitate the tasks of building metabolic flux ratios from Gas Chromatography–Mass Spectrometry (GC-MS) data. This tool will allow to compute the flux ratios and export them, so they can be used in other tools for flux analysis (e.g. the one developed in the first part of this thesis). The tool is be able to:

- perform the aforementioned analysis of flux ratios for any user inputted carbon transitions model and measured fragments;
- predict the calculable ratios for a given fragmentation, without the need for introducing the actual experimental data;
- show every step result in the pipeline analysis, so inconsistencies can be detected and evaluated in any step of the analysis.

1.3 Thesis organization

This thesis is organized in 7 chapters. The current chapter gives a brief introduction to the topics and provides the problem statement, putting forward the main aims of this work.

The second chapter refers to the state of the art mainly related to metabolic engineering, constraint-based modeling, omics data, stoichiometric MFA and ^{13}C MFA.

Chapter 3 contains the methods for the tasks of constraint-based flux analysis. The extensions to the traditional approaches for flux analysis in terms of exploring the benefits of using additional constraints obtained from experimental insights are introduced. In chapter 4, the software for performing these methods is presented. The software architecture and implementation are discussed, also presenting the user interaction and some case studies.

The algorithms for performing ^{13}C metabolic flux ratio analysis are described in the chapter 5, while the software to perform such tasks is introduced in chapter 6. The main features of the software are presented, as well as its architecture, implementation and some case studies to validate the functionalities.

Finally, some final conclusions and ideas for future work are given in chapter 7.

BIBLIOGRAPHY

[Alm et al., 2005] Alm, E. J., Huang, K. H., Price, M. N., Koche, R. P., Keller, K., Dubchak, I. L., and Arkin, A. P. (2005). The MicrobesOnline Web site for comparative genomics. *Genome research*, 15(7):1015–1022.

[Antoniewicz et al., 2007] Antoniewicz, M. R., Kelleher, J. K., and Stephanopoulos, G. (2007). Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions. *Metabolic engineering*, 9(1):68–86.

[Blank et al., 2005] Blank, L. M., Lehmbeck, F., and Sauer, U. (2005). Metabolic-flux and network analysis in fourteen hemiascomycetous yeasts. *FEMS yeast research*, 5(6-7):545–58.

[Blum and Stein, 1982] Blum, J. J. and Stein, R. B. (1982). On the analysis of metabolic networks. *Biological regulation and development*, 3:99–125.

[Chance et al., 1983] Chance, E. M., Seeholzer, S. H., Kobayashi, K., and Williamson, J. R. (1983). Mathematical analysis of isotope labeling in the citric acid cycle with applications to 13C NMR studies in perfused rat hearts. *Journal of Biological Chemistry*, 258(22):13785–13794.

[Christensen and Nielsen, 1999] Christensen, B. and Nielsen, J. (1999). Isotopomer analysis using GC-MS. *Metabolic Engineering*, 1(4):282–290.

[Edwards et al., 2001] Edwards, J. S., Ibarra, R. U., and Palsson, B. O. (2001). In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nature biotechnology*, 19(2):125–130.

[Emmerling et al., 2002] Emmerling, M., Dauner, M., Ponti, A., Fiaux, J., Hochuli, M., Szyperski, T., Wüthrich, K., Bailey, J. E., and Sauer, U. (2002). Metabolic Flux Responses to Pyruvate Kinase Knockout in Escherichia coli. *Society*, 184(1):152–164.

[Fischer and Sauer, 2003] Fischer, E. and Sauer, U. (2003). Metabolic flux profiling of Escherichia coli mutants in central carbon metabolism using GC-MS. *European Journal of Biochemistry*, 270(5):880–891.

[Griffiths, 2001] Griffiths, M. (2001). The application of biotechnology to industrial sustainability. Technical report, Organization for Economic Cooperation and Development (OECD), Paris.

[Henry et al., 2010] Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., and Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology*, 28(9):977–982.

[Heux et al., 2012] Heux, S., Fuchs, T. J., Buhmann, J., Zamboni, N., and Sauer, U. (2012). A high-throughput metabolomics method to predict high concentration cytotoxicity of drugs from low concentration profiles. *Metabolomics*, 8(3):433–443.

[Ideker et al., 2001] Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: systems biology. *Annual review of genomics and human genetics*, 2(1):343–372.

[Jørgensen et al., 1995] Jørgensen, H., Nielsen, J., Villadsen, J., and Møllgaard, H. (1995). Metabolic flux distributions in penicillium chrysogenum during fed-batch cultivations. *Biotechnology and Bioengineering*, 46(2):117–131.

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

[Maaheimo et al., 2001] Maaheimo, H., Fiaux, J., Cakar, Z. P., Bailey, J. E., Sauer, U., and Szyperski, T. (2001). Central carbon metabolism of Saccharomyces cerevisiae explored by biosynthetic fractional 13C labeling of common amino acids. *European Journal of Biochemistry*, 268(8):2464–2479.

[Mahadevan et al., 2003] Mahadevan, R., Schilling, C. H., and Others (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic engineering*, 5(4):264.

[Malloy et al., 1988] Malloy, C. R., Sherry, A. D., and Jeffrey, F. M. (1988). Evaluation of carbon flux and substrate selection through alternate pathways involving the citric acid cycle of the heart by 13C NMR spectroscopy. *Journal of Biological Chemistry*, 263(15):6964–6971.

[McKinlay and Harwood, 2010] McKinlay, J. B. and Harwood, C. S. (2010). Carbon dioxide fixation as a central redox cofactor recycling mechanism in bacteria. *Proceedings of the National Academy of Sciences*, 107(26):11669–11675.

[Nielsen, 1998] Nielsen, J. (1998). Metabolic engineering: techniques for analysis of targets for genetic manipulations. *Biotechnology and bioengineering*, 58(2-3):125–132.

[Nissen et al., 1997] Nissen, T. L., Schulze, U., Nielsen, J., and Villadsen, J. (1997). Flux distributions in anaerobic, glucose-limited continuous cultures of Saccharomyces cerevisiae. *Microbiology*, 143(1):203–218.

[Nöh et al., 2007] Nöh, K., Grönke, K., Luo, B., Takors, R., Oldiges, M., and Wiechert, W. (2007). Metabolic flux analysis at ultra short time scale: isotopically non-stationary 13C labeling experiments. *Journal of biotechnology*, 129(2):249–67.

[Rantanen et al., 2008] Rantanen, A., Rousu, J., Jouhten, P., Zamboni, N., Maaheimo, H., and Ukkonen, E. (2008). An analytic and systematic framework for estimating metabolic flux ratios from 13C tracer experiments. *BMC bioinformatics*, 9:266.

[Rocha et al., 2010] Rocha, I., Maia, P., Evangelista, P., Vilaça, P., Soares, S. a., Pinto, J., Nielsen, J., Patil, K., Ferreira, E., and Rocha, M. (2010). OptFlux: an open-source software platform for in silico metabolic engineering. *BMC systems biology*, 4(1):45.

[Sauer, 2004] Sauer, U. (2004). High-throughput phenomics: experimental methods for mapping fluxomes. *Current opinion in biotechnology*, 15(1):58–63.

[Sauer, 2006] Sauer, U. (2006). Metabolic networks in motion: 13C-based flux analysis. *Molecular systems biology*, 2:62.

[Sauer et al., 1997] Sauer, U., Hatzimanikatis, V., Bailey, J. E., Hochuli, M., Szyperski, T., and Wüthrich, K. (1997). Metabolic fluxes in riboflavin-producing Bacillus subtilis. *Natire Publishing Group*, 15:448–452.

[Sauer et al., 1999] Sauer, U., Lasko, D. R., Fiaux, J., Hochuli, M., Glaser, R. W., Szyperski, T., Wüthrich, K., and Bailey, J. E. (1999). Metabolic flux ratio analysis of genetic and environmental modulations of Escherichia coli central carbon metabolism. *Journal of bacteriology*, 181(21):6679–88.

[Schmidt et al., 1997] Schmidt, K., Carlsen, M., Nielsen, J., and Villadsen, J. (1997). Modeling isotopomer distributions in biochemical networks using isotopomer mapping matrices. *Biotechnology and Bioengineering*, 55(6):831–840.

[Stephanopoulos et al., 1998] Stephanopoulos, G., Aristidou, A. A., Nielsen, J. H. i., and Nielsen, J. (1998). *Metabolic engineering: principles and methodologies*. Academic Press.

[Szyperski, 1995a] Szyperski, T. (1995a). Biosynthetically Directed Fractional 13C-labeling of Proteinogenic Amino Acids. *European Journal of Biochemistry*, 232(2):433–448.

[Szyperski, 1995b] Szyperski, T. (1995b). Biosynthetically directed fractional 13C-labeling of proteinogenic amino acids. An efficient analytical tool to investigate intermediary metabolism.

[Szyperski et al., 1999] Szyperski, T., Glaser, R. W., Hochuli, M., Fiaux, J., Sauer, U., Bailey, J. E., and Wüthrich, K. (1999). Bioreaction network topology and metabolic flux ratio analysis by biosynthetic fractional 13C labeling and two-dimensional NMR spectroscopy. *Metabolic engineering*, 1(3):189–197.

[Tang et al., 2009] Tang, Y. J., Martin, H. G., Myers, S., Rodriguez, S., Baidoo, E. E. K., and Keasling, J. D. (2009). Advances in analysis of microbial metabolic fluxes via 13C isotopic labeling. *Mass Spectrometry Reviews*, 28(2):362–375.

[Tringe and Rubin, 2005] Tringe, S. G. and Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nature reviews genetics*, 6(11):805–814.

[Vallino and Stephanopoulos, 1993] Vallino, J. J. and Stephanopoulos, G. (1993). Metabolic flux distributions in Corynebacterium glutamicum during growth and lysine overproduction. *Biotechnology and Bioengineering*, 41(6):633–646.

[van Winden et al., 2002] van Winden, W. A., Heijnen, J. J., and Verheijen, P. J. T. (2002). Cumulative bondomers: a new concept in flux analysis from 2D [13C, 1H] COSY NMR data. *Biotechnology and bioengineering*, 80(7):731–745.

[Varma and Palsson, 1994] Varma, A. and Palsson, B. O. (1994). Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. *Nature biotechnology*, 12:994–998.

[Walker et al., 1982] Walker, T. E., Han, C. H., Kollman, V. H., London, R. E., and Matwiyoff, N. A. (1982). 13C nuclear magnetic resonance studies of the biosynthesis by Microbacterium ammoniaphilum of L-glutamate selectively enriched with carbon-13. *Journal of Biological Chemistry*, 257(3):1189–1195.

[Walsh and Koshland, 1984] Walsh, K. and Koshland, D. E. (1984). Determination of flux through the branch point of two metabolic cycles. The tricarboxylic acid cycle and the glyoxylate shunt. *Journal of Biological Chemistry*, 259(15):9646–9654.

[Warnecke et al., 2007] Warnecke, F., Luginbühl, P., Ivanova, N., Ghassemian, M., Richardson, T. H., Stege, J. T., Cayouette, M., McHardy, A. C., Djordjevic, G., Aboushadi, N., and Others (2007). Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*, 450(7169):560–565.

[Wiechert, 2001] Wiechert, W. (2001). 13C metabolic flux analysis. *Metabolic engineering*, 3(3):195–206.

[Wiechert and de Graaf, 1996] Wiechert, W. and de Graaf, A. A. (1996). In vivo stationary flux analysis by 13C labeling experiments. *Advances in biochemical engineering/biotechnology*, 54:109–54.

[Wiechert and de Graaf, 1997] Wiechert, W. and de Graaf, A. A. (1997). Bidirectional reaction steps in metabolic networks: I. Modeling and simulation of carbon isotope labeling experiments. *Biotechnology and bioengineering*, 55(1):101–17.

[Wiechert et al., 1999] Wiechert, W., Möllney, M., Isermann, N., Wurzel, M., and de Graaf, A. A. (1999). Bidirectional reaction steps in metabolic networks: III. Explicit solution and analysis of isotopomer labeling systems. *Biotechnology and bioengineering*, 66(2):69–85.

[Wiechert et al., 2001] Wiechert, W., Möllney, M., Petersen, S., and de Graaf, A. A. (2001). A universal framework for 13C metabolic flux analysis. *Metabolic engineering*, 3(3):265–83.

[Wiechert et al., 1997] Wiechert, W., Siefke, C., de Graaf, A. A., and Marx, A. (1997). Bidirectional reaction steps in metabolic networks: II. Flux estimation and statistical analysis. *Biotechnology and bioengineering*, 55(1):118–35.

[Young et al., 2008] Young, J. D., Walther, J. L., Antoniewicz, M. R., Yoo, H., and Stephanopoulos, G. (2008). An elementary metabolite unit (EMU) based method of isotopically nonstationary flux analysis. *Biotechnology and bioengineering*, 99(3):686–699.

[Zelle et al., 2008] Zelle, R. M., de Hulster, E., van Winden, W. A., de Waard, P., Dijkema, C., Winkler, A. A., Geertman, J.-M. A., van Dijken, J. P., Pronk, J. T., and van Maris, A. J. A. (2008). Malic acid production by Saccharomyces cerevisiae: engineering of pyruvate carboxylation, oxaloacetate reduction, and malate export. *Applied and environmental microbiology*, 74(9):2766–2777.

[Zupke and Stephanopoulos, 1994] Zupke, C. and Stephanopoulos, G. (1994). Modeling of isotope distributions and intracellular fluxes in metabolic networks using atom mapping matrices. *Biotechnology Progress*, 10(5):489–498.

CHAPTER 2

STATE OF THE ART

2.1 Metabolic engineering: concepts and historical notes

The idea that cells could be engineered to achieve the overproduction of several interesting chemical and pharmaceutical products has gained attention since Cohen and Boyer introduced external genes into *Escherichia coli* [Cohen et al., 1973]. Nevertheless, while the production of insulin, which is a very complex molecule, is obtained by overexpressing a single gene, the production of much simpler compounds, as ethanol, involves controlling the overexpression of several genes. Therefore, the attention from the biotechnology community was more focused in medical applications, whereas other targets were left aside [Woolston et al., 2013].

In the 1980s, the academic community has started to investigate how substrates could be converted to a target product through a set of steps (reactions), composing paths inside a network. This network is defined by a set of nodes (metabolites) and edges connecting the nodes (reactions). The investigation of alternative pathways connecting substrates to products and their thermodynamic properties, the understanding of how kinetic control is exerted, or the schematization of genetic circuits that originate a certain pattern of gene expression and protein synthesis have resulted in published works from Stephanopoulos and Valino, and Bailey [Stephanopoulos et al., 1991, Bailey, 1991].

The progress in techniques for DNA recombination has been of great importance to implement pathway modifications. Through genetic engineering it has become possible to modify particular enzymatic reactions that participate in desired metabolic pathways. In this context, the concept of Metabolic Engineering (ME) emerges, which can be defined as the induced improvement of the production of target compounds or cellular properties, through the use of recombinant DNA technology to modify specific biochemical reactions, or even introduce new ones [Stephanopoulos, 1999]. This management of biochemical reactions is made in terms of amplification, inhibition, deletion, transfer, or deregulation of the corresponding genes or enzymes.

The rational design of microorganisms may encompass metabolic or genetic regulation, the removal of feedback inhibition or the elucidation of rate-limiting reactions. Therefore, microorganisms have been modified through ME to improve the yield and productivity of native products. In the 1990s, biosynthetic genes on *Corynebacterium glutamicum* have been amplified, where the recombinant strains showed higher activities of homoserine dehydrogenase, homoserine kinase, and threonine kinase [Eikmanns et al., 1991]. The same organism was engineered in [Ikeda and Katsumata, 1992], where its enzymes were overexpressed to produce tyrosine or phenylalanine in abundance. In [Colon et al., 1995], the overprocuction of isoleucine has been induced by the overexpression of the gene encoding threonine dehydratase in a *Corynebacterium lactofermentum* threonine producer.

ME can also be applied to spread the substrates used by the microorganisms. For instance, *Zymomonas mobilis* was modified so that it could grew with the pentose sugar xylose as substrate and efficiently ferment it to ethanol [Zhang et al., 1995].

In [Kumar et al., 1992], the authors designed a strain of *Saccharomyces cerevisiae* that uses lactose and expresses the cDNA for a secreted, thermostable lacA gene from *Aspergillus niger*.

Bacterial genes from *Alcaligenes eutrophus* have been introduced into *Arabidopsis thaliana* with the objective to improve the accumulation of poly-hydroxybutyrate (PHB) in higher plants [Nawrath et al., 1994]. However, biopolymer synthesis has been previously explored in organisms other than plants, including *Escherichia coli* by cloning genes from *Alcaligenes eutrophus* [Schubert et al., 1988, Slater et al., 1988] and *Corynebacterium glutamicum* [Peoples et al., 1988].

More recently, a comprehensive range of applications in ME have been developed. For instance, in Medicine, *Escherichia coli* and *Saccharomyces cerevisiae* have been manipulated to improve the production of natural products, and precursors, with remarkable properties, such as the anti-cancer taxol [Ajikumar et al., 2010, Engels et al., 2008], the antiparasitic artemisinin [Ro et al., 2006, Westfall et al., 2012], the anti-oxidant tanshinones [Dai et al., 2012, Zhou et al., 2012], among others [Dai et al., 2013].

Since the introduction of ME, biotechnological processes have been adopted for the industrial production of valuable compounds. Several advantageous materials can be industrially produced through fermentation of engineered microorganisms including amino acids, organic acids, alcohols, and vitamins, which could be used, for example, as food ingredients or drugs [OECD, 2011]. A key concept in industrial biotechnology is the sustainability of its processes, which refers to the energy and raw materials used, the stability and safety of the processes, the quality of the final product, and waste production. Taking into consideration such aspects, the competitiveness can be augmented through reduced costs of produc-

tion [Schmid et al., 2002]. Several examples of the application of biotechnology to the sustainability of industrial processes are summarized in published works [Griffiths, 2001, Gavrilescu and Chisti, 2005].

2.2 Genome sequencing and omics data

After developing a methodology for determining DNA sequences by primer synthesis with DNA polymerase [Sanger and Coulson, 1975], Sanger and co-authors, have accomplished the first fully sequenced genome in 1976, by sequencing the genetic material of the bacteriophage ϕX174 (with 5386 nucleotides) [Sanger et al., 1977]. After that, the 16.5 kb human mitochondrial genome has also been sequenced [Anderson et al., 1981].

In 1995, the bacteria *Haemophilus influenzae* had its genome sequenced [Fleischmann et al., 1995], which was the first free-living organism to have its genome fully sequenced [Hall, 2007]. In the years after, several other microbes were sequenced, such as *Escherichia coli* [Blattner et al., 1997], *Archaeoglobus fulgidus* (the first sequenced archaeon) [Klenk et al., 1997], and *Mycobacterium tuberculosis* [Cole et al., 1998].

Nowadays, high-throughput genome sequencing technologies have been providing for several genome sequencing projects. The genome of the yeast *Saccharomyces cerevisiae* was completely sequenced [Goffeau et al., 1996], as the fungal pathogen of humans *Candida Albicans* [Butler et al., 2009]. Moreover, genome sequencing of mammals including human [Lander et al., 2001], mouse [Chinwalla et al., 2002] and the humans' closest living relatives, chimpanzees [Sequencing et al., 2005], have been performed.

Genome sequences specify the blueprint for cellular function, where all genes comprehended in the genome can be thought to define the functional

space of the corresponding organism [Smedsgaard and Nielsen, 2005]. Also, knowledge on the function of all proteins, and on the genes that are expressed in different environmental conditions, is also needed to define the organism's functional space [Nielsen et al., 2007]. Since the cost of genome sequencing has drastically decreased, and consequently, a high number of sequencing projects has been put forward, the focus has been directed to the investigation of the function of individual genes.

This understanding about gene functions, named as functional genomics, uses bioinformatics as a key resource, for instance to provide the translation of the coded functions coming from high number of sequenced genomes and homology searches against sequences of known function.

In more recent years, other experimental techniques have evolved to answer questions about the different components of living cells. Nowadays, several high-throughput techniques exist, generating the so-called ''omics" data. The primary aim of "omics" technologies, such as transcriptomics, proteomics and metabolomics, is the nontargeted identification and quantification of all transcripts, proteins, and metabolites, respectively, present in a specific biological sample. By their nature, these technologies reveal unexpected properties of biological systems. A second and more challenging aspect of "omics" technologies is the refined analysis of quantitative dynamics in biological systems.

Through the combination of the existing ''omics" techniques, and the advances of bioinformatics and computational tools, functional genomics may identify the function of a large number of genes for a given species. Table 2.1 gives a brief description of the main ''omics" technologies.

Table 2.1: Summary of the datasets generated by the main "omic" technologies.

genomics	comprises all the nucleotides in the genetic material of a given organism.
transcriptomics	encompasses the full genetic information that is expressed, i.e. transcribed to mRNA; mRNA can be measured using techniques as microarrays or RNA sequencing.
proteomics	comprehensive list of the proteins (and their quantification), e.g using mass spectrometry or 2D electrophoresis.
metabolomics	comprehensive set of small molecules and their quantification, using for instance nuclear magnetic resonance, or gas/liquid chromatography coupled with mass spectrometry.
fluxomics	set of *in vivo* carbon fluxes obtained from quantitative analysis that refers to the intracellular activities of enzymes and pathways that are responsible for the interconversion of metabolites.
interactomics	interactions between the different components of the cell, such as, protein-protein, protein-dna and protein-metabolite interactions, among others.

2.3 Metabolic modeling

The use of software tools has received attention in the field of metabolic research since long time ago. Chance has modeled single enzymes with analog computers [Chance, 1943, Chance et al., 1952], and the dynamics of metabolic pathways using digital computers [Chance et al., 1960]. Data obtained from the different ‘’omics” data introduced in the last section can be used to identify the components in a network representing the events occurring in the cell, and the interaction between these components can thus be inspected.

2.3.1 Systems biology

Biological systems are complex, and their complexity is both on the level of the network, and of the individual components that form the network. The function of an individual element cannot be isolated, so that a cell can be characterized as a symbiotic system [Kitano, 2002a]. Systems biology supports its foundations in a system-level understanding [Kitano and Others, 2001], and it does not examine genes and proteins in isolation. On the contrary, it explores the behavior and relationships of the entire set of elements in the target system simultaneously, while it is functioning [Ideker et al., 2001]. It is an integrated process that encompasses computational modeling, system analysis, technology development and quantitative experiments [Kitano, 2002b].

Systems biology has provided a number of approaches to the simulation of the phenotype of wild type or mutant strains, such as Flux Balance Analysis (FBA) [Kauffman et al., 2003]. However, the underlying models still suffer from some limitations and generally lack kinetic and regulatory information, thus limiting the accuracy of simulation methods that are restricted in their applicability (e.g. FBA assumes steady state conditions).

To overcome these limitations, global approaches on gene and protein expression measurement, the two most popular of the "omics" data (transcriptomics and proteomics, respectively), have been generating large amounts of data increasingly used for biological modeling [Dunn and Ellis, 2005,Pandey and Mann, 2000].

The systematic analyses of the transcriptome and metabolome, and the correlation of the expression pattern of genes with the accumulation pattern of metabolites, have been productive ways to deduce the functions of genes [Oksman-Caldentey and Saito, 2005]. There are examples of soft-

ware applications that integrate data derived from transcriptomics and metabolomics, projecting them onto maps that support these systematic analyses [Thimm et al., 2004, Tokimatsu et al., 2005].

2.3.2 Genome-scale metabolic models

Current biochemical knowledge and the information collected from the annotation of the numerous genome sequencing projects allowed the development of metabolic models for several organisms [Feist et al., 2009, Henry et al., 2010].

The major aim is to develop computational models of biological systems so that their response to perturbations can be predicted, thus enabling important benefits to research efforts, for instance in drug discovery and Metabolic Engineering [Nielsen, 2001].

Metabolites are linked via anabolic and catabolic reaction networks. By using metabolic reconstruction based on genome annotation and biochemical knowledge, researchers can investigate metabolic networks and their structural features [Fell and Wagner, 2000, Jeong et al., 2000, Schuster et al., 2000].

The first Genome-Scale Metabolic Model (GSMM) was reconstructed, in 1999, for the organism *Haemophilus influenzae* [Edwards and Palsson, 1999], which consisted in set of 488 reactions connecting 343 metabolites, where 296 genes were considered.

Since then, GSMMs have been built for many well curated organisms. For *Escherichia coli*, the first GSMM has been released by Edwards and Palsson, named *i*JE660 containing 660 metabolic genes, 627 reactions and 438 metabolites [Edwards and Palsson, 2000]. This model was extended to originate the version *i*JR904 with 904 genes, 931 biochemical reactions,

and 625 metabolites [Reed et al., 2003]. This version also contains Gene-Protein-Reaction (GPR) associations that indicate which genes are needed for each reaction. In 2007, the model iAF1260 has been constructed containing perisplasmic reactions and metabolites, resulting in 1260 genes, a much higher number of reactions, 2077, and 1039 metabolites [Feist et al., 2007]. The version *i*JO1366 has been published in 2011, accounting for new biosynthetic pathways for cofactors, and updates for the biomass reaction and for the parameters relating to growth and non-growth maintenance [Orth et al., 2011]. This GSMM encompasses 1366 genes, 2251 reactions and 1136 metabolites. Also in 2011, the model *i*CA1273 has been published for the strain *E. coli* W (ATCC 9637) including 1273 genes, 2477 reactions and 1111 metabolites [Archer et al., 2011].

Concerning *Saccharomyces cerevisiae*, since 2003, several GSMM reconstructions have been made available. The first model, *i*FF708, contains 708 genes, 1175 reactions and 584 metabolites [Förster et al., 2003a]. In 2008, a consensus metabolic network reconstruction has been published [Herrgård et al., 2008]. From the 6183 genes in the yeast genome, 832 were incorporated in the generated model that contains 1857 reactions and 1168 metabolites. The GSMM *i*MM904 has been published in 2009 [Mo et al., 2009], containing 904 genes, 1412 reactions and 713 metabolites, it was based in the previous *i*ND750 [Duarte et al., 2004]. More recently, the version *i*TO977 has been released addressing 977 genes [Österlund et al., 2013]. This model expanded the consensus model and the version iIN800 [Nookaew et al., 2008] by applying gap-filling methods and introducing new reactions based on public databases and literature. The final version resulted in 1566 reactions and 1353 metabolites in the reconstructed model. Some published reconstructed genome-scale models are listed in Table 2.2.

Table 2.2: Summary of some published genome-scale metabolic models.

Organism	Name	Year	Genes	Reactions	Metabolites	Reference
			Bacteria			
Dinoroseobacter shibae	*i*Dsh827	2013	827	1488	1614	[Rex et al., 2013]
Escherichia coli	*i*JE660	2000	660	627	438	[Edwards and Palsson, 2000]
	*i*JR904	2003	904	931	625	[Reed et al., 2003]
	*i*AF1260	2007	1260	2077	1039	[Feist et al., 2007]
	*i*JO1366	2011	1366	2251	1136	[Orth et al., 2011]
	*i*CA1273	2011	1273	2477	1111	[Archer et al., 2011]
Haemophilus influenzae	*i*JE296	1999	296	488	343	[Edwards and Palsson, 1999]
	*i*CS400	2000	400	461	451	[Schilling and Palsson, 2000]
Helicobacter pylori	*i*CS291	2002	291	388	340	[Schilling et al., 2002]
	*i*IT341	2005	341	476	485	[Thiele et al., 2005]
			Eukaryota			
Homo sapiens	Recon 1	2007	1496	3311	2766	[Duarte et al., 2007]
	Recon 2	2013	1789	7440	5063	[Thiele et al., 2013]
Kluyveromyces lactis	*i*OD907	2014	907	1867	1476	[Dias et al., 2014]
Mus musculus	-	2005	473	1220	872	[Sheikh et al., 2005]
	-	2008	1399	2037	2179	[Nielsen, 2008]
	-	2009	724	1494	1287	[Selvarasu et al., 2009]

Continued on next page

Table 2.2 – continued from previous page

Organism	Name	Year	Genes	Reactions	Metabolites	Reference
	*i*MM1415	2010	1415	3726	2774	[Sigurdsson et al., 2010]
Saccharomyces cerevisiae	*i*FF708	2003	708	1175	584	[Förster et al., 2003a]
	*i*ND750	2004	750	1149	646	[Duarte et al., 2004]
	*i*LL672	2005	672	1038	636	[Kuepfer et al., 2005]
	*i*MH805775	2008	832	1857	1168	[Herrgård et al., 2008]
	*i*IN800	2008	800	1446	1013	[Nookaew et al., 2008]
	*i*TO977	2008	977	1566	1353	[Österlund et al., 2013]
	*i*MM904	2009	904	1412	713	[Mo et al., 2009]
Archaea						
Methanosarcina acetivorans	*i*VS941	2011	941	705	708	[Kumar et al., 2011]
	*i*MB745	2011	745	756	715	[Benedict et al., 2012]
Methanosarcina barkeri	*i*AF692	2006	692	619	558	[Feist et al., 2006]
	*i*MG746	2013	746	741	719	[Gonnerman et al., 2013]

Currently, there are many protocols to support GSMM reconstructions [Thiele and Palsson, 2010, Rocha et al., 2008]. However, the reconstruction of a metabolic model is a demanding task, and this was a motivation for developing several computational tools to assist the model curation and automatize some essential steps of the reconstruction process. Examples are Model SEED [Henry et al., 2010], Pathway tools [Karp et al., 2009], MicrobesFlux [Feng et al., 2012], RAVEN toolbox [Agren et al., 2013] and merlin [Dias et al., 2015].

The information used to guide the GSMM the reconstruction process is available in public repositories. The accessible data ranges from genome sequences to the annotation of functional capabilities of proteins and their interaction in metabolic pathways, stored in specific databases such as the ones listed bellow:

- NCBI, a set of databases for biomedical, genomic and other biological data [Wheeler et al., 2007];
- BioCyc, collects Pathway/Genome Databases (PGDBs) [Caspi et al., 2008];
- BRENDA, contains functional and molecular information of enzymes [Schomburg et al., 2002];
- KEGG, collection of information on genes, metabolites, reactions and pathways [Kanehisa and Goto, 2000];
- TCDB, responsible for maitaining the classification system for membrane transporter proteins [Saier et al., 2006];
- Uniprot, with functional information on proteins [Consortium and Others, 2011];

- MetaCyc, contains pathways involved in primary and secondary metabolism, and the associated compounds, enzymes, and genes [Caspi et al., 2008];

- GOLD, with information about genome and metagenome sequencing projects [Pagani et al., 2012].

For some of the repositories, the stored information is manually validated such as SwissProt (part of the UniProt database) and Metacyc.

To visualize reconstructed networks in a graphical fashion, tools like CellDesigner [Funahashi et al., 2006], Cytoscape [Shannon et al., 2003] or the metabolic network visualization plugin from *OptFlux* [Noronha et al., 2014], are available. The majority of the tools in systems biology such as the ones for reconstructing networks, executing phenotype simulation methods and visualizing networks, are compatible with the Systems Biology Markup Language (SBML) format [Hucka et al., 2003], that is an open and free XML-based format for representing biochemical reaction networks.

However, predictions using GSMMs fall short because regulatory events are precluded (although protein modifications, such as phosphorylation reactions can be included into the network). Thus, the search for the metabolic network that comprises the real cell state, the instantaneous active pathway, in a complex theoretical network under certain conditions, is still ongoing.

2.3.3 Constrained-based modeling

A number of factors may influence cell's functioning, since it relies on integrated mechanisms, such as metabolism, transcriptional (and other types) of regulation and signal transduction. To achieve confident predictive mod-

els, the understanding of how these mechanisms interact with each other is of great importance.

Although several approaches to model cell behaviour as a whole have been put forward [Tomita, 2001], the underlying models still suffer from some limitations and generally lack kinetic and regulatory information. The former type of information offers a comprehensive dynamical and quantitative characterization of the systems, based on biochemical knowledge, but until now, there are no dynamic models in a genome scale that can be applied to support ME strategies, due to the high difficulty in gathering kinetic data [Smallbone et al., 2010, Smallbone et al., 2013]. Moreover, for such scale, their simulations would be computationally very expensive [Ederer et al., 2010].

Constraint-Based Modeling (CBM) (Figure 2.1) emerged as an alternative and has been extensively applied in the investigation of biochemical reaction networks [Bordbar et al., 2014]. In this field, several efforts have been put forward in the development of phenotype prediction methods that can take as inputs both genetic and environmental conditions. One of such methods is the traditional Flux Balance Analysis [Papoutsakis, 1984, Savinell and Palsson, 1992b, Savinell and Palsson, 1992a, Kauffman et al., 2003]. Through the combination of these constrained-based methods with reconstructed GSMMs, powerful tools are made available to significantly support the rational design of optimized strains.

The central postulation that supports CBM methods is that, despite the fact that the concentration being not constant to every intracellular metabolites at all times, the turnover of the pools of most metabolites is extremely fast, so their concentrations quickly adapt to new levels, even under high environmental changes imposed to the cells. Therefore, it is not

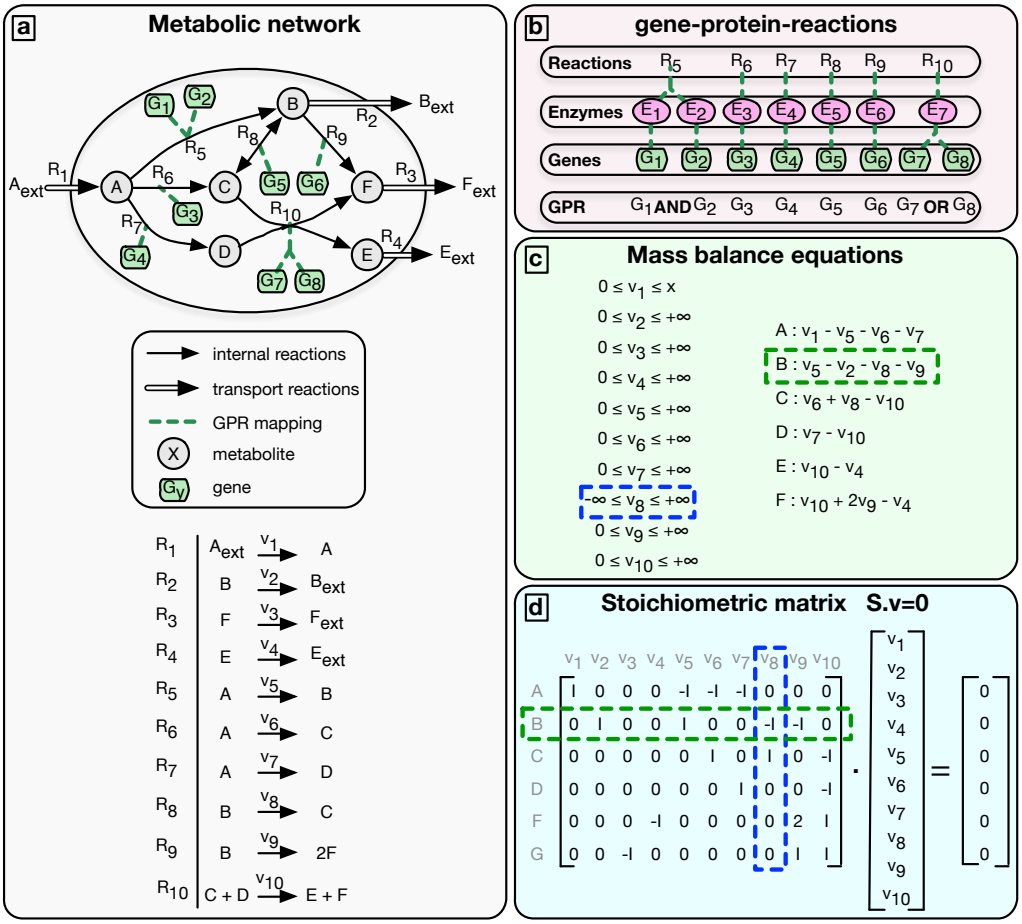


Figure 2.1: Illustration of a pseudo constraint-based metabolic genome-scale model. The example refers to the common elements in this kind of modeling including: a) a network model where the reversibility of each reaction is indicated by an arrow; for each reaction R_i, a flux variable v_i is defined; b) Gene-Protein-Rules, defining relationships between reactions, enzymes and their encoding genes; and c) the constraints defined over flux values (a represents the limiting uptake rate for the substrate A) and the steady-state mass balance equations for each intracellular metabolite; d) stoichiometric matrix for the system.

far from reality to assume that intracellular metabolites are in a stationary state. This assumption implies that there is no accumulation of metabolites, and thus all fluxes leading to their formation and degradation are mass-balanced [Varma and Palsson, 1994]. This principle can be translated in the form of the following equation:

$$S.v = 0 \tag{2.1}$$

where S is an $m \times n$ stoichiometric matrix, for a set of m metabolites and a set of n reactions, and v is the vector of n reaction fluxes (Figure 2.1 d). For each reaction of a metabolic network, bounds for their minimum and maximum values can be configured to reflect thermodynamic feasibility, i.e., reaction directionality, and flux capacity of the reactions (Figure 2.1 c).

The information on GPR associations provided in GSMMs (Figure 2.1 b) allows to build more accurate methods for both phenotype prediction. The traditional representation of such rules is made through the use of Boolean logic expressions, with permits to model how the reactions are encoded by their genes through a set of ANDs and ORs operations. Through these expressions, it is possible to represent, for instance, protein complexes and iso-enzymes.

2.3.4 Phenotype constrained-based simulation

The available biological constraints, such as steady-state, reaction reversibility or environmental conditions, configure a flux cone of admissible flux distriburions [Kauffman et al., 2004], which represents the underdetermination of the system. Due to the high number of degrees of freedom in metabolic systems, especially when dealing with GSSMs, the systems are normally

transformed in an optimization problem, and the objective function of such problems is moulded based on biological assumptions. A common approach is to assume that organisms evolved in such a way that favours some particular objectives.

The most known method, FBA, describes these assumptions by configuring an objective function that maximizes growth [Edwards et al., 2001, Ibarra et al., 2002,Segre et al., 2002] or the production of a certain metabolite [Schuetz et al., 2007], minimizing the global energy expenditures of the cell. The linearity of the constraints and the objective function allow the system to be formulated as a Linear Programming (LP) problem.

However, FBA does not account for some relevant biological phenomena, such as the consequences of regulatory constraints under certain environmental and media conditions. Nonetheless, there are variants such as regulatory FBA (rFBA) [Covert et al., 2001] and steady-state FBA (SR-FBA) [Shlomi et al., 2007] which consider transcriptional regulation through appropriate additional constraints.

The objective of maximizing growth can be seen as credible when dealing with wild type strains, but for mutant strains this is not the case. In this context, Regulatory On/Off Minimization (ROOM) has been developed [Shlomi et al., 2005], which aims at minimizing the number of significantly changed fluxes, obtained from genetic modifications, with respect to a reference flux distribution. Since this method utilizes binary terms in its objective function, a Mixed-Integer Linear Programming (MILP) problem is configured, which results in increased complexity.

Based on the same motivation as ROOM, Minimization of Metabolic Adjustment (MOMA) deals with the weight of shifting between operating regions [Segre et al., 2002]. Instead of maximizing growth, MOMA relies on

the minimization of the sum of the squared differences between the reference (typically wild type) and mutant flux distributions through Quadratic Programming (QP). An alternative implementation, but with the same purpose, is Linear MOMA (LMOMA) [Becker et al., 2007b], which replaces the quadratic configuration of MOMA by a linear objective function, by minimizing the sum of the absolute values of the difference between the fluxes in the wild type and mutant strain.

With the objective to overcome some of the limitations in MOMA, Minimization of Metabolites Balance (MiMBl) has been introduced [Brochado et al., 2012]. The target of this approach is not to optimize fluxes, but rather, metabolite turnovers. This methodology allows suppressing the problems originating from the sensitivity of the solutions to the stoichiometric representations.

All the aforementioned methods share the issue that multiple solutions may exist for the same optimal objective value. In this context, the Parsimonious enzyme usage Flux Balance Analysis (pFBA) method has been developed [Lewis et al., 2010]. The method is based on the assumption that the strains with minimal enzyme usage will have a competitive advantage.

A different methodology aiming at characterizing the constrained system has been developed, named Flux Variability Analysis (FVA) [Mahadevan et al., 2003]. It can be used for both computing tight bounds for the fluxes in the model, and to check how a specific flux may vary in optimal or sub-optimal solutions. Its application is useful, for instance, to check the robustness of a flux distribution in a mutant strain simulation regarding its production capability of a given metabolite.

The mathematical formulations of the phenotype methods are omitted here, since these provided in detailed in chapter 3.

2.4 Stoichiometric metabolic flux analysis

Metabolic Flux Analysis (MFA) aims at calculating (or estimating) *in vivo* fluxes from metabolite uptake and secretion rates, through the utilization of a system of linear equations obtained from reaction stoichiometry [Papoutsakis, 1984]. Supplementary to the extracellular fluxes information, other physical, chemical and biological information can be added to the analysis, such as, thermodynamic directionality, enzymatic capacity, gene regulation, among others, to constrain the feasible fluxes for a given physiological condition.

Additionally, MFA can provide insights on the identification of branch point control in cellular pathways, identify alternative metabolic pathways and calculate maximum theoretical yields from a certain substrate [Stephanopoulos et al., 1998]. Moreover, metabolic flux split ratios might be fixed to guide this type of analysis.

Details about the concepts and formulations underlying stoichiometric MFA will be given in the next chapter, where algebraic methods for resolving determined and overdetermined systems, and methods based on linear and quadratic programming, are introduced.

2.4.1 Thermodynamics-based metabolic flux analysis

Since traditional stoichiometric MFA does not account for the thermodynamics of reactions, flux values can be assigned to several reactions even if they are not thermodynamically feasible [UkáKim et al., 2008]. Additional constraints, based on the first and second laws of thermodynamics can be added to the mass balances, so the appearance of unfeasible reactions and metabolic loops in the returned flux distributions disappear. This has given

rise to thermodynamics-based metabolic flux analysis (TMFA) [Henry et al., 2007] which allows to calculate flux distributions with additional information on thermodynamics in the form of constraints, using MILP formulations.

2.4.2 Hybrid metabolic flux analysis

Hybrid modeling can be applied to investigate steady-state metabolic models to compensate missing information, such as, incomplete stoichiometry, absence of knowledge about how central metabolic activities are linked with recombinant gene expression, or how metabolic regulation is exerted by environmental stimuli [Carinhas et al., 2012].

A recent approach, named Hybrid MFA [Carinhas et al., 2011], extends a previous work [Bernal et al., 2009] for rational design for *Baculovirus* production in insect cells. It combines stoichiometric MFA with multivariate statistical analysis to fulfill the absence, or ill-defined, mechanistic detail related to cellular functions in metabolic networks, such as, the synthesis of a complex recombinant product.

For a set of measured fluxes, MFA is performed to estimate the flux distribution which, together with target measurements, e.g. productivity, is then used for finding a multivariate statistical regression sub-model. The result is a vector of regression coefficients that shows the correlation between each flux and cell productivity. This allows for further productivity predictions after metabolic manipulations.

2.4.3 Software tools for stoichiometric metabolic flux analysis

Numerous tools for general-purpose constraint-based analysis have been developed in the last few years. In Additional file 1, in supplementary material, the main features of these tools are summarized in a comparative table.

All these tools offer the possibility to perform FBA, however few provide flexible and easy-to-use tools for defining additional constraints as those coming from measured or otherwise known fluxes or metabolic flux ratios. For example, in this GSMMs' era, it is important not only to provide users with the capacity to compute flux distributions from experimental data for determined systems, but also to evaluate the impact of the measurements for systems that, although remaining underdetermined, see their solution space reduced. Moreover, it is important to be able to combine different types of experimental data or other available knowledge, such as flux measurements or ratios for intracellular fluxes with exchange fluxes or to easily perform these calculations for both wild-type and mutant strains.

Not many tools exploit the various derivations of FBA that enable the application of tailored approaches to specific problems. *COBRA Toolbox v2.0* [Schellenberger et al., 2011] is a tool that contains many relevant features, but requires the commercial software Matlab. *CellNetAnalyser* [Klamt et al., 2007], the successor of *FluxAnalyzer* [Klamt et al., 2003], provides key features for MFA, such as the classification of models according to (in)determinacy and redundancy. However, it is also built over Matlab and is itself a commercial tool, although it allows an academic license.

A few software tools, like *MetaFluxNet* [Latendresse et al., 2012a], which is incorporated in *Pathway tools* [Karp et al., 2009], and *SBRT* [Wright

and Wagner, 2008], execute matrix operations to calculate flux distributions without the application of optimization methods. Also, importing metabolic flux ratios to constrain MFA methods is a feature not supported by existing tools, although *SBRT* and COBRA accept the introduction of linear combinations of reactions to constrain the systems.

2.4.4 Applications of stoichiometric metabolic flux analysis

Stoichiometric MFA, as it stands, can be successfully applied to different case studies where the number of equations is enough to solve the systems, and also for the cases where the systems are undetermined and no additional data from tracer experiments is available [Zamorano et al., 2010]. However, the advantages of such experiments will be discussed in the some of the next sections of this chapter.

Least-squares MFA based methods, and its variants, have been used for both determined and overdetermined systems, as the analysis of hybridoma cells [Sidorenko et al., 2008] and human embryonic kidney cells [Henry et al., 2005, Nadeau et al., 2000]. Also, the fluxes in *Penicillium chrysogenum* have been calculated through the use of substrate uptake rates and formation rates of products and macromolecules during fed-batch cultivations [Jørgensen et al., 1995].

In 2006, the metabolism of lactate in chinese hamster ovary (CHO) cells has been investigated using both glucose and galactose as carbon sources, where specific consumption and production rates of the main compounds were used [Altamirano et al., 2006]. In the same year, measured fluxes obtained in batch cultivation conditions for numerous mutant strains of *Escherichia coli* for the production of succinate were used to estimate their

intracellular fluxes and pinpoint critical flux split ratios [Sánchez et al., 2006].

The metabolic flux distribution in *Clostridium butyricum* has been investigated during fermentative hydrogen production where measured rates from glucose uptake and from the production of soluble compounds have been used to constrain the simulations in *MetaFluxNet* software [Cai et al., 2010].

MFA based on linear programming optimization has been applied to GSMMs for several organisms, such as, *Mus musculus* [Sheikh et al., 2005]. *Saccharomyces cerevisiae* and CHO cells. For instance, in *Saccharomyces cerevisiae* [Bro et al., 2006], ME strategies were investigated to improve bioethanol production by fixing the growth rate to a fixed value, or limiting the glucose and xylose uptake rates and manipulating internal fluxes (setting them to zero or changing their bounds). Regarding CHO cells, a study has been performed, where consumption and production rates of extracellular glucose, amino acids, lactate, ammonia and oxygen were used to constrain FBA simulations [Martínez et al., 2013].

2.5 Metabolomics

Through the combination of metabolic networks with metabolome analysis, it is possible to create links between the different metabolites in the metabolic network. This can also be used to calculate the fluxes through different parts of the metabolism. It is, hereby, possible to correlate metabolite levels and fluxes, which enables identification of key control points in the metabolism.

Metabolomics is a rapidly emerging field in postgenomic research. It is

the ambitious analytic approach aiming at the unbiased identification and quantification of all low molecular weight compounds present in a specific biological sample from an organism, grown under given conditions [Villas-Boas et al., 2007, Fiehn et al., 2000]. This is different from metabolic fingerprinting, which focuses on collecting and analyzing data from crude metabolite mixtures to rapidly classify samples instead of separating individual metabolites by physical parameters. Thus, metabolomics is best suited for the investigation of metabolic networks via the quantification of individual metabolites, without a bias concerning the choice of targets to be analyzed [Weckwerth, 2003].

The complete set of small molecules, named metabolome, represents the ultimate phenotype of cells, and is the response to the perturbation of gene expression and the modulation of protein functions, which are caused by the environment or genetic mutations. Moreover, the metabolome can influence gene expression and protein function. Therefore, metabolomics plays a key role in understanding cellular systems, recently emerging as an important tool in the functional annotation of genes and enzymes [Bino et al., 2004].

In addition, metabolomics has gained importance in biotechnology applications, as exemplified by quantitative loci analysis, prediction of food quality, and evaluation of genetically modified crops [Saito and Matsuda, 2010]. It has a number of industrial applications such as the identification of metabolic engineering targets in *Escherichia coli*, yeast and algae [Park et al., 2005] and fermentation process improvement [Kern et al., 2007]. Systems biology driven by metabolome data will aid in deciphering the secrets of various cell systems and their application to biotechnology.

Metabolome analysis provides a number of advantages over transcriptomics and proteomics, since it represents the information that more closely

reflects the phenotype, i.e. the activity of the cell.

As compared with genomics (nucleic acids) and proteomics (peptides and proteins), metabolomics deals with a range of chemically more diverse molecules. The large variations in the relative concentrations of metabolites also make metabolite analysis more complicated. For instance, the number of metabolites present in the plant kingdom is estimated to exceed 200000, an enormous number derived from the high compound structural diversity [Weckwerth, 2003].

The amount of data generated by metabolomics experiments has demanded efforts from the Bioinformatics community, in the tasks of data storage and organization, as well as in the development of efficient data analysis methods [Mendes, 2002]. Further advancements in metabolomics technologies are needed to cover the wide array of phytochemicals and increase the spatial and temporal resolution of metabolome analysis. Strengthening existing spectral databases and sophisticated informatics for more accurate annotation of metabolite signals is necessary, and these efforts must be internationally harmonized and publicly available.

2.5.1 Analytical techniques

Towards the use of metabolome analysis, and the integration of these data with metabolic models, there is a need for an accurate quantitative analysis of specific metabolites obtained under well-defined conditions, not only in terms of measurement of relative levels, but also measurement of actual concentrations of the different metabolites [Nielsen and Oliver, 2005]. To accomplish such objectives, it is of the major importance to define appropriate data standards, develop standard analytical methods and appropriate libraries of mass spectra to be used by analytical methods [Villas-Boas et al.,

2007].

Also, there are a number of analytical tools that can be used to perform the comprehensive identification and, particularly, the accurate quantification of metabolites, including gas and liquid chromatography or capillary electrophoresis, coupled with mass spectrometry (GC, LC or CE-MS) and nuclear magnetic resonance spectroscopy (NMR). Such methods have attained a technical robustness that is comparable or even better than conventional mRNA or protein profiling technologies [Fiehn et al., 2000,Fiehn, 2002].

The first developments in metabolomics were achieved with the use of NMR. For instance, ^{13}C-NMR spectroscopy is useful for metabolic studies, which requires knowledge on the covalent structure of compounds and the corresponding unambigous resonance assignments of ^{13}C, and in some cases both the isotopes ^{13}C and ^{1}H as in two-dimensional nuclear magnetic resonance spectroscopy (2D NMR). NMR spectroscopy is a powerful method to unravel the covalent structures of organic molecules [Szyperski, 1998], and is often used with other techniques to achieve such goal [Croasmun and Carlson, 1994]. Its elucidation capabilities are higher than MS techniques and its profiles might contain qualitative and quantitative information about a variety of small molecules in a given sample.

However, MS has higher sensitivity than NMR [Lee and Kerns, 1999,Wu, 2000] and can measure all isotopes, while NMR only allows for detecting isotopes with an odd number of nucleons such as ^{1}H, ^{13}C, ^{1}5N and ^{3}1P [Klein and Heinzle, 2012]. In the case of GC-MS, it has gained some popularity given its stable retention time, robust protocols for sample preparation and ability to identify a broad set of metabolites such as amino acids, organic acids, and sugars with good sensitivity by employing a chemical deriva-

tization of hydrophilic metabolites [Dunn and Ellis, 2005]. Moreover, a good reproducibility of GC-MS analyses in terms of the sharp separation of metabolites by GC, combined with the stable ionization achieved by electron impact (EI) ionization is preferable for analyzing complex samples extracts. Metabolomics based on GC-MS has been widely used in different type of matrices, such as, microbial [Birkemeyer et al., 2005], fungi [Nilsson et al., 1996], plants [Chen et al., 1998,Fan et al., 2001], among others [Villas-Bôas et al., 2005, Koek et al., 2011].

A technical challenge in GC-MS metabolomics has been processing complex raw GC-MS chromatogram data [Kopka, 2006]. While the creation of multiple fragment ions by EI is convenient for metabolite identification, this makes it difficult to separate each metabolite signal from overlapping peaks in the GC-MS chromatogram. Currently, several peak-picking software packages equipped with sophisticated peak deconvolution functions are available such as AMDIS for GC-MS data interpretation [Davies, 1998, Stein, 1999], MET-IDEA for data extraction of GC-MS-based metabolomics [Broeckling et al., 2006]. and MetAlign that performs full scan LC-MS and GC-MS comparisons [Lommen and Kools, 2012].

2.5.2 Data repositories and analysis

Various types of metabolomics data are generated daily and accumulated in databases of metabolomics laboratories. The appearance of comprehensive databases for metabolomics data, to make these data available among the laboratories and the transcriptomics community, has been limited by the several distinct experimental platforms and the lack of a recognized standard for the experiments. Ideally, all the metabolic profiling analysis should be performed through standard methods with the same instruments

and under the same conditions. But a high compatibility of metabolome data among laboratories is impractical, since the majority of data have been acquired with different types of machines and in customized conditions.

Nevertheless, recently some efforts have been put forward [Kopka, 2006] and even a human metabolome database has been created (http://www.hmdb.ca). In the case of GC-MS, there is an effort to share standard compound data among laboratories [Scholz and Fiehn, 2006] and relating to NMR data, there are databases of chemical shifts of standard compounds and data repositories, such as the spectral database at the Platform for RIKEN Metabolomics [Akiyama et al., 2008].

In terms of data analysis, the route followed in the metabolomics field is quite similar to the one taken on the fields of DNA microarray or proteomic data analysis. In fact, the structure of the data and the analysis problems bear many similarities. The main efforts in data analysis have concentrated in methods for data pre-processing (in many cases dependent on the platform used), unsupervised techniques (principal components analysis, hierarchical and other clustering algorithms) and supervised machine learning algorithms (e.g discriminant analysis, partial least squares, neural networks, tree or rule induction, evolutionary algorithms) [Goodacre et al., 2004,Kell, 2004]. In the last few years, a number of platforms have become available to perform these analyses, such as MeltDB or MetaboAnalyst [Neuweger et al., 2008, Xia et al., 2009].

Over the last few years, metabolomics has been mainly used as a functional genomics tool and the field has been mostly driven by the study of applications with plants (higher plants have the ability to synthesize a vast array of compounds that differ in chemical complexity and biological activity, playing indispensable roles in chemical defenses against biotic and

abiotic stresses). However, metabolome analysis is able to bring important contributions to modeling tasks in Systems Biology [Kell, 2004, Weckwerth, 2003] and can also be successfully applied in microbial studies [van der Werf et al., 2005]. Indeed, it has been shown that metabolomics can provide a very useful tool in metabolic model reconstruction [May et al., 2008] and it can be combined with methods from pathway analysis [Förster et al., 2002] for model validation.

2.6 Fluxomics

Intracellular metabolic rates (fluxes) represent integrative information of cell phenotype, being a function of gene expression, translation, protein modifications and protein-metabolite interaction [Nielsen, 2003]. The quantitative analysis of *in vivo* carbon fluxes in the intracellular activities of enzymes and pathways is a valuable resource for examining cell's phenotypes and mechanisms of cellular regulation [Bailey, 1999, Hellerstein, 2003, Wittmann, 2007].

Several studies with fluxome analyses have been made, such as providing rational alternatives to improve lysine production in *Corynebacterium glutamicum* [Becker et al., 2005, Becker et al., 2007a]. Essentially, it has been applied since long time ago to biomedicine [Lapidot and Nissim, 1980, Nissim et al., 1981, Tsalikian et al., 1984, Nissim and Lapidot, 1986, Shen et al., 1999], and now it is applied broadly in several different fields in physiology and biotechnology [Wittmann and Heinzle, 1999, Christensen and Nielsen, 2000a, Dauner et al., 2001a, Shimizu, 2002, Fischer and Sauer, 2003b, Wittmann, 2007, Sengupta et al., 2011, Xiong et al., 2010]

Although fluxome analysis by itself is not so accurate to understand

gene functions due to its integrative character, which should be combined with other analysis such as transcriptomics, proteomics or metabolomics, it can give important insights in the field of functional analysis, since, for example, it is possible to investigate the response of a metabolic network fluxes to gene deletions [dos Santos et al., 2003, Förster et al., 2003b].

External fluxes, i.e., fluxes of substrates going inside the cell and fluxes originating products secreted from the cell, can be measured and used in flux analysis [Vallino and Stephanopoulos, 1993, Jørgensen et al., 1995]. However, intracellular fluxes cannot be directly measured, since there is no straight method to detect them and it is not possible to assess the reaction rates directly from metabolite concentrations, and the opposite is also not possible to be done.

A molecule can be configured as an isotopic tracer, if some of the atoms belonging to its chemical structure are replaced by a different isotope, which can be either stable, such as, ^{13}C, ^{2}H, ^{18}O and ^{15}N, or radiative, such as, ^{14}C and ^{3}H.

If cells are fed with labeled substrates, e.g, glucose labeled with the stable isotope of carbon (^{13}C), the ^{13}C enrichment pattern tracked throughout the cell's metabolic network and metabolic fluxes can be determined through a number of computational procedures . The target analyses in fluxomics are the same as metabolomics, however in a distinct fashion, since the attention is focused in the isotopic distribution of the metabolites, and not on their concentration [Zamboni, 2007]. The fluxes between metabolites provide time-dependent information on how cells operate. When it comes to the order of seconds to minutes, the fluxes response to environmental perturbations is extremely influenced by kinetic regulation, which by itself is a function of metabolite concentrations and the kinetic properties

of the related enzymes. For longer periods, such as growth under a given environment or genetic variants, the fluxes reflect the effects of genetic and metabolic regulation [Zamboni and Sauer, 2005].

2.7 ^{13}C metabolite enrichment

2.7.1 Labeling and metabolic stationary state

Although there are several developments in instationary ^{13}C-MFA [Wiechert and Nöh, 2005, Nöh et al., 2006, Nöh and Wiechert, 2006, Nöh et al., 2007, Young et al., 2011, Wiechert and Nöh, 2013, Young, 2014, Kajihata et al., 2014], the most common scenario is to perform ^{13}C-MFA under metabolic and isotopic steady-state conditions, that is, metabolic fluxes and the labeling of metabolites and amino acids do not change along time. These requisites impose that the duration of the experiment must be much longer than the inverse of the growth rate [Nöh and Wiechert, 2006, Wiechert and Nöh, 2005]. Such conditions can be achieved in continuous cultures over three generation times [Tang et al., 2007].

Therefore, the length of the experiment is highly critical to execute an accurate measurement of the data. After the feeding with a ^{13}C substrate, the labeling patterns of the intracellular metabolite pools switch to a transitory phase until the isotopic distribution is in steady-state as the metabolic fluxes are. When the system contains metabolite pools with very large sizes, such as the glutamate pool in *Corynebacterium glutamicum* [Tesch et al., 1999], the stationary time might be equivalent to the bioreactor residence time [Wiechert, 2001].

Generally, chemostat culture is the preferred one to guarantee the metabolic and isotopically stationary state. For instance, to analyze microbial behav-

ior when there are changes in the environment, e.g. substrate concentration, or in in the specific growth rate, chemostate is actually the best mode to operate. Nevertheless, in industry, the production processes are performed in batch or fed-batch, where the cell switches its state from the growth phase to the stationary phase [Iwatani et al., 2008].

For macromolecular biomass constituents, such as, DNA, RNA, proteins and cell wall, the time span to reach an isotopic steady-state is much higher when compared with intracellular intermediate metabolites. But, normally, the intracellular intermediates are not measured, and, therefore, their labeling information is deduced from biomass components that are synthetized by these precursors [Szyperski, 1995,Eisenreich et al., 1993]. To execute this retrobiosynthetic labeling deducing approach, the isotopically equilibrium must be obtained for both intermediate metabolites and macromolecules. But not every microorganism takes too long to achieve the wash-out effect, that is the replacement of the original unlabeled cells by cells with the macromolecular compounds in isotopic steady-state [van Winden et al., 2001], where the time span can range from two to four bioreactor residence times [Wiechert, 2001].

There are scenarios where it is not possible to achieve an isotopic equilibrium. For example, with industrially relevant media, it has been shown that it was not possible to achieve the isotopic steady-state for amino acids and lactic acid in Chinese hamster ovary cell cultures [Deshpande et al., 2009], due to the exchange between intracellular and large extracellular amino acid pools [Sengupta et al., 2011]. Consequently, it was not possible to use the labeling patterns of amino acids to infer the labeling pattern of phosphorylated sugars, as typically done for the tasks of stationary ^{13}C-MFA [Schmidt et al., 1997, Wiechert et al., 2001].

2.7.2 Measuring the labeling pattern of amino acids and central intermediates

In ^{13}C tracer experiments, proteinogenic amino acids are usually used since they are much more abundant than intermediate metabolites, are stable, and allow to obtain the labeling pattern of their precursors. Sixteen out of the twenty existing amino acids can be retrieved after biomass hydrolysis [Dauner and Sauer, 2000]. The biomass hydrolysis is executed after the cell mass is harvested by centrifugation, and the macromolecular constituents are isolated from the broth next to cell disrupting [Wiechert, 2001].

Since the relationship between precursor metabolites and amino acids are known, it is possible to infer the labeling patterns of the precursors through the measured labeling in the amino acids [Gombert et al., 2001, Maaheimo et al., 2001, dos Santos et al., 2003, Yuan et al., 2010].

Measuring the intermediate metabolites in the central carbon metabolism offer a significant extension on the application of ^{13}C-MFA. A disadvantage when dealing with the measurenment of amino acids is that a minimal media is necessary to prevent the influence of a bias in the isotopomer measurements. But, in many cases, there is the need for nutrient supplements and only the proteinogenic amino acids produced in the central metabolism, which are not excluded from the medium can be analyzed in ^{13}C-MFA [Christiansen et al., 2002].

The labeling enrichment of central intermediate metabolites can be measured through rapid sampling and quenching of biomass, and afterwards its extraction, separation and highly sensitive detection by LC-MS [Lange et al., 2001, van Dam et al., 2002]. However, the labeling distribution throughout the central metabolites was hampered due to their extreme turnover rates and low concentrations [Szyperski, 1998, Dauner and Sauer,

2000].

Nevertheless, there are efforts to measure intermediate metabolites, such as in [Winden et al., 2005], where a new method based on LC-MS for determining the ^{13}C distribution on intermediate metabolites of the central metabolism has been implemented. The method also allows determining the labeling pattern of intermediates that are not precursors of biomass constituents. Moreover, it does not rely on assumptions about the presence or absence of alternative paths to a certain amino acid or the compartmentalization of pathways in eukaryote species.

Unlike proteinogenic amino acids, intracellular metabolites are constantly renewed, reflecting the actual metabolic flux environment, and are mostly used in non stationary ^{13}C-MFA rather in the traditional stationary analysis [Grotkjæ r et al., 2004].

2.7.3 Strategies for substrate labeling

The adopted strategy in terms of substrate labeling in ^{13}C tracing experiments is of the most relevance (as shown in Figure 2.2), since it highly influences the determination of the metabolic flux distribution [Möllney et al., 1999].

The type of labeling strategy varies from a single labeled substrate, such as [1,2-^{13}C] glucose, or a mixture of tracers from the same substrate like [1-^{13}C] mixed with [U-^{13}C] glucose, to a combination of different labeled compounds, such as [U-^{13}C] glucose with [U-^{13}C] glutamine. For the last case, an example happens in mammalian species, where the use of multiple substrates is required, thus the selected tracer will extremely influence the quality of the metabolic fluxes calculations [Metallo et al., 2009].

The position of the labeled carbon of a molecule is not discriminated in

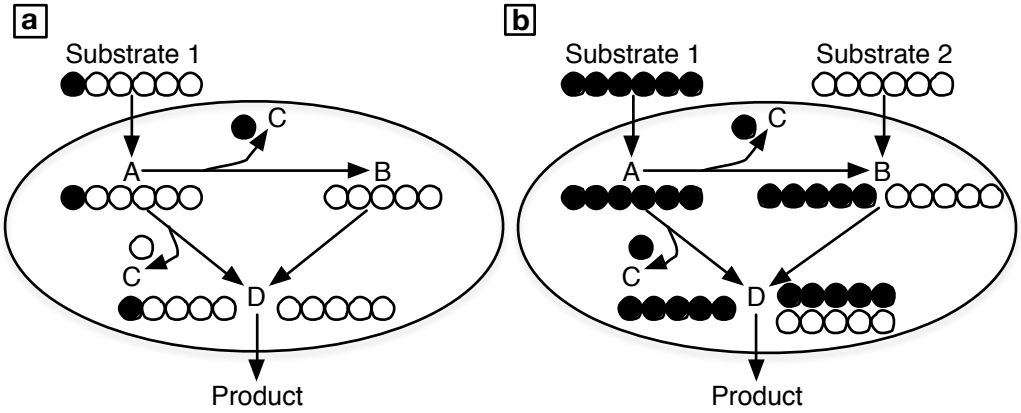


Figure 2.2: Illustration of two substrate labeling strategies. a) The labeled carbon of the intracellular metabolite D can be obtained only by the left-hand pathway linking the substrate to it. The labeling pattern in the pool of D will determine the fraction of flux deriving its carbon skeleton from the distinct pathways. b) Since substrate 1 is fully labeled, the positions of the labeled atoms are not useful to trace how the labeling is transferred to metabolite D, however if a non-labeled substrate is added, the fraction of labeling between the different pathways can be determined.

terms of mass isotopomers, i.e., they vary in the number of heavy atoms but are not dependent of which carbon positions is labeled. The use of a mixture of unlabeled glucose with [U-^{13}C] is especially useful to examine the fluxes downstream phosphoenolpyruvate and some exchange fluxes resulting in C-C bond cleavages [Fischer et al., 2004].

Relating to positional isotopomers, the number and positions of the ^{13}C atoms in the molecules are precisely known. This kind of labeling is particular useful for investigating the labeling distribution in the upper part of metabolism, such as, glycolysis, pentose phosphate pathway and Entner-Doudoroff pathway. For example, using a positional labeled glucose in its first carbon atom, but not with a fully labeled glucose, the reactions in the pentose phosphate pathway and Entner-Doudoroff pathway could be accurately differentiated [Fischer et al., 2004, Suthers et al., 2007a].

When it comes to use a positional labeled substrate in a ^{13}C tracer experiment, although NMR has low sensitivity, it is more advantageus since

it allows to determine positional labeling information on the isotopomers by discriminating between identical molecules [Stephanopoulos, 1999, Winder et al., 2011]. On the other hand, GC-MS is high sensitive and allows to obtain the fraction of a certain molecular fragment contain that contains a certain number of labels [Nanchen et al., 2007].

However, a better metabolic flux resolution may be obtained by using both uniform and positional labeled substrates either combined as a substrate mixture in the same experiment [Wittmann and Heinzle, 2002, Kiefer et al., 2004, Fischer et al., 2004], or separately in different experiments [Fischer et al., 2004, Krömer et al., 2004]. In [Araúzo-Bravo and Shimizu, 2003], it was also demonstrated that the best labeling strategy for *Synechocystis* sp. PCC6803 is a mixture of 70% unlabeled, 10% [U-^{13}C] and 20% [1,2-$^{13}C_2$] glucose to achieve a most reliable flux distribution.

2.7.4 Parallel labeling experiments

Parallel isotope labeling experiments refers to the activity of conducting more than one tracer experiment in parallel, under the same conditions, but with different feeding in terms of labeled substrate. In order to minimize biological variability, all the experiments begin with the same seed culture, so that biological variability can be minimized [Ahn and Antoniewicz, 2013]. Some advantages arise from performing this approach of tracer experiments, as reducing the time stamp of tracer experiments through the use of various entry points of isotopes, shaping the experiments to better elucidate certain metabolic fluxes, validating reconstructed network models, and also enhancing the tasks of ^{13}C-MFA in the cases where there are few measurements [Crown and Antoniewicz, 2013a].

Recently, works using complementary parallel labeling experiments tech-

nique for metabolic flux analysis (COMPLETE-MFA) [Leighty and Antoniewicz, 2013] have been published. Leighty and co-authors applied the parallel labeling strategy to *Escherichia coli* to validate stated assumptions for the wild type [Leighty and Antoniewicz, 2012]. They used different configurations of mixtures of [U-^{13}C] glucose and unlabeled glucose, ranging the percentage of the uniform labeled glucose from 0% to 100%, where all the experiments were performed at isotopic steady-state with the same inoculum as starting point.

Different isotopic tracers have been used to determine the fluxes in the Chinese hamster ovary (CHO) cells, where [1,2-^{13}C]glucose and [U-^{13}C] glutamine were used [Ahn and Antoniewicz, 2013]. Cells of *Clostridium acetobutylicum* ATCC 824 have also been grown in parallel cultures, but with the tracers [U-^{13}C] glucose and [1-^{13}C]glucose to systematically quantify its intracellular fluxes [Au et al., 2014]. In [Crown et al., 2015], the authors have performed a massive set of parallel experiments to compare the performance of several isotopic tracer configurations for *Escherichia coli*. Additionally to the common isotopic tracers configurations, such as the mixture of [1-^{13}C]glucose and [U-^{13}C]glucose, and [1,2-^{13}C] glucose, they have used novel tracers as [2,3-^{13}C]glucose, [4,5,6-^{13}C]glucose, [2,3,4,5,6-^{13}C]glucose and a mixture of [1-^{13}C]glucose with [4,5,6-^{13}C]glucose.

2.8 Stationary ^{13}C Metabolic flux analysis

The configuration of a proper objective function in optimization-based constraint methods, such as FBA, is not straightforward [Schuetz et al., 2007]. LP based approaches can have multiple solutions, i.e. different flux distributions that satisfy all the constraints and have the same optimal value

for the objective function [Mahadevan et al., 2003]. Additional limitations arise from the reliance on balances for cofactors, such as NADH and NADPH [Christensen and Nielsen, 1999] and from the presence of reversible reactions and futile cycles [Stephanopoulos et al., 1998, Wiechert and de Graaf, 1997].

Some of these problems can be attenuated by the experimental measurement of certain metabolic fluxes, such as the rates of formation / consumption of some compounds (e.g. substrates as glucose and O_2, or products as CO_2, acetate or lactate), which is the basis of the traditional approaches for performing stoichiometric MFA, but these are rarely able to turn the system into a determined one, allowing the unique determination of all fluxes in the model.

Moreover, although constrained-based modeling analysis may propose clues about the best set of knockouts to achieve a desired objective, there is a variety of possibilities in the actual response of the cell in terms of metabolic flux distribution, ranging from virtually no change, to overall unpredicted side effects on the control mechanisms [Copeland et al., 2012]. Also, cellular metabolism may operate in a suboptimal mode, and the optimal solutions generated by methods based on objective functions may be different of the real state of the cell metabolism [Palsson et al., 2003, Fischer and Sauer, 2005].

An alternative approach is to feed cells with carbon labeled substrates. Stable isotopes have been used for analysing metabolism since 1935, where recently developed deuterium labeling methods have been used to investigate metabolic processes [Schoenheimer et al., 1940]. In the early 50s, carbon fragments in the metabolism of rat liver were elucidated using ^{14}C radiative isotopes [Weinman et al., 1950, Strisower et al., 1951]. However,

^{13}C has been proven to be preferable to perform such analysis, rather than ^{14}C. In this context, emerges the concept of ^{13}C Metabolic Flux Analysis, which through the use of ^{13}C tracer experiments, performs more reliable analyses on cell intracellular fluxes [Blum and Stein, 1982, Malloy et al., 1988, Chance et al., 1983, Walsh and Koshland, 1984, Walker et al., 1982].

In the current days, the protocols to perform tasks of ^{13}C-MFA are quite well established [Zamboni et al., 2009, Crown and Antoniewicz, 2013b], and through its application it is possible to correctly resolve metabolic fluxes for reversible reactions [Yoo et al., 2008], parallel reactions and metabolic cycles in a model network [Kleijn et al., 2006]. Moreover, compartment fluxes are also accurately calculated [Wahrheit et al., 2011, Kruger et al., 2012, O'Grady et al., 2012], as it happens with stereochemistry-specific fluxes [Crown et al., 2011].

2.8.1 Traditional pipeline in ^{13}C metabolic flux analysis

The first important decision to make when performing ^{13}C tracer experiments (Figure 2.3) is related to the kind of labeling strategy that will be followed, namely, whether isolated or parallel labeling experiments will be executed, if positional or uniform labeled substrates will be used, and if substrate mixtures will the fed to the cultures. This rational design of labeling experiments is highly dependent of the target analysis, and several works have been published in this subject [Wittmann and Heinzle, 2001, Crown et al., 2012, Antoniewicz, 2013, Crown et al., 2015]. Afterwards, cell culture with the labeling tracers is employed and then the labeling enrichments in intracellular and extracellular metabolites are measured through the use of analytical techniques, such as, GC-MS, LC-MS, tandem MS and NMR.

Subsequently, an iterative procedure, usually by least-squares regression, is implemented, where the labeling patterns of metabolites are estimated and compared to the measurements. The labeling enrichment of the metabolites is computed based on the simulation of the labeling experiment by using the inputted substrate isotopomer fractions and the guessed flux distribution in the metabolic model, until a best fitting is obtained. In the first iteration, initial flux values are introduced, and in the forward iterations, the predicted discrepancy is accounted for a systematic variation of the generated flux distributions, through the utilization of an optimization method, that can be a meta-heuristic method as simulated annealing or evolutionary algorithms, or a more traditional approach based on Newton-like algorithms, among other options [Wiechert, 2001].

2.8.2 ^{13}C metabolic network model

The metabolic model used for the tasks of ^{13}C-MFA is of the main importance, since it can have a great influence over the design of tracer experiments [Antoniewicz, 2013]. The models for ^{13}C-MFA are not common metabolic models, since there is the need to provide information to allow the investigation on how the labeled atoms transit throughout the network.

In early 90s, Mavrovouniotis and co-authors developed an algorithm for biochemical pathway reconstruction which aimed to generate all the possible pathways between a set of metabolites [Mavrovouniotis et al., 1990]. However, the generation of the atomic transitions in the reactions belonging to the metabolic pathways was not on the scope of the method. This is actually the most tedious task when building ^{13}C metabolic networks, although there are tools to facilitate the access of the information revealing how the carbon atoms are mapped from substrates to products in the network reac-

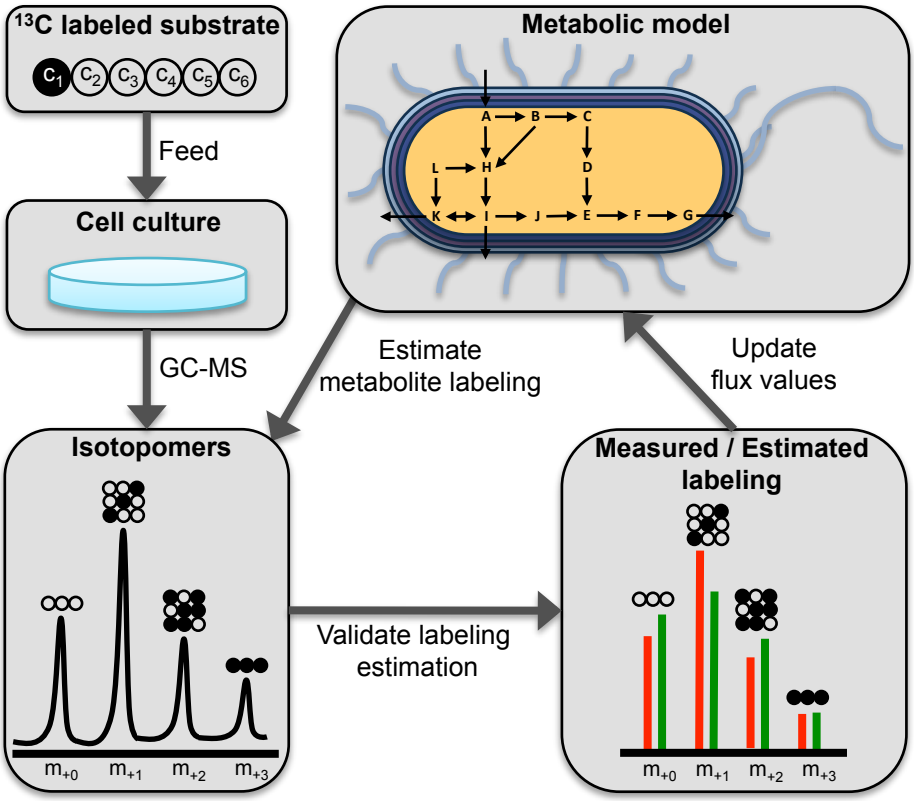


Figure 2.3: Traditional pipeline of ^{13}C metabolic flux analysis. Cells are fed with a ^{13}C substrate and the labeling pattern of metabolites is measured by an analytical technique (e.g. GC-MS). A metabolic model network is used to simulate labeling patterns of the intracellular metabolites, providing values for the fluxes of the reactions in the network. These last steps are repeated in consecutive iterations, until the flux values are properly adjusted to fit the calculated labeling to the measured data.

tions, including the ARM project [Arita, 2003], KEGG RPAIR [Kanehisa et al., 2006] or MetaCyc [Latendresse et al., 2012b]. In [Pitkänen et al., 2009], the authors aimed to improve the ARM method, building a graph-based approach that is able to reveal branching pathways that transfer the maximum number of atoms from the precursors to a target product.

Typically, the models used for the application of ^{13}C-MFA are far smaller than GSMMs, usually containing few reactions [Antoniewicz et al., 2007b, Suthers et al., 2007a, UkáKim et al., 2008]. However, recent developments

have been made in order to incorporate information needed for ^{13}C-MFA in a genome-scale view. A new graph algorithm based on the concept of pattern recognition for compound matching has been proposed [Ravikirthi et al., 2011] to build a genome-scale atom mapping for *Escherichia coli*, which maps all the non-hydrogen atoms in every reaction of the previously published GSMM *i*AF1260 [Feist et al., 2007]. Recently, an effort has been put forward to use information about atoms mapping to assist the task of genome-scale network reconstruction [Pitkänen et al., 2014].

All factors that can influence the labeling pattern of metabolites must be included in a ^{13}C-MFA network model, including metabolic cycles, reactions' reversibility and rotationally symmetric compounds, as fumarate or succinate, that can originate scrambling of isotopic labeling [Ravikirthi et al., 2011]. Moreover, details on reaction mechanisms are valuable sources of information, since their exclusion may hamper the correctness of ^{13}C-MFA calculations [Antoniewicz, 2013].

2.8.3 Isotopic labeling modeling and representation

The labeling data obtained from analytical techniques, as GC-MS or NMR, must be mapped onto metabolic fluxes, so their elucidation can be achieved. The metabolic fluxes inference is achieved by mathematical models that represent the propagation of the labeled atoms from substrates to products.

The isotopic labeling distribution throughout a metabolic network has been manually modelled, through mathematical equations, and solved a long time ago [Heath, 1968]. Since then, many models have been developed that are based on matrix operations, where the matrices are configured by vectors representing the isotopic enrichment of metabolites.

2.8.3.1 Isotopomers

Isotopomers are isomers of a metabolite for which the labeling vary in only one atom, for instance, having one atom of carbon differing from ^{12}C to ^{13}C. So, if each atom variation corresponds to a distinct isotopomer, a molecule with n atoms that can be in a binary state, i.e., labeled or not labeled, may originate 2^n isotopomers [Malloy et al., 1988]. The traditional representation of an isotopomer is a sequence of bits, where each position indicates if the atom in the corresponding position is either labeled or not.

Based on this representation, an isotopomer distribution vector (IDV) can be built. For an n-sized compound, the associated IDV will contain a position for each of its isotopomers. In Figure 2.4 a, the first position of the IDV of a 3-atom molecule corresponds to the isotopomer with no labeled atom, the second position refers to the isotopomer with the third atom labeled, and so on, until the last vector position that corresponds to a fully labeled representation of the molecule. An IDV of a molecule with three atoms can be represented as follows,

$$IDV = \begin{bmatrix} I_{000} \\ I_{001} \\ I_{010} \\ I_{011} \\ I_{100} \\ I_{101} \\ I_{110} \\ I_{111} \end{bmatrix} \tag{2.2}$$

where each position represents the molar fraction of the isotopomer corresponding to that labeling pattern of the molecule, and the sum of all isotopomers in the IDV equals to one.

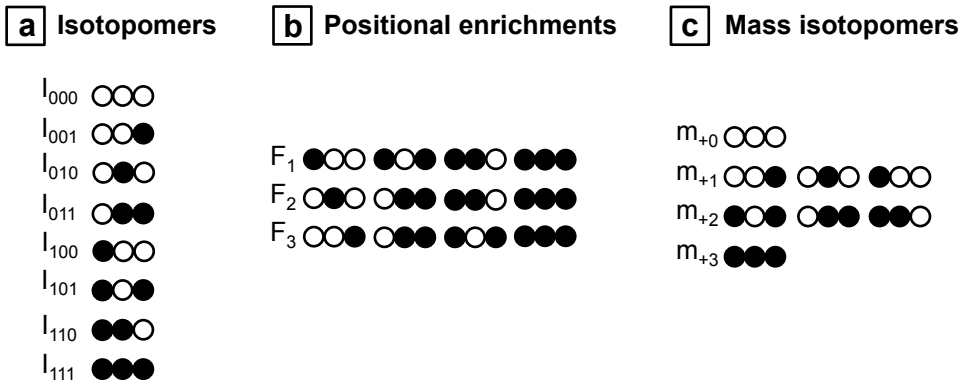


Figure 2.4: Vector representation of the isotopic content of a molecule. a) For a molecule with 3-carbon atoms there will be 8 possible isotopomers in a carbon-labeling study; b) in terms of positional enrichments, 3 fractional enrichment exist, one for each carbon position; c) each mass isotopomer is the sum of all isotopomers for containing the number of labeled atoms, independent of the atoms positions.

Isotopomer representation has been used in a broad set of approaches for flux distribution inference, as will be described in further sections of this document.

2.8.3.2 Positional enrichments

This representation allows to translate the labeling fraction originating from a specific atom position, and is useful when dealing with data originating from NMR techniques. Each positional fractional enrichment i will be the sum of the isotopomers containing at least one labeled atom in the same position i (Figure 2.4 b). For example, a molecule containing 3 atoms will have 3 positional fractional enrichments, where the first one corresponds to all isotopomers that contains the first atom labeled, independent of the atom position. Therefore, the vector representing the fractional enrichments can

be seen as a liner relationship with the isotopomer fractions [Antoniewicz, 2006]:

$$FE = \begin{bmatrix} F_1 \\ F_2 \\ F_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} I_{000} \\ I_{001} \\ I_{010} \\ I_{011} \\ I_{100} \\ I_{101} \\ I_{110} \\ I_{111} \end{bmatrix} \tag{2.3}$$

For instance, in [Sonntag et al., 1993] the authors used a very short model of the Gram-positive *Corynebacterium glutamicum* based on positional enrichments, for the reaction sequence of lysine formation from aspartate and pyruvate. The positional enrichments were built from data obtained by ^{13}C- and ^{1}H-NMR spectroscopy, and a simple system of linear equations was used to calculate the flux distribution.

2.8.3.3 Mass isotopomers

Mass isotopomer Distribution Vector (MDV)s are used for data originating from MS, since in this type of analytical techniques, the molecules are separated according to their mass-to-charge ratio (m/z). For a given mass spectrum, the intensity of the same ion peak is influenced by all the isotopomers with a fixed number of labeled atoms corresponding to that peak. A MDV, containing a mass isotopomer for all the possible number of labeled atoms of the molecule, including the mass of its lightest isotopomer,

can also be obtained from the IDVs, as the following equation shows:

$$MDV = \begin{bmatrix} m_{+0} \\ m_{+1} \\ m_{+2} \\ m_{+3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} . \begin{bmatrix} I_{000} \\ I_{001} \\ I_{010} \\ I_{011} \\ I_{100} \\ I_{101} \\ I_{110} \\ I_{111} \end{bmatrix} \tag{2.4}$$

2.8.3.4 Atom Mapping Matrices

Based on the isotopomers representation, the concept of Atom Mapping Matrices (AMM)s has been introduced, which is a mathematical formulation to describe the transfer of atoms between substrates and products [Zupke and Stephanopoulos, 1994]. For a certain reaction with two reactants and two products, four AMMs are generated, two tracking the transfer of carbons from one reactant to the two products, and other two referring to the other reactant to the same reaction products. Each AMM is constructed in such a way that the multiplication of the reactant fractional enrichments by the AMM, indicates the fractional enrichments of the corresponding product.

2.8.3.5 Isotopomer mapping matrices

Afterwards, the concept of AMM has been generalized to the form of Isotopomer Mapping Matrices (IMM) [Schmidt et al., 1997], which map all isotopomer mass balances of metabolite pools into a model of nonlinear

algebraic equations. Since the number of isotopomers belonging to each metabolite pool is very high, the authors introduced the IMMs to sum up all pairs of reactant isotopomers that originate the corresponding product Isotopomer Distribution Vector (IDV) isotopomers. For each network reaction, there will be an IMM defined for every pair of reactants and associated product. The resulting model, encompasses variables related to both IDVs and metabolic fluxes.

These models were applied to investigate the central metabolism of *Escherichia coli* [Schmidt et al., 1999]. However, this approach has some limitations, since even for a given flux distribution, the set of equations generated to the identification of the associated IDVs are still nonlinear [Suthers et al., 2007b].

2.8.3.6 Cumomer labeling balances

Due to the excessive number of isotopomer balance equations in the previous approach, and to its nonlinear configuration, the concept of cumomers has been created [Wiechert et al., 1999]. Through a transformer matrix, the IDVs are converted to cumomers and the resulting equations are made simpler, able to be solved analytically, without any loss of information. The transformed variables, named cumomer fractions, which is a short version for cumulative isotopomer fraction, refers to the sum of the isotopomer fractions of the corresponding metabolite with the same labeling pattern of the cumomer.

For the three 1-cumomer fractions, the equations can be written as

follows:

$$C_{1xx} = \sum_{j,k=0}^{1} C_{1,j,k} \ , \ C_{x1x} = \sum_{i,k=0}^{1} C_{i,1,k} \ , \ C_{xx1} = \sum_{i,j=0}^{1} C_{i,j,1} \tag{2.5}$$

where x is a binary digit, meaning that the atom is either labeled or unlabeled in the corresponding position.

The cumomers are computed by solving a cascade of linear equations, although the nonlinear coupling of metabolic fluxes with the IDVs remains [Zomorrodi et al., 2012].

2.8.3.7 Isotopomer path tracing concept

Forbes and co-authors have developed the isotopomer path tracing method that reduces the number of isotopomer variables in previous configurations to the number of experimentally measured isotopomers, and identifies all isotopomer paths that produce the observed data [Forbes et al., 2001]. The carbon atoms tracing is accomplished through the use of AMMs, using Simulated Annealing (SA) and the Newton-Raphson method to solve the nonlinear components of the system.

2.8.3.8 Cumulative bondomers

A linear fragmented molecule comprising n carbons atoms in its backbone has $n-1$ C-C bounds. Based on this number of C-C bonds, a molecule can be said to have 2^{n-1} bondomers, which are entities that only vary in terms of positions and number of C-C bonds that persist intact, after the labeled substrate has entered in the cell's metabolism [van Winden et al., 2002].

The bondomer distribution of the molecules can be simulated similarly as the approaches for simulating isotopomers, and its balancing relies on the concept of a Bondomer Mapping Matrix (BMM) that specify which bondomers of the reactants in the model reactions originate the products' bondomers. The drawback of this approach is that it is designed only for uniformly labeled substrates.

2.8.3.9 Elementary Metabolite Units

A framework for modeling isotopomer tracer systems has introduced the concept of Elementary Metabolite Unit (EMU)s [Antoniewicz et al., 2007a], that are functional units generated by an efficient decomposition algorithm that aims at identifying the minimal information that is relevant to simulate the labeling content over a metabolic reaction. The EMUs of a given compound are designated as the distinct subsets of the compound's atoms. If the compound has n atoms, it will have a maximum of $2^n - 1$ EMUs.

As in the isotopomer and cumomer based methods, the EMUs are the basis of a system of equations linking the metabolic fluxes with the isotopomer abundances, where the computed abundances by this approach are equivalent to the mentioned methods, but with a substantial reduced number of variables. Moreover, the framework allows to analyze the labeling distribution from multiple isotopic tracers.

2.8.3.10 Flux coupling with elementary metabolite units

In [Suthers et al., 2010], the authors combined the EMU framework with flux coupling to an *Escherichia coli* model, where the mass isotopomer distribution measurements and atom mapping model were adapted from [Suthers et al., 2007a]. The second framework allows to efficiently determine

if any two fluxes are: directionally coupled, when a non-zero flux implies a non-zero flux in another reaction, but the reverse is not mandatory; partially coupled fluxes, for the cases where a non-zero flux in one reaction implies a non-zero value, but variable, of the flux in another reaction; and also fully coupled, when a non-zero flux in a reaction implies a non-zero fixed value in the other reaction [Burgard et al., 2004]. This combined approach, allowed to reduce the number of variables in a previous approach OptMeas [Chang et al., 2008], which is a formulation based in MILP and uses IDVs for mathematically representing the labeling propagation.

2.8.3.11 Fluxomers

In 2011, a new concept has been released named fluxomers [Srour et al., 2011], which aimed to overcome the presence of unstable local minima originated from the non-convex nature of the objective function in the optimization problems, and the complex mathematical systems and computational methods implemented in the previous approaches.

Contrary to other methods, fluxomers do not consider the metabolic fluxes and isotopomers as distinct variables, rather these are combined in composite variables, referred as ''isotopically labeled fluxes" meaning that they are rate at which a metabolic reaction transfers the isotopic content of substrates to the products. This results in a simpler system with reduced complexity and nonlinearity. Consequently, this simplification facilitates the using of traditional optimization algorithms.

2.8.3.12 Strongly connected components

More recently, the equations in the (quasi-) linear model of both comomer and EMU strategies were assembled and through the use of graph decom-

position and path tracing algorithms, smaller linear subsystems were generated, originating the Strongly Connected Component (SCC)s by exploiting the connectivity properties of the cumomer and EMU models [Weitzel et al., 2007]. The reduced system were then translated in a cascade of symbolic equation systems, which allows to obtain either a numerical solution, or exact solutions based on arbitrary precision arithmetic [Weitzel et al., 2013].

2.8.3.13 Non-linear least squares with inequality constraints

More recently, a new algorithm that combines Non-Linear Least Squares with Inequality Constraints (NLSIC) has been introduced [Sokol et al., 2012]. The NLSIC algorithm aims at solving the problems inherent to ^{13}C-MFA approaches, enhancing the numerical accuracy of the final solutions with good convergence robustness, but not losing convergence speed. In the tested examples, the authors have not found problems with local minima, and argue that the developed algorithm is suitable to solve constrained least-squares problems with ill-conditioned or rank-deficient Jacobian. For instance, this approach has been successfully applied to *Escherichia coli* [Millard et al., 2014].

2.9 Methods of inference in ^{13}C tracer experiments

Different approaches to exploit the rich information obtained from ^{13}C tracer experiments to infer metabolic fluxes have been developed. These can be classified in two main groups: the global and local paradigms.

2.9.1 Global methods of inference in ^{13}C-MFA

The global approaches for ^{13}C tracer experiments include all the aforementioned concepts for isotope balancing, which is the natural extension of the stoichiometric balancing approach to include labeling data [Zamboni, 2007]. This extension originates a system containing a metabolic model that serves as the backbone for the balance equations, and, simultaneously, encompassing extracellular fluxes and labeling patterns. The difference here is that, on contrary of the traditional stoichiometric balancing, where the systems are configured with an unique mass balance for each metabolite in the network, the balances are made by drawing one equation for each isotopomer [Zupke et al., 1997, Schmidt et al., 1997, Klapa et al., 1999, Dauner et al., 2001b]. Every network topology can be assessed by this class of isotopomer balancing, as well as every substrate labeling distributions originating from the different analytical techniques [Marx et al., 1996, Shen et al., 1999, Schmidt et al., 1999, Christensen and Nielsen, 2000b, Petersen et al., 2000].

However, in this type of systems, the equations translating the isotopomer balances are nonlinear. Also, since the number of isotopomers is very large, several equations are added to the balance systems, but the number of variables linking metabolites to the isotopomers also increases, which still make the system underdetermined. Therefore, the problem has to be solved through the use of iterative algorithms that in each iteration, generate candidate flux distributions so the fluxes are fitted to minimize the difference between the simulated and experimental isotopic data (Figure 2.3) [Zupke and Stephanopoulos, 1994, Schmidt et al., 1997, Wiechert et al., 1999, Wiechert et al., 2001, van Winden et al., 2002, Antoniewicz et al., 2007a, Nöh et al., 2007, Young et al., 2008]. Nevertheless, often the optimization algorithms applied in such methodologies do not reach a global

optimum [Ghosh et al., 2005].

The costs of computation of such equation systems are too high and the followed approaches, as cumomers, that allows for an explicit solution on systems of large isotopomer networks, or the EMU framework, that improve the estimations by a systematic simplification of an isotopomer network, have given rise to much more efficient concepts as described above. These improvements have enabled to efficiently decrease the number of variables in this kind of global inference methods.

Currently, recent advances in the global approaches have boosted the appearance of new resources for performing global inference in a much more efficient manner, and the time of computation in the optimization processes has decreased considerably [Weitzel et al., 2007, Weitzel et al., 2013, Hanke et al., 2013, Niedenführ et al., 2015]

2.9.2 Local paradigm for ^{13}C-MFA

The global approaches aim at comprehending and integrating all the available data so they can explore the maximum possible information. However, the estimation of flux distributions is much influenced by inaccurate components in the metabolic models, or even in the physiological data. Thus, the entire flux distribution estimation can be drastically penalized. The complex formulations in the global approaches might hamper the interpretation of the results and no direct evidence to a particular flux is given since the flux distribution is obtained by fitting the problem to all data [Emmerling et al., 2002].

As an alternative, an unsupervised local approach has been put forward, named Metabolic Flux Ratio (METAFoR) analysis [Szyperski, 1995, Sauer et al., 1999, Fiaux et al., 1999], which uses probabilistic equations to in-

terpret the ^{13}C-labeling pattern of metabolites, by calculating the ratios of fluxes that produce a given metabolite through the different pathways in the metabolic network. The ratios probabilistic equations have been originality obtained for NMR data, derived from relating the intensities of the multiplet components to the relative abundance of compound fragments [Szyperski, 1995, Szyperski et al., 1996, Szyperski, 1998, Szyperski et al., 1999] The calculation of metabolic flux ratios is independent of the flux distribution in the network [Sauer et al., 1999, Fischer and Sauer, 2003c, Blank and Sauer, 2004], where no knowledge is needed either on the rates of uptake and formation of external metabolites, or on the biomass composition [Ebert et al., 2012, Rantanen et al., 2008].

In fact, flux ratios by itself can be used to pinpoint novel pathways or unexpected cross activity [Fischer and Sauer, 2003a], characterize network topologies [Fuhrer et al., 2005], express rigidity and suboptimal performance in metabolism [Fischer and Sauer, 2005] or examine the adaptative evolution for *in silico* designed strains [Hua et al., 2006].

Relevant independent metabolic flux ratios have been reported for the central carbon metabolism of bacteria, yeast or plants from ^{13}C tracer experiments, using MS techniques as in the cases of *Escherichia coli*, *Arthrobacter* sp. CGMCC 3584 [Niu et al., 2013], *Bacillus subtilis* [Zamboni et al., 2004], the fungal *Yarrowia lipolytica* [Blank et al., 2005], or the green alga *Chlorella protothecoides* [Xiong et al., 2010], and NMR in the cases of *Escherichia coli* [Szyperski, 1995], *Saccharomyces cerevisiae* [Maaheimo et al., 2001], and *Pichia stipites* [Fiaux et al., 2003].

From the metabolic flux ratios, constraints can be configured to reduce the degrees of freedom of the systems in stationary MFA, and thus, conducting the analysis to solutions that are more likely to reflect the actual phys-

iological state of the cell. In Table 2.3, some applications of the METAFoR approach are summarized, referring the used analytical technique and the substrate labeling strategy for their ^{13}C tracer experiments.

In 2007, a study has shown that for some branchpoints in the central metabolism, the concept of bondomer abundances can also be used to directly calculate fluxes through an approach similar to the METAFoR analysis [Sriram et al., 2007].

2.9.3 Fluxome profiling

A particular field of interest is also fluxome profiling, that, without the need of using a metabolic model, can give great insights for discriminating mutants or other conditions, and assess their metabolic impact. Similarly to other omics, where data mining methods are applied, ^{13}C labeling patterns can be investigated through the use of multivariate analysis [Zamboni and Sauer, 2005]. This approach can be applied in complex media, and also to any isotopic tracer, e.g., ^{13}C, ^{2}H or ^{15}N.

For instance, the flux fingerprints of 10 *Bacillus subtilis* mutants from ^{13}C and ^{2}H experiments have been compared through the application of unsupervised multivariate statistics [Zamboni and Sauer, 2004]. However, fluxomics profiling studies need a large number of samples or replicates. In a recent work, a high-throughput fluxomics study has been applied to 80 cultivations of *Escherichia coli* mutants, through a fully integrated platform that combines a robotic system for ^{13}C tracer experiments and automated data interpretation from NMR spectra [Heux et al., 2014].

Table 2.3: Summary of applications with ratios constrained metabolic flux analysis.

Organism	Year	Technique	Substrate labeling	Reference
Arthrobacter	2013	GC-MS	100% [1-^{13}C] glucose / 20% [U-^{13}C] x 80% natural glucose	[Niu et al., 2013]
Aspergillus niger	2012	GC-MS	100% [1-^{13}C] glucose / 50% [U-^{13}C] x 50% natural glucose	[Driouch et al., 2012]
Bacillus subtilis	2004	GC-MS	30% [1-^{13}C] x 30% [U-^{13}C] x 40% natural glucose	[Zamboni et al., 2004]
Chlorella protothecoides	2010	GC-MS	100% [1-^{13}C] glucose / 10% [U-^{13}C] x 90% natural glucose	[Xiong et al., 2010]
Escherichia coli	1995	2D [^{13}C,^{1}H]-COSY NMR	10% [U-^{13}C] x 90% natural glucose	[Szyperski, 1995]
	1999	2D [^{13}C,^{1}H]-COSY NMR	10% [U-^{13}C] x 90% natural glucose	[Fiaux et al., 1999]
	2002	2D [^{13}C,^{1}H]-COSY NMR	10% [U-^{13}C] x 90% natural glucose	[Emmerling et al., 2002]
	2003	GC-MS	100% [1-^{13}C] glucose / 20% [U-^{13}C] x 80% natural glucose	[Fischer and Sauer, 2003b]
Pichia pastoris	2011	GC-MS	20% [U-^{13}C] x 80% natural glucose	[Heyland et al., 2011]

Continued on next page

Table 2.3 – continued from previous page

Organism	Year	Technique	Substrate labeling	Reference
Pseudomonas denitrificans	2012	GC-MS	100% [1-^{13}C] glucose / 20% [U-^{13}C] x 80% natural glucose	[Wang et al., 2012]
Pseudomonas putida	2007	GC-MS	100% [1-^{13}C] glucose / 20% [U-^{13}C] x 80% natural glucose	[del Castillo et al., 2007]
	2013	GC-MS	100% [1-^{13}C] glucose / 20% [U-^{13}C] x 80% natural glucose	[Silva-Rocha et al., 2013]
Rhodobacter sphaeroides	2012	GC-MS	100% [1-^{13}C] glucose / 30% [U-^{13}C] x 70% natural glucose	[Tao et al., 2012]
Saccharomyces cerevisiae	2001	2D [^{13}C,^{1}H]-COSY NMR	10% [U-^{13}C] x 90% natural glucose	[Maaheimo et al., 2001]
	2004	GC-MS	20% [U-^{13}C] x 80% natural glucose	[Blank and Sauer, 2004]
Trichoderma reesei	2009	^{13}C-HSQC NMR	10% [U-^{13}C] x 90% natural glucose / 10% [U-^{13}C] x 90% natural sorbitol	[Jouhten et al., 2009]

2.9.4 Software tools for ^{13}C metabolic flux analysis

A number of software tools for helping the tasks of ^{13}C labeling tracer experiments have been put forward (Figure 2.5). The majority are based on global methods of inference, and are distinguished in terms of the concepts their isotopomer balances are built on, the methods they use for solving the iterative fitting problems, or the analytical techniques they accept, among other features. Here, a brief description about the main tools is given.

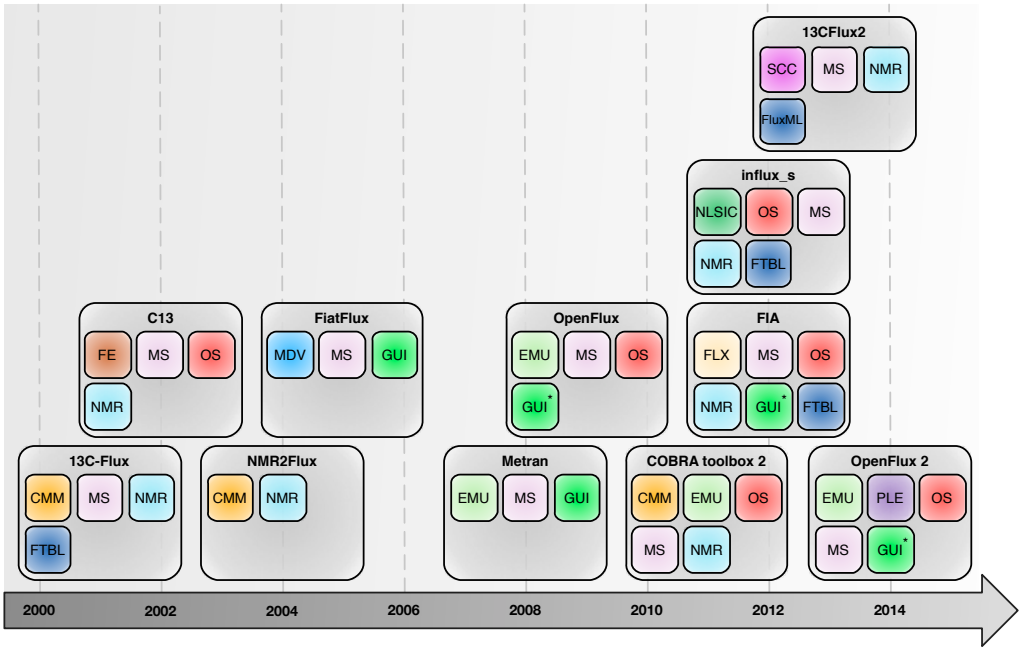


Figure 2.5: Software tools for stationary ^{13}C-MFA. For each software, its isotopic labeling representation and the type of analytical technique that it supports are represented in the small squared boxes inside the software box. CMM, comomers; FTBL, formatted file for ^{13}C-MFA experiments; FE, fractional enrichment; OS, open-source; FLX, fluxomers; PLE, parallel labeling experiments. Fluxml is a recent extension of the old FTBL format. For some of the tools, the GUI is available only for some basic configurations, and for the case of 13CFLUX2 it is integrated with the Omix visualization tool.

13C-FLUX has been widely used for ^{13}C-MFA [McKinlay et al., 2007, Zamboni et al., 2009, Jordà et al., 2014]. It was originally released in 2001 and was based in the comomer concept, dealing with data from both MS

and NMR. Twelve years later, its second version [Weitzel et al., 2013] has been published, where the strongly connected components are implemented leading to a high increase of its performance. However, both versions are commercial, although academic licenses might be acquired under request. Another drawback is that it runs only in Linux platforms. In terms of visualization, it does include Graphical User Interface (GUI)s for its core functionalities, but it is integrated with *Omix* [Droste et al., 2013] through plug-ins that extend *Omix* to a comprehensive modeling front end for *13C-FLUX*2.

C13 is part of the *BioMet Toolbox* [Cvijovic et al., 2010,Garcia-Albornoz et al., 2014], and uses fractional enrichments data obtained from MS and NMR, while *NMR2FLUX* [Sriram et al., 2004] works over cumomers for NMR data.

Currently, the only available software that calculates metabolic flux distributions based on the local METAFoR approach is *FiatFlux* [Zamboni et al., 2005]. The software has been used for several ^{13}C tracer experiments [Fuhrer et al., 2005,Fong et al., 2006,del Castillo et al., 2007,Heyland et al., 2011]. It uses previous manually derived flux ratio equations [Sauer et al., 1999,Fischer and Sauer, 2003b] to calculate flux ratio values based on the ^{13}C patterns detected by GC-MS in protein-bound amino acids, which are then used in established flux ratio constraints [Fischer et al., 2004].

Metran [Yoo et al., 2008] was the first released software accounting for the elementary metabolite units framework. Afterwards, the open-source *OpenFLUX* has been published [Quek et al., 2009], also using the EMU framework. Although it has a component written in Java to parse the input data and generate the metabolic model, which is made available through GUIs, the core analysis functionalities are all implemented in MATLAB,

with no GUI made available. The *COBRA Toolbox v2.0* also has tools for supporting the tasks of ^{13}C-MFA, and uses both the concepts of cumomers and EMUs for MS and NMR data [Schellenberger et al., 2011].

In 2012, when the fluxomers approach was introduced, the authors also developed the software *FIA* to facilitate the executions of their implemented method, through a basic front-end interface for the main command line script [Srour et al., 2011]. In the same year the *influx_s* software was released [Sokol et al., 2012], and is based on the Non-linear least squares with inequality constraints approach, where both NMR and MS data can be inputed in the FTBL format, which is a file format metabolic and isotope network specification designed by *13C-FLUX*.

More recently, *OpenFLUX2* has been released [Shupletsov et al., 2014], adding features to its previous version, such as, the calculation of parallel labeling experiments. However, its dependency on MATLAB still remains in this version.

BIBLIOGRAPHY

[Agren et al., 2013] Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I., and Nielsen, J. (2013). The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. *PLoS computational biology*, 9(3):e1002980.

[Ahn and Antoniewicz, 2013] Ahn, W. S. and Antoniewicz, M. R. (2013). Parallel labeling experiments with [1, 2-13 C] glucose and [U-13 C] glutamine provide new insights into CHO cell metabolism. *Metabolic engineering*, 15:34–47.

[Ajikumar et al., 2010] Ajikumar, P. K., Xiao, W.-H., Tyo, K. E. J., Wang, Y., Simeon, F., Leonard, E., Mucha, O., Phon, T. H., Pfeifer, B., and Stephanopoulos, G. (2010). Isoprenoid pathway optimization for Taxol precursor overproduction in Escherichia coli. *Science*, 330(6000):70–74.

[Akiyama et al., 2008] Akiyama, K., Chikayama, E., Yuasa, H., Shimada, Y., Tohge, T., Shinozaki, K., Hirai, M. Y., Sakurai, T., Kikuchi, J., and Saito, K. (2008). PRIMe: a Web site that assembles tools for metabolomics and transcriptomics. *In silico biology*, 8:339–345.

[Altamirano et al., 2006] Altamirano, C., Illanes, A., Becerra, S., Cairó, J. J., and Gòdia, F. (2006). Considerations on the lactate consumption by CHO cells in the presence of galactose. *Journal of Biotechnology*, 125(4):547–556.

[Anderson et al., 1981] Anderson, S., Bankier, A. T., Barrell, B. G., De Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., and Others (1981). Sequence and organization of the human mitochondrial genome.

[Antoniewicz, 2006] Antoniewicz, M. R. (2006). *Comprehensive analysis of metabolic pathways through the combined use of multiple isotopic tracers*. PhD thesis, Massachusetts Institute of Technology.

[Antoniewicz, 2013] Antoniewicz, M. R. (2013). (13)C metabolic flux analysis: optimal design of isotopic labeling experiments. *Current opinion in biotechnology*, null(null).

[Antoniewicz et al., 2007a] Antoniewicz, M. R., Kelleher, J. K., and Stephanopoulos, G. (2007a). Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions. *Metabolic engineering*, 9(1):68–86.

[Antoniewicz et al., 2007b] Antoniewicz, M. R., Kraynie, D. F., Laffend, L. A., González-Lergier, J., Kelleher, J. K., and Stephanopoulos, G. (2007b). Metabolic flux analysis in a nonstationary system: fed-batch fermentation of a high yielding strain of E. coli producing 1, 3-propanediol. *Metabolic engineering*, 9(3):277–292.

[Araúzo-Bravo and Shimizu, 2003] Araúzo-Bravo, M. J. and Shimizu, K. (2003). An improved method for statistical analysis of metabolic flux analysis using isotopomer mapping matrices with analytical expressions. *Journal of Biotechnology*, 105(1):117–133.

[Archer et al., 2011] Archer, C. T., Kim, J. F., Jeong, H., Park, J. H., Vickers, C. E., Lee, S. Y., and Nielsen, L. K. (2011). The genome sequence of E. coli W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of E. coli. *BMC genomics*, 12(1):9.

[Arita, 2003] Arita, M. (2003). In silico atomic tracing by substrate-product relationships in Escherichia coli intermediary metabolism. *Genome Research*, 13(11):2455–2466.

[Au et al., 2014] Au, J., Choi, J., Jones, S. W., Venkataramanan, K. P., and Antoniewicz, M. R. (2014). Parallel labeling experiments validate Clostridium acetobutylicum metabolic network model for 13 C metabolic flux analysis. *Metabolic engineering*, 26:23–33.

[Bailey, 1991] Bailey, J. E. (1991). Toward a science of metabolic engineering. *Science*, 252(5013):1668–1675.

[Bailey, 1999] Bailey, J. E. (1999). Lessons from metabolic engineering for functional genomics and drug discovery. *Nature biotechnology*, 17(7):616–618.

[Becker et al., 2007a] Becker, J., Klopprogge, C., Herold, A., Zelder, O., Bolten, C. J., and Wittmann, C. (2007a). Metabolic flux engineering of L-lysine production in Corynebacterium glutamicum—over expression and modification of G6P dehydrogenase. *Journal of biotechnology*, 132(2):99–109.

[Becker et al., 2005] Becker, J., Klopprogge, C., Zelder, O., Heinzle, E., and Wittmann, C. (2005). Amplified expression of fructose 1, 6-bisphosphatase in Corynebacterium glutamicum increases in vivo flux through the pentose phosphate pathway and lysine production on different carbon sources. *Applied and environmental microbiology*, 71(12):8587–8596.

[Becker et al., 2007b] Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. O., and Herrgård, M. J. (2007b). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature protocols*, 2(3):727–738.

[Benedict et al., 2012] Benedict, M. N., Gonnerman, M. C., Metcalf, W. W., and Price, N. D. (2012). Genome-scale metabolic reconstruction and hypothesis testing in the methanogenic archaeon Methanosarcina acetivorans C2A. *Journal of bacteriology*, 194(4):855–865.

[Bernal et al., 2009] Bernal, V., Carinhas, N., Yokomizo, A. Y., Carrondo, M. J. T., and Alves, P. M. (2009). Cell density effect in the baculovirus-insect cells system: A quantitative analysis of energetic metabolism. *Biotechnology and bioengineering*, 104(1):162–180.

[Bino et al., 2004] Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B. J., Mendes, P., Roessner-Tunali, U., Beale, M. H., Trethewey, R. N., Lange, B. M., Wurtele, E. S., and Sumner, L. W. (2004). Potential of metabolomics as a functional genomics tool. *Trends in plant science*, 9(9):418–25.

[Birkemeyer et al., 2005] Birkemeyer, C., Luedemann, A., Wagner, C., Erban, A., and Kopka, J. (2005). Metabolome analysis: the potential of¡ i¿ in vivo¡/i¿ labeling with stable isotopes for metabolite profiling. *Trends in biotechnology*, 23(1):28–33.

[Blank et al., 2005] Blank, L. M., Lehmbeck, F., and Sauer, U. (2005). Metabolic-flux and network analysis in fourteen hemiascomycetous yeasts. *FEMS yeast research*, 5(6-7):545–558.

[Blank and Sauer, 2004] Blank, L. M. and Sauer, U. (2004). TCA cycle activity in Saccharomyces cerevisiae is a function of the environmentally determined specific growth and glucose uptake rates. *Microbiology*, 150(4):1085–1093.

[Blattner et al., 1997] Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., and Others (1997). The complete genome sequence of Escherichia coli K-12. *science*, 277(5331):1453–1462.

[Blum and Stein, 1982] Blum, J. J. and Stein, R. B. (1982). On the analysis of metabolic networks. *Biological regulation and development*, 3:99–125.

[Bordbar et al., 2014] Bordbar, A., Monk, J. M., King, Z. A., and Palsson, B. O. (2014). Constraint-based models predict metabolic and associated cellular functions. *Nature Reviews Genetics*, 15(2):107–120.

[Bro et al., 2006] Bro, C., Regenberg, B., Förster, J., and Nielsen, J. (2006). In silico aided metabolic engineering of Saccharomyces cerevisiae for improved bioethanol production. *Metabolic engineering*, 8(2):102–11.

[Brochado et al., 2012] Brochado, A. R., Andrejev, S., Maranas, C. D., and Patil, K. R. (2012). Impact of stoichiometry representation on simulation of

genotype-phenotype relationships in metabolic networks. *PLoS computational biology*, 8(11):e1002758.

[Broeckling et al., 2006] Broeckling, C. D., Reddy, I. R., Duran, A. L., Zhao, X., and Sumner, L. W. (2006). MET-IDEA: data extraction tool for mass spectrometry-based metabolomics. *Analytical chemistry*, 78(13):4334–4341.

[Burgard et al., 2004] Burgard, A. P., Nikolaev, E. V., Schilling, C. H., and Maranas, C. D. (2004). Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome research*, 14(2):301–312.

[Butler et al., 2009] Butler, G., Rasmussen, M. D., Lin, M. F., Santos, M. A. S., Sakthikumar, S., Munro, C. A., Rheinbay, E., Grabherr, M., Forche, A., Reedy, J. L., and Others (2009). Evolution of pathogenicity and sexual reproduction in eight Candida genomes. *Nature*, 459(7247):657–662.

[Cai et al., 2010] Cai, G., Jin, B., Saint, C., and Monis, P. (2010). Metabolic flux analysis of hydrogen production network by Clostridium butyricum W5: Effect of pH and glucose concentrations. *International Journal of Hydrogen Energy*, 35(13):6681–6690.

[Carinhas et al., 2011] Carinhas, N., Bernal, V., Teixeira, A. P., Carrondo, M. J. T., Alves, P. M., and Oliveira, R. (2011). Hybrid metabolic flux analysis: combining stoichiometric and statistical constraints to model the formation of complex recombinant products. *BMC systems biology*, 5(1):34.

[Carinhas et al., 2012] Carinhas, N., Oliveira, R., Alves, P. M., Carrondo, M. J. T., and Teixeira, A. P. (2012). Systems biotechnology of animal cells: the road to prediction. *Trends in biotechnology*, 30(7):377–385.

[Caspi et al., 2008] Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S. Y., Shearer, A. G., Tissier, C., and Others (2008). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research*, 36(suppl 1):D623—-D631.

[Chance, 1943] Chance, B. (1943). The kinetics of the enzyme-substrate compound of peroxidase. *Journal of Biological Chemistry*, 151(2):553–577.

[Chance et al., 1960] Chance, B., Garfinkel, D., Higgins, J., and Hess, B. (1960). Metabolic control mechanisms. V. A solution for the equations representing interactions between glycolysis and respiration in ascites tumor cells. *J. biol. Chem*, 235:2426–2439.

[Chance et al., 1952] Chance, B., Greenstein, D. S., Higgins, J., and Yang, C. C. (1952). The mechanism of catalase action. II. Electric analog computer studies. *Archives of biochemistry and biophysics*, 37(2):322–339.

[Chance et al., 1983] Chance, E. M., Seeholzer, S. H., Kobayashi, K., and Williamson, J. R. (1983). Mathematical analysis of isotope labeling in the citric acid cycle with applications to 13C NMR studies in perfused rat hearts. *Journal of Biological Chemistry*, 258(22):13785–13794.

[Chang et al., 2008] Chang, Y., Suthers, P. F., and Maranas, C. D. (2008). Identification of optimal measurement sets for complete flux elucidation in metabolic flux analysis experiments. *Biotechnology and bioengineering*, 100(6):1039–1049.

[Chen et al., 1998] Chen, Z., Landman, P., Colmer, T. D., and Adams, M. A. (1998). Simultaneous Analysis of Amino and Organic Acids in Extracts of Plant Leaves astert-Butyldimethylsilyl Derivatives by Capillary Gas Chromatography. *Analytical biochemistry*, 259(2):203–211.

[Chinwalla et al., 2002] Chinwalla, A. T., Cook, L. L., Delehaunty, K. D., Fewell, G. A., Fulton, L. A., Fulton, R. S., Graves, T. A., Hillier, L. W., Mardis, E. R., McPherson, J. D., and Others (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562.

[Christensen and Nielsen, 1999] Christensen, B. and Nielsen, J. (1999). Isotopomer analysis using GC-MS. *Metabolic Engineering*, 1(4):282–290.

[Christensen and Nielsen, 2000a] Christensen, B. and Nielsen, J. (2000a). Metabolic network analysis. A powerful tool in metabolic engineering. *Advances in biochemical engineering/biotechnology*, 66:209–31.

[Christensen and Nielsen, 2000b] Christensen, B. and Nielsen, J. (2000b). Metabolic network analysis of penicillium chrysogenum using 13c-labeled glucose. *Biotechnology and bioengineering*, 68(6):652–659.

[Christiansen et al., 2002] Christiansen, T., Christensen, B., and Nielsen, J. (2002). Metabolic network analysis of Bacillus clausii on minimal and semirich medium using 13 C-labeled glucose. *Metabolic engineering*, 4(2):159–169.

[Cohen et al., 1973] Cohen, S. N., Chang, A. C. Y., Boyer, H. W., and Helling, R. B. (1973). Construction of biologically functional bacterial plasmids in vitro. *Proceedings of the National Academy of Sciences*, 70(11):3240–3244.

[Cole et al., 1998] Cole, S., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E. r., and Others (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, 393(6685):537–544.

[Colon et al., 1995] Colon, G. E., Nguyen, T. T., Jetten, M. S. M., Sinskey, A. J., and Stephanopoulos, G. (1995). Production of isoleucine by overexpression

of ilvA in a Corynebacterium lactofermentum threonine producer. *Applied microbiology and biotechnology*, 43(3):482–488.

[Consortium and Others, 2011] Consortium, U. and Others (2011). Ongoing and future developments at the Universal Protein Resource. *Nucleic acids research*, 39(suppl 1):D214—-D219.

[Copeland et al., 2012] Copeland, W. B., Bartley, B. A., Chandran, D., Galdzicki, M., Kim, K. H., Sleight, S. C., Maranas, C. D., and Sauro, H. M. (2012). Computational tools for metabolic engineering. *Metabolic engineering*, 14(3):270–280.

[Covert et al., 2001] Covert, M. W., Schilling, C. H., and Palsson, B. (2001). Regulation of gene expression in flux balance models of metabolism. *Journal of theoretical biology*, 213(1):73–88.

[Croasmun and Carlson, 1994] Croasmun, W. R. and Carlson, R. M. K. (1994). *Two-dimensional NMR spectroscopy: applications for chemists and biochemists*, volume 15. John Wiley & Sons.

[Crown et al., 2012] Crown, S. B., Ahn, W. S., and Antoniewicz, M. R. (2012). Rational design of 13C-labeling experiments for metabolic flux analysis in mammalian cells. *BMC systems biology*, 6(1):43.

[Crown and Antoniewicz, 2013a] Crown, S. B. and Antoniewicz, M. R. (2013a). Parallel labeling experiments and metabolic flux analysis: past, present and future methodologies. *Metabolic engineering*, 16:21–32.

[Crown and Antoniewicz, 2013b] Crown, S. B. and Antoniewicz, M. R. (2013b). Publishing 13 C metabolic flux analysis studies: a review and future perspectives. *Metabolic engineering*, 20:42–48.

[Crown et al., 2011] Crown, S. B., Indurthi, D. C., Ahn, W. S., Choi, J., Papoutsakis, E. T., and Antoniewicz, M. R. (2011). Resolving the TCA cycle and pentose-phosphate pathway of Clostridium acetobutylicum ATCC 824: Isotopomer analysis, in vitro activities and expression analysis. *Biotechnology journal*, 6(3):300–305.

[Crown et al., 2015] Crown, S. B., Long, C. P., and Antoniewicz, M. R. (2015). Integrated 13 C-metabolic flux analysis of 14 parallel labeling experiments in Escherichia coli. *Metabolic engineering*.

[Cvijovic et al., 2010] Cvijovic, M., Olivares-Hernández, R., Agren, R., Dahr, N., Vongsangnak, W., Nookaew, I., Patil, K. R., and Nielsen, J. (2010). BioMet Toolbox: genome-wide analysis of metabolism. *Nucleic acids research*, 38(suppl 2):W144—-W149.

[Dai et al., 2012] Dai, Z., Liu, Y., Huang, L., and Zhang, X. (2012). Production of miltiradiene by metabolically engineered Saccharomyces cerevisiae. *Biotechnology and bioengineering*, 109(11):2845–2853.

[Dai et al., 2013] Dai, Z., Liu, Y., Zhang, X., Shi, M., Wang, B., Wang, D., Huang, L., and Zhang, X. (2013). Metabolic engineering of Saccharomyces cerevisiae for production of ginsenosides. *Metabolic engineering*, 20:146–156.

[Dauner et al., 2001a] Dauner, M., Bailey, J. E., and Sauer, U. (2001a). Metabolic flux analysis with a comprehensive isotopomer model in Bacillus subtilis. *Biotechnology and bioengineering*, 76(2):144–56.

[Dauner and Sauer, 2000] Dauner, M. and Sauer, U. (2000). GC-MS analysis of amino acids rapidly provides rich information for isotopomer balancing. *Biotechnology progress*, 16(4):642–9.

[Dauner et al., 2001b] Dauner, M., Storni, T., and Sauer, U. (2001b). Bacillus subtilis metabolism and energetics in carbon-limited and excess-carbon chemostat culture. *Journal of bacteriology*, 183(24):7308–7317.

[Davies, 1998] Davies, T. (1998). The new automated mass spectrometry deconvolution and identification system (AMDIS). *Spectroscopy Europe*, 10(3):24–27.

[del Castillo et al., 2007] del Castillo, T., Ramos, J. L., Rodríguez-Herva, J. J., Fuhrer, T., Sauer, U., and Duque, E. (2007). Convergent peripheral pathways catalyze initial glucose catabolism in Pseudomonas putida: genomic and flux analysis. *Journal of bacteriology*, 189(14):5142–5152.

[Deshpande et al., 2009] Deshpande, R., Yang, T. H., and Heinzle, E. (2009). Towards a metabolic and isotopic steady state in CHO batch cultures for reliable isotope-based metabolic profiling. *Biotechnology journal*, 4(2):247–263.

[Dias et al., 2014] Dias, O., Pereira, R., Gombert, A. K., Ferreira, E. C., and Rocha, I. (2014). iOD907, the first genome-scale metabolic model for the milk yeast Kluyveromyces lactis. *Biotechnology journal*, 9(6):776–790.

[Dias et al., 2015] Dias, O., Rocha, M., Ferreira, E., and Rocha, I. (2015). Reconstructing genome-scale metabolic models with merlin. *Nucleic Acids Research*, In Press.

[dos Santos et al., 2003] dos Santos, M. M., Gombert, A. K., Christensen, B., Olsson, L., and Nielsen, J. (2003). Identification of in vivo enzyme activities in the cometabolism of glucose and acetate by Saccharomyces cerevisiae by using 13C-labeled substrates. *Eukaryotic Cell*, 2(3):599–608.

[Driouch et al., 2012] Driouch, H., Melzer, G., and Wittmann, C. (2012). Integration of in vivo and in silico metabolic fluxes for improvement of recombinant protein production. *Metabolic engineering*, 14(1):47–58.

[Droste et al., 2013] Droste, P., Nöh, K., and Wiechert, W. (2013). Omix–A Visualization Tool for Metabolic Networks with Highest Usability and Customizability in Focus. *Chemie Ingenieur Technik*, 85(6):849–862.

[Duarte et al., 2007] Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srivas, R., and Palsson, B. O. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6):1777–1782.

[Duarte et al., 2004] Duarte, N. C., Herrgård, M. J., and Palsson, B. O. (2004). Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome research*, 14(7):1298–1309.

[Dunn and Ellis, 2005] Dunn, W. B. and Ellis, D. I. (2005). Metabolomics: current analytical platforms and methodologies. *TrAC Trends in Analytical Chemistry*, 24(4):285–294.

[Ebert et al., 2012] Ebert, B. E., Lamprecht, A.-L., Steffen, B., and Blank, L. M. (2012). Flux-P: Automating Metabolic Flux Analysis. *Metabolites*, 2(4):872–890.

[Ederer et al., 2010] Ederer, M., Schlatter, R., Witt, J., Feuer, R., Bona-Lovasz, J., Henkel, S., and Sawodny, O. (2010). An introduction to kinetic, constraint-based and Boolean modeling in systems biology. In *Control Applications (CCA), 2010 IEEE International Conference on*, pages 129–134. IEEE.

[Edwards et al., 2001] Edwards, J. S., Ibarra, R. U., and Palsson, B. O. (2001). In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nature biotechnology*, 19(2):125–130.

[Edwards and Palsson, 1999] Edwards, J. S. and Palsson, B. O. (1999). Systems properties of the Haemophilus influenzaeRd metabolic genotype. *Journal of Biological Chemistry*, 274(25):17410–17416.

[Edwards and Palsson, 2000] Edwards, J. S. and Palsson, B. O. (2000). The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences*, 97(10):5528–5533.

[Eikmanns et al., 1991] Eikmanns, B. J., Metzger, M., Reinscheid, D., Kircher, M., and Sahm, H. (1991). Amplification of three threonine biosynthesis genes

inCorynebacterium glutamicum and its influence on carbon flux in different strains. *Applied microbiology and biotechnology*, 34(5):617–622.

[Eisenreich et al., 1993] Eisenreich, W., Strauss, G., Werz, U., Fuchs, G., and Bacher, A. (1993). Retrobiosynthetic analysis of carbon fixation in the phototrophic eubacterium Chloroflexus aurantiacus. *European Journal of Biochemistry*, 215(3):619–632.

[Emmerling et al., 2002] Emmerling, M., Dauner, M., Ponti, A., Fiaux, J., Hochuli, M., Szyperski, T., Wüthrich, K., Bailey, J. E., and Sauer, U. (2002). Metabolic Flux Responses to Pyruvate Kinase Knockout in Escherichia coli. *Society*, 184(1):152–164.

[Engels et al., 2008] Engels, B., Dahm, P., and Jennewein, S. (2008). Metabolic engineering of taxadiene biosynthesis in yeast as a first step towards Taxol (Paclitaxel) production. *Metabolic engineering*, 10(3):201–206.

[Fan et al., 2001] Fan, T. W.-M., Lane, A. N., Shenker, M., Bartley, J. P., Crowley, D., and Higashi, R. M. (2001). Comprehensive chemical profiling of gramineous plant root exudates using high-resolution NMR and MS. *Phytochemistry*, 57(2):209–221.

[Feist et al., 2007] Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., and Palsson, B. O. (2007). A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular systems biology*, 3(1).

[Feist et al., 2009] Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., and Palsson, B. O. (2009). Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology*, 7(2):129–143.

[Feist et al., 2006] Feist, A. M., Scholten, J., Palsson, B. O., Brockman, F. J., and Ideker, T. (2006). Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. *Molecular systems biology*, 2(1).

[Fell and Wagner, 2000] Fell, D. A. and Wagner, A. (2000). The small world of metabolism. *Nature biotechnology*, 18(11):1121–1122.

[Feng et al., 2012] Feng, X., Xu, Y., Chen, Y., and Tang, Y. J. (2012). MicrobesFlux: a web platform for drafting metabolic models from the KEGG database. *BMC systems biology*, 6(1):94.

[Fiaux et al., 1999] Fiaux, J., Andersson, C. I. J., Holmberg, N., Bülow, L., Kallio, P. T., Szyperski, T., Bailey, J. E., and Wüthrich, K. (1999). 13C NMR

flux ratio analysis of Escherichia coli central carbon metabolism in microaerobic bioprocesses. *Journal of the American Chemical Society*, 121(6):1407–1408.

[Fiaux et al., 2003] Fiaux, J., Çakar, Z. P., Sonderegger, M., Wüthrich, K., Szyperski, T., and Sauer, U. (2003). Metabolic-flux profiling of the yeasts Saccharomyces cerevisiae and Pichia stipitis. *Eukaryotic cell*, 2(1):170–180.

[Fiehn, 2002] Fiehn, O. (2002). Metabolomics–the link between genotypes and phenotypes. *Plant molecular biology*, 48(1-2):155–171.

[Fiehn et al., 2000] Fiehn, O., Kopka, J., Trethewey, R. N., and Willmitzer, L. (2000). Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Analytical chemistry*, 72(15):3573–3580.

[Fischer and Sauer, 2003a] Fischer, E. and Sauer, U. (2003a). A novel metabolic cycle catalyzes glucose oxidation and anaplerosis in hungry Escherichia coli. *Journal of Biological Chemistry*, 278(47):46446–46451.

[Fischer and Sauer, 2003b] Fischer, E. and Sauer, U. (2003b). Metabolic flux profiling of Escherichia coli mutants in central carbon metabolism using GC-MS. *European Journal of Biochemistry*, 270(5):880–891.

[Fischer and Sauer, 2003c] Fischer, E. and Sauer, U. (2003c). Metabolic flux profiling of Escherichia coli mutants in central carbon metabolism using GC-MS. *European Journal of Biochemistry*, 270(5):880–891.

[Fischer and Sauer, 2005] Fischer, E. and Sauer, U. (2005). Large-scale in vivo flux analysis shows rigidity and suboptimal performance of Bacillus subtilis metabolism. *Nature genetics*, 37(6):636–640.

[Fischer et al., 2004] Fischer, E., Zamboni, N., Sauer, U., and Others (2004). High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry derived 13C constraints. *Analytical biochemistry*, 325(2):308.

[Fleischmann et al., 1995] Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., and Others (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223):496–512.

[Fong et al., 2006] Fong, S. S., Nanchen, A., Palsson, B. O., and Sauer, U. (2006). Latent pathway activation and increased pathway capacity enable Escherichia coli adaptation to loss of key metabolic enzymes. *Journal of Biological Chemistry*, 281(12):8024–8033.

[Forbes et al., 2001] Forbes, N. S., Clark, D. S., and Blanch, H. W. (2001). Using isotopomer path tracing to quantify metabolic fluxes in pathway models containing reversible reactions. *Biotechnology and Bioengineering*, 74:196–211.

[Förster et al., 2003a] Förster, J., Famili, I., Fu, P., Palsson, B. O., and Nielsen, J. (2003a). Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome research*, 13(2):244–253.

[Förster et al., 2003b] Förster, J., Famili, I., Palsson, B. O., and Nielsen, J. (2003b). Large-scale evaluation of in silico gene deletions in Saccharomyces cerevisiae. *OMICS A Journal of Integrative Biology*, 7(2):193–202.

[Förster et al., 2002] Förster, J., Gombert, A. K., and Nielsen, J. (2002). A functional genomics approach using metabolomics and in silico pathway analysis. *Biotechnology and Bioengineering*, 79(7):703–712.

[Fuhrer et al., 2005] Fuhrer, T., Fischer, E., and Sauer, U. (2005). Experimental Identification and Quantification of Glucose Metabolism in Seven Bacterial Species †. *Society*, 187(5):1581–1590.

[Funahashi et al., 2006] Funahashi, A., Matsuoka, Y., Jouraku, A., Kitano, H., and Kikuchi, N. (2006). CellDesigner: a modeling tool for biochemical networks. In *Proceedings of the 38th conference on Winter simulation*, pages 1707–1712. Winter Simulation Conference.

[Garcia-Albornoz et al., 2014] Garcia-Albornoz, M., Thankaswamy-Kosalai, S., Nilsson, A., Väremo, L., Nookaew, I., and Nielsen, J. (2014). BioMet Toolbox 2.0: genome-wide analysis of metabolism and omics data. *Nucleic acids research*, page gku371.

[Gavrilescu and Chisti, 2005] Gavrilescu, M. and Chisti, Y. (2005). Biotechnology—a sustainable alternative for chemical industry. *Biotechnology Advances*, 23(7):471–499.

[Ghosh et al., 2005] Ghosh, S., Zhu, T., Grossmann, I. E., Ataai, M. M., and Domach, M. M. (2005). Closing the loop between feasible flux scenario identification for construct evaluation and resolution of realized fluxes via NMR. *Computers & chemical engineering*, 29(3):459–466.

[Goffeau et al., 1996] Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., and Others (1996). Life with 6000 genes. *Science*, 274(5287):546–567.

[Gombert et al., 2001] Gombert, A. K., dos Santos, M. M., Christensen, B., and Nielsen, J. (2001). Network identification and flux quantification in the central metabolism of Saccharomyces cerevisiae under different conditions of glucose repression. *Journal of bacteriology*, 183(4):1441–1451.

[Gonnerman et al., 2013] Gonnerman, M. C., Benedict, M. N., Feist, A. M., Metcalf, W. W., and Price, N. D. (2013). Genomically and biochemically accurate metabolic reconstruction of Methanosarcina barkeri Fusaro, iMG746. *Biotechnology journal*, 8(9):1070–1079.

[Goodacre et al., 2004] Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., and Kell, D. B. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. *TRENDS in Biotechnology*, 22(5):245–252.

[Griffiths, 2001] Griffiths, M. (2001). The application of biotechnology to industrial sustainability. Technical report, Organization for Economic Cooperation and Development (OECD), Paris.

[Grotkjæ r et al., 2004] Grotkjæ r, T., Å kesson, M., Christensen, B., Gombert, A. K., and Nielsen, J. (2004). Impact of transamination reactions and protein turnover on labeling dynamics in 13C-labeling experiments. *Biotechnology and bioengineering*, 86(2):209–216.

[Hall, 2007] Hall, N. (2007). Advanced sequencing technologies and their wider impact in microbiology. *Journal of Experimental Biology*, 210(9):1518–1525.

[Hanke et al., 2013] Hanke, T., Nöh, K., Noack, S., Polen, T., Bringer, S., Sahm, H., Wiechert, W., and Bott, M. (2013). Combined fluxomics and transcriptomics analysis of glucose catabolism via a partially cyclic pentose phosphate pathway in Gluconobacter oxydans 621H. *Applied and environmental microbiology*, 79(7):2336–2348.

[Heath, 1968] Heath, D. F. (1968). The redistribution of carbon label by the reactions involved in glycolysis, gluconeogenesis and the tricarboxylic acid cycle in rat liver. *Biochem. J*, 110:313–335.

[Hellerstein, 2003] Hellerstein, M. K. (2003). In vivo measurement of fluxes through metabolic pathways: the missing link in functional genomics and pharmaceutical research. *Annual review of nutrition*, 23(1):379–402.

[Henry et al., 2007] Henry, C. S., Broadbelt, L. J., and Hatzimanikatis, V. (2007). Thermodynamics-based metabolic flux analysis. *Biophysical journal*, 92(5):1792–1805.

[Henry et al., 2010] Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., and Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology*, 28(9):977–982.

[Henry et al., 2005] Henry, O., Perrier, M., and Kamen, A. (2005). Metabolic flux analysis of HEK-293 cells in perfusion cultures for the production of adenoviral vectors. *Metabolic engineering*, 7(5-6):467–76.

[Herrgård et al., 2008] Herrgård, M. J., Swainston, N., Dobson, P., Dunn, W. B., Arga, K. Y., Arvas, M., Blüthgen, N., Borger, S., Costenoble, R., Heinemann, M., et al. (2008). A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature biotechnology*, 26(10):1155–1160.

[Heux et al., 2014] Heux, S., Poinot, J., Massou, S., Sokol, S., and Portais, J.-C. (2014). A novel platform for automated high-throughput fluxome profiling of metabolic variants. *Metabolic engineering*, 25:8–19.

[Heyland et al., 2011] Heyland, J., Fu, J., Blank, L. M., and Schmid, A. (2011). Carbon metabolism limits recombinant protein production in Pichia pastoris. *Biotechnology and bioengineering*, 108(8):1942–1953.

[Hua et al., 2006] Hua, Q., Joyce, A. R., Fong, S. S., and Palsson, B. O. (2006). Metabolic analysis of adaptive evolution for in silico-designed lactate-producing strains. *Biotechnology and bioengineering*, 95(5):992–1002.

[Hucka et al., 2003] Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., and Others (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531.

[Ibarra et al., 2002] Ibarra, R. U., Edwards, J. S., and Palsson, B. O. (2002). Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, 420(6912):186–189.

[Ideker et al., 2001] Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: systems biology. *Annual review of genomics and human genetics*, 2(1):343–372.

[Ikeda and Katsumata, 1992] Ikeda, M. and Katsumata, R. (1992). Metabolic engineering to produce tyrosine or phenylalanine in a tryptophan-producing Corynebacterium glutamicum strain. *Applied and environmental microbiology*, 58(3):781–785.

[Iwatani et al., 2008] Iwatani, S., Yamada, Y., and Usuda, Y. (2008). Metabolic flux analysis in biotechnology processes. *Biotechnology letters*, 30(5):791–9.

[Jeong et al., 2000] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654.

[Jordà et al., 2014] Jordà, J., de Jesus, S. S., Peltier, S., Ferrer, P., and Albiol, J. (2014). Metabolic flux analysis of recombinant Pichia pastoris growing on

different glycerol/methanol mixtures by iterative fitting of NMR-derived 13 C-labelling data from proteinogenic amino acids. *New biotechnology*, 31(1):120–132.

[Jørgensen et al., 1995] Jørgensen, H., Nielsen, J., Villadsen, J., and Møllgaard, H. (1995). Metabolic flux distributions in penicillium chrysogenum during fed-batch cultivations. *Biotechnology and Bioengineering*, 46(2):117–131.

[Jouhten et al., 2009] Jouhten, P., Pitkänen, E., Pakula, T., Saloheimo, M., Penttilä, M., and Maaheimo, H. (2009). 13C-metabolic flux ratio and novel carbon path analyses confirmed that Trichoderma reesei uses primarily the respirative pathway also on the preferred carbon source glucose. *BMC systems biology*, 3(1):104.

[Kajihata et al., 2014] Kajihata, S., Furusawa, C., Matsuda, F., and Shimizu, H. (2014). OpenMebius: An open source software for isotopically nonstationary 13 C-based metabolic flux analysis. *BioMed research international*, 2014.

[Kanehisa and Goto, 2000] Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

[Kanehisa et al., 2006] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic acids research*, 34(suppl 1):D354—-D357.

[Karp et al., 2009] Karp, P. D., Paley, S. M., Krummenacker, M., Latendresse, M., Dale, J. M., Lee, T. J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., and Others (2009). Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in bioinformatics*, page bbp043.

[Kauffman et al., 2003] Kauffman, K. J., Prakash, P., and Edwards, J. S. (2003). Advances in flux balance analysis. *Current Opinion in Biotechnology*, 14(5):491–496.

[Kauffman et al., 2004] Kauffman, S., Peterson, C., Samuelsson, B., and Troein, C. (2004). Genetic networks with canalyzing Boolean rules are always stable. *Proceedings of the National Academy of Sciences of the United States of America*, 101(49):17102–17107.

[Kell, 2004] Kell, D. B. (2004). Metabolomics and systems biology: making sense of the soup. *Current opinion in microbiology*, 7(3):296–307.

[Kern et al., 2007] Kern, A., Tilley, E., Hunter, I. S., Legiša, M., and Glieder, A. (2007). Engineering primary metabolic pathways of industrial micro-organisms. *Journal of biotechnology*, 129(1):6–29.

[Kiefer et al., 2004] Kiefer, P., Heinzle, E., Zelder, O., and Wittmann, C. (2004). Comparative metabolic flux analysis of lysine-producing Corynebacterium glutamicum cultured on glucose or fructose. *Applied and environmental microbiology*, 70(1):229–239.

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

[Kitano, 2002b] Kitano, H. (2002b). Systems biology: a brief overview. *Science*, 295(5560):1662–1664.

[Kitano and Others, 2001] Kitano, H. and Others (2001). *Foundations of systems biology*. MIT press Cambridge, MA;.

[Klamt et al., 2007] Klamt, S., Saez-Rodriguez, J., and Gilles, E. D. (2007). Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC systems biology*, 1(1):2.

[Klamt et al., 2003] Klamt, S., Stelling, J., Ginkel, M., and Gilles, E. D. (2003). FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps. *Bioinformatics*, 19(2):261–269.

[Klapa et al., 1999] Klapa, M. I., Park, S. M., Sinskey, A. J., and Stephanopoulos, G. (1999). Metabolite and isotopomer balancing in the analysis of metabolic cycles: I. Theory. *Biotechnology and bioengineering*, 62(4):375–391.

[Kleijn et al., 2006] Kleijn, R. J., van Winden, W. A., Ras, C., van Gulik, W. M., Schipper, D., and Heijnen, J. J. (2006). 13C-labeled gluconate tracing as a direct and accurate method for determining the pentose phosphate pathway split ratio in Penicillium chrysogenum. *Applied and environmental microbiology*, 72(7):4743–4754.

[Klein and Heinzle, 2012] Klein, S. and Heinzle, E. (2012). Isotope labeling experiments in metabolomics and fluxomics. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 4(3):261–272.

[Klenk et al., 1997] Klenk, H.-P., Clayton, R. A., Tomb, J.-F., White, O., Nelson, K. E., Ketchum, K. A., Dodson, R. J., Gwinn, M., Hickey, E. K., Peterson, J. D., and Others (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus. *Nature*, 390(6658):364–370.

[Koek et al., 2011] Koek, M. M., Jellema, R. H., van der Greef, J., Tas, A. C., and Hankemeier, T. (2011). Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. *Metabolomics*, 7(3):307–328.

[Kopka, 2006] Kopka, J. (2006). Current challenges and developments in GC-MS based metabolite profiling technology. *Journal of biotechnology*, 124(1):312–22.

[Krömer et al., 2004] Krömer, J. O., Sorgenfrei, O., Klopprogge, K., Heinzle, E., and Wittmann, C. (2004). In-depth profiling of lysine-producing Corynebacterium glutamicum by combined analysis of the transcriptome, metabolome, and fluxome. *Journal of bacteriology*, 186(6):1769–1784.

[Kruger et al., 2012] Kruger, N. J., Masakapalli, S. K., and Ratcliffe, R. G. (2012). Strategies for investigating the plant metabolic network with steady-state metabolic flux analysis: lessons from an Arabidopsis cell culture and other systems. *Journal of experimental botany*, 63(6):2309–2323.

[Kuepfer et al., 2005] Kuepfer, L., Sauer, U., and Blank, L. M. (2005). Metabolic functions of duplicate genes in Saccharomyces cerevisiae. *Genome research*, 15(10):1421–1430.

[Kumar et al., 1992] Kumar, V., Ramakrishnan, S., Teeri, T. T., Knowles, J. K. C., and Hartley, B. S. (1992). Saccharomyces cerevisiae cells secreting an Aspergillus niger β-galactosidase grow on whey permeate. *Nature Biotechnology*, 10(1):82–85.

[Kumar et al., 2011] Kumar, V. S., Ferry, J. G., and Maranas, C. D. (2011). Metabolic reconstruction of the archaeon methanogen Methanosarcina Acetivorans. *BMC systems biology*, 5(1):28.

[Lander et al., 2001] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., and Others (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[Lange et al., 2001] Lange, H. C., Eman, M., Van Zuijlen, G., Visser, D., Van Dam, J. C., Frank, J., De Mattos, M. J., and Heijnen, J. J. (2001). Improved rapid sampling for in vivo kinetics of intracellular metabolites in Saccharomyces cerevisiae. *Biotechnology and Bioengineering*, 75(4):406–415.

[Lapidot and Nissim, 1980] Lapidot, A. and Nissim, I. (1980). Regulation of pool sizes and turnover rates of amino acids in humans: 15 N-glycine and 15 N-alanine single-dose experiments using gas chromatography—Mass spectrometry analysis. *Metabolism*, 29(3):230–239.

[Latendresse et al., 2012a] Latendresse, M., Krummenacker, M., Trupp, M., and Karp, P. D. (2012a). Construction and completion of flux balance models from pathway databases. *Bioinformatics*, 28(3):388–396.

[Latendresse et al., 2012b] Latendresse, M., Malerich, J. P., Travers, M., and Karp, P. D. (2012b). Accurate Atom-Mapping Computation for Biochemical Reactions. *Journal of chemical information and modeling*, 52(11):2970–2982.

[Lee and Kerns, 1999] Lee, M. S. and Kerns, E. H. (1999). LC/MS applications in drug development. *Mass Spectrometry Reviews*, 18(3-4):187–279.

[Leighty and Antoniewicz, 2012] Leighty, R. W. and Antoniewicz, M. R. (2012). Parallel labeling experiments with [U-13 C] glucose validate E. coli metabolic network model for 13 C metabolic flux analysis. *Metabolic engineering*, 14(5):533–541.

[Leighty and Antoniewicz, 2013] Leighty, R. W. and Antoniewicz, M. R. (2013). COMPLETE-MFA: complementary parallel labeling experiments technique for metabolic flux analysis. *Metabolic engineering*, 20:49–55.

[Lewis et al., 2010] Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. A., Charusanti, P., Polpitiya, A. D., Adkins, J. N., Schramm, G., Purvine, S. O., Lopez-Ferrer, D., and Others (2010). Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Molecular systems biology*, 6(1).

[Lommen and Kools, 2012] Lommen, A. and Kools, H. J. (2012). MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware. *Metabolomics*, 8(4):719–726.

[Maaheimo et al., 2001] Maaheimo, H., Fiaux, J., Cakar, Z. P., Bailey, J. E., Sauer, U., and Szyperski, T. (2001). Central carbon metabolism of Saccharomyces cerevisiae explored by biosynthetic fractional 13C labeling of common amino acids. *European Journal of Biochemistry*, 268(8):2464–2479.

[Mahadevan et al., 2003] Mahadevan, R., Schilling, C. H., and Others (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic engineering*, 5(4):264.

[Malloy et al., 1988] Malloy, C. R., Sherry, A. D., and Jeffrey, F. M. (1988). Evaluation of carbon flux and substrate selection through alternate pathways involving the citric acid cycle of the heart by 13C NMR spectroscopy. *Journal of Biological Chemistry*, 263(15):6964–6971.

[Martínez et al., 2013] Martínez, V. S., Dietmair, S., Quek, L.-E., Hodson, M. P., Gray, P., and Nielsen, L. K. (2013). Flux balance analysis of CHO cells before and after a metabolic switch from lactate production to consumption. *Biotechnology and bioengineering*, 110(2):660–6.

[Marx et al., 1996] Marx, A., de Graaf, A. A., Wiechert, W., Eggeling, L., and Sahm, H. (1996). Determination of the fluxes in the central metabolism of Corynebacterium glutamicum by nuclear magnetic resonance spectroscopy combined with metabolite balancing. *Biotechnology and Bioengineering*, 49(2):111–129.

[Mavrovouniotis et al., 1990] Mavrovouniotis, M. L., Stephanopoulos, G., and Stephanopoulos, G. (1990). Computer-aided synthesis of biochemical pathways. *Biotechnology and bioengineering*, 36(11):1119–1132.

[May et al., 2008] May, P., Wienkoop, S., Kempa, S., Usadel, B., Christian, N., Rupprecht, J., Weiss, J., Recuenco-Munoz, L., Ebenhöh, O., Weckwerth, W., and Others (2008). Metabolomics-and proteomics-assisted genome annotation and analysis of the draft metabolic network of Chlamydomonas reinhardtii. *Genetics*, 179(1):157–166.

[McKinlay et al., 2007] McKinlay, J. B., Shachar-Hill, Y., Zeikus, J. G., and Vieille, C. (2007). Determining¡ i¿ Actinobacillus succinogenes¡/i¿ metabolic pathways and fluxes by NMR and GC-MS analyses of¡ sup¿ 13¡/sup¿ C-labeled metabolic product isotopomers. *Metabolic engineering*, 9(2):177–192.

[Mendes, 2002] Mendes, P. (2002). Emerging bioinformatics for the metabolome. *Briefings in bioinformatics*, 3(2):134–145.

[Metallo et al., 2009] Metallo, C. M., Walther, J. L., and Stephanopoulos, G. (2009). Evaluation of 13 C isotopic tracers for metabolic flux analysis in mammalian cells. *Journal of biotechnology*, 144(3):167–174.

[Millard et al., 2014] Millard, P., Massou, S., Wittmann, C., Portais, J.-C., and Létisse, F. (2014). Sampling of intracellular metabolites for stationary and non-stationary 13 C metabolic flux analysis in Escherichia coli. *Analytical biochemistry*, 465:38–49.

[Mo et al., 2009] Mo, M. L., Palsson, B. O., and Herrgård, M. J. (2009). Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC systems biology*, 3(1):37.

[Möllney et al., 1999] Möllney, M., Wiechert, W., Kownatzki, D., and de Graaf, A. A. (1999). Bidirectional reaction steps in metabolic networks: IV. Optimal design of isotopomer labeling experiments. *Biotechnology and bioengineering*, 66(2):86–103.

[Nadeau et al., 2000] Nadeau, I., Sabatié, J., Koehl, M., Perrier, M., and Kamen, A. (2000). Human 293 cell metabolism in low glutamine-supplied culture: interpretation of metabolic changes through metabolic flux analysis. *Metabolic engineering*, 2(4):277–92.

[Nanchen et al., 2007] Nanchen, A., Fuhrer, T., and Sauer, U. (2007). Determination of Metabolic Flux Ratios From 13C-Experiments and Gas Chromatography-Mass Spectrometry Data. In *Metabolomics*, pages 177–197. Springer.

[Nawrath et al., 1994] Nawrath, C., Poirier, Y., and Somerville, C. (1994). Targeting of the polyhydroxybutyrate biosynthetic pathway to the plastids of Arabidopsis thaliana results in high levels of polymer accumulation. *Proceedings of the National Academy of Sciences*, 91(26):12760–12764.

[Neuweger et al., 2008] Neuweger, H., Albaum, S. P., Dondrup, M., Persicke, M., Watt, T., Niehaus, K., Stoye, J., and Goesmann, A. (2008). MeltDB: a software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics*, 24(23):2726–2732.

[Niedenführ et al., 2015] Niedenführ, S., Wiechert, W., and Nöh, K. (2015). How to measure metabolic fluxes: a taxonomic guide for 13 C fluxomics. *Current opinion in biotechnology*, 34:82–90.

[Nielsen, 2001] Nielsen, J. (2001). Metabolic engineering. *Applied Microbiology and Biotechnology*, 55(3):263–283.

[Nielsen, 2003] Nielsen, J. (2003). It is all about metabolic fluxes. *Journal of Bacteriology*, 185(24):7031–7035.

[Nielsen and Oliver, 2005] Nielsen, J. and Oliver, S. (2005). The next wave in metabolome analysis. *Trends in biotechnology*, 23(11):544–546.

[Nielsen et al., 2007] Nielsen, J., Villas-Bôas, S. G., Roessner, U., Hansen, M. A. E., and Smedsgaard, J. r. (2007). Metabolomics in functional genomics and systems biology. *Metabolome Analysis: An Introduction*, pages 1–14.

[Nielsen, 2008] Nielsen, L. K. (2008). On the reconstruction of the Mus musculus genome-scale metabolic network model. *Genome Informatics*, 21:253.

[Nilsson et al., 1996] Nilsson, T., Larsen, T. O., Montanarella, L., and Madsen, J. r. O. g. (1996). Application of head-space solid-phase microextraction for the analysis of volatile metabolites emitted by Penicillium species. *Journal of Microbiological Methods*, 25(3):245–255.

[Nissim and Lapidot, 1986] Nissim, I. and Lapidot, A. (1986). Dynamic aspects of amino acid metabolism in alloxan-induced diabetes and insulin-treated rabbits: In vivo studies with 15 N and gas chromatography-mass spectrometry. *Biochemical medicine and metabolic biology*, 35(1):88–100.

[Nissim et al., 1981] Nissim, I., Yudkoff, M., Yang, W., Terwilliger, T., and Segal, S. (1981). Rapid gas chromatographic-mass spectrometric analysis of [15 N] urea: Application to human metabolic studies. *Clinica Chimica Acta*, 109(3):295–304.

[Niu et al., 2013] Niu, H., Chen, Y., Yao, S., Liu, L., Yang, C., Li, B., Liu, D., Xie, J., Chen, X., Wu, J., and Others (2013). Metabolic flux analysis of¡

i¿ Arthrobacter¡/i¿ sp. CGMCC 3584 for cAMP production based on¡ sup¿ 13¡/sup¿ C tracer experiments and gas chromatography–mass spectrometry. *Journal of Biotechnology*.

[Nöh et al., 2007] Nöh, K., Grönke, K., Luo, B., Takors, R., Oldiges, M., and Wiechert, W. (2007). Metabolic flux analysis at ultra short time scale: isotopically non-stationary 13C labeling experiments. *Journal of biotechnology*, 129(2):249–67.

[Nöh et al., 2006] Nöh, K., Wahl, A., and Wiechert, W. (2006). Computational tools for isotopically instationary 13 C labeling experiments under metabolic steady state conditions. *Metabolic engineering*, 8(6):554–577.

[Nöh and Wiechert, 2006] Nöh, K. and Wiechert, W. (2006). Experimental design principles for isotopically instationary 13C labeling experiments. *Biotechnology and bioengineering*, 94(2):234–251.

[Nookaew et al., 2008] Nookaew, I., Jewett, M. C., Meechai, A., Thammarongtham, C., Laoteng, K., Cheevadhanarak, S., Nielsen, J., and Bhumiratana, S. (2008). The genome-scale metabolic model iIN800 of Saccharomyces cerevisiae and its validation: a scaffold to query lipid metabolism. *BMC Systems Biology*, 2(1):71.

[Noronha et al., 2014] Noronha, A., Vilaça, P., and Rocha, M. (2014). An integrated network visualization framework towards metabolic engineering applications. *BMC bioinformatics*, 15(1):6594.

[OECD, 2011] OECD (2011). *Future Prospects for Industrial Biotechnology*. Paris, France: OECD Publishing.

[Oksman-Caldentey and Saito, 2005] Oksman-Caldentey, K.-M. and Saito, K. (2005). Integrating genomics and metabolomics for engineering plant metabolic pathways. *Current opinion in biotechnology*, 16(2):174–179.

[Orth et al., 2011] Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M., and Palsson, B. O. (2011). A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011. *Molecular systems biology*, 7(1).

[Österlund et al., 2013] Österlund, T., Nookaew, I., Bordel, S., and Nielsen, J. (2013). Mapping condition-dependent regulation of metabolism in yeast through genome-scale modeling. *BMC systems biology*, 7(1):36.

[O'Grady et al., 2012] O'Grady, J., Schwender, J., Shachar-Hill, Y., and Morgan, J. A. (2012). Metabolic cartography: experimental quantification of metabolic fluxes from isotopic labelling studies. *Journal of experimental botany*, page ers032.

[Pagani et al., 2012] Pagani, I., Liolios, K., Jansson, J., Chen, I.-M. A., Smirnova, T., Nosrat, B., Markowitz, V. M., and Kyrpides, N. C. (2012). The Genomes OnLine Database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata. *Nucleic acids research*, 40(D1):D571—-D579.

[Palsson et al., 2003] Palsson, B. O., Price, N. D., and Papin, J. A. (2003). Development of network-based pathway definitions: the need to analyze real metabolic networks. *TRENDS in Biotechnology*, 21(5):195–198.

[Pandey and Mann, 2000] Pandey, A. and Mann, M. (2000). Proteomics to study genes and genomes. *Nature*, 405(6788):837–846.

[Papoutsakis, 1984] Papoutsakis, E. T. (1984). Equations and calculations for fermentations of butyric acid bacteria. *Biotechnology and bioengineering*, 26(2):174–187.

[Park et al., 2005] Park, S. J., Lee, S. Y., Cho, J., Kim, T. Y., Lee, J. W., Park, J. H., and Han, M.-J. (2005). Global physiological understanding and metabolic engineering of microorganisms based on omics studies. *Applied microbiology and biotechnology*, 68(5):567–579.

[Peoples et al., 1988] Peoples, O. P., Liebl, W., Bodis, M., Maeng, P. J., Follettie, M. T., Archer, J. A., and Sinskey, A. J. (1988). Nucleotide sequence and fine structural analysis of the Corynebacterium glutamicum hom-thrB operon. *Molecular microbiology*, 2(1):63–72.

[Petersen et al., 2000] Petersen, S., de Graaf, A. A., Eggeling, L., Möllney, M., Wiechert, W., and Sahm, H. (2000). In vivo quantification of parallel and bidirectional fluxes in the anaplerosis of Corynebacterium glutamicum. *Journal of Biological Chemistry*, 275(46):35932–35941.

[Pitkänen et al., 2014] Pitkänen, E., Jouhten, P., Hou, J., Syed, M. F., Blomberg, P., Kludas, J., Oja, M., Holm, L., Penttilä, M., Rousu, J., and Others (2014). Comparative Genome-Scale Reconstruction of Gapless Metabolic Networks for Present and Ancestral Species. *PLoS computational biology*, 10(2):e1003465.

[Pitkänen et al., 2009] Pitkänen, E., Jouhten, P., and Rousu, J. (2009). Inferring branching pathways in genome-scale metabolic networks. *BMC systems biology*, 3(1):103.

[Quek et al., 2009] Quek, L.-E., Wittmann, C., Nielsen, L. K., and Krömer, J. O. (2009). OpenFLUX: efficient modelling software for 13C-based metabolic flux analysis. *Microbial cell factories*, 8(1):25.

[Rantanen et al., 2008] Rantanen, A., Rousu, J., Jouhten, P., Zamboni, N., Maaheimo, H., and Ukkonen, E. (2008). An analytic and systematic framework for estimating metabolic flux ratios from 13C tracer experiments. *BMC bioinformatics*, 9:266.

[Ravikirthi et al., 2011] Ravikirthi, P., Suthers, P. F., and Maranas, C. D. (2011). Construction of an E. Coli genome-scale atom mapping model for MFA calculations. *Biotechnology and bioengineering*, 108(6):1372–1382.

[Reed et al., 2003] Reed, J. L., Vo, T. D., Schilling, C. H., Palsson, B. O., and Others (2003). An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biol*, 4(9):R54.

[Rex et al., 2013] Rex, R., Bill, N., Schmidt-Hohagen, K., and Schomburg, D. (2013). Swimming in light: a large-scale computational analysis of the metabolism of Dinoroseobacter shibae. *PLoS computational biology*, 9(10):e1003224.

[Ro et al., 2006] Ro, D.-K., Paradise, E. M., Ouellet, M., Fisher, K. J., Newman, K. L., Ndungu, J. M., Ho, K. A., Eachus, R. A., Ham, T. S., Kirby, J., and Others (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, 440(7086):940–943.

[Rocha et al., 2008] Rocha, I., Förster, J., and Nielsen, J. (2008). Design and application of genome-scale reconstructed metabolic models. In *Microbial Gene Essentiality: Protocols and Bioinformatics*, pages 409–431. Springer.

[Saier et al., 2006] Saier, M. H., Tran, C. V., and Barabote, R. D. (2006). TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic acids research*, 34(suppl 1):D181—-D186.

[Saito and Matsuda, 2010] Saito, K. and Matsuda, F. (2010). Metabolomics for functional genomics, systems biology, and biotechnology. *Annual review of plant biology*, 61:463–89.

[Sánchez et al., 2006] Sánchez, A. M., Bennett, G. N., and San, K.-Y. (2006). Batch culture characterization and metabolic flux analysis of succinate-producing Escherichia coli strains. *Metabolic engineering*, 8(3):209–26.

[Sanger et al., 1977] Sanger, F., Air, G., Barrell, B., Brown, N., Coulson, A., Fiddes, C., Hutchison, C., Slocombe, P., and Smith, M. (1977). Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature*, 265:687–695.

[Sanger and Coulson, 1975] Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3):441–448.

[Sauer et al., 1999] Sauer, U., Lasko, D. R., Fiaux, J., Hochuli, M., Glaser, R. W., Szyperski, T., Wüthrich, K., and Bailey, J. E. (1999). Metabolic flux ratio analysis of genetic and environmental modulations of Escherichia coli central carbon metabolism. *Journal of bacteriology*, 181(21):6679–88.

[Savinell and Palsson, 1992a] Savinell, J. M. and Palsson, B. O. (1992a). Optimal selection of metabolic fluxes for¡ i¿ in vivo¡/i¿ measurement. II. Application to¡ i¿ Escherichia coli¡/i¿ and hybridoma cell metabolism. *Journal of theoretical biology*, 155(2):215–242.

[Savinell and Palsson, 1992b] Savinell, J. M. and Palsson, B. O. (1992b). Optimal selection of metabolic fluxes for in vivo measurement. I. Development of mathematical methods. *Journal of theoretical biology*, 155(2):201–214.

[Schellenberger et al., 2011] Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., and Others (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2. 0. *Nature protocols*, 6(9):1290–1307.

[Schilling et al., 2002] Schilling, C. H., Covert, M. W., Famili, I., Church, G. M., Edwards, J. S., and Palsson, B. O. (2002). Genome-scale metabolic model of Helicobacter pylori 26695. *Journal of bacteriology*, 184(16):4582–4593.

[Schilling and Palsson, 2000] Schilling, C. H. and Palsson, B. O. (2000). Assessment of the metabolic capabilities of Haemophilus influenzae Rd through a genome-scale pathway analysis. *Journal of theoretical biology*, 203(3):249–283.

[Schmid et al., 2002] Schmid, A., Hollmann, F., Park, J. B., and Bühler, B. (2002). The use of enzymes in the chemical industry in Europe. *Current Opinion in Biotechnology*, 13(4):359–366.

[Schmidt et al., 1997] Schmidt, K., Carlsen, M., Nielsen, J., and Villadsen, J. (1997). Modeling isotopomer distributions in biochemical networks using isotopomer mapping matrices. *Biotechnology and bioengineering*, 55(6):831–40.

[Schmidt et al., 1999] Schmidt, K., Nielsen, J., and Villadsen, J. (1999). Quantitative analysis of metabolic fluxes in Escherichia coli, using two-dimensional NMR spectroscopy and complete isotopomer models. *Journal of Biotechnology*, 71(1):175–189.

[Schoenheimer et al., 1940] Schoenheimer, R., Rittenberg, D., and Others (1940). The Study of IntermediaryMetabolism of Animals with the Aid of Isotopes. *Physiological Reviews*, 20:218–248.

[Scholz and Fiehn, 2006] Scholz, M. and Fiehn, O. (2006). SetupX–a public study design database for metabolomic projects. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 169–180.

[Schomburg et al., 2002] Schomburg, I., Chang, A., and Schomburg, D. (2002). BRENDA, enzyme data and metabolic information. *Nucleic acids research*, 30(1):47–49.

[Schubert et al., 1988] Schubert, P., Steinbüchel, A., and Schlegel, H. G. (1988). Cloning of the Alcaligenes eutrophus genes for synthesis of poly-beta-hydroxybutyric acid (PHB) and synthesis of PHB in Escherichia coli. *Journal of Bacteriology*, 170(12):5837–5847.

[Schuetz et al., 2007] Schuetz, R., Kuepfer, L., and Sauer, U. (2007). Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Molecular systems biology*, 3(1).

[Schuster et al., 2000] Schuster, S., Fell, D. A., and 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.

[Segre et al., 2002] Segre, D., Vitkup, D., and Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, 99(23):15112–15117.

[Selvarasu et al., 2009] Selvarasu, S., Karimi, I. A., Ghim, G.-H., and Lee, D.-Y. (2009). Genome-scale modeling and in silico analysis of mouse cell metabolic network. *Molecular BioSystems*, 6(1):152–161.

[Sengupta et al., 2011] Sengupta, N., Rose, S. T., and Morgan, J. A. (2011). Metabolic flux analysis of CHO cell metabolism in the late non-growth phase. *Biotechnology and bioengineering*, 108(1):82–92.

[Sequencing et al., 2005] Sequencing, T. C., Consortium, A., and Others (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87.

[Shannon et al., 2003] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.

[Sheikh et al., 2005] Sheikh, K., Förster, J., and Nielsen, L. K. (2005). Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of Mus musculus. *Biotechnology progress*, 21(1):112–21.

[Shen et al., 1999] Shen, J., Petersen, K. F., Behar, K. L., Brown, P., Nixon, T. W., Mason, G. F., Petroff, O. A. C., Shulman, G. I., Shulman, R. G., and Rothman, D. L. (1999). Determination of the rate of the glutamate/glutamine cycle in the human brain by in vivo 13C NMR. *Proceedings of the National Academy of Sciences*, 96(14):8235–8240.

[Shimizu, 2002] Shimizu, A. (2002). A review on metabolic pathway analysis with emphasis on isotope labeling approach. *Biotechnology and Bioprocess Engineering*, 7(5):237–251.

[Shlomi et al., 2005] Shlomi, T., Berkman, O., and Ruppin, E. (2005). Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7695–7700.

[Shlomi et al., 2007] Shlomi, T., Eisenberg, Y., Sharan, R., and Ruppin, E. (2007). A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Molecular systems biology*, 3(1):101.

[Shupletsov et al., 2014] Shupletsov, M. S., Golubeva, L. I., Rubina, S. S., Podvyaznikov, D. A., Iwatani, S., and Mashko, S. V. (2014). OpenFLUX2 : 13 C-MFA modeling software package adjusted for the comprehensive analysis of single and parallel labeling experiments. pages 1–25.

[Sidorenko et al., 2008] Sidorenko, Y., Wahl, A., Dauner, M., Genzel, Y., and Reichl, U. (2008). Comparison of metabolic flux distributions for MDCK cell growth in glutamine- and pyruvate-containing media. *Biotechnology progress*, 24(2):311–20.

[Sigurdsson et al., 2010] Sigurdsson, M. I., Jamshidi, N., Steingrimsson, E., Thiele, I., and Palsson, B. O. (2010). A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. *BMC systems biology*, 4(1):140.

[Silva-Rocha et al., 2013] Silva-Rocha, R., Chavarría, M., Kleijn, R. J., Sauer, U., and de Lorenzo, V. (2013). The IHF regulon of exponentially growing Pseudomonas putida cells. *Environmental microbiology*, 15(1):49–63.

[Slater et al., 1988] Slater, S. C., Voige, W. H., and Dennis, D. E. (1988). Cloning and expression in Escherichia coli of the Alcaligenes eutrophus H16 poly-beta-hydroxybutyrate biosynthetic pathway. *Journal of bacteriology*, 170(10):4431–4436.

[Smallbone et al., 2013] Smallbone, K., Messiha, H. L., Carroll, K. M., Winder, C. L., Malys, N., Dunn, W. B., Murabito, E., Swainston, N., Dada, J. O., Khan, F., and Others (2013). A model of yeast glycolysis based on a consistent kinetic characterisation of all its enzymes. *FEBS letters*, 587(17):2832–2841.

[Smallbone et al., 2010] Smallbone, K., Simeonidis, E., Swainston, N., and Mendes, P. (2010). Towards a genome-scale kinetic model of cellular metabolism. *BMC systems biology*, 4(1):6.

[Smedsgaard and Nielsen, 2005] Smedsgaard, J. r. and Nielsen, J. (2005). Metabolite profiling of fungi and yeast: from phenotype to metabolome by MS and informatics. *Journal of experimental botany*, 56(410):273–286.

[Sokol et al., 2012] Sokol, S., Millard, P., and Portais, J. C. (2012). Influx_s: Increasing numerical stability and precision for metabolic flux analysis in isotope labelling experiments. *Bioinformatics*, 28(5):687–693.

[Sonntag et al., 1993] Sonntag, K., Eggeling, L., Graaf, A. A., and Sahn, H. (1993). Flux partitioning in the split pathway of lysine synthesis in Corynebacterium glutamicum. *European journal of biochemistry*, 213(3):1325–1331.

[Sriram et al., 2004] Sriram, G., Fulton, D. B., Iyer, V. V., Peterson, J. M., Zhou, R., Westgate, M. E., Spalding, M. H., and Shanks, J. V. (2004). Quantification of compartmented metabolic fluxes in developing soybean embryos by employing biosynthetically directed fractional 13C labeling, two-dimensional [13C, 1H] nuclear magnetic resonance, and comprehensive isotopomer balancing. *Plant Physiology*, 136(2):3043–3057.

[Sriram et al., 2007] Sriram, G., Fulton, D. B., and Shanks, J. V. (2007). Flux quantification in central carbon metabolism of Catharanthus roseus hairy roots by 13 C labeling and comprehensive bondomer balancing. *Phytochemistry*, 68(16):2243–2257.

[Srour et al., 2011] Srour, O., Young, J. D., and Eldar, Y. C. (2011). Fluxomers: A new approach for 13C metabolic flux analysis. *BMC systems biology*, 5(1):129.

[Stein, 1999] Stein, S. E. (1999). An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry*, 10(8):770–781.

[Stephanopoulos, 1999] Stephanopoulos, G. (1999). Metabolic fluxes and metabolic engineering. *Metabolic engineering*, 1(1):1–11.

[Stephanopoulos et al., 1998] Stephanopoulos, G., Aristidou, A. A., Nielsen, J. H. i., and Nielsen, J. (1998). *Metabolic engineering: principles and methodologies*. Academic Press.

[Stephanopoulos et al., 1991] Stephanopoulos, G., Vallino, J. J., for the Advancement of Science, A. A., and Others (1991). Network rigidity and metabolic engineering in metabolite overproduction. *Science(Washington)*, 252(5013):1675–1681.

[Strisower et al., 1951] Strisower, E. H., Chaikoff, I. L., and Weinman, E. O. (1951). Conversion of C14-palmitic acid to glucose. I. Normal and diabetic rats. *Journal of Biological Chemistry*, 192(2):453–463.

[Suthers et al., 2007a] Suthers, P. F., Burgard, A. P., Dasika, M. S., Nowroozi, F., Van Dien, S., Keasling, J. D., and Maranas, C. D. (2007a). Metabolic flux elucidation for large-scale models using 13 C labeled isotopes. *Metabolic engineering*, 9(5):387–405.

[Suthers et al., 2007b] Suthers, P. F., Burgard, A. P., Dasika, M. S., Nowroozi, F., Van Dien, S., Keasling, J. D., and Maranas, C. D. (2007b). Metabolic flux elucidation for large-scale models using 13C labeled isotopes. *Metabolic engineering*, 9(5-6):387–405.

[Suthers et al., 2010] Suthers, P. F., Chang, Y. J., and Maranas, C. D. (2010). Improved computational performance of MFA using elementary metabolite units and flux coupling. *Metabolic engineering*, 12(2):123–8.

[Szyperski, 1995] Szyperski, T. (1995). Biosynthetically directed fractional 13C-labeling of proteinogenic amino acids. An efficient analytical tool to investigate intermediary metabolism.

[Szyperski, 1998] Szyperski, T. (1998). 13C-NMR, MS and metabolic flux balancing in biotechnology research. *Q. Rev. Biophys*, 31(1):41–31106.

[Szyperski et al., 1996] Szyperski, T., Bailey, J. E., and Wüthrich, K. (1996). Detecting and dissecting metabolic fluxes using biosynthetic fractional 13 C labeling and two-dimensional NMR spectroscopy. *Trends in Biotechnology*, 14(12):453–459.

[Szyperski et al., 1999] Szyperski, T., Glaser, R. W., Hochuli, M., Fiaux, J., Sauer, U., Bailey, J. E., and Wüthrich, K. (1999). Bioreaction network topology and metabolic flux ratio analysis by biosynthetic fractional 13C labeling and two-dimensional NMR spectroscopy. *Metabolic engineering*, 1(3):189–197.

[Tang et al., 2007] Tang, Y. J., Hwang, J. S., Wemmer, D. E., and Keasling, J. D. (2007). Shewanella oneidensis MR-1 fluxome under various oxygen conditions. *Applied and environmental microbiology*, 73(3):718–729.

[Tao et al., 2012] Tao, Y., Liu, D., Yan, X., Zhou, Z., Lee, J. K., and Yang, C. (2012). Network identification and flux quantification of glucose metabolism in Rhodobacter sphaeroides under photoheterotrophic H2-producing conditions. *Journal of bacteriology*, 194(2):274–283.

[Tesch et al., 1999] Tesch, M., De Graaf, A. A., and Sahm, H. (1999). In vivo fluxes in the ammonium-assimilatory pathways in Corynebacterium glutamicum studied by15N nuclear magnetic resonance. *Applied and environmental microbiology*, 65(3):1099–1109.

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

[Thiele et al., 2013] Thiele, I., Swainston, N., Fleming, R. M. T., Hoppe, A., Sahoo, S., Aurich, M. K., Haraldsdottir, H., Mo, M. L., Rolfsson, O., Stobbe, M. D., and Others (2013). A community-driven global reconstruction of human metabolism. *Nature biotechnology*, 31(5):419–425.

[Thiele et al., 2005] Thiele, I., Vo, T. D., Price, N. D., and Palsson, B. O. (2005). Expanded metabolic reconstruction of Helicobacter pylori (iIT341 GSM/GPR): an in silico genome-scale characterization of single-and double-deletion mutants. *Journal of bacteriology*, 187(16):5818–5830.

[Thimm et al., 2004] Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L. A., Rhee, S. Y., and Stitt, M. (2004). mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal*, 37(6):914–939.

[Tokimatsu et al., 2005] Tokimatsu, T., Sakurai, N., Suzuki, H., Ohta, H., Nishitani, K., Koyama, T., Umezawa, T., Misawa, N., Saito, K., and Shibata, D. (2005). KaPPA-View. A web-based analysis tool for integration of transcript and metabolite data on plant metabolic pathway maps. *Plant physiology*, 138(3):1289–1300.

[Tomita, 2001] Tomita, M. (2001). Whole-cell simulation: a grand challenge of the 21st century. *Trends in biotechnology*, 19(6):205–210.

[Tsalikian et al., 1984] Tsalikian, E., Howard, C., Gerich, J. E., and Haymond, M. W. (1984). Increased leucine flux in short-term fasted human subjects: evidence for increased proteolysis. *American Journal of Physiology-Endocrinology And Metabolism*, 247(3):E323—-E327.

[UkáKim et al., 2008] UkáKim, H., YongáKim, T., and YupáLee, S. (2008). Metabolic flux analysis and metabolic engineering of microorganisms. *Molecular BioSystems*, 4(2):113–120.

[Vallino and Stephanopoulos, 1993] Vallino, J. J. and Stephanopoulos, G. (1993). Metabolic flux distributions in Corynebacterium glutamicum during growth and lysine overproduction. *Biotechnology and Bioengineering*, 41(6):633–646.

[van Dam et al., 2002] van Dam, J. C., Eman, M. R., Frank, J., Lange, H. C., van Dedem, G. W. K., and Heijnen, S. J. (2002). Analysis of glycolytic intermediates in Saccharomyces cerevisiae using anion exchange chromatography and electrospray ionization with tandem mass spectrometric detection. *Analytica Chimica Acta*, 460(2):209–218.

[van der Werf et al., 2005] van der Werf, M. J., Jellema, R. H., and Hankemeier, T. (2005). Microbial metabolomics: replacing trial-and-error by the unbiased selection and ranking of targets. *Journal of Industrial Microbiology and Biotechnology*, 32(6):234–252.

[van Winden et al., 2001] van Winden, W., Verheijen, P., and Heijnen, S. (2001). Possible pitfalls of flux calculations based on 13 C-labeling. *Metabolic engineering*, 3(2):151–162.

[van Winden et al., 2002] van Winden, W. A., Heijnen, J. J., and Verheijen, P. J. T. (2002). Cumulative bondomers: a new concept in flux analysis from 2D [13C, 1H] COSY NMR data. *Biotechnology and bioengineering*, 80(7):731–745.

[Varma and Palsson, 1994] Varma, A. and Palsson, B. O. (1994). Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. *Nature biotechnology*, 12:994–998.

[Villas-Bôas et al., 2005] Villas-Bôas, S. G., Mas, S., Akesson, M., Smedsgaard, J. r., and Nielsen, J. (2005). Mass spectrometry in metabolome analysis. *Mass spectrometry reviews*, 24(5):613–46.

[Villas-Boas et al., 2007] Villas-Boas, S. G., Nielsen, J., Smedsgaard, J., Hansen, M. A. E., and Roessner-Tunali, U. (2007). *Metabolome analysis: an introduction*, volume 24. John Wiley & Sons.

[Wahrheit et al., 2011] Wahrheit, J., Nicolae, A., and Heinzle, E. (2011). Eukaryotic metabolism: measuring compartment fluxes. *Biotechnology journal*, 6(9):1071–1085.

[Walker et al., 1982] Walker, T. E., Han, C. H., Kollman, V. H., London, R. E., and Matwiyoff, N. A. (1982). 13C nuclear magnetic resonance studies of the biosynthesis by Microbacterium ammoniaphilum of L-glutamate selectively enriched with carbon-13. *Journal of Biological Chemistry*, 257(3):1189–1195.

[Walsh and Koshland, 1984] Walsh, K. and Koshland, D. E. (1984). Determination of flux through the branch point of two metabolic cycles. The tricarboxylic acid cycle and the glyoxylate shunt. *Journal of Biological Chemistry*, 259(15):9646–9654.

[Wang et al., 2012] Wang, Z.-J., Wang, P., Liu, Y.-W., Zhang, Y.-M., Chu, J., Huang, M.-z., Zhuang, Y.-P., and Zhang, S.-L. (2012). Metabolic flux analysis of the central carbon metabolism of the industrial vitamin B 12 producing strain Pseudomonas denitrificans using 13 C-labeled glucose. *Journal of the Taiwan Institute of Chemical Engineers*, 43(2):181–187.

[Weckwerth, 2003] Weckwerth, W. (2003). Metabolomics in systems biology. *Annual review of plant biology*, 54(1):669–689.

[Weinman et al., 1950] Weinman, E. O., Chaikoff, I. L., Dauben, W. G., Gee, M., and Entenman, C. (1950). Relative rates of conversion of the various carbon atoms of palmitic acid to carbon dioxide by the intact rat. *Journal of Biological Chemistry*, 184(2):735–744.

[Weitzel et al., 2013] Weitzel, M., Nöh, K., Dalman, T., Niedenführ, S., Stute, B., and Wiechert, W. (2013). 13CFLUX2—high-performance software suite for 13C-metabolic flux analysis. *Bioinformatics*, 29(1):143–145.

[Weitzel et al., 2007] Weitzel, M., Wiechert, W., and Nöh, K. (2007). The topology of metabolic isotope labeling networks. *BMC bioinformatics*, 8(1):315.

[Westfall et al., 2012] Westfall, P. J., Pitera, D. J., Lenihan, J. R., Eng, D., Woolard, F. X., Regentin, R., Horning, T., Tsuruta, H., Melis, D. J., Owens, A., and Others (2012). Production of amorphadiene in yeast, and its conversion to dihydroartemisinic acid, precursor to the antimalarial agent artemisinin. *Proceedings of the National Academy of Sciences*, 109(3):E111—-E118.

[Wheeler et al., 2007] Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., and Others (2007). Database resources of the national center for biotechnology information. *Nucleic acids research*, 35(suppl 1):D5—-D12.

[Wiechert, 2001] Wiechert, W. (2001). 13C metabolic flux analysis. *Metabolic engineering*, 3(3):195–206.

[Wiechert and de Graaf, 1997] Wiechert, W. and de Graaf, A. A. (1997). Bidirectional reaction steps in metabolic networks: I. Modeling and simulation of carbon isotope labeling experiments. *Biotechnology and bioengineering*, 55(1):101–17.

[Wiechert et al., 1999] Wiechert, W., Möllney, M., Isermann, N., Wurzel, M., and de Graaf, A. A. (1999). Bidirectional reaction steps in metabolic networks: III. Explicit solution and analysis of isotopomer labeling systems. *Biotechnology and bioengineering*, 66(2):69–85.

[Wiechert et al., 2001] Wiechert, W., Möllney, M., Petersen, S., and de Graaf, A. A. (2001). A universal framework for 13C metabolic flux analysis. *Metabolic engineering*, 3(3):265–83.

[Wiechert and Nöh, 2005] Wiechert, W. and Nöh, K. (2005). From Stationary to Instationary Metabolic Flux Analysis. *Adv Biochem Eng Biotechnol*, 92:145 –172.

[Wiechert and Nöh, 2013] Wiechert, W. and Nöh, K. (2013). Isotopically non-stationary metabolic flux analysis: complex yet highly informative. *Current opinion in biotechnology*.

[Winden et al., 2005] Winden, W. A., Dam, J. C., Ras, C., Kleijn, R. J., Vinke, J. L., Gulik, W. M., and Heijnen, J. J. (2005). Metabolic-flux analysis of Saccharomyces cerevisiae CEN. PK113-7D based on mass isotopomer measurements of 13C-labeled primary metabolites. *FEMS yeast research*, 5(6-7):559–568.

[Winder et al., 2011] Winder, C. L., Dunn, W. B., and Goodacre, R. (2011). TARDIS-based microbial metabolomics: time and relative differences in systems. *Trends in microbiology*, 19(7):315–322.

[Wittmann, 2007] Wittmann, C. (2007). Fluxome analysis using GC-MS. *Microbial cell factories*, 6:6.

[Wittmann and Heinzle, 1999] Wittmann, C. and Heinzle, E. (1999). Mass spectrometry for metabolic flux analysis. *Biotechnology and bioengineering*, 62(6):739–750.

[Wittmann and Heinzle, 2001] Wittmann, C. and Heinzle, E. (2001). Modeling and experimental design for metabolic flux analysis of lysine-producing Corynebacteria by mass spectrometry. *Metabolic engineering*, 3(2):173–191.

[Wittmann and Heinzle, 2002] Wittmann, C. and Heinzle, E. (2002). Genealogy profiling through strain improvement by using metabolic network analysis: metabolic flux genealogy of several generations of lysine-producing corynebacteria. *Applied and environmental microbiology*, 68(12):5843–5859.

[Woolston et al., 2013] Woolston, B. M., Edgar, S., and Stephanopoulos, G. (2013). Metabolic engineering: past and future. *Annual review of chemical and biomolecular engineering*, 4:259–288.

[Wright and Wagner, 2008] Wright, J. and Wagner, A. (2008). The systems biology research tool: evolvable open-source software. *BMC systems biology*, 2(1):55.

[Wu, 2000] Wu, Y. (2000). The use of liquid chromatography–mass spectrometry for the identification of drug degradation products in pharmaceutical formulations. *Biomedical Chromatography*, 14(6):384–396.

[Xia et al., 2009] Xia, J., Psychogios, N., Young, N., and Wishart, D. S. (2009). MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic acids research*, 37(suppl 2):W652—-W660.

[Xiong et al., 2010] Xiong, W., Liu, L., Wu, C., Yang, C., and Wu, Q. (2010). 13C-tracer and gas chromatography-mass spectrometry analyses reveal metabolic flux distribution in the oleaginous microalga Chlorella protothecoides. *Plant physiology*, 154(2):1001–1011.

[Yoo et al., 2008] Yoo, H., Antoniewicz, M. R., Stephanopoulos, G., and Kelleher, J. K. (2008). Quantifying reductive carboxylation flux of glutamine to lipid in a brown adipocyte cell line. *Journal of Biological Chemistry*, 283(30):20621–20627.

[Young, 2014] Young, J. D. (2014). INCA: a computational platform for isotopically non-stationary metabolic flux analysis. *Bioinformatics*, 30(9):1333–1335.

[Young et al., 2011] Young, J. D., Shastri, A. A., Stephanopoulos, G., and Morgan, J. A. (2011). Mapping photoautotrophic metabolism with isotopically nonstationary 13 C flux analysis. *Metabolic engineering*, 13(6):656–665.

[Young et al., 2008] Young, J. D., Walther, J. L., Antoniewicz, M. R., Yoo, H., and Stephanopoulos, G. (2008). An elementary metabolite unit (EMU) based method of isotopically nonstationary flux analysis. *Biotechnology and bioengineering*, 99(3):686–699.

[Yuan et al., 2010] Yuan, Y., Yang, T. H., and Heinzle, E. (2010). 13C metabolic flux analysis for larger scale cultivation using gas chromatography-combustion-isotope ratio mass spectrometry. *Metabolic engineering*, 12(4):392–400.

[Zamboni, 2007] Zamboni, N. (2007). Toward metabolome-based 13C flux analysis: a universal tool for measuring in vivo metabolic activity. *Metabolomics*, 18(April).

[Zamboni et al., 2009] Zamboni, N., Fendt, S.-M., Rühl, M., and Sauer, U. (2009). 13C-based metabolic flux analysis. *Nature protocols*, 4(6):878–892.

[Zamboni et al., 2005] Zamboni, N., Fischer, E., and Sauer, U. (2005). FiatFlux–a software for metabolic flux analysis from 13C-glucose experiments. *BMC bioinformatics*, 6(1):209.

[Zamboni et al., 2004] Zamboni, N., Maaheimo, H., Szyperski, T., Hohmann, H.-P., and Sauer, U. (2004). The phosphoenolpyruvate carboxykinase also catalyzes C 3 carboxylation at the interface of glycolysis and the TCA cycle of Bacillus subtilis. *Metabolic engineering*, 6(4):277–284.

[Zamboni and Sauer, 2004] Zamboni, N. and Sauer, U. (2004). Model-independent fluxome profiling from 2H and 13C experiments for metabolic variant discrimination. *Genome Biol*, 5(12):R99.

[Zamboni and Sauer, 2005] Zamboni, N. and Sauer, U. (2005). Fluxome profiling in microbes. In *Metabolome Analyses: Strategies for Systems Biology*, pages 307–322. Springer.

[Zamorano et al., 2010] Zamorano, F., Wouwer, a. V., and Bastin, G. (2010). A detailed metabolic flux analysis of an underdetermined network of CHO cells. *Journal of biotechnology*, 150(4):497–508.

[Zhang et al., 1995] Zhang, M., Eddy, C., Deanda, K., Finkelstein, M., and Picataggio, S. (1995). Metabolic engineering of a pentose metabolism pathway in ethanologenic Zymomonas mobilis. *Science*, 267(5195):240–243.

[Zhou et al., 2012] Zhou, Y. J., Gao, W., Rong, Q., Jin, G., Chu, H., Liu, W., Yang, W., Zhu, Z., Li, G., Zhu, G., and Others (2012). Modular pathway engineering of diterpenoid synthases and the mevalonic acid pathway for miltiradiene production. *Journal of the American Chemical Society*, 134(6):3234–3241.

[Zomorrodi et al., 2012] Zomorrodi, A. R., Suthers, P. F., Ranganathan, S., and Maranas, C. D. (2012). Mathematical optimization applications in metabolic networks. *Metabolic Engineering*.

[Zupke and Stephanopoulos, 1994] Zupke, C. and Stephanopoulos, G. (1994). Modeling of isotope distributions and intracellular fluxes in metabolic networks using atom mapping matrices. *Biotechnology Progress*, 10(5):489–498.

[Zupke et al., 1997] Zupke, C., Tompkins, R., Yarmush, D., and Yarmush, M. (1997). Numerical isotopomer analysis: estimation of metabolic activity. *Analytical biochemistry*, 247(2):287–293.

CHAPTER 3

METHODS FOR CONSTRAINT-BASED FLUX ANALYSIS

3.1 Introduction

The determination of metabolic fluxes is a fundamental resource to obtain insights regarding which enzymes participate in metabolic networks and also to simulate the cell's behavior under distinct types of genetic and environmental perturbations. Metabolic flux analysis (MFA) is widely used to calculate intra-cellular fluxes revealing the degree of pathway commitment in different metabolic processes during steady-state conditions, operating as a valuable tool in the detection of physiological alterations and to describe cell phenotypes.

Since intra-cellular metabolic fluxes cannot be directly measured, it is useful to have methodologies to estimate and calculate fluxes by applying mass balances around intra-cellular metabolites and from experimentally determined nutrient uptake and product secretion rates that can be readily obtained with standard methodologies. However, an accurate estimation of intracellular fluxes is not straightforward as metabolic networks have different alternative metabolic pathways leading to the production of the same metabolites [Bonarius et al., 1997].

^{13}C-isotopic tracing is a technique that can also be used to measure fluxes, where cells are grown on a ^{13}C-labeled carbon substrate, and the ^{13}C-labeling pattern in their proteinogenic amino acids can be determined through nuclear magnetic resonance or mass spectrometry. Through carbon

labeling experiments and other knowledge on the relative pathway activity of an organism under certain conditions, it is also possible to obtain flux relationships in the form of ratios between sets of reaction rates [Sauer et al., 1999, McAnulty et al., 2012, Fischer et al., 2004].

This chapter describes methods regarding the application of MFA to metabolic models, making use of experimental data on the form of flux values for a set of the reactions in the model or, alternatively, flux ratios between sets of reaction fluxes. It is important to notice that this chapter does not provide any methods for the processing of fluxomics or metabolomics data in any specific format. Rather, it assumes the availability of general purpose data in the forms of flux values or flux ratios, regardless of their origin, and defines methods to work with those data.

The methods described here are implemented in the software tool *CBFA* that is described in the next chapter. This chapter provides the description and mathematical formulation of the methods and some algorithmic details.

3.2 Basics of constraint-based modeling

All methods described here work over a stoichiometric model that includes information on the set of reactions and metabolites occurring in the cell (or in a given system to be modeled), as described in the previous chapter in the section devoted to constraint-based modeling.

3.2.1 Metabolite balance

From a stoichiometric metabolic model, a system of balance equations around the intra-cellular metabolites of the cell is built. The balance is made by setting the input flow to be equal to the output flow plus the ac-

cumulation of matter. Thus, the variation of the concentration in time for a given metabolite is given by the difference between the input flow in the metabolite and the output flow from it:

$$\frac{dX}{dt} = S.v \tag{3.1}$$

where X is a vector that contains the concentration of the m metabolites, v is an n dimensional vector with the metabolic fluxes and S is the m x n stoichiometric matrix of the model.

If the cell is assumed to be in a pseudo-steady state condition, then there is no accumulation of matter and, therefore, the variation of the concentration of the metabolites will be zero along time and, as such [Heinrich and Schuster, 1996]:

$$S.v = 0 \tag{3.2}$$

The previous expression defines a set of m algebraic equations, where the fluxes are the n variables of the system. Usually, n is higher than m, specially when dealing with genome-scale metabolic models. This implies that trying to solve the underlying system usually leads to an under-determined system. An alternative is to define appropriate objective functions and use the constraints to restrict the solution space of the system (Figure 3.1).

3.2.2 Other constraints

To reduce the solution space of the system, thus approaching determined systems, it is possible to enhance the system with additional constraints originating from different sources (Figure 3.2). This information can be used to constrain directly the domain of the variables (fluxes), by defin-

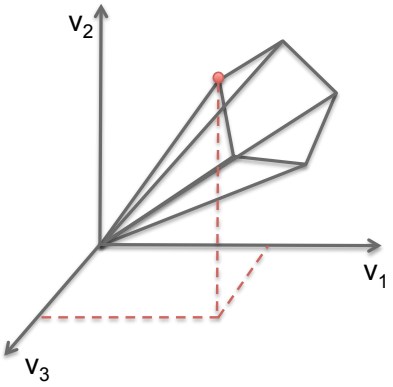


Figure 3.1: Allowable solution space of constraint-based modeling. The identified point represents the optimal solution.

ing bounds determined from reversibility (thermodynamics), experimental data, genetic and/ or environmental conditions. Alternatively, we can provide relationships between the reaction rates by setting up equalities (or inequalities) involving reaction flux values, typically from fluxomics data.

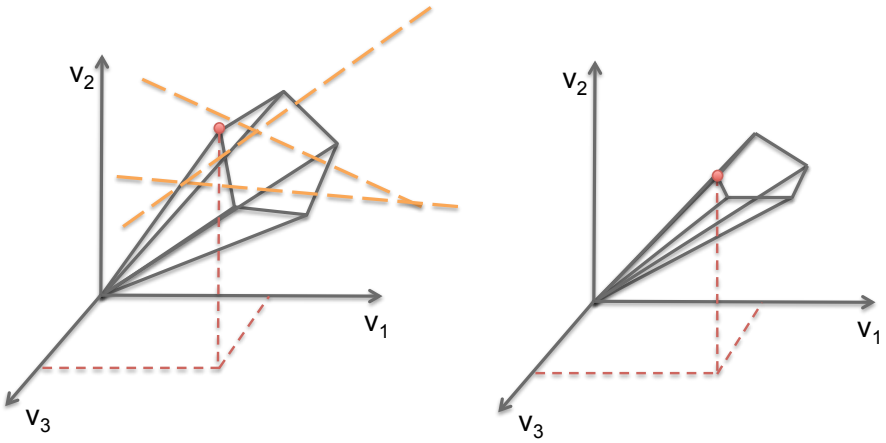


Figure 3.2: Solution space reduced with additional constraints.

Solution space reduced with additional constraints. The allowable solution space is reduced with the constraints derived from the input data and the optimal solution may be shifted in the new flux cone.

From the stoichiometric model and thermodynamics knowledge about reversibility, upper and lower limits can be defined for the fluxes of each re-

action. However, if there is information about the environmental conditions of a simulation (e.g. nutrients in the media), this can be used to reduce the domain for the fluxes (Figure 3.3.B).

More precise bounds can be found from experimental data (e.g. fluxomics) providing measurements for the rates of formation and consumption of some compounds involved in the network, such as extra-cellular substrates (e.g. glucose) and O_2, as well as products (such as CO_2, acetate and lactate). This kind of data can be used to set the fluxes to have a certain value, or to respect a range for specific phenotype simulations (Figure 3.3.C). In such case, if a set of replicates for the measurements are given for a specific flux, these values can be used to calculate the mean and the margin of error of the flux value and, consequently, reduce the domain of the flux as follows:

$$-E_i + \bar{v}_{m,i} \leq v_i \leq \bar{v}_{m,i} + E_i \quad \forall i \in \mathcal{M} \tag{3.3}$$

where $\bar{v}_{m,i}$ is the mean of a set of measurements of the flux i, E_i is the margin of error for the flux i and $\mathcal{M}$ is the set of indexes of the measured fluxes in the model.

When simulating the phenotype of a mutant strain, it is common to define the knockout of a set of genes (or reactions). In this case, the reactions encoded by the genes for which the knockouts were defined (or directly the selected reactions) are constrained to have a flux equal to zero (Figure 3.3.D).

Another type of constraints that can be incorporated in the system is to define a relationship between the fluxes in the model. These relationships are constructed as flux ratio equations, where it is possible to model the

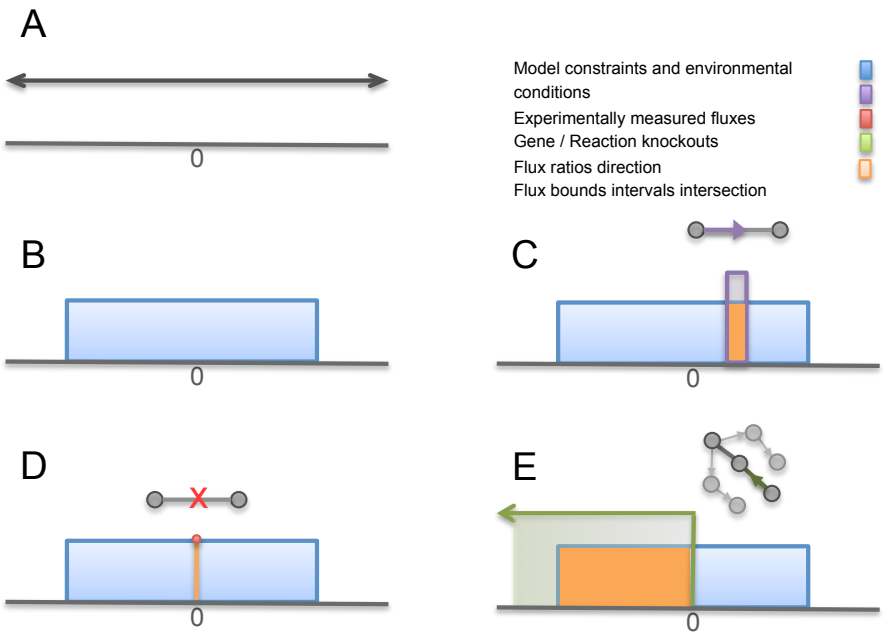


Figure 3.3: Flux bounds obtained from different type constraints. ***A****: the flux has no constraint either on the upper or lower bound.* ***B****: the domain of the flux is defined by the model or environmental conditions.* ***C****: the bounds of the flux are set from experimentally measurements of the flux.* ***D****: the flux is set to have a zero value since it corresponds to a knockout (gene or reaction).* ***E****: the bounds are configured so that the flux respects the direction (relating to the stoichiometric matrix) in which it appears in the flux ratio constraint*

amount of a certain flux (or a set of fluxes) with respect to other fluxes. These metabolic flux ratios constraints are in the form:

$$\frac{\sum_{i=1}^{n} \kappa_i v_i}{\sum_{j=1}^{n} \kappa_j v_j} = \tau \tag{3.4}$$

where κ_i, $\kappa_j \in \mathbb{R}$ are constants, v_i and v_j are fluxes, and $\tau \in \mathbb{R}\backslash\{0\}$ is the ratio between the numerator and denominator fluxes. In this case, the operator that relates the fluxes is the equality operator, but it is possible to define inequalities, where the ratio represents a upper or lower bound.

To add this type of constraints to the system, the ratio equations are

converted to its linear form:

$$\sum_{i=1}^{n} \kappa_i v_i - \tau \sum_{j=1}^{n} \kappa_j v_j = 0 \tag{3.5}$$

Some problems arise from this kind of equations. For instance, in a simple case, where a ratio between two fluxes (e.g. $v_1/v_2 = 0.5$) is set, it is expressed that v_1 is constrained to have a value that is half the value of v_2. However, if v_1 is the flux of a reversible reaction, and this reaction is occurring in the negative direction with respect to the stoichiometric matrix, mathematically this ratio will force the flux v_2 to be also negative. Because the ratio is a positive number, both the numerator and denominator of the equation must have the same signal. So, this kind of expressions has an influence in the signal of the fluxes of the solution space. Moreover, if we are dealing with inequality constraints, in the conversion of the expressions to its linear form, if the signal of the flux is not known, it is not possible to determine if the inequality operator will need to be changed or not.

Therefore, either the expression is changed to take into account the absolute value of the fluxes or the direction of the flux has to be set in the expression. If we assume that the direction of the flux is set, this kind of expression will add two types of constraints to the system. Firstly, the expression that constrains the fluxes absolute values is added to the system, and afterwards the bounds of the fluxes are updated to respect the direction of the flux in the expression (Figure 3.3.D). These bounds are configured in the form:

$$v_i \geq 0 \quad \forall i \in \mathcal{R}_{pos} \tag{3.6}$$

where the flux i is in the positive direction and $\mathcal{R}_{pos}$ is the set of indexes of the fluxes in the model that are set to be positive in the metabolic flux ratio constraints.

$$v_i \leq 0 \quad \forall i \in \mathcal{R}_{neg} \tag{3.7}$$

where the flux i is in the negative direction and $\mathcal{R}_{neg}$ is the set of indexes of the fluxes in the model that are set to be negative in the metabolic flux ratio constraints.

If there is a non empty intersection between the two previous sets, it is necessary to define how to deal with this case in terms of flux bounds. A priority list is defined such that its order defines which constraint will be assumed, if it is not possible to take into account both constraints, set as follows:

1. Gene/ reaction knockouts;
2. Measured fluxes;
3. Ratio fluxes direction;
4. Environmental conditions;
5. Model constraints.

If an overlap of a constraint into another occurs, and it is possible to consider both of them, the domain of the flux is updated with this consideration (orange region in Figure 3.3.E where the lower bound is defined by the stoichiometry, thermodynamics or environmental conditions and the

upper bound is set to zero meaning that the flux is negative in the ratio expression it belongs to).

3.3 System determination

Since the number of reactions is typically greater than the number of metabolites in the model, there is a certain degree of freedom in the set of algebraic equations related to the metabolite balance. Usually, these models are, therefore, under-determined even with the additional constraints obtained as it was introduced in the previous chapter. However, in some cases, this might not be the case, i.e. there may be enough constraints to allow solving the system or even to over-constrain it. In order to determine the type of system, it is necessary to take into account the several inputs and related constraints. Firstly, from the stoichiometric model, the number of reactions and the number of mass balance equations (balanced metabolites) are retrieved (Figure 3.4).

Since the balance equations have the reaction fluxes as its unknown variables, depending on the number of the equations in the system, it might not be possible to algebraically solve the system if the number of linearly independent equations is less than the number of variables. Indeed, this difference between the number of reactions and the number of mass balance equations defines the number of degrees of freedom in the model:

$$\#\begin{matrix}\text{model degrees}\\ \text{of freedom}\end{matrix} = \#\,reactions - \#\begin{matrix}\text{linearly independent}\\ \text{balance equations}\end{matrix} \tag{3.8}$$

However, it is possible to reduce this number through the consideration

of the other constraints mentioned in the previous section. If experimentally measured fluxes are given with exact values, i.e., there is no margin of error from the measurements replicates, each flux can be set to the measured value, thus possibly reducing the number of degrees of freedom, as the number of unknowns decreases.

If a margin of error exists for a flux, the measurements are still used but they do not set a fixed value for the flux, but instead a range of possible values (Eq.3.3). In this case, the flux still has to be calculated, and the number of degrees of freedom is not reduced (although the new constraints will limit the feasible space).

It is also possible to decrease the number of degrees of freedom by adding gene/ reaction knockouts. In these cases, the flux values for the affected reactions are set to zero, consequently decreasing the number of unknown variables of the system by a value equal to the number of affected reactions.

When a set of metabolic flux ratios is given, the equality ratios define new linear constraints to add to the ones coming from the model and, therefore, they are likely to reduce the degrees of freedom of the model (by a number equal to the number of linearly independent expressions they add to the system).

As it happens with the experimentally measured fluxes, the inequality equations are not used to reduce the degrees of freedom, though they are used to constrain the solution space of the system.

As it will be explained on a posterior section of this document (section 3.4.2), to solve the system algebraically, the stoichiometric matrix of the model is extended with the flux ratios and partitioned, obtaining a smaller matrix, S_c corresponding to the unknown fluxes, i.e. fluxes that are neither measured nor deleted through the application of knockouts in the case of

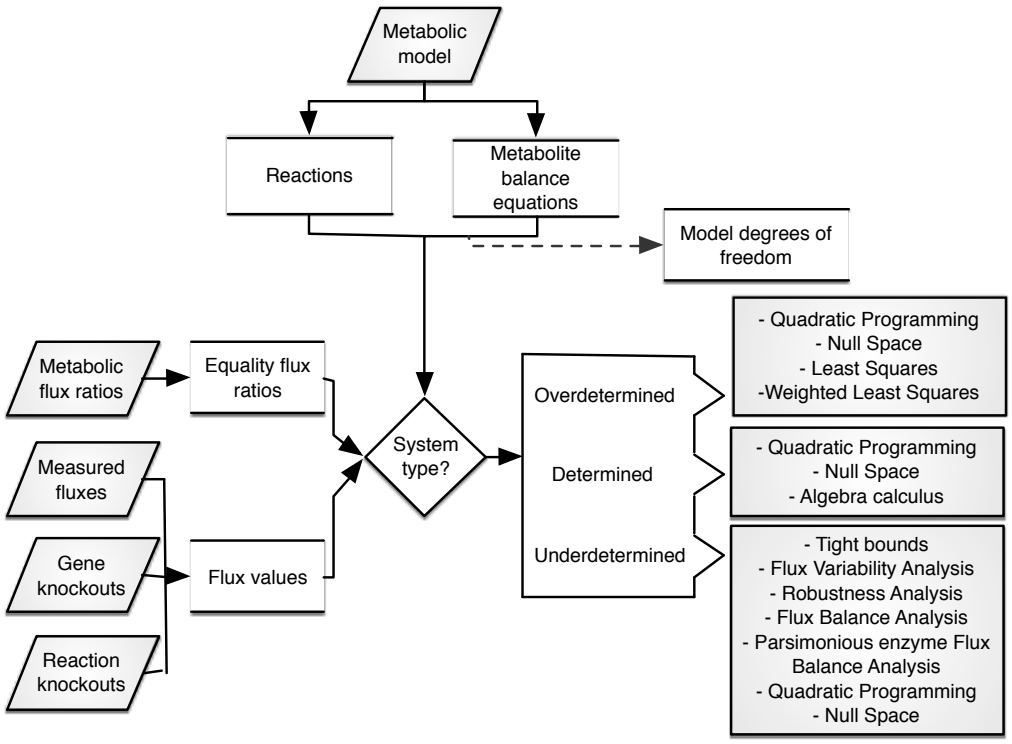


Figure 3.4: Model degrees of freedom. The type of the system is inferred from the degrees of freedom of the model and the constraints given to the system. The parallelogram shapes represent the inputs of the system and the light grey rectangles are the methods that can be applied to each type of system.

mutant strains. For now, it is important to mention that this matrix is the one that is used to calculate the real degrees of freedom of the configured system from the input data. Therefore, the degrees of freedom of the system can be obtained as:

$$\#\begin{matrix}\text{system degrees}\\ \text{of freedom}\end{matrix} = \#\,unknown\ reactions - r_{S_c} \tag{3.9}$$

where r_{S_c} stands for the rank of the matrix S_c.

Although the existence of flux measurements and knockouts may reduce the number of unknowns of the system, because they depend on the r_{S_c}

matrix, there is not a proportional decrease, i.e. the decrease in the degrees of freedom may be less than the number of equality constraints introduced. This happens since the elimination of the columns (corresponding either to measured fluxes or to deleted reactions) from the stoichiometric matrix may result on the appearance of new linearly dependent rows, thus limiting the reduction in the overall degrees of freedom.

When considering all possible constraints for a given simulation, the number of degrees of freedom is calculated taking into account the original model, the environmental constraints, the knockouts defined (if any) and all constraints coming from measured data. If the overall number of degrees of freedom is greater than zero, the system is said to be underdetermined. This is the most common case, typically obtained when dealing with large-scale models (e.g. genome-scale). On the other hand, in the particular case where the number of degrees of freedom equals zero, the information in the system is enough to calculate all the unknowns, and the system is called determined. In the more general case in which the additional constraints reduce the number of degrees of freedom by a value larger than the original one (i.e. the overall calculation of the degrees of freedom would be negative), the system is said to over-determined and the additional data can be used to better fit the unknowns. The overall process of system determination is provided in Figure 3.4.

3.4 Methods and formulations

Depending on the type of the designed system, different approaches can be used to determine the unknown fluxes. When the system is under-determined, optimization methods to search for an optimum solution in

the constrained solution space are used. In the cases where the system is determined, algebraic methods can be directly used to solve the system and calculate the unknowns (flux values). If the system is over-determined, algebraic methods are used to fit the variables of the system to the data. Although there is this separation between the methods that are available depending on the type of the system, the null space method and quadratic programming optimization are formulated to be applied for all types of systems.

From the different inputs of the system, it is possible to organize the data in classes of data types (Figure 3.5). The stoichiometric matrix is built from the metabolic model where every row of the matrix is related to the mass balance of an intra-cellular metabolite. The flux bounds include the set of lower and upper limits of reaction rates given by the model, the flux limits related to environmental conditions, and also the intervals defined by the experimentally measured fluxes according to the mean and margin of error of the given replicates. The set of flux values comprehends the fluxes that must have zero values originating from the knockouts and also the fixed measured fluxes, i.e., when the margin of error is not considered. Finally, the list of metabolic flux ratios is split into a set for the equality equations and another set for the inequality equations. Figure 3.5 provides a schematic representation of the data types and available methods.

3.4.1 Optimization methods

This section describes the formulation of the optimization methods, which estimate the flux values through the definition of an objective function and using all the inputs of the system to constrain the space of admissible solutions.

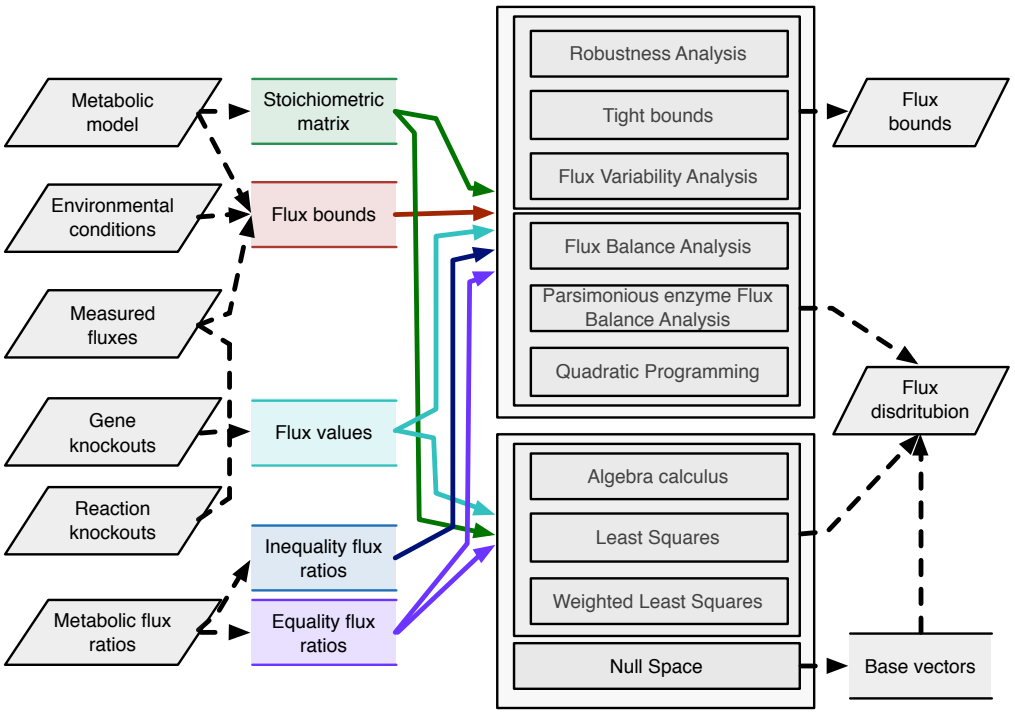


Figure 3.5: Methods inputs and outputs. The inputs of the system are partitioned into different classes of data types that are used by the MFA methods. The parallelogram shapes on the left and on the right represent the inputs and outputs, respectively, of the system. The colored rectangles denote the classes of data types.

3.4.1.1 Flux balance analysis

Flux balance analysis (FBA) [Papoutsakis, 1984] is a mathematical modeling approach widely used to quantitatively simulate the values of metabolic fluxes of a certain organism under a set of physiochemical conditions. It is formulated as a linear programming (LP) problem that maximizes (or minimizes) a configured objective function, such as the biomass flux (an artificial flux representing the growth rate). This approach searches for the maximum (or minimum) value of the objective function that respects the constraints that are configured from the input data:

$$
\begin{aligned}
\text{maximize/minimize} \quad & v_{objective} \\
\text{subject to:} \quad & S.v = 0 \\
& \textstyle\sum_{i=1}^{n} \kappa_i v_i - \tau \sum_{j=1}^{n} \kappa_j v_j = 0 \\
& Lb_i \leq v_i \leq Ub_i \qquad \forall i \notin \mathcal{M} \cup \mathcal{K} \\
& -E_i + \bar{v}_{m,i} \leq v_i \leq \bar{v}_{m,i} + E_i \qquad \forall i \in \mathcal{M} \\
& v_i = 0 \qquad \forall i \in \mathcal{K} \\
& v_i \geq 0 \qquad \forall i \in \mathcal{R}_{pos} \\
& v_i \leq 0 \qquad \forall i \in \mathcal{R}_{neg}
\end{aligned}
$$

where S is the m x n stoichiometric matrix of the model, v is an n dimensional vector with the metabolic fluxes where v_i is the i^{th} flux of the model, Lb_i and Ub_i are the upper and lower bounds, respectively, of the fluxes, $\bar{v}_{m,i}$ is the mean of a set of measurements of the flux i (or a fixed value if there is only one measurement of the flux), E_i is the margin of error for the flux i (or zero if there is only one measurement of the flux), $\mathcal{M}$ is the set of indexes of the measured fluxes in the model, $\mathcal{K}$ is the set of indexes of the fluxes with knockouts and $\mathcal{R}_{pos}$ and $\mathcal{R}_{neg}$ are the set of indexes of the fluxes that are set be positive and negative, respectively, in the metabolic flux ratio constraints.

The lower and upper bounds of the variables are obtained as previously described and added to the problem, where the equality ones are configured to have the same value for the lower and upper bounds. With respect to the equations that translate the metabolic flux ratios, here both the equality and inequality constraints are used, by adding the linear constraints to the problem.

3.4.1.2 Parsimonious enzyme usage flux balance analysis

In the parsimonious enzyme usage flux balance analysis (pFBA), there is the assumption that, under growth pressure, the strains that process the growth rate most rapidly and efficiently, with the minimum enzyme usage, are selected. It is possible to approximate this theory through the optimization of the growth rate using FBA, and afterwards minimizing the metabolic fluxes related to all reactions in the model [Lewis et al., 2010]. This assumption has been the motivation for the formulation of an MFA method that searches for the minimal sum of all fluxes in the metabolic network that contributes for an outlined objective function, such as the maximization of the biomass, or another target compound.

This is accomplished in two sequential steps: first an optimization of the objective function is performed through the application of the FBA-based MFA method, outlined in the previous section, and in the second step the obtained objective value is fixed (as an additional constraint) and a new linear programming problem is configured to minimize the sum of all fluxes, keeping all other constraints unchanged. The overall formulation is therefore the following:

$$
\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{n} |v_i| \\
\text{subject to:} \quad & v_{objective} = \alpha \max/\min v_{objective} && \alpha \in \mathbb{R}^+ \\
& S.v = 0 \\
& \textstyle\sum_{i=1}^{n} \kappa_i v_i - \tau \sum_{j=1}^{n} \kappa_j v_j = 0 \\
& Lb_i \leq v_i \leq Ub_i && \forall i \notin \mathcal{M} \cup \mathcal{K} \\
& -E_i + \bar{v}_{m,i} \leq v_i \leq \bar{v}_{m,i} + E_i && \forall i \in \mathcal{M} \\
& v_i = 0 && \forall i \in \mathcal{K} \\
& v_i \geq 0 && \forall i \in \mathcal{R}_{pos} \\
& v_i \leq 0 && \forall i \in \mathcal{R}_{neg}
\end{aligned}
$$

where $v_{objective}$ is the optimized objective flux, α is a relaxing coefficient to be applied to the optimized objective and all other symbols take the meaning explained in the previous section.

Because the formulated objective function is non-linear, due to the absolute values of the fluxes, it has to be rearranged in order to perform linear programming to solve the problem. Therefore, each flux variable is decomposed in two different variables, and the following set of constraints are added to the problem:

$$
\begin{aligned}
& v_i + v_i^- - v_i^+ \\
& v_i^+ \geq 0 \\
& v_i^- \geq 0
\end{aligned}
\tag{3.10}
$$

where v_i^+ and v_i^- correspond to the positive and negative direction of the flux v_i, respectively. To eliminate the absolute values of the objective function, each of it is replaced as follows

$$|v_i| = v_i^+ + v_i^- \tag{3.11}$$

Since the formulation is a minimization problem, the simplex algorithm will ensure that at least one of the directions of each flux will be equal to zero, and the values of the two directions of the flux v_i are obtained as:

$$\begin{cases} v_i > 0, v_i^+ = v_i \ \ and \ \ v_i^- = 0 \\ v_i < 0, v_i^+ = 0 \ \ and \ \ v_i^- = v_i \\ v_i = 0, v_i^+ = 0 \ \ and \ \ v_i^- = 0 \end{cases} \tag{3.12}$$

3.4.1.3 Quadratic programming

When a set of flux measurements exists, it is possible to formulate a problem that estimates the network fluxes in order to minimize the difference between the measured values and the calculated ones, keeping the other constraints of the system active.

This approach is similar to the variant of FBA called minimization of metabolic adjustment (MOMA) [Segre et al., 2002], which has the assumption that mutant strains show flux distributions that are as close as possible to the wild-type optimum. Here, instead of the approximation to a reference flux distribution, the fluxes that have been measured are selected to be approximated.

Thus, the objective function of this optimization problem can be defined as the minimization of the sum of the squares of the differences between

the measured and the calculated fluxes, as given next:

$$\text{minimize} \sum_{i=1}^{n} (v_{c,i} - \bar{v}_{m,i})^2 \tag{3.13}$$

where n is the number of measured fluxes, $v_{c,i}$ is the i^{th} calculated flux and $\bar{v}_{m,i}$ is the mean of a set of measurements of the flux i or a fixed value if there is only one measurement of the flux.

Since all constraints are linear and the objective function is quadratic, this configures a problem of quadratic programming (QP) optimization, where the objective function can be converted to the form:

$$\text{minimize } \frac{1}{2} v_c^T H v_c - 2\bar{v}_m^T v_c + \sum_{i=1}^{n} \bar{v}_{m,i}^2 \tag{3.14}$$

where v_c is an n-dimensional row vector with the calculated fluxes, $\bar{v}_m$ is an n-dimensional row vector with the mean of the measured fluxes (or fixed values if there is only one measurement of the fluxes), H is an $(n \times n)$ symmetric matrix, being n is the number of measured fluxes. The scalar multiplication $-2\bar{v}_m$ represents the coefficients of the linear terms in the objective function and H describes the coefficients of the quadratic terms. In this optimization problem, the H matrix is the identity matrix multiplied by 2:

$$H = 2.I \tag{3.15}$$

where I is an $(n \times n)$ identity matrix. Therefore, the objective function can be simplified by removing the scalar $1/2$ and the H matrix from the

expression, and the equation 3.14 can be rewritten as:

$$\text{minimize } v_c^T v_c - 2\bar{v}_m^T v_c + \sum_{i=1}^{n} \bar{v}_{m,i}^2 \tag{3.16}$$

Regarding the constraints, its formulation is similar to the previous methods, with a slight difference since the measured fluxes are not added to the constraints rows (they are used instead in the objective function in this case). The overall formulation is the following:

$$\begin{aligned}
\text{minimize} \quad & v_c^T v_c - 2v_m^T v + \sum_{i=1}^{n} \bar{v}_{m,i}^2 \\
\text{subject to:} \quad & S.v = 0 \\
& \textstyle\sum_{i=1}^{n} \kappa_i v_i - \tau \sum_{j=1}^{n} \kappa_j v_j = 0 \\
& Lb_i \leq v_i \leq Ub_i \qquad \forall i \notin \mathcal{M} \cup \mathcal{K} \\
& v_i = 0 \qquad \forall i \in \mathcal{K} \\
& v_i \geq 0 \qquad \forall i \in \mathcal{R}_{pos} \\
& v_i \leq 0 \qquad \forall i \in \mathcal{R}_{neg}
\end{aligned}$$

where all the relevant symbols have been presented above. In this formulation there is a constant term in the objective function related to the sum of the squares of the flux measurements, that can be dropped from the formulation when the problem is set up to be solved.

3.4.1.4 Tight bounds calculation

The tight bounds method uses a different approach from the previous ones, since it does not return a flux distribution that better suits an objective function, but instead it computes the lower and upper limits for every non

free flux of the system and thus characterizes the boundaries of the system under certain conditions.

The FBA-based MFA method allows configuring optimization problems optimizing specific fluxes. This feature is used here, for every flux of a reaction in the model that has neither been experimentally measured nor has an associated knockout, to calculate both its minimum and maximum values under the same constraints. The result of each minimization and maximization optimization process is taken as the lower and upper bound, respectively, of this flux.

The formulation for all minimization and maximization problems uses the same constraints as defined in the FBA and pFBA-based method, given by:

$$
\begin{aligned}
t_{Lb,i} = \text{minimize} \quad & v_i && \forall i \notin \mathcal{M} \cup \mathcal{K} \\
t_{Ub,i} = \text{maximize} \quad & v_i && \forall i \notin \mathcal{M} \cup \mathcal{K} \\
\text{subject to:} \quad & S.v = 0 && \\
& \textstyle\sum_{i=1}^{n} \kappa_i v_i - \tau \sum_{j=1}^{n} \kappa_j v_j = 0 && \\
& Lb_i \leq v_i \leq Ub_i && \forall i \notin \mathcal{M} \cup \mathcal{K} \\
& -E_i + \bar{v}_{m,i} \leq v_i \leq \bar{v}_{m,i} + E_i && \forall i \in \mathcal{M} \\
& v_i = 0 && \forall i \in \mathcal{K} \\
& v_i \geq 0 && \forall i \in \mathcal{R}_{pos} \\
& v_i \leq 0 && \forall i \in \mathcal{R}_{neg}
\end{aligned}
$$

where $t_{Lb,i}$ and $t_{Ub,i}$ are the tight bounds of the flux i and the other symbols take the meaning defined in the previous cases.

Flux variability analysis

The approaches based on LP problems can have multiple solutions that satisfy all the constraints and, nevertheless, have the same optimal value for the objective function. To achieve the characterization of alternative solutions that satisfy all the constraints and allow for the same optimal objective value, flux variability analysis (FVA) can be applied to determine the admissible ranges of flux values under optimality [Mahadevan et al., 2003].

This method is a slight variation of the tight bounds method. First, it is performed a simulation related to the objective function through the application of the FBA method. From this simulation, the solution with the flux distribution is kept and a flux is selected to be fixed in the calculation of the tight bounds. It is also possible to configure the value of the fixed flux to be a percentage of its value (typically varying from 0 to 100%) in the wild-type strain. The only difference in respect to the tight bounds formulation, is the addition of the following constraint:

$$v_f = \rho \max/\min v_f \quad , \quad \rho \in \mathbb{R}^+ \tag{3.17}$$

where f is index of the fixed flux in the objective solution and ρ is the percentage of the fixed flux value.

3.4.1.5 Robustness Analysis

Robustness analysis is a method that enables to investigate how an objective function changes in response to variations in the optimal fluxes of specific reactions of the metabolic model [Edwards and Palsson, 2000, Palsson,] .

A first simulation is performed to determine the optimal flux distribution for a defined optimization problem (e.g. maximization of the growth rate).

The result of this simulation is kept as the control distribution. For each reaction selected to perform the robustness analysis, its flux is varied from 0 to 100% of its value in the control distribution, and the optimization problem is solved with this additional constraint. The result is a set, for each selected flux, containing the resulting objective values for the control flux variation.

The formulation to maximize/minimize an objective function for each control flux variation is given as follows:

$$
\begin{aligned}
\text{maximize/minimize} \quad & v_{objective} & \\
& v_i = \rho\, v_{control,i} & 0 \leq \rho \leq 100, i \in \mathcal{C} \\
& S.v = 0 & \\
& \textstyle\sum_{i=1}^{n} \kappa_i v_i - \tau \sum_{j=1}^{n} \kappa_j v_j = 0 & \\
& Lb_i \leq v_i \leq Ub_i & \forall i \notin \mathcal{M} \cup \mathcal{K} \\
& -E_i + \bar{v}_{m,i} \leq v_i \leq \bar{v}_{m,i} + E_i & \forall i \in \mathcal{M} \\
& v_i = 0 & \forall i \in \mathcal{K} \\
& v_i \geq 0 & \forall i \in \mathcal{R}_{pos} \\
& v_i \leq 0 & \forall i \in \mathcal{R}_{neg}
\end{aligned}
$$

where $\mathcal{C}$ is the set of indexes of the fluxes that are selected for the robustness analysis, $v_{control,i}$ is the control flux for which the problem is formulated and ρ is the percentage to vary the control flux.

3.4.2 Algebraic methods

In the methods described in this section, the metabolic flux ratio equations are added to the ones coming from the stoichiometric matrix and mass balance equations. However, bound constraints are not incorporated in the rows of the stoichiometric matrix, and therefore only equality equations are considered in the following methods. Therefore, flux ratios given by inequalities are ignored in these methods.

Equation 3.5 is used to convert the ratios to its linear form and these are added as new rows of the stoichimetric matrix, where the coefficients of the fluxes in the ratio equations, and its product to the negative of the ratio in the equation, are added to the corresponding reaction column (Fig 3.6). The balance equation in Eq. (3.2) is redefined accordingly:

$$S_\tau.v = 0 \tag{3.18}$$

where S_τ is an $(m + r)$ x n extended stoichiometric matrix with the last r rows corresponding to the ratio equations linear form, and v is an n dimensional vector with the metabolic reaction rates.

3.4.2.1 Algebraic calculations

If the extended matrix has no linearly dependent rows ant the number of flux measurements equals to the difference between the number of reactions and the number of balance equations, the solution of the system is unique and it can be obtained by solving the linear system of equation 3.18. This is accomplished through the separation of the measured fluxes from the non-measured ones in the vector v. For this purpose, the coefficients in the extended stoichiometric matrix S_τ have to be partitioned as well. Therefore,

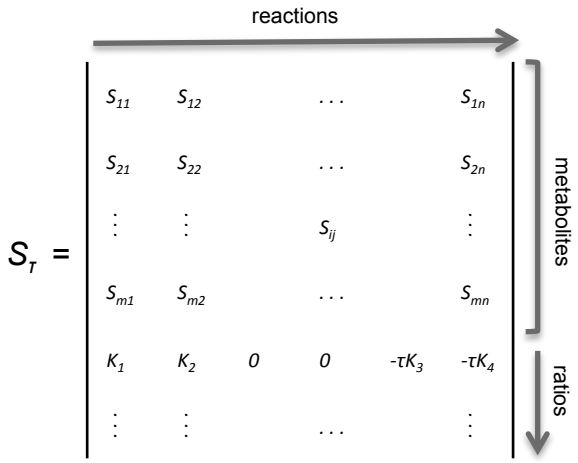


Figure 3.6: Extended stoichiometric matrix. The stoichiometric matrix is extended with the linear ratio equations. The columns represent the reactions of the model, the first rows correspond to metabolite balance equations of the system and the last rows represents the coefficients of the equality metabolic flux ratios.

equation 3.18 can be reformulated as:

$$S_m v_m + S_c v_c = 0 \tag{3.19}$$

where S_m and S_c represents the partitioning of the matrix S_τ with the columns corresponding to the stoichiometric coefficients of the fluxes that have been measured and the columns of the non-measured fluxes, respectively, and v_m and v_c are the vectors of the measured and unknown fluxes, respectively.

If S_c is an n-by-n square matrix, n equals to the number of metabolites plus the number of equality ratio constraints, and also equals to the number of unknown fluxes. In such case, if S_c is not a singular matrix, the matrix can be inverted and the previous equation can be rewritten such that the vector of the unknown reactions rate can be obtained from the stoichiomet-

ric matrix rearrangement and the experimentally measured fluxes:

$$v_c = -{S_c}^{-1} S_m v_m \tag{3.20}$$

However, if S_c is an m-by-n non-square matrix, it cannot be inverted and the system can not be solved from the above equation. Therefore, the solution of the equation 3.20 is rather found by the application of the Moore-Penrose pseudoinverse of S_c [Penrose, 1955]:

$$v_c = -{S_c}^{+} S_m v_m \tag{3.21}$$

where ${S_c}^{+}$ is the Moore-Penrose pseudoinverse of S_c.

Because the number of equations in S_c (m) will be less than the number of the unknown fluxes (n), the Moore-Penrose pseudoinverse of S_c is obtained through the calculation of the right inverse of S_c. Thus, equation (3.21) can be rewritten in the form:

$$v_c = -\left(S_c ({S_c}^{T} S_c)^{-1} \right) S_m v_m \tag{3.22}$$

3.4.2.2 Least squares

If the system is overdetermined, the additional information can be used to test the consistency of the overall balances and the accuracy of the flux measurements [Stephanopoulos et al., 1998]. In this scenario, equation 3.21 can also be used. However, since in this case the number of rows in the S_c matrix is greater than the number of its columns, the Moore-Penrose pseudoinverse of S_c is obtained through the calculation of the left inverse

of S_c, and equation 3.21 can be rephrased as:

$$v_c = -\left(\left(S_c {S_c}^T\right)^{-1} S_c\right) S_m v_m \tag{3.23}$$

This equation represents the least square estimation of the unknown reactions fluxes.

3.4.2.3 Weighted least squares

Since experimentally measured data usually have associated errors, the previous method is not as accurate as it could be, because it does not assume that there is a variance for the residuals of the measured fluxes. If information about the variance on the flux measurements exists, this variance can be used to weight the least squares fitting. Consequently, the best weighted estimation of the measured fluxes on an over-determined system can be obtained by:

$$\hat{v_m} = \left(I - E(\sigma\sigma^T) {R_r}^T \Psi^{-1} R_r\right) v_m \tag{3.24}$$

where I is the identity matrix, Ψ is the m x m variance-covariance matrix for the residuals of the measured fluxes and R_r is a reduced matrix of the redundancy matrix R. [Wang and Stephanopoulos, 1983] show that the estimates for the measured reaction rates results in a smaller standard deviation, compared to the the direct use of the measurements, making the estimates likely to be more reliable to calculate the non measured rates. The variance-covariance matrix can be obtained as:

$$\Psi = R_r E(\sigma\sigma^T) R_r^T \tag{3.25}$$

where $E(\sigma\sigma^T)$ is an m x m diagonal matrix, called measured flux vari-

ance matrix, such that the element in the i-th row and in the i-th column correspond to the variance of the flux i [Madron et al., 1977]:

$$E(\sigma\sigma^T) = \begin{bmatrix} \sigma_0^2 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_m^2 \end{bmatrix} \tag{3.26}$$

Moreover, R is the redundancy matrix that conveys the relations between all the measured reaction fluxes. A zero column in R indicates that the other measured rates do not have significant expression for the corresponding measured flux, i.e., it is not balanceable. Thus, the columns of matrix R can be inspected to identify the columns of the matrix R that can be used to distinguish between the balanceable reaction rates from the non-balanceable ones. Hereupon, the redundancy matrix can be expressed as [der Heijden et al., 1994]:

$$R = S_m - S_c S_c^{+} S_m \tag{3.27}$$

However, the redundancy matrix contains redundant equations related to the measured reaction rates. If the linearly dependent rows of R are removed, a reduced redundancy matrix (R_r) is obtained with rank(R) independent equations, that must be satisfied to calculate the nom-measured rates from the measured rates.

Finally, the non-measured reaction rates can be estimated through the substitution of $\hat{v_m}$ from equation 3.24 into equation 3.23:

$$v_c = -((S_c {S_c}^T)^{-1} S_c) S_m (S_c \Psi^{-1} {S_c}^T)^{-1} S_c \Psi^{-1} v_m \tag{3.28}$$

3.4.2.4 Null space

This method handles all the types of systems that can be configured from the stoichiometric metabolic model and the input data. It is based on the nullspace of the extended stoichiometry matrix (S_τ) that corresponds to the coefficients of the reactions on the metabolite balance equations of the metabolites and the coefficients of the reactions on the equality metabolic flux ratios.

As it has been shown on Eq 3.2 the metabolite balance is a system of linear equations that forces the rates of consumption and production of a metabolite to be equal. The most trivial solution to this system is the zero vector which would represent the thermodynamic equilibrium. However, this is not the interesting solution when analysing flux distributions. Since the cell must have degrees of freedom due to the much higher number of reactions in its biochemical networks representation relating to the nodes (metabolites), several different flux distributions usually fulfil the system of Eq 3.2.

Algebraically, all possible solutions of the system are encompassed in the null space of S_τ [Strang, 1980]. Eq. 3.2 has $n - rank(S)$ linearly independent solutions (which define the dimension of the null space) that can be organized in a kernel matrix (K) that encompasses the base vectors that span the null space of S_τ. Therefore, the flux distributions into the null space of S_τ can be configured by linear combinations (b) of the columns in (K) as shown in:

$$v = K\beta \tag{3.29}$$

where K is an $n \times r$ matrix, where n defines the number of fluxes in the model

and r is equal to the difference between n and the rank of the stoichiometric matrix. Also, β defines a column vector with r positions, each corresponding to the free variables of the system.

To find the null space of S_τ, the reduced echelon form of S_τ ($ref(S_\tau)$) is obtained. The solution for $ref(S_\tau)\ v = 0$ is the same as for the systems $S_\tau v = 0$ and $Uv = 0$ (U is the upper triangular factor obtained from LU decomposition). Here, the free variables are defined as the variables that do not correspond to columns in $ref(S_\tau)$ with pivot entries (the first non zero entries in their rows). To assign general solutions to $ref(S_\tau)\ v = 0$ (and consequently, to the Eq. 3.2), one may assign arbitrary values to the free variables. Accordingly, there are r-infinity solutions where the complete solution set is a combination of the special solutions corresponding to the free and independent variables:

$$v = v_{f,1} \begin{bmatrix} c_{1,0} \\ c_{1,1} \\ \vdots \\ c_{1,n} \end{bmatrix} + v_{f,2} \begin{bmatrix} c_{2,0} \\ c_{2,1} \\ \vdots \\ c_{2,n} \end{bmatrix} + \cdots + v_{f,k} \begin{bmatrix} c_{k,0} \\ c_{k,1} \\ \vdots \\ c_{k,n} \end{bmatrix} \tag{3.30}$$

where v is an $n \times 1$ column vector corresponding to the reaction fluxes, $v_{f,i}$ is the free variable i (the ith position on the vector β in Eq. 3.29) and each vector c_i multiplied by the free variables corresponds to the i special solution, i.e., the ith base vector of the null space of S_τ.

Moreover, usually the kernel matrix is not unique and not every qualitative different flux distribution might be represented by the base vectors in K [Szallasi et al., 2006]. Nevertheless, the kernel matrix can be used to detect strictly detailed balanced reactions, if the corresponding rows have

only zero values [Heinrich and Schuster, 1996]. Also, the occurrence of coupled/correlated reaction sets, that represent a set of reactions that operate with a fixed ratio in their rates [Pfeiffer et al., 2001], such as reactions in the same linear pathway, can be identified from the null-space matrix, since the rows of the reactions will only differ by a scalar. Moreover, in the presence of a system whose reactions are completely disconnected or its fluxes are not related to the other fluxes in the network, sub-networks can be identified if the kernel is block-diagnosable [Heinrich and Schuster, 1996, Szallasi et al., 2006].

Thus, although a single point in the null space of S_τ is commonly not possible to be obtained to translate a solution that reflects the flux distribution of a given experiment, due to the more likely occurrence of undetermined systems, the null space of the stoichiometric matrix can be used to characterize the configured systems.

After obtaining the null space of the stoichiometric matrix (K), the number of unknowns of the system is said to be equal to the number of base vectors in K (number of free variables in Eq. 3.29). A new matrix (K_m^T) is obtained where its columns refer to the lines in K that correspond to measured fluxes. Therefore, the number of rows and columns in K_m^T will be equal to the number of unknowns and measurements, respectively. This matrix will be decomposed with LU decomposition and its upper triangular factor is retrieved and stored in a matrix U and its lower triangular factor is stored in a matrix L. Since the decomposition might need to perform some row and column permutations, these are stored to map the rows and columns to the original indexes.

The rank of K_m^T (r_{K_m}) is always less or equal to the number of measurements and less or equal to the number of unknowns. Some conclusions can

be made if r_{K_m} is not equal to the number of unknowns:

- the matrix K_m^T has a null space;
- K_m^T is rank deficient, and its rank deficiency is equal to the difference between the number of unknowns and its rank;
- there exists a number of linearly independent solutions equal to the rank deficiency plus one.

Principal flux distribution

To deal with the rank deficient case, a new smaller system that contains a unique solution is defined. The r_{K_m}-first rows of the upper factor of K_m^T are copied to a new matrix U_s, and an invertible leading matrix ($U_{leading}$) is defined by the r_{K_m}-first linearly independent rows of the transposed matrix of U_s, and the vector β of the free fluxes in Eq. 3.29 is obtained after solving the following systems:

$$U_{leading}\gamma_s = v_{ms} \tag{3.31}$$

$$L^T y = \gamma \tag{3.32}$$

where v_{ms} is a vector with the r_{K_m}-first measured fluxes, with the original indexes permuted to the K_m^T after its LU decomposition. The vector γ_s in Eq 3.31 is obtained through forward substitution, solving the lower triangular system, and y is obtained from back substitution by solving the upper triangular system. The number of elements of γ is equal to the number of unknowns and its r_{K_m}-first positions are obtained from the vector v_{ms}. The vector β is finally obtained by indexing the elements in y with respect to the permutations performed during the LU decomposition of K_m^T. Note

that the indexes of the rows in K_m^T matches the row indexing in K and therefore, Eq. 3.29 can be used to compute the values of the fluxes.

However, before computing the vector y, a verification is done relating to the vector γ_s. From this vector, the measured fluxes are computed from the above equation bellow, and afterwards compared with the actual measured fluxes (v_m).

$$v_{m,computed} = U_s^T.\gamma_s \tag{3.33}$$

If the sum of the module of the difference between the elements in $v_{m,computed}$ and the elements in v_m is greater than a given value ϵ, the vector γ_s is recomputed by least squares approximation:

$$U_s U_s^T \gamma_s = U_s v_m \tag{3.34}$$

Alternative flux distributions

If K_m^T is rank deficient, alternative solutions are computed. Therefore, the number of unknowns and the rank of the matrix K_m^T (r_{K_m}) will dictate the number of alternative solutions of the system. As it was stated before, the difference between these values reflects the rank deficiency of the system and the method will return rank deficiency+1 alternative solutions. To calculate these solutions, an iterative process is performed by initializing a new vector of zeros for γ and giving the value one to each position, starting from the index r_{K_m}. Thus, in each iteration γ, we will not have a zero value in only one position, and this is used to obtain the β vector and posteriorly to calculate the alternative flux distribution, as in the calculation of the principal flux distribution, through the equations 3.32 and 3.29. Each value obtained in this step is afterwards summed to the corresponding flux value in the previously obtained principal solution.

Symbols used in this chapter

Table 3.1: Symbols used in the equations and formulations of the methods.

Symbol	Description
X	vector with metabolite concentrations
v	vector with the metabolic fluxes
S	stoichiometric matrix of the model
v_i	i^{th} metabolic flux of the model
$\bar{v}_{m,i}$	mean of a set of measurements of the flux i
E_i	margin of error for the flux i
$\mathcal{M}$	set of indexes of the measured fluxes in the model
τ	metabolic flux ratio
$\mathcal{R}_{pos}$	set of indexes of the fluxes in the model that are set to be positive in the metabolic flux ratio constraints
$\mathcal{R}_{neg}$	set of indexes of the fluxes in the model that are set to be negative in the metabolic flux ratio constraints
$v_{objective}$	identifies the flux that has been selected to be maximized/ minimized
Lb_i	lower bound of flux i
Ub_i	upper bound of flux i
$\mathcal{K}$	set of indexes of the fluxes with knockouts
α	relaxing coefficient to be applied to the optimized objective
v_i^+	positive direction of the flux v_i, with respect to the stoichiometric matrix
v_i^-	negative direction of the flux v_i, with respect to the stoichiometric matrix
$v_{c,i}$	the i^{th} calculated flux
v_c	row vector with the calculated fluxes
$\bar{v}_m$	row vector with the mean of the measured fluxes
H	an $(n \times n)$ symmetric matrix, being n is the number of measured fluxes, describing the coefficients of the quadratic terms
I	identity matrix
$t_{Lb,i}$	lower tight bound for flux i

Continued on next page

Table 3.1 – continued from previous page

Symbol	**Description**
$t_{Ub,i}$	upper tight bound for flux i
v_f	fixed flux in the objective solution
ρ	percentage of the fixed flux value
$v_{control,i}$	control flux i
ρ	percentage of the control flux value
$\mathcal{C}$	set of indexes of the fluxes for the robustness analysis
S_τ	extended stoichiometric matrix, with the last rows corresponding to the ratio equations in linear form
S_m	partitioning of S_τ, with the columns corresponding to the stoichiometric coefficients of the measured fluxes
S_c	partitioning of S_τ, with the columns corresponding to the stoichiometric coefficients of fluxes that are neither measured nor deleted through the application of knockouts
r_{S_c}	rank of the matrix S_c
${S_c}^+$	Moore-Penrose pseudoinverse of S_c
$\hat{v_m}$	weighted estimation of the measured fluxes
Ψ	variance-covariance matrix for the residuals of the measured fluxes
R	redundancy matrix that conveys the relations between all the measured reaction fluxes
R_r	reduced matrix of the redundancy matrix
$E(\sigma\sigma^T)$	measured flux variance matrix
σ_i^2	variance of the flux i
K	Kernel matrix that encompasses the base vectors that span the null space of S_τ
β	column vector with the free variables of the system
$ref(S_\tau)$	reduced echelon form of S_τ
U	upper triangular factor obtained from LU decomposition
L	lower triangular factor obtained from LU decomposition
$v_{f,i}$	free variable i
c_i	vector that corresponds to the ith special solution (the ith base vector of the null space of S_τ

Continued on next page

Table 3.1 – continued from previous page

Symbol	Description
K_m^T	matrix where its columns refer to the lines in K that correspond to measured fluxes
r_{K_m}	rank of K_m^T
U_s	matrix with the r_{K_m}-first rows of the upper factor of K_m^T
$U_{leading}$	invertible leading matrix
v_{ms}	vector with the r_{K_m}-first measured fluxes
γ_s	solution of forward substitution, by solving the lower triangular system in 3.31
y	solution of back substitution by solving the upper triangular system in 3.32
$v_{m,computed}$	vector with computed fluxes that have been measured

BIBLIOGRAPHY

[Bonarius et al., 1997] Bonarius, H. P. J., Schmid, G., and Tramper, J. (1997). Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends in Biotechnology*, 15(8):308–314.

[der Heijden et al., 1994] der Heijden, R., Heijnen, J. J., Hellinga, C., Romein, B., and Luyben, K. (1994). Linear constraint relations in biochemical reaction systems: I. Classification of the calculability and the balanceability of conversion rates. *Biotechnology and bioengineering*, 43(1):3–10.

[Edwards and Palsson, 2000] Edwards, J. S. and Palsson, B. O. (2000). Robustness analysis of the escherichiacoli metabolic network. *Biotechnology Progress*, 16(6):927–939.

[Fischer et al., 2004] Fischer, E., Zamboni, N., Sauer, U., and Others (2004). High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry derived 13C constraints. *Analytical biochemistry*, 325(2):308.

[Heinrich and Schuster, 1996] Heinrich, R. and Schuster, S. (1996). *The regulation of cellular systems*, volume 416. Chapman & Hall New York.

[Lewis et al., 2010] Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. A., Charusanti, P., Polpitiya, A. D., Adkins, J. N., Schramm, G., Purvine, S. O., Lopez-Ferrer, D., and Others (2010). Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Molecular systems biology*, 6(1).

[Madron et al., 1977] Madron, F., Veverka, V., and Vaněček, V. (1977). Statistical analysis of material balance of a chemical reactor. *AIChE Journal*, 23(4):482–486.

[Mahadevan et al., 2003] Mahadevan, R., Schilling, C. H., and Others (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic engineering*, 5(4):264.

[McAnulty et al., 2012] McAnulty, M. J., Yen, J. Y., Freedman, B. G., and Senger, R. S. (2012). Genome-scale modeling using flux ratio constraints to enable metabolic engineering of clostridial metabolism in silico. *BMC systems biology*, 6(1):42.

[Palsson,] Palsson, B. O. Systems biology: properties of reconstructed networks. 2006.

[Papoutsakis, 1984] Papoutsakis, E. T. (1984). Equations and calculations for fermentations of butyric acid bacteria. *Biotechnology and bioengineering*, 26(2):174–187.

[Penrose, 1955] Penrose, R. (1955). A generalized inverse for matrices. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 51, pages 406–413.

[Pfeiffer et al., 2001] Pfeiffer, T., Schuster, S., and Bonhoeffer, S. (2001). Cooperation and competition in the evolution of ATP-producing pathways. *Science*, 292(5516):504–507.

[Sauer et al., 1999] Sauer, U., Lasko, D. R., Fiaux, J., Hochuli, M., Glaser, R. W., Szyperski, T., Wüthrich, K., and Bailey, J. E. (1999). Metabolic flux ratio analysis of genetic and environmental modulations of Escherichia coli central carbon metabolism. *Journal of bacteriology*, 181(21):6679–88.

[Segre et al., 2002] Segre, D., Vitkup, D., and Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, 99(23):15112–15117.

[Stephanopoulos et al., 1998] Stephanopoulos, G., Aristidou, A. A., Nielsen, J. H. i., and Nielsen, J. (1998). *Metabolic engineering: principles and methodologies*. Academic Press.

[Strang, 1980] Strang, G. (1980). Linear Algebra and its Applications.

[Szallasi et al., 2006] Szallasi, Z., Stelling, J., and Periwal, V. (2006). *System modelling in cellular biology*. MIT Press, Cambridge, MA.

[Wang and Stephanopoulos, 1983] Wang, N. S. and Stephanopoulos, G. (1983). Application of macroscopic balances to the identification of gross measurement errors. *Biotechnology and bioengineering*, 25(9):2177–2208.

CHAPTER 4

CBFA: A SOFTWARE TOOL FOR CONSTRAINT-BASED FLUX ANALYSIS

4.1 Introdution

In this chapter, *Constraint-Based Flux Analysis (CBFA)*, a general-purpose and flexible application to conduct metabolic flux analysis operations will be proposed. This application is independent of the type of data available and their processing, implementing the methods described and formulated in the previous chapter. *CBFA* is able to work with regular constraint-based models, without information about atom transitions between precursors and products in the model reactions (as required by other tools). Independence from carbon atom mapping facilitates user interaction, since this kind of information is not easily compiled for genome-scale models, though efforts to accomplish such task have been reported [Ravikirthi et al., 2011,Pitkänen et al., 2014].

In *CBFA*, input data are not fluxomics measurements generated by analytical techniques, such as Mass Spectrometry (MS) or Nuclear Magnetic Resonance (NMR), rather known or measured metabolic flux ratios, which can be easily used to reflect the portions of flux that generate a certain metabolite originating from different pathways [Sauer et al., 1999, Fischer and Sauer, 2003] and that have already been used in the context of genome-scale models [McAnulty et al., 2012]. Other data, such as known or measured exchange or internal fluxes can also be directly inputted.

The aim was to provide a simple to use software tool focused on the straightforward application of several state-of-the-art flux analysis methods and the simulation of the behaviour of wild type or mutant strains. Thus, the development of tools to generate metabolic flux ratios from fluxomics data (Gas Chromatography–Mass Spectrometry (GC-MS)) is addressed in the complementary software *MiNeRAl*, that will be addressed in posterior chapters.

The main features of this tool are the following:

- Open-source – it allows all users to use the tool freely and invites contribution of other researchers by providing the source code to the community;

- User-friendly – facilitates its use by users with no/little background in modeling/ programming;

- Simple – needs only a few steps to perform flux analysis methods and obtain flux distributions, simplifying the use of experimental information in the simulations;

- Modular – as it is incorporated in the *OptFlux* software [Rocha et al., 2010], it follows a plug-in based architecture, facilitating the addition of specific features;

- Compatible with standards – compatibility with the Systems Biology Markup Language (SBML) [Hucka et al., 2003], the Mathematical Markup Language (MathML) [Ausbrooks et al., 2009] and several layout standards for visualization (SBGN-ML, XGMML, SBML Layout) rendering interoperability with other tools easier.

In its current version, *CBFA* accommodates several tools and algorithms that have been developed for the manipulation of metabolic models:

- methods for phenotype simulation, such as Flux Balance Analysis (FBA) and its variant Parsimonious enzyme usage Flux Balance Analysis (pFBA) [Lewis et al., 2010];
- a method to minimize the error between calculated and measured fluxes, through a Quadratic Programming (QP) formulation;
- methods for system characterization:
  - through the calculation of tight bounds of fluxes in the model;
  - through the analysis of the null space of the stoichiometric matrix.
- a method for robustness analysis of a configured objective function;
- methods to calculate fluxes when the non-measured fluxes can be calculated from the input constraints, and exploiting redundancies when additional data are available;
- a suitable model visualization tool to facilitate the interpretation of results.

In Additional file 1, in supplementary material, some of the features of this tool are highlighted, when compared to some of the applications, which were approached in chapter 2. It also lists, in a different tab, the novel features of this tool in the framework of the *OptFlux* application.

The main concepts used in the development of this tool and its main functionalities are presented in the next sections.

4.2 Software functionalities

An overview of the overall workflow implemented in *CBFA* is provided in Figure 4.1, where the different inputs, the generated constraints, the system types, the supported methods and their outputs are shown.

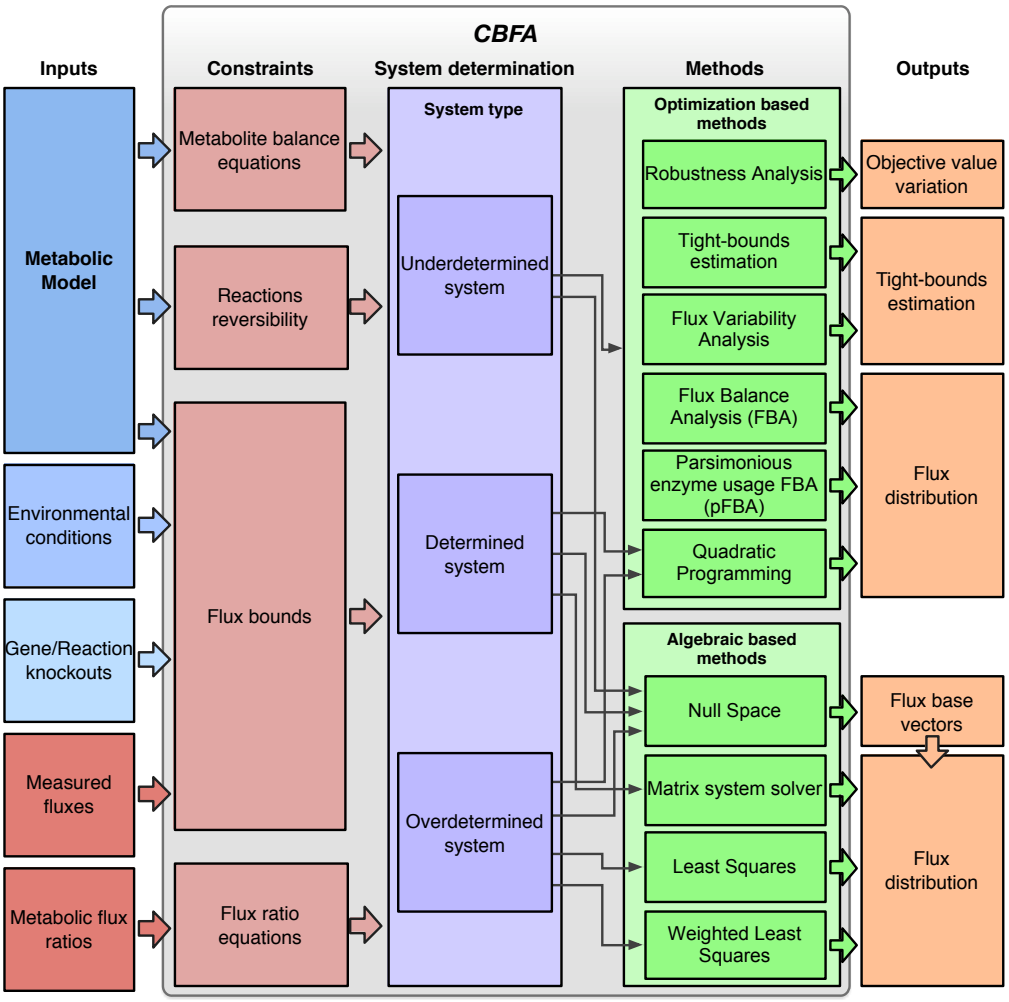


Figure 4.1: Overview of the application. On the left side, the inputs are represented; the centre box contains the different functional blocks of the application: the types of constraints, the determination of the system type and the supported methods; the right boxes represent the outputs of the methods.

In this section, the main steps and alternatives offered by the pipeline to execute the flux analysis methods are briefly explained. The detailed

description of the implemented methods and complete mathematical formulations were presented in the previous chapter.

Additionally, a Beginner's tutorial with illustrations for all the steps needed to perform the software tasks is available to help first-time users (Additional file 2 and online documentation on the site).

4.2.1 System configuration

Starting with a metabolic model that can be loaded in different formats (e.g. SBML or CSV files), the user can configure inputs to flux analysis methods, including environmental and genetic conditions (gene/reaction deletions), measured or otherwise known fluxes and/ or flux ratios given as expressions in the form:

$$\frac{\sum_{i=1}^{n} \kappa_i v_i}{\sum_{j=1}^{n} \kappa_j v_j} = \tau \tag{4.1}$$

where κ_i, $\kappa_j \in \mathbb{R}$ and $\tau \in \mathbb{R}\backslash\{0\}$ are user-defined real numbers and v_i and v_j are fluxes.

This allows the definition of constraints of different origins, e.g. coming from calculations over experimental data, for example from ^{13}C experiments, or deriving from existing knowledge about a biological system. The specific calculations to reach flux ratios from experimental data in different formats are not part of this software.

Given this information, degrees of freedom of the system are calculated from the properties of the original stoichiometric matrix of the model, the defined knockouts, the experimental values defined for some fluxes, and the

provided flux ratios. The accurate number of degrees of freedom is obtained by the difference between the number of unknown variables (fluxes) of the system and the number of linearly independent equations of the system.

However, since the purpose of this work is mainly to perform flux analysis tasks upon GSMMs, the typical configurations will lead to large numbers of degrees of freedom. It is not relevant to calculate the exact number of degrees of freedom if this is too high. Therefore, the costly algebraic operations that are performed towards the exact calculation of the degrees of freedom are not always made, rather an approximation is firstly calculated, given by the difference between the unknowns and the equations of the system, without taking into account the linear dependency between them. If this approximation gets close to a certain threshold, the proper number of degrees of freedom is calculated and presented to the user.

To reduce the degrees of freedom, the user can add new flux measurements, gene/ reaction deletions or equality flux ratios. If the type of system changes, the set of available methods is updated accordingly.

4.2.2 Available methods for flux analysis

Two approaches are used to determine flux distributions: algebraic or optimization based methods (linear or quadratic programming). Depending on the degrees of freedom of the system, different methods can be applied Figure 4.1. If the system is either determined or over-determined it can be algebraically solved and, therefore, fluxes can be calculated. In the latter case, the additional information can be used to calculate fluxes by regular or weighted least squares fitting.

However, with genome-scale models, systems are more commonly underdetermined, even when experimental information is added. Thus, op-

timization methods are used to obtain flux values, by defining an objective function, while respecting the defined constraints. FBA and pFBA can be selected, maximizing a given objective function (e.g. the biomass flux). An alternative is to use quadratic programming to minimize the difference between the values of a subset of measured fluxes and their calculated values. The resulting formulation is mathematically similar to the minimization of metabolic adjustment Minimization of Metabolic Adjustment (MOMA) [Segre et al., 2002].

The null space approach can be used to calculate a unique solution to the non- measured fluxes, when the system is either determined or overdetermined. It can also be used when the system is underdetermined, but in this case returning the admissible flux space represented by the generating base vectors. These vectors can be used to generate valid flux distributions.

In order to characterize the system, other approaches can also be used like the computation of tight bounds of the system under certain conditions. Here, the lower and the upper admissible bounds for all fluxes, which have neither been measured nor have associated knockouts, are obtained through the application of an optimization-based method that maximizes/ minimizes the flux of interest under the same set of constraints. This method is an approach based on Flux Variability Analysis (FVA) [Mahadevan et al., 2003] where no constraints are defined for the objective function.

Different solutions leading to the same optimal objective value can be achieved by the methods based on LP problems. To characterize the alternative solutions that satisfy the imposed set of constraints and have the same objective value, FVA is again used, this time to calculate the admissible ranges of flux values under optimality [Mahadevan et al., 2003].

To analyze the changes of the objective function in response to variations

of a specific flux, a method for robustness analysis is also available. For a selected set of fluxes, their values in a prior optimization for the given objective are retrieved. After that, each flux is varied in different percentages of its prior value and the effects of these changes on the objective function can be studied [Edwards and Palsson, 2000].

Furthermore, all the aforementioned methods can be used to investigate how a set of knockouts is reflected on the flux distribution. This is accomplished by the configuration of knockouts as additional constraints to the system. These knockouts can be inserted as reactions or as genes by using the gene-reaction rules of the model (when these are available).

The results and configurations used by the methods described above can be exported to a PDF file, including the model details, the method that was used, all the constraints that were selected, and the result of the simulation in terms of objective function and net conversions. Moreover, it reports the bounds that were actually used in the simulations when constraints overlap with each other.

4.2.3 User interaction

The aim of *CBFA* is to provide the community with a tool to perform flux analysis tasks that is simple, intuitive and with high usability. The interaction with this software is based on three main concepts:

- *Datatypes*: Encapsulate the different types of objects of the application, holding relevant data such as models, flux measurements, flux ratios, simulation results, etc. The application can keep several instances (objects) of the same datatype;

- *Views*: To check the contents of a datatype, panel views are defined

that present to the user the most relevant information that it encompasses. Since a single datatype might contain different types of information, it can have multiple views shown as different panels (tabs);

- *Operations*: All the functionalities are invoked in the form of operations, which are units with well-defined sets of inputs and outputs (datatypes). Whenever an operation is called, a dialog is presented to the user to define the set of input objects. The most common scenario is to create a new instance of a certain datatype (e.g. the result of a simulation) or to modify existing instances (e.g. add flux measurements to an existing set).

Figure 4.2 provides snapshots of some views and interfaces of the operations invoked when performing simulations and constraints configuration, showing a typical workflow when working with *CBFA*.

A metabolic model is loaded defining a working project, which is the root datatype in the hierarchical organization of the clipboard (left-hand side panel). The visualization of a selected object is presented on the right side. The structure and content of the visualization depends on the specific datatype, and on the object's specific information. The hierarchical tree of the clipboard keeps all the results of the operations that are performed during the software utilization, namely datatypes for environmental conditions, flux measurements, flux ratios, and simulation results. Results are clustered in the clipboard according to the method used.

All the operations are performed on the scope of a certain project that keeps a single instance of a metabolic model. The other datatypes, such as flux measurements, can have multiple instances, under the same project, and the user can perform different tasks using alternative instances. More-

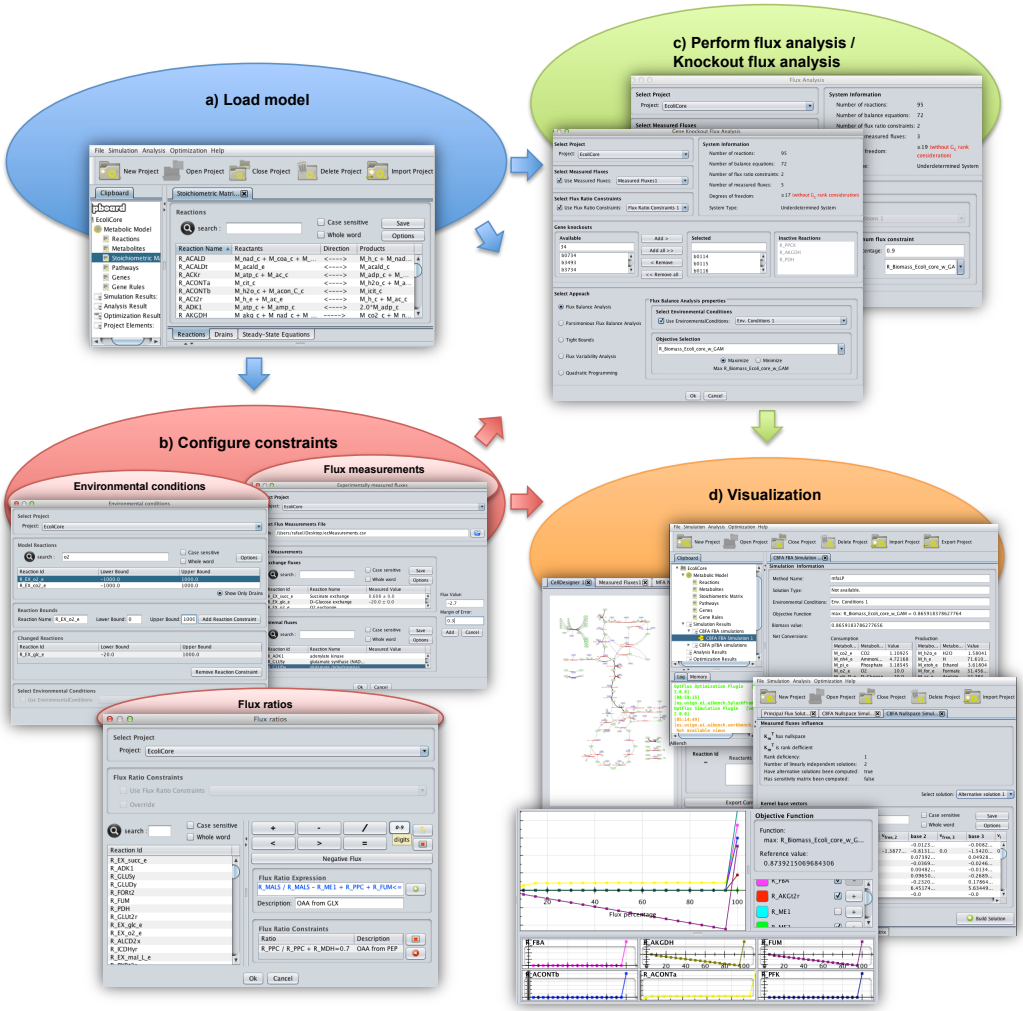


Figure 4.2: Screenshots of CBFA. a) Load model – the first step is to load a metabolic model and a new project will be created and made available on the clipboard of the application. Afterwards, it is possible to visualize their information through the available views; b) Configure constraints – different type of constraints can be configured through the use of graphical interfaces that enable to set the parameters to create the datatypes to be used on further operations; c) Perform flux analysis/knockout flux analysis – the user can select different constraints to perform several flux analysis methods using a metabolic model. It is also possible to configure a set of knockouts and perform flux analysis over the mutant with the same constraints; d) Visualization – the results of the operations create/update datatypes on the clipboard, and the output can always be checked through the views that are defined for the datatypes.

over, in the same running application, it is possible to have more than one project in the clipboard.

Operations are made available either on the top menu bar of the application or in a popup menu that is triggered when a right click event occurs upon a clipboard item. In the last case, only the operations that have the selected type as one of its input argument types are shown. Detailed information on how the main operations can be performed and to interact with the software is provided in the supplementary material, Additional file 2, that provides the software's tutorial.

4.3 Case studies

To illustrate some of the main features of *CBFA*, two case studies were considered. First, the yeast *Saccharomyces cerevisiae* was investigated through the use of a simplified metabolic model for its growth [Förster et al., 2002], while in the second one, a genome-scale model of the microorganism *Escherichia coli* [Reed et al., 2003] was used to perform flux analysis tasks based on optimization methods.

4.3.1 *Saccharomyces cerevisiae* (over)determined configured system

In the first case, the aim was to show how the implemented methods for (over)determined systems could be exploited. The model encompasses 45 reactions and 49 metabolites. The study was performed to analyse the phenotype response obtained by performing a deletion of the fumarate hydratase 1 reaction, through a knockout of the fum1 gene [Blank et al., 2005], under aerobic conditions with glucose as the carbon source.

Without adding constraints to the model, the system has 9 degrees of freedom. After setting the fum1 knockout, and adding the uptake and excretion rates given in [Blank et al., 2005], the configured system remained underdetermined. So, some flux ratios, such as the fraction of phosphoenolpyruvate originating from oxaloacetate and the fraction of oxaloacetate deriving from pyruvate [Blank et al., 2005], were added. In this way, the system was made determined. Thus, it was possible to calculate the intracellular fluxes (see the full results in the Additional file 3 in supplementary material) through algebraic methods.

As an illustration of the results, in Figure 4.3, the distribution of the central carbon metabolism fluxes in *Saccharomyces cerevisiae* is illustrated by overlaying fluxes over the network topology. This figure also serves as an example of the visualization capabilities of the tool.

4.3.2 Large-scale constrained-based flux analysis in *Escherichia coli*

In a second case study, to demonstrate how optimization-based methods can be used to investigate flux distributions in GSMMs, an analysis was made using the *i*JR904 model for *Escherichia coli* MG1655 [Reed et al., 2003]. This model contains 932 reactions, (including a biomass equation), 618 internal metabolites and 143 external metabolites (with respective drain fluxes) are present in the model.

The original model has 332 degrees of freedom, and, therefore, the system was underdetermined and appropriate for the utilization of optimization-based methods, such as the ones based on linear programming approaches. In this case, the high number of degrees of freedom precludes the use of methods for determined or over-determined systems, as the amount of ex-

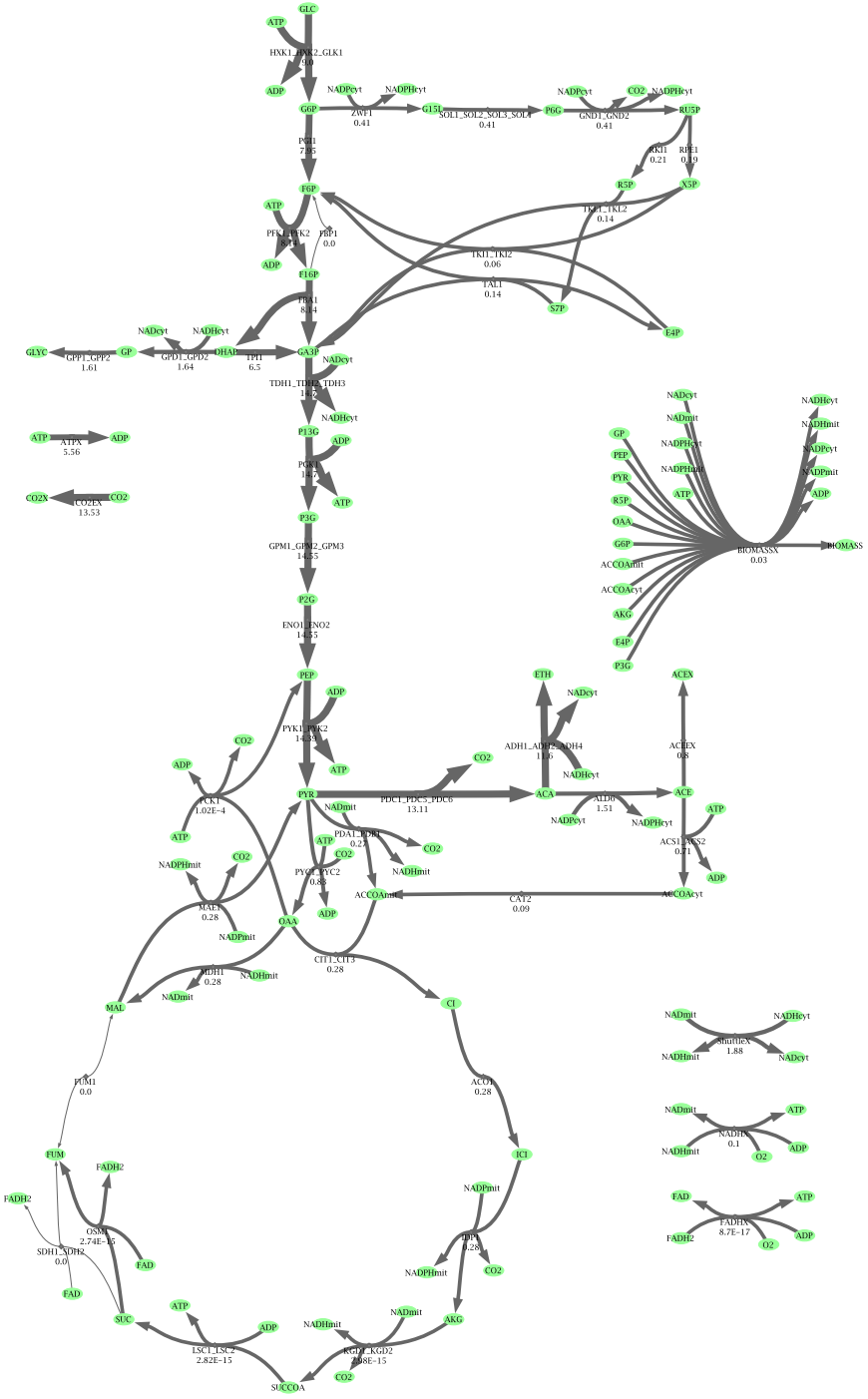


Figure 4.3: Visualization of flux analysis results for Saccharomyces cerevisiae. The flux values were obtained through least squares, since the configured system was determined. The thickness of the arrows is proportional to the flux value of the corresponding reactions, after overlaying the fluxes over the network layout. Thin light grey arrows represent reactions with no flux value.

tra constraints to be added is very large.

To configure the experimental inputs to the analysis, the work from [Fischer et al., 2004] was considered. Here, to study the influence of the inputs, the metabolic flux distribution was examined by incrementally adding constraints to the system. Analyses were made under aerobic conditions with glucose as the carbon source.

The first step was to perform an FBA-based method. By analysing the flux distribution, the absolute values for the reactions R_ADK1 and R_ADK3 were considered to be too high. R_ADK1 was also occurring on the reverse direction, making a cycle with R_ADK3. After performing an FVA analysis, the two reactions had as lower and upper bounds the two extremes of their domains. Therefore, a pFBA-based method was performed, to minimize the total sum of the flux absolute values and remove this futile cycle. It was possible to confirm that the values for these two reactions changed in comparison to the FBA simulation, such that R_ADK1 changed its direction and R_ADK3 was set to zero. For the simulations performed in this first step, the only constraint that has been added was an environmental condition to set the glucose uptake rate to 11 mmol/gDW/hr. The objective was to maximize growth and the obtained value for the biomass flux was 0.97 hr^{-1}.

In order to approximate the distribution according to [Fischer et al., 2004], a second step was done, adding external flux measurements for O_2, CO_2 and acetate, which resulted in a decrease of the biomass flux to 0.81 hr^{-1}. However, there were still some differences in the internal fluxes with respect to the published values. Therefore, some flux ratios were added and changes were detected in the internal flux distributions, such as the flux of phosphoenolpyruvate carboxylase and ATP:oxaloacetate carboxy-lyase.

Figure 4.4 shows the comparison of the flux distribution in the central metabolism between the simulations without and with the incorporation of the metabolic flux ratios. The detailed results (flux distributions) are given in Additional file 4 in supplementary material, together with some notes on the conditions.

4.4 Implementation

CBFA is fully implemented in the Java language, and thus available for all platforms. The only non-Java parts consist of the linear/ quadratic programming solvers[1], where several interfaces are implemented such as the *GNU Linear Programming Kit* [Makhorin, 2008] and *IBM ILOG CPLEX*.

In order to explore several features for model handling and visualization, and to enable the interaction with different tools, *CBFA* was built as part of *OptFlux*. *OptFlux* is an open-source and modular Metabolic Engineering (ME) software platform developed by the host group.

4.4.1 *CBFA* software architecture

All the Java classes in *CBFA* are separated in two main modules: the Core module with the implementations of all the developed simulation methods and associated type dependencies, and the Controller module that encompasses the graphical component of the software (Figure 4.5). Both modules are integrated with the same level modules of *OptFlux*. The Core module is completely autonomous of the Controller module and its Application Programming Interface (API) is designed to allow the use of the implemented

[1]Quadratic programming problems can be solved with some specific solvers. Please refer to the tutorial document (available on `http://www.optflux.org/cbfa`) to check the solvers that are supported in *CBFA*.

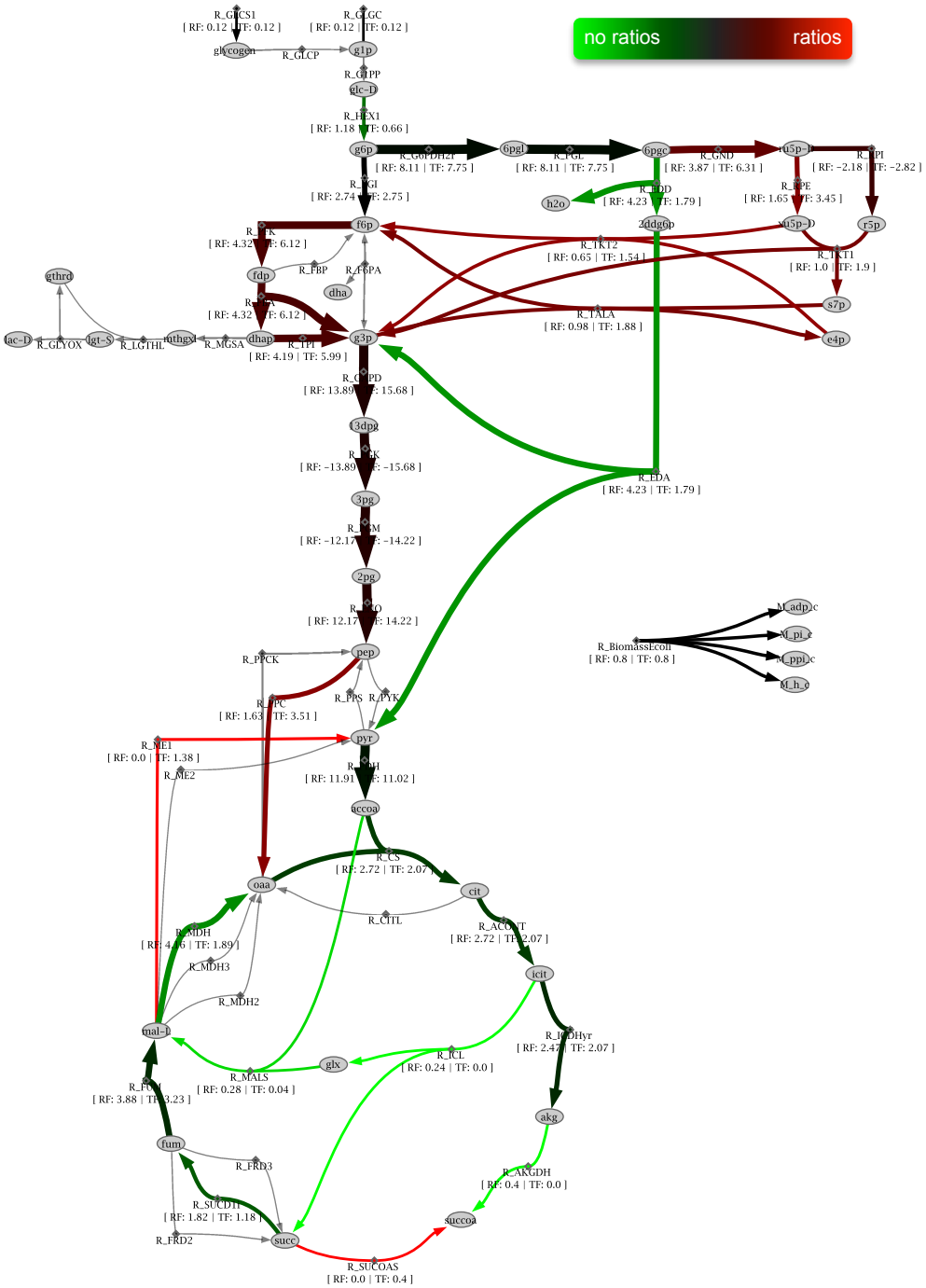


Figure 4.4: Visualization of flux analysis results for Escherichia coli. The central carbon metabolism of Escherichia coli is shown with a comparison between pFBA-based simulations before and after adding metabolic flux ratio constraints. Here, grey arrows indicate reactions where there is no flux in both simulations, while red and green arrows represent reactions for which the simulations with and without flux ratios, respectively, returned flux values. The darker the arrows are, the nearer the fluxes in both simulations.

flux analysis methods by external computational tools, with no need to add dependencies to the module containing the Graphical User Interface (GUI) components of the software.

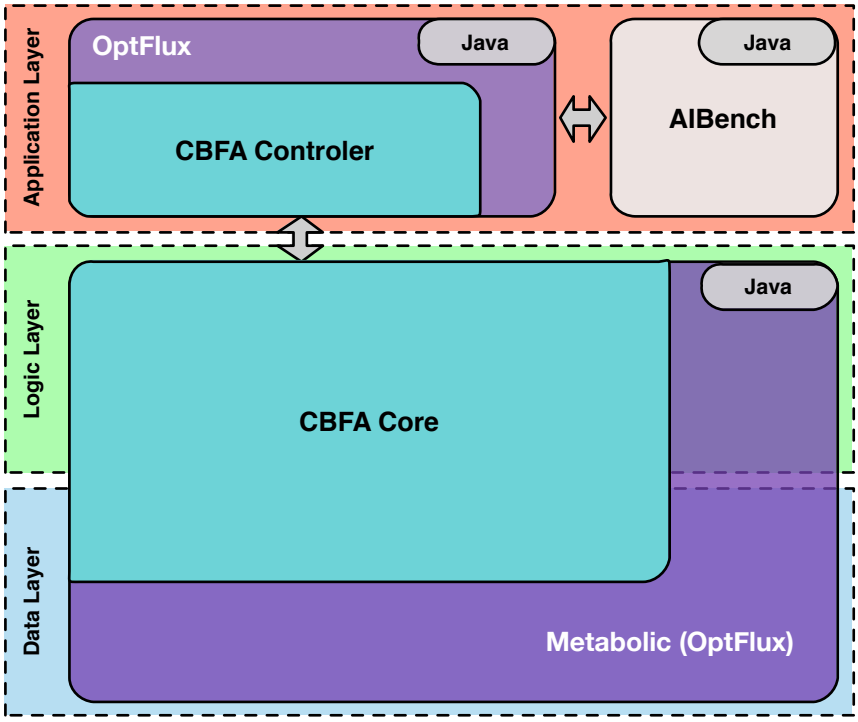


Figure 4.5: CBFA layers. The application layer is the top layer, composed by the Controller module, OptFlux application module and AIBench, which are all connected since CBFA and OptFlux follow the AIBench architecture. The implementations in the data and logic layers belong to the Core module which extends the API of OptFlux relating to metabolic simulation tasks. The Core is completely independent from the Controller module. All the layers implementations are developed in Java.

However, since the Core of *CBFA* explores *OptFlux* by using its API to access metabolic models and extending its simulation resources, to use the module, programmers need to make their Java projects dependent of the different pieces in the *OptFlux*'s workbench. This is not an issue due to the fact that already different projects have been defined for the Metabolic related components, and to the application part.

OptFlux is implemented following the *AIBench* paradigm, which is a

plug-in based application framework [Glez-Peña et al., 2010]. This characteristic facilitates the incorporation of new features and pieces of software in the application layer as new plug-ins, as it is the case of *CBFA*.

The entire application layer follows the Model-View-Controller (MVC) design pattern, incorporating three well-defined concepts: operations, datatypes and datatype views (Figure 4.6). This design pattern allows the combination and re-use of different units of work and facilitates the continuous development of new features.

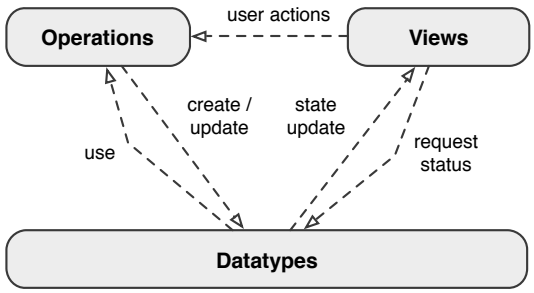


Figure 4.6: Application layer MVC architecture. Operations can use datatypes as inputs, but also can produce, or update them, as the results of its corresponding tasks. The datatypes are observed by the views, which show their contents. It is possible to apply some changes in the datatypes, from the views, but always by calling the responsible operations.

Every functionality in *CBFA* is made available in the Controler in the form of an *AIBench* operation. Through a set of Java annotations, an operation is structured in terms of input and output ports, i.e., datatypes that work as wrappers for Core module classes. All the datatypes are generated through such operations, and by the time they are instantiated, they become visible in the software's clipboard (left panel in the main frame). These datatypes are observed by the views that are registered to account for their events. Any time an operation updates a datatype, the responsible views that are activated will reflect the changes. There is no direct

interaction between a view and a datatype, although it is possible to call operations over the datatype from a view.

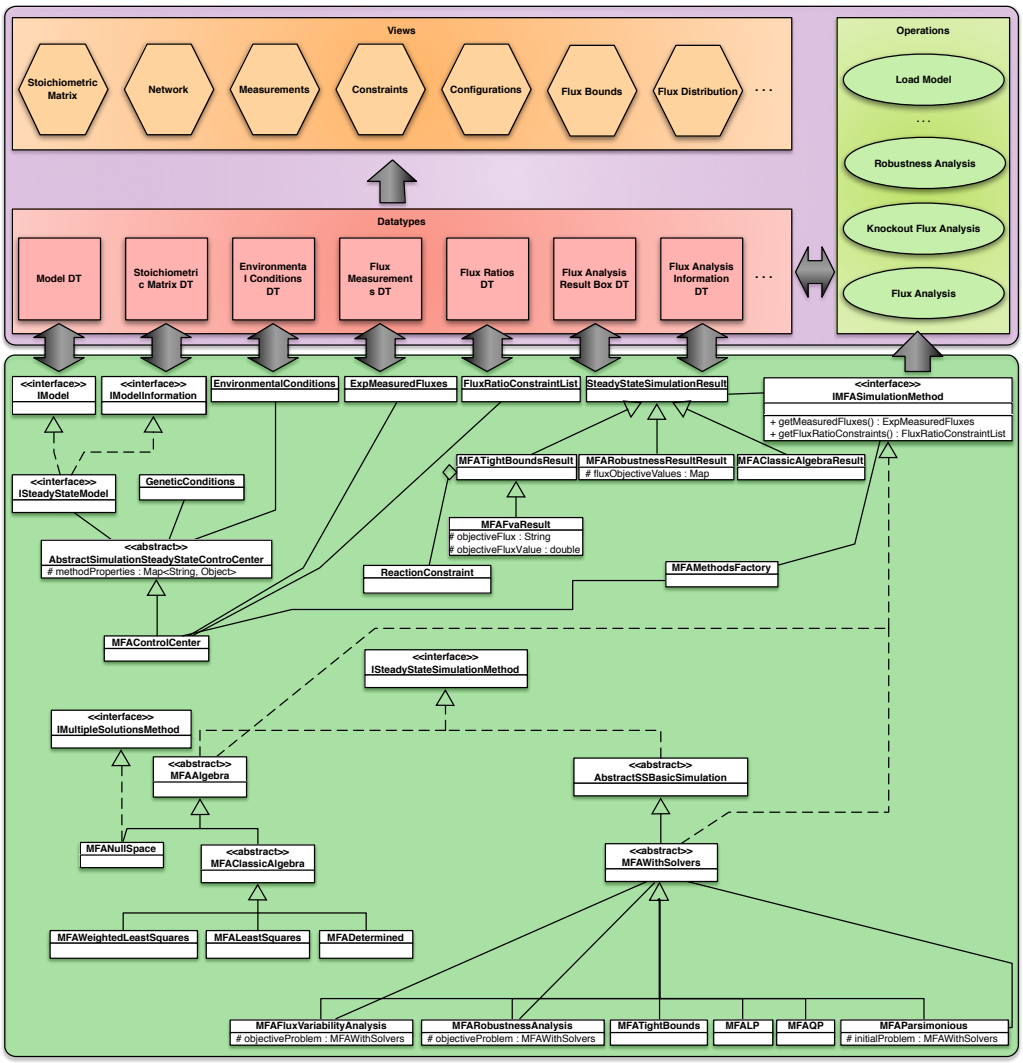


Figure 4.7: Functional modules of the CBFA application, including the developed datatypes, views and operations. The bottom box shows an overview of the classes implemented in the core layer.

The concepts of datatypes, views and operations are, therefore, used to build the application layer and to construct graphical user interfaces, which intermediate between the user and all the core methods implemented. In Figure 4.7, the architecture of the application is illustrated showing the

main datatypes, views and operations, as well as the relationship with the core classes utilized in the implementation.

4.4.2 Performing metabolic flux analysis simulations

All the implemented methods for performing flux analysis in *CBFA* are registered in an entity named *MFAControlCenter* (Figure 4.7). An instance of this class will be responsible to invoke the appropriate method based on the user inputs. When calling the simulation method from this entity, a sequence of steps is performed to: configure the linear/quadratic problem to be solved (for the cases where an optimization problem has been chosen); override the model reactions constraints based on the imputed measured fluxes, environmental and genetic conditions, and metabolic flux ratios; execute the simulation problem; and, finally, construct the final solution.

In Figure 4.8, an Unified Modeling Language (UML) sequence diagram is shown with the main steps for the configuration of a pFBA method execution, chosen here as an example. Although some steps are omitted to simplify the diagram, the main steps for understanding how the process works are illustrated. The diagram is continued in Figure 4.9, where the main instructions to simulate the problem are presented.

4.4.3 Availability and requirements

The software is made available, together with other resources, in the *Opt-Flux* home page. In Table 4.1, the relevant technical information about *CBFA* is summarized.

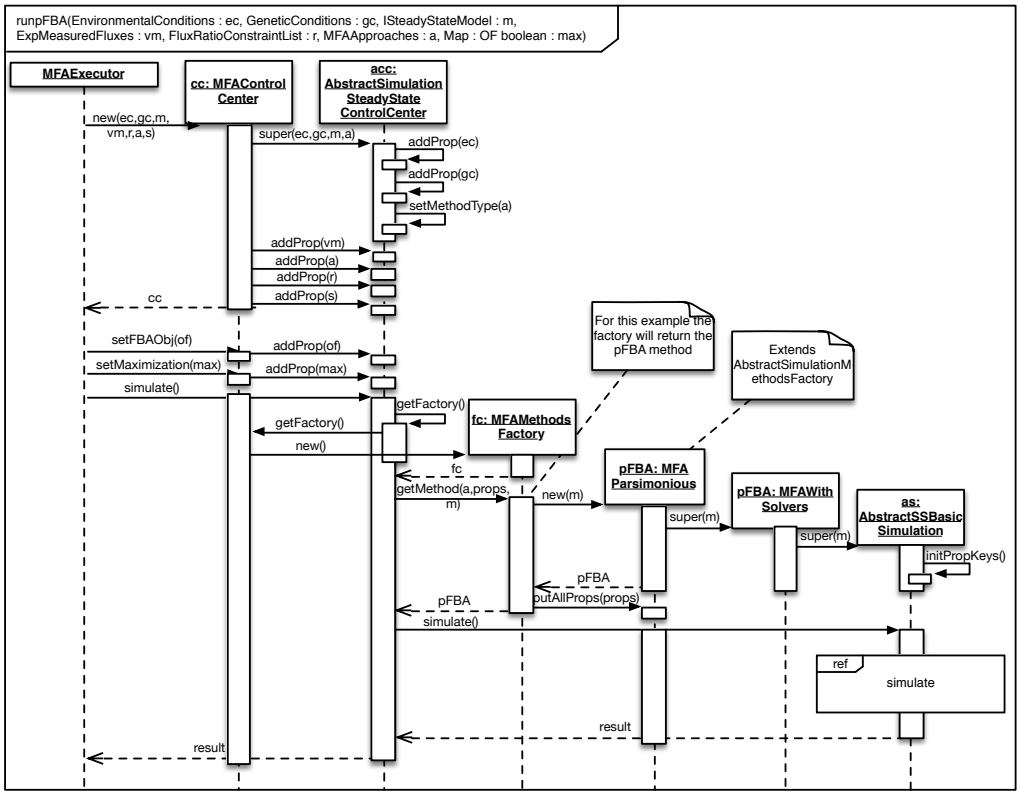


Figure 4.8: UML sequence diagram for the configuration of an execution of the pFBA method. The MFAControlCenter encompasses the user constraints to be added to the LP problems. Here, an initial MFA problem is instantiated to perform a simulation to the main objective, afterwards the same constraints are used for the pFBA minimization problem.

Table 4.1: CBFA technical information.

Software name	*CBFA* plugin for *OptFlux*
Project home page	http://www.optflux.org
Application tutorial	http://www.optflux.org/cbfa (Tutorial file in supplementary material, as Additional file 2)
Operating system(s)	Platform independent
Programming languages	Java
Other requirements	Java JRE 1.7.x (for Mac OS users the installation of JDK 1.7 is recommended), GLPK
License	GNU-GPL, version 3

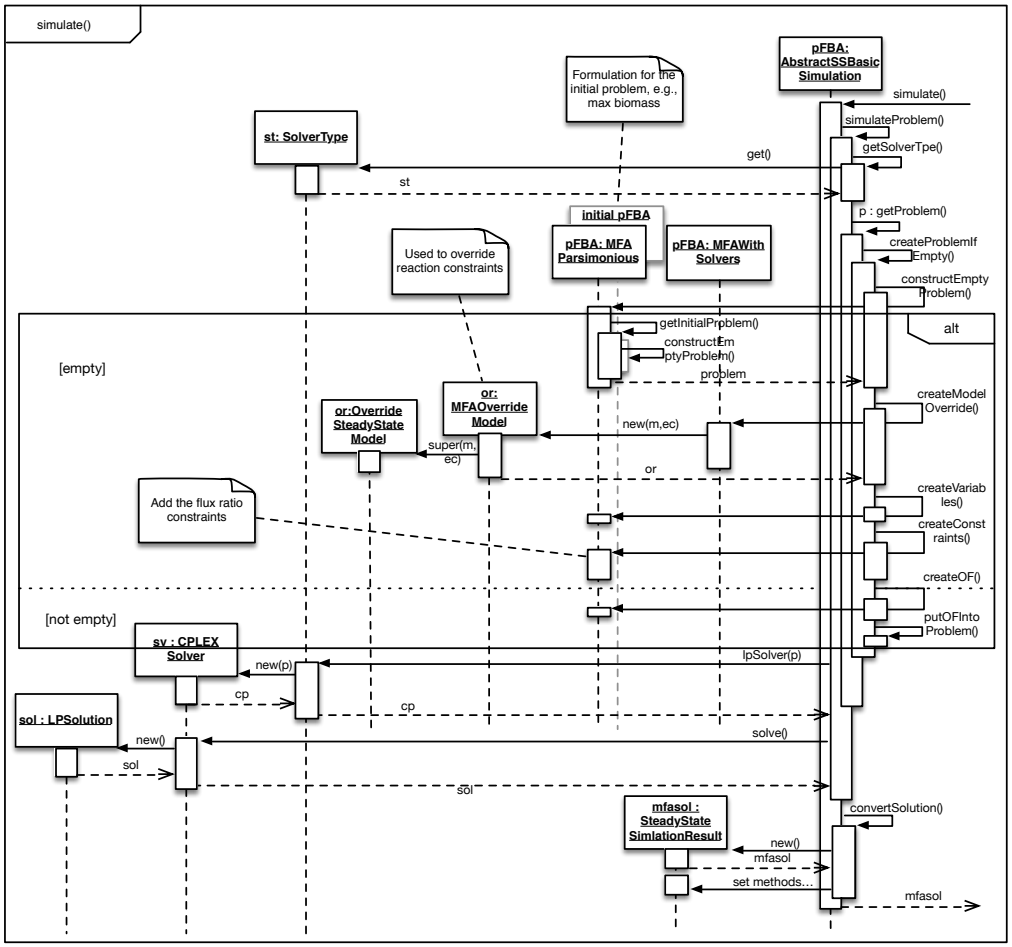


Figure 4.9: Continuation of the UML sequence diagram for executing a pFBA method. The steps for the simulations are shown, where the CPLEX solver is chosen to solve the LP problems.

BIBLIOGRAPHY

[Ausbrooks et al., 2009] Ausbrooks, R., Buswell, S., Carlisle, D., Chavchanidze, G., Dalmas, S., Devitt, S., Diaz, A., Dooley, S., Hunter, R., Ion, P., and Others (2009). Mathematical Markup Language (MathML) Version 3.0. W3C Candidate Recommendation of 15 December 2009. *World Wide Web Consortium*, 13.

[Blank et al., 2005] Blank, L. M., Kuepfer, L., and Sauer, U. (2005). Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome biology*, 6(6):R49.

[Edwards and Palsson, 2000] Edwards, J. S. and Palsson, B. O. (2000). Robustness analysis of the escherichiacoli metabolic network. *Biotechnology Progress*, 16(6):927–939.

[Fischer and Sauer, 2003] Fischer, E. and Sauer, U. (2003). Metabolic flux profiling of Escherichia coli mutants in central carbon metabolism using GC-MS. *European Journal of Biochemistry*, 270(5):880–891.

[Fischer et al., 2004] Fischer, E., Zamboni, N., and Sauer, U. (2004). High-throughput metabolic flux analysis based on gas chromatography–mass spectrometry derived 13C constraints. *Analytical Biochemistry*, 325(2):308–316.

[Förster et al., 2002] Förster, J., Gombert, A. K., and Nielsen, J. (2002). A functional genomics approach using metabolomics and in silico pathway analysis. *Biotechnology and bioengineering*, 79(7):703–12.

[Glez-Peña et al., 2010] Glez-Peña, D., Reboiro-Jato, M., Maia, P., Rocha, M., Díaz, F., and Fdez-Riverola, F. (2010). AIBench: A rapid application development framework for translational research in biomedicine. *Computer methods and programs in biomedicine*, 98(2):191–203.

[Hucka et al., 2003] Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., and Others (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531.

[Lewis et al., 2010] Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. A., Charusanti, P., Polpitiya, A. D., Adkins, J. N., Schramm, G., Purvine, S. O.,

Lopez-Ferrer, D., and Others (2010). Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Molecular systems biology*, 6(1).

[Mahadevan et al., 2003] Mahadevan, R., Schilling, C. H., and Others (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic engineering*, 5(4):264.

[Makhorin, 2008] Makhorin, A. (2008). GLPK (GNU linear programming kit).

[McAnulty et al., 2012] McAnulty, M. J., Yen, J. Y., Freedman, B. G., and Senger, R. S. (2012). Genome-scale modeling using flux ratio constraints to enable metabolic engineering of clostridial metabolism in silico. *BMC systems biology*, 6(1):42.

[Pitkänen et al., 2014] Pitkänen, E., Jouhten, P., Hou, J., Syed, M. F., Blomberg, P., Kludas, J., Oja, M., Holm, L., Penttilä, M., Rousu, J., and Others (2014). Comparative Genome-Scale Reconstruction of Gapless Metabolic Networks for Present and Ancestral Species. *PLoS computational biology*, 10(2):e1003465.

[Ravikirthi et al., 2011] Ravikirthi, P., Suthers, P. F., and Maranas, C. D. (2011). Construction of an E. Coli genome-scale atom mapping model for MFA calculations. *Biotechnology and bioengineering*, 108(6):1372–1382.

[Reed et al., 2003] Reed, J. L., Vo, T. D., Schilling, C. H., Palsson, B. O., and Others (2003). An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biol*, 4(9):R54.

[Rocha et al., 2010] Rocha, I., Maia, P., Evangelista, P., Vilaça, P., Soares, S. a., Pinto, J., Nielsen, J., Patil, K., Ferreira, E., and Rocha, M. (2010). OptFlux: an open-source software platform for in silico metabolic engineering. *BMC systems biology*, 4(1):45.

[Sauer et al., 1999] Sauer, U., Lasko, D. R., Fiaux, J., Hochuli, M., Glaser, R. W., Szyperski, T., Wüthrich, K., and Bailey, J. E. (1999). Metabolic flux ratio analysis of genetic and environmental modulations of Escherichia coli central carbon metabolism. *Journal of bacteriology*, 181(21):6679–88.

[Segre et al., 2002] Segre, D., Vitkup, D., and Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, 99(23):15112–15117.

CHAPTER 5

ALGORITHMS TO INFER METABOLIC FLUX RATIOS FROM FLUXOMICS DATA

5.1 Introduction

In this chapter, a novel automated approach for determining metabolic flux ratios from experimental data is proposed, that is independent of any assumptions on the metabolic network and on the reversibility of the reactions. This framework generates all the possible equations to calculate metabolic flux ratios from a given Genome-Scale Metabolic Model (GSMM), taken as input a set of intact carbon fragments of internal metabolites for which relative abundances have been measured (or determined from the abundance of the amino acids that are biosynthesized by them [Szyperski, 1995]). Moreover, when not enough labeling measurements are given to the system, equations are built to estimate ratio bounds for the metabolic fluxes. In the cases that the generated equations are exact and linear, the ratios can be algebraically calculated. On the other cases, a meta-heuristic method is configured to estimate the ratios. Thus, the obtained ratios can be used to set up metabolic flux ratio constraints for existing computational methods designed to perform constraint-based flux analysis and Metabolic Flux Ratio (METAFoR) analysis [Zamboni et al., 2005, McAnulty et al., 2012].

5.2 Definitions and problem statement

5.2.1 Mass isotopomer distribution vectors

When ^{13}C-labeled substrates are fed to cells, the ^{13}C labeling of the substrate is propagated throughout the metabolic pathways, and consequently the compounds that are interconnected by these pathways show a labeling pattern that is the result of how labeled carbon atoms flow throughout these pathways. A molecule that contains n carbon atoms, can have 2^n different labeling patterns, due to the different possible combinations of the labeled carbon atoms and their positions in the molecule. Each labeling pattern defines a positional isotopomer of the molecule, while the set of isotopomers with the same number of labeled atoms defines the mass isotopomers of the molecule. Based on these mass isotopomers, for an n carbon-sized molecule, it is possible to build a probability vector, named Mass isotopomer Distribution Vector (MDV), with n+1 positions, where the i-th position of the vector corresponds to the probability of the molecule being labeled on i carbon atoms.

Using experimental data from Mass Spectrometry (MS), the probabilities are assigned from the fractional abundance of the m_{+i} mass of the fragment, where the m_{+0} corresponds to the mass of the lightest isotopomer, and the remaining are the abundance of molecules with higher masses [Fis-

cher and Sauer, 2003]:

$$MDV_x = \begin{bmatrix} m_{+0} \\ m_{+1} \\ \vdots \\ m_{+n} \end{bmatrix}, \sum_{i=1}^{n} m_{+i} = 1 \tag{5.1}$$

The experimental data is often obtained from techniques that detect fragments instead of entire molecules. The most common scenario is the use of Gas Chromatography–Mass Spectrometry (GC-MS), where a set of ion fragments from different parts of the carbon skeleton of the molecules are detected. It is, therefore, important to introduce the representation of a fragment as $\mathrm{A}_{i,j}$, stating that it represents a fragment of compound A that starts in its i-th and ends in the j-th carbon atoms.

5.2.1.1 Naturally occurring isotopes

When the measured data correspond to amino acids, some operations over the MDV of the fragments are necessary to correct the vectors for their natural isotope abundances [van Winden et al., 2002]. The correction is made for all the biologically relevant elements belonging to the compound fragments. In the case of the methyl chloroformate (MCF) derivatization [Smart et al., 2010], the derivatives will contain atom elements from carbon (C), oxygen (O), nitrogen (N), hydrogen (H), Silicon (Si) and (Sulfur), so for a given MCF fragment, a correction matrix configured for all these elements will be used to correct its MDV, according to equation 5.2.

$$MDV_x^{corr} = C_{x,CONHSiS} \,.\, MDV_x \tag{5.2}$$

The correction matrix $C_{x,CONHSiS}$ is obtained through the following equation

$$C_{x,CONHSiS} = \prod_{\substack{\text{all elements s of} \\ \text{CONHSiS}}} C_{x,s} \tag{5.3}$$

where $C_{x,s}$ is an $m \times m$ correction matrix for the element s in the fragment x, and m is the length of the MDV of fragment x.

The first column of $C_{x,s}$ is composed by the abundances of the isopologues (usually identified by their chemical formula), which are molecular species that although differ in isotopic content, have the same elemental and chemical compositions [Hellerstein and Neese, 1999]. As an example, for the case of the fragment C_2-C_4 of Aspartate, which contains four atoms of oxygen, the correction matrix is represented as in Equation 5.4.

$$C_{Acetate,O} = \begin{bmatrix} a(^{16}O_4) & 0 & 0 & 0 \\ a(^{16}O_3\,^{17}O) & a(^{16}O_4) & 0 & 0 \\ a(^{16}O_2\,^{17}O_2 + ^{16}O_3\,^{18}O) & a(^{16}O_3\,^{17}O) & a(^{16}O_4) & 0 \\ a(^{16}O^{17}O_3 + ^{16}O_2\,^{17}O^{18}O) & a(^{16}O_2\,^{17}O_2 + ^{16}O_3\,^{18}O) & a(^{16}O_3\,^{17}O) & a(^{16}O_4) \end{bmatrix} \tag{5.4}$$

where each element of the matrix $(a_{i,j})$ is 0 if no mass isotopomer (m_{+i}) exists and each element $a_{i,2}$, …, $a_{i,m}$ is equal to $a_{i-j+1,1}$, if $i \geq j$, or equal to 0, if $i < j$.

The abundance for the isotopologue j of the atom element s can be obtained as follows [van Winden et al., 2002]:

$$a(isotopologue_{s,j}) = \left(\sum_{i=1}^{n} f(I_i)\right)! \; . \; \prod_{i=1}^{n} \frac{x(I_i)^{f(I_i)}}{f(I_i)!} \tag{5.5}$$

where n is the number of naturally occurring isotopes, I_i is the ith isotope, $f(I_i)$ is the frequency of I_i within the isotopologue and $x(I_i)$ is the natural abundance of I_i. The sum of the frequencies of the isotopes for each isotopologue has to be equal to the number of atoms of s in the considered compound fragment.

After correcting the amino acids MDVs, it is possible to obtain the same type of vectors for their precursors [Fischer and Sauer, 2003]. Therefore, the fragments of the internal metabolites will depend on the analytical technique to obtain the data, and also on the network topology, as it will be shown in Section 6.2.8.3.

In the following, $\mathcal{F}$ will denote the set of fragments that have been measured, i.e., that contain an associated MDV. Note that the proposed framework is designed to work over this set, while other fragments and metabolites are not measured and their MDVs can only be inferred (if enough data are available).

5.2.2 Carbon atoms mapping

Currently, to perform the tasks of ^{13}C-MFA, models with additional information on how the carbon atoms are transferred from the substrates to the products, in each biochemical reaction, (Figure 5.1) are a valuable resource [Ravikirthi et al., 2011]. These models contain the carbon atom mapping of the biochemical reactions, which consists in a bijection of the

reactant's to the product's carbon atoms [Latendresse et al., 2012], enabling to elucidate how the labeling patterns in the carbon skeleton of the substrates are propagated through all the metabolites on the network. From this carbon mapping, it is possible to determine which carbons are contributing to the labeling of a specific fragment of a certain metabolite, even when the metabolite is synthesized by two or more pathways.

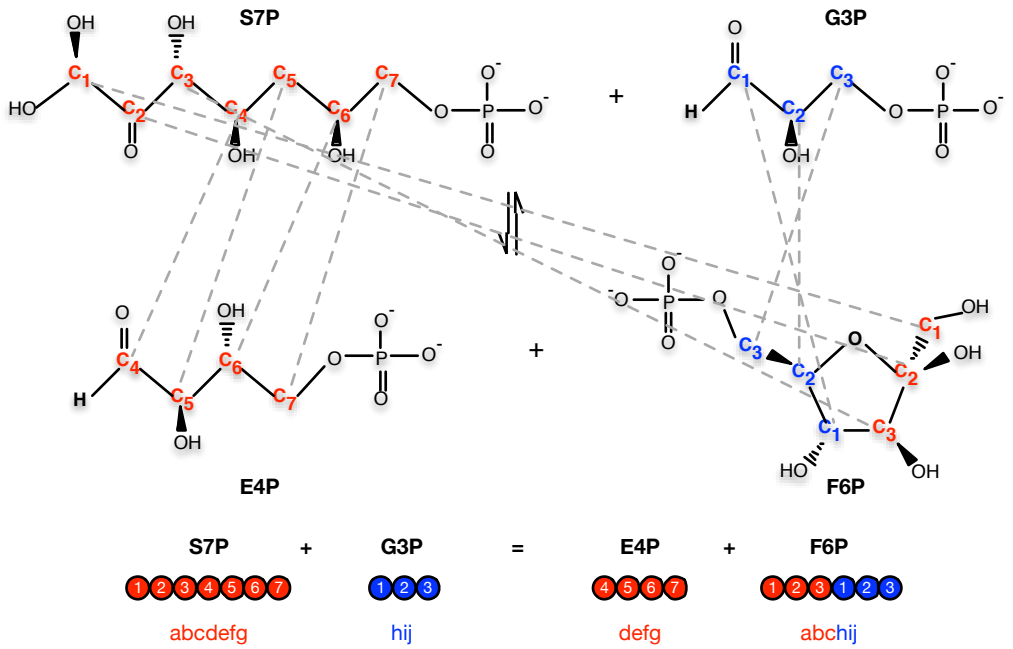


Figure 5.1: Carbon atom transitions between the metabolites in the transaldolase reaction. The circles represent the ordered carbon atoms for each species, and the letters are the convention that is inputted to the algorithms. S7P, sedoheptulose 7-phosphate; G3P, D-glyceraldehyde 3-phosphate; E4P, D-erythrose 4-phosphate; F6P, D-fructose 6-phosphate.

The model enriched with carbon atom mapping can be represented as a graph $\mathcal{N} = (\mathcal{M}, \mathcal{T}, \mathcal{I})$, where $\mathcal{M}$ is the set of nodes associated to the metabolites, $\mathcal{T}$ is the set of edges representing carbon transitions between the reactants and products of the corresponding reactions and $\mathcal{I}$ indicates the set of transitions that are irreversible.

5.2.3 Inferring mass isotopomer distribution vectors

Through the use of $\mathcal{N}$, it is possible to navigate over the network to search for the metabolite fragments in $\mathcal{F}$ that originate a certain target compound (or fragment). This work aims at determining ratios that reflect the contribution of different pathways to the ^{13}C-labeling distribution of a set of metabolite fragments. To do that, each measured MDV in $\mathcal{F}$ is selected as a candidate to be a product of different pathways.

Here, a path p to a product fragment is defined as $\rho_p = (\mathcal{T}_p, \mathcal{F}_p)$ with $\mathcal{T}_p \subseteq \mathcal{T}$ and $\mathcal{F}_p \subseteq \mathcal{F}$, which, through a series of transitions $\mathcal{T}_p$ associated to a set of intermediate metabolites in $\mathcal{M}$, connects a set of measured fragments $\mathcal{F}_p$ to the product (also in $\mathcal{F}$).

Consider $\mathcal{P}$ as the set of paths originating the carbon backbone of a metabolite. For each candidate fragment of that metabolite (Z), if $\mathcal{P} \geq 2$, an expression to infer the MDV of Z can be built using the algorithms that will be explained in the next sections. Since there are at least two alternative pathways contributing to Z, the expression will contain ratio variables that define the fraction of the labeling contribution of each path. From the expression, the MDV of Z is inferred after obtaining the values for the ratio variables. This is compared to the measured one, to adjust the best values for the ratios, which will minimize the difference between the inferred and measured MDVs, as will be detailed next.

5.3 Linking source fragments to products

Through the use of $\mathcal{N}$, it is possible to navigate over the network to search for the metabolite fragments in $\mathcal{F}$ that originate a certain target compound (or fragment) Z. This process will visit every metabolite in $\mathcal{M}$ that is a

vertex in a transition in $\mathcal{T}$ between the matched fragments in $\mathcal{F}$ and Z.

5.3.1 Labeling pattern of combined precursor fragments

In this task, different scenarios may occur. If $\mathcal{P}$ is the set of pathways originating a carbon backbone of a metabolite (Z), the simplest case occurs when $|\mathcal{P}| = 1$, i.e., there is only one converging pathway to Z (ρ_1), and $|\mathcal{F}_1| = 1$ (Figure 5.2 A). Therefore, to trace the carbons that contribute to Z, we just need to directly associate the carbon transfers throughout the series of transitions in ρ_1. The MDV of Z is assumed to be equal to the MDV of the single source fragment.

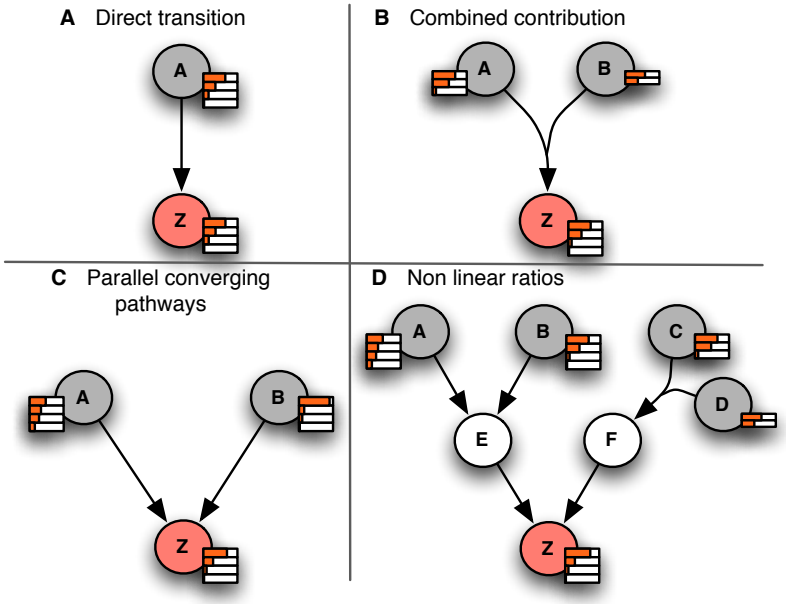


Figure 5.2: Possible scenarios of carbon labeling contribution of the precursors (grey circles) to their products (magenta circles). The arrows represent the carbon atom transitions between the reagents and the products of each reaction, and each orange bar represents the probabilities of the fragment to be labeled in a different number of carbon atoms.

When a single pathway ρ_1 contributes to the labeling pattern of the

product fragment and $|\mathcal{F}_1| > 1$ (i.e. at least one carbon transition in $\mathcal{T}_p$ has more than one substrate), the combination of the labeling pattern of all the fragments of $\mathcal{F}_1$ must be considered (Figure 5.2 B).

For instance, let us consider a fragment of a metabolite Z with three carbon atoms, which originates from a fragment of metabolite A, with two carbons, and a fragment of B, that has a single carbon atom. The four positions of the MDV_A are determined as follows

$$\begin{cases} MDV_Z(0) = MDV_A(0)MDV_B(0) \\ MDV_Z(1) = MDV_A(0)MDV_B(1) + MDV_A(1)MDV_B(0) \\ MDV_Z(2) = MDV_A(2)MDV_B(0) + MDV_A(1)MDV_B(1) \\ MDV_Z(3) = MDV_A(2)MDV_B(1) \end{cases} \tag{5.6}$$

In this case, the probability of Z being not labeled (Z_0) equals to the probability of A being not labeled (A_0) times the probability of B being not labeled (B_0), similarly Z_1 equals to A_1 times B_0 plus A_0 times B_1, Z_2 equals A_2 times B_0 plus A_1 times B_1 and, finally, the probability of Z having all its carbons labeled (Z_3) is equal to A_2 times B_1.

Please note that, in the following expressions, the MDV of a metabolite fragment Z is simply given by the representation of the fragment and $Z(i)$ is used to denote the value of the MDV in the position i, i.e. the probability of Z being labeled at i carbon atoms.

The previous operation can be extended to more than two fragments, and define the MDV of a fragment that is a combination of all the fragments in $\mathcal{F}_1$=$\{S^1, S^2, \ldots, S^n\}$. Generalizing, the MDV of a fragment with m carbon atoms, that is obtained by a combination of two or more fragments, is

a column vector of m+1 positions, where each position k from 0 to m is a value $Z(k)$ obtained as:

$$Z(k) = \sum_{\substack{\text{all combinations} \\ \text{where } \sum_{i=1}^{m} l_i = k}} \left(\prod_{i=1}^{m} S^i(l_i) \right), k = 0, \ldots n \tag{5.7}$$

where l_i is the labeling of S^i in the combination.

5.3.2 Contribution of different pathways to the same product fragment

In the case where $|\mathcal{P}| > 1$, i.e. more than one path originates the target fragment (Figure 5.2 C), the labeling pattern of the carbons of the fragment product Z might be influenced by the fraction τ_i of each converging pathway ρ_i. These fractions define split ratios that are related to the activity of the different pathways with respect to their contributions to the production of the product metabolite. In such scenario, the proportion of each pathway can be calculated through the use of the MDV of the product fragment and the MDVs that represent the sources $\mathcal{F}_i$ of each ρ_i:

$$Z = \sum_{i=1}^{|\mathcal{P}|} \tau_i . MDV_{\rho_i}, \sum_{i=1}^{|\mathcal{P}|} \tau_i = 1 \tag{5.8}$$

where τ_i is the ratio of Z that is derived from the pathway ρ_i that has a measured (or calculated) MDV given by MDV_{ρ_i}.

However, the scenario becomes more complex when, in the search for a branch with a converging ratio, it is not possible to obtain the MDV of a fragment source of the pathway straight from the set of measurements,

i.e., from the direct association of one source to one product fragment, or from the combination operation of the fragments that originate the source (Figure 5.2 D). In such cases, where a split ratio is followed by another ratio in at least one of its branches, the unknown ratios will be multiplied by each other and, therefore, the scenario results in a nonlinear ratios product. For instance, if the fragment from the carbons two to four of the metabolite Z originates from two different pathways, where the source of the pathway from F is obtained from the combination of C and D, and the pathway that comes from E drives the search to another split ratio from the sources A and B, the expression that enables to calculate the MDV of B is given by:

$$Z = \tau_1.\tau_{1.1}.A \ + \ \tau_1.\tau_{1.2}.B + \tau_2.(C \times D) \tag{5.9}$$

where τ_1, τ_2, $\tau_{1.1}$ and $\tau_{1.2}$ are the ratios of the converging pathways to Z, $\tau_1 + \tau_2 = 1$, $\tau_{1.1} + \tau_{1.2} = 1$ and $\times$ represents the operator of the combination of source fragments, defined by expression 5.7 above.

5.3.3 Labeling influence of carbon dioxide

The labeling contribution of certain paths throughout the network may originate from the combination of compound fragments with the carbon atom of CO_2. An example is phosphoenolpyruvate carboxylase, where the carbon of CO_2 is coupled with the three carbons of phosphoenolpyruvate to form the carbon skeleton of oxoloacetate. If the labeling of a product fragment $Z_{1,4}$ is obtained from two routes, where in the first it originates from the fragment $B_{2,5}$, while in the second one of its carbons is obtained from CO_2 and the other three carbons from $A_{1,3}$, the expression to calculate

the MDV of $Z_{1,4}$ would be:

$$Z = \tau.(A \times CO_2) + (1 - \tau).C \tag{5.10}$$

However, the ^{13}C labeling of CO_2 is not determined and may be lower than the input substrate [Fischer and Sauer, 2003]. Therefore, it has to be considered as an additional unknown in the equation above. If the labeling of CO_2 is identified by σ_{co2}, than its MDV can be written as:

$$CO_2 = \begin{bmatrix} 1 - \sigma_{co2} \\ \sigma_{co2} \end{bmatrix} \tag{5.11}$$

So, by applying Equation 5.7 to obtain the combination of the MDVs of $A_{1,3}$ and CO_2, the following matrix is achieved:

$$A_{1,3} \times B = \begin{bmatrix} A_{1,3}(0) - \sigma_{co2}.A_{1,3}(0) \\ A_{1,3}(1) - \sigma_{co2}.A_{1,3}(1) + \sigma_{co2}.A_{1,3}(0) \\ A_{1,3}(2) - \sigma_{co2}.A_{1,3}(2) + \sigma_{co2}.A_{1,3}(1) \\ A_{1,3}(3) - \sigma_{co2}.A_{1,3}(3) + \sigma_{co2}.A_{1,3}(2) \\ \sigma_{co2}.A_{1,3}(3) \end{bmatrix} \tag{5.12}$$

Using Equation 5.12 to replace the combination of $A_{1,3}$ with CO_2 in

Equation 5.10, the ratios equation for $\mathrm{Z}_{1,4}$ is written as:

$$\begin{cases} Z_{1,4}(0) = \tau.\left(A_{1,3}(0) - \sigma_{co2}.A_{1,3}(0)\right) + (1-\tau).C_{2,5}(0) \\ Z_{1,4}(1) = \tau.\left(A_{1,3}(1) + \sigma_{co2}.A_{1,3}(0) - \sigma_{co2}.A_{1,3}(1)\right) + (1-\tau).C_{2,5}(1) \\ Z_{1,4}(2) = \tau.\left(A_{1,3}(2) - \sigma_{co2}.A_{1,3}(2) + \sigma_{co2}.A_{1,3}(1)\right) + (1-\tau).C_{2,5}(2) \\ Z_{1,4}(3) = \tau.\left(A_{1,3}(3) - \sigma_{co2}.A_{1,3}(3) + \sigma_{co2}.A_{1,3}(2)\right) + (1-\tau).C_{2,5}(3) \\ Z_{1,4}(4) = \tau.\sigma_{co2}.A_{1,3}(3) + (1-\tau).C_{2,5}(4) \end{cases} \tag{5.13}$$

Each equation in the system above, represents the expression to obtain the fractional abundance of the mass of the i-th isotopomer of $\mathrm{A}_{1,3}$. Rearranging the equations, the following is obtained:

$$\begin{cases} Z_{1,4}(0) = \tau.\left(A_{1,3}(0) - C_{2,5}(0)\right) + \tau\sigma_{co2}.\left(-A_{1,3}(0)\right) + C_{2,5}(0) \\ Z_{1,4}(1) = \tau.\left(A_{1,3}(1) - C_{2,5}(1)\right) + \tau\sigma_{co2}.\left(A_{1,3}(0) - A_{1,3}(1)\right) + C_{2,5}(1) \\ Z_{1,4}(2) = \tau.\left(A_{1,3}(2) - C_{2,5}(2)\right) + \tau\sigma_{co2}.\left(A_{1,3}(1) - A_{1,3}(2)\right) + C_{2,5}(2) \\ Z_{1,4}(3) = \tau.\left(A_{1,3}(3) - C_{2,5}(3)\right) + \tau\sigma_{co2}.\left(A_{1,3}(2) - A_{1,3}(3)\right) + C_{2,5}(3) \\ Z_{1,4}(4) = \tau.\left(-C_{2,5}(4)\right) + \tau\sigma_{co2}.\left(A_{1,3}(3)\right) + C_{2,5}(4) \end{cases} \tag{5.14}$$

The system of equations is equivalent to the next matrix notation (note that the fragments of the MDVs were omitted to simplify their representa-

tion):

$$\begin{bmatrix} Z(0) \\ Z(1) \\ Z(2) \\ Z(3) \\ Z(4) \end{bmatrix} = \tau. \left(\begin{bmatrix} A(0) \\ A(1) \\ A(2) \\ A(3) \\ 0 \end{bmatrix} - \begin{bmatrix} B(0) \\ B(1) \\ B(2) \\ B(3) \\ B(4) \end{bmatrix} \right) + \tau\sigma_{co2}. \left(\begin{bmatrix} 0 \\ A(0) \\ A(1) \\ A(2) \\ A(2) \end{bmatrix} - \begin{bmatrix} A(0) \\ A(1) \\ A(2) \\ A(3) \\ 0 \end{bmatrix} \right) + \begin{bmatrix} B(0) \\ B(1) \\ B(2) \\ B(3) \\ B(4) \end{bmatrix} \tag{5.15}$$

The Equation 5.15 can be generalized to calculate the MDV of a product fragment (Z) originating from a different number of fragment sources, either combined with CO_2, or not. The general equation can be written as:

$$Z = \sum_{i=1}^{|\mathcal{S}^{co2}|} \left(\tau_i. \left(\begin{bmatrix} S_i \\ 0 \end{bmatrix} - S_n \right) + \sigma_{co2}\tau_i. \left(\begin{bmatrix} 0 \\ S_i \end{bmatrix} - \begin{bmatrix} S_i \\ 0 \end{bmatrix} \right) \right) + \sum_{j=1}^{|\mathcal{S}^{\sim co2}|} \tau_j.(S_j - S_n) + S_n \tag{5.16}$$

where $\mathcal{S}^{co2} \subseteq \mathcal{F}$ is the set of fragment sources that are combined with CO_2, $\mathcal{S}^{\sim co2} \subseteq \mathcal{F}$ the set of sources not combined with CO_2 and S_n is the MDV of the last considered source fragment. If $\mathcal{S}^{\sim co2} = \emptyset$, then $S_n = \emptyset$, on the other hand, If $\mathcal{S}^{co2} = \emptyset$, the equation reduces to the generic equation for calculating flux ratios without CO_2.

5.4 Determining amino acids precursors

Although the central carbon metabolism is highly conserved between different organisms, there are key differences in their constituent pathways. In fact, even the way precursors distribute their carbon atoms among amino acids might change. For instance, for *Saccharomyces cerevisiae* the carbon backbone of lysine originates from different precursor fragments, when comparing to *Escherichia coli*, since its composition is made by bonds between the five atoms arising from 2-oxoglutarate and the second carbon of acetyl-Coa (Figure 5.3).

Thus, it is not realistic to assume that the precursor metabolites of amino acids are alike in all organisms. This is the motivation for the method developed here, that aims at calculating the MDVs of intermediate metabolites that synthesize amino acids in a generic fashion. Based on a set of amino acids fragments, the algorithm uses the carbon atoms graph to determine which metabolite fragments are used to form the amino acids fragments.

When deriving the expressions to calculate the amino acid precursors, the algorithm does not consider converging pathways. Since the labeling content of the amino acids is considered to be derived strictly from its precursors,either in a one-to-one fashion, or from the combination of different fragments, no ratios from different pathways are allowed when determining the vectors of the precursors.

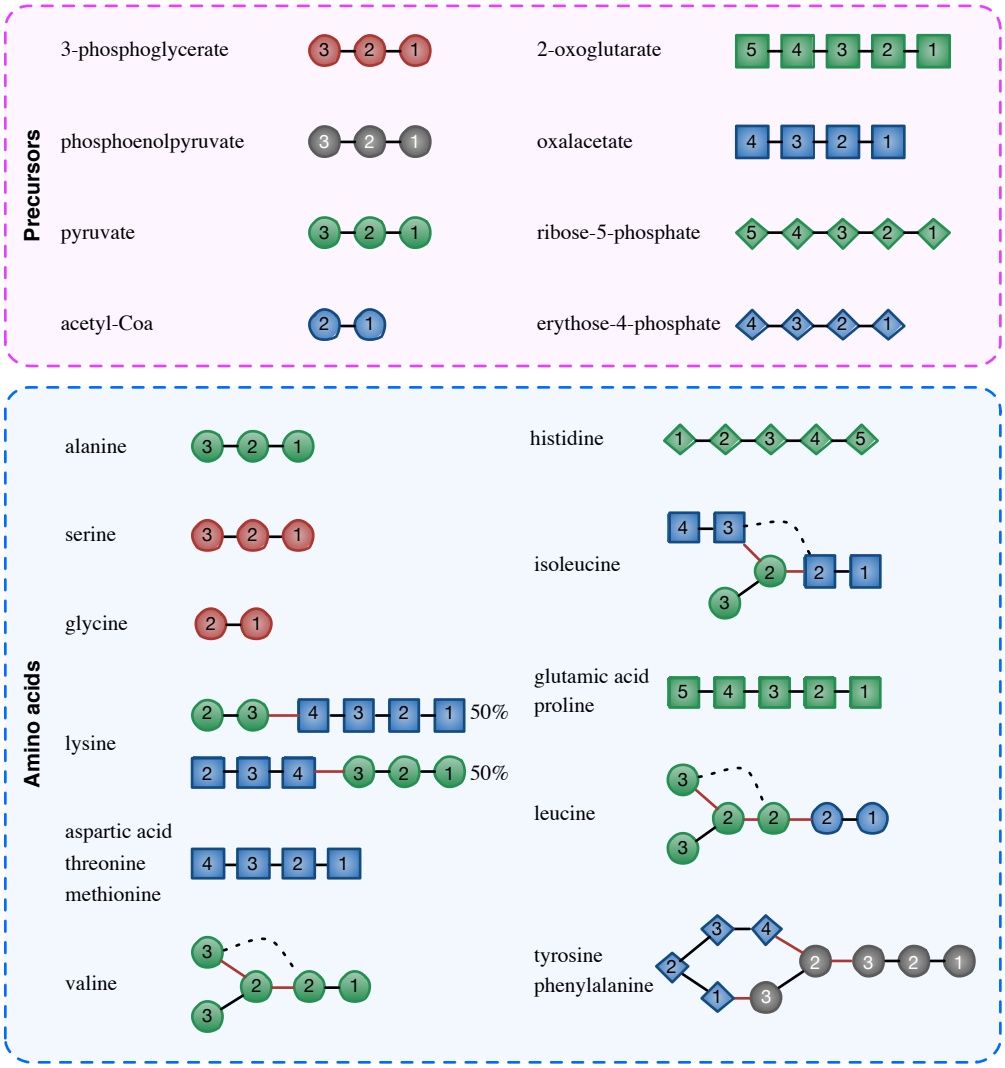


Figure 5.3: Schematization of how the amino acid carbon skeletons derive from their precursors in Escherichia coli. Carbon atoms represented by circles, squares and diamonds, refer to precursors originating from glycolysis, TCA cycle and pentose phosphate pathway, respectively. Dark lines indicate carbon-carbon connections in intact fragments originating from the same molecule. Dashed lines symbolize connections between carbons in the same source molecule, but not directly connected in its carbon skeleton. Red lines represent carbon bonds between fragments originating from different precursors. This image was adapted from [Szyperski, 1995].

5.4.1 Possible scenarios when searching for precursors

Hereupon, without the consideration of converging metabolic flux ratios to intermediate metabolites, the following cases may arise when searching for the amino acids precursors (Figure 5.4):

- 1-to-1 association, the labeling content of the amino acid is equal to its single metabolite precursor. This is the simplest case, where no calculations have to be done and the MDV of the precursors can be assumed to be equal to the measured amino acid MDV. So, in Figure 5.4 a) if the fragment C_2-C_5 of glutamine obtained in the MCF derivatization is measured, then $AKG_{2,5}$ can be obtained as

$$AKG_{2,5} = Glu_{2,5} \tag{5.17}$$

- 2-to-1 association, to form the carbon backbone of an amino acid, two precursors are used. In Figure 5.4 b), the carbons of lysine are obtained from two precursor fragments $PYR_{2,3}$ and $OAA_{1,4}$, so its MDV is obtained from the combination of the MDVs of each precursor:

$$Lys_{1,9} = PYR_{2,3} \times OAA_{1,4} \tag{5.18}$$

- n-to-1 association, the carbon atoms of an amino acid may arise from more than two different fragments. As a matter of fact, the same precursor metabolite can contribute to the amino acid with different fragments through different metabolic routes as it is shown in Figure 5.4 c). The fragment C_1-C_9 of tyrosine is derived from the fragments

C_1-C_3 and C_2-C_3 of phosphoenolpyruvate, and from the entire carbon skeleton of erythrose 4-phosphate, ant its MDV can be obtained according to the expression:

$$Tyr_{1,9} = PEP_{1,3} \times E4P_{1,4} \times PEP_{2,3} \tag{5.19}$$

- 1-to-1 rearrangement, the carbon atoms that build up some intermediate metabolite can be rearranged to form the amino acid fragment. When this happens, the search for the fragments in the metabolite reactant, that originates the product fragment, may not find a consecutive list of carbons in the reactant. So, different non-consecutive fragments of the reactant, in the context of the same atom transitions reaction, are responsible to form the target carbons fragment, and, consequently, are retrieved by the method. For instance, when searching for the precursor of the fragment C_3-C_9 of phenylalanine, it can be noticed that the carbon atoms of chorismate are rearranged to form the phenylalanine fragment in such a way that its two fragments C_2-C_7 and C_10 are used (Figure 5.4 d), and therefore used in the next iterations of the search for the amino acid precursors. The expression that translates how the fragment $Phe_{3,9}$ is derived from its precursors is written as:

$$Phe_{3,9} = PEP_3 \times E4P_{1,4} \times PEPYR_{2,3} \tag{5.20}$$

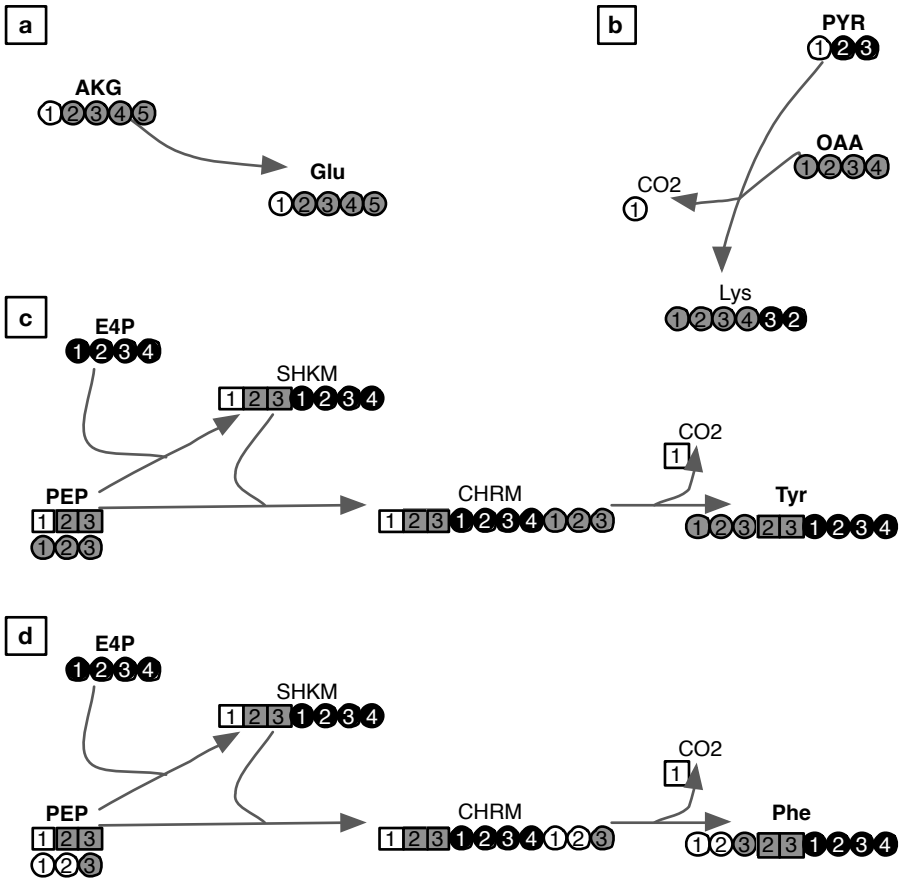


Figure 5.4: Examples of how carbon atoms can be transferred between intermediate metabolites to amino acids. a) Simplest scenario, where all carbons originate from the same precursor; b) The carbon skeleton of an amino acid can be obtained from the combination of two precursors. c) The same metabolite can contribute to the carbons composition of an amino acid with different fragments, through different routes. d) The carbon atoms that configures an intermediate metabolite can be rearranged to originate the amino acid fragment. AKG, α-Ketoglutarate ; Glu, Glutamine; Pyr, Pyruvate; OAA, Oxaloacetate; Lys, Lysine; E4P, Erythrose 4-phosphate; PEP, Phosphoenolpyruvate; SHKM, Shikimate; CHRM, Chorismate; CO2, Carbon dioxide; Tyr, Tyrosine; Phe, Phenylalanine.

5.4.2 Meta-heuristics to compute precursors mass isotopomer distribution vectors

In the cases where MDVs of the precursors need to be computed, an optimization approach can be used to solve the non linear equations that are obtained for each mass isotopomer of the target amino acid. Here, global optimization methods, such as Evolutionary Algorithm (EA), have been implemented.

For all developed meta-heuristic methods, the solution is a vector, where each of its positions corresponds to a mass isotopomer of the considered precursors' fragments. For the case c) in Figure 5.4, the solution is represented as follows:

$$\begin{bmatrix} PEP_{1,3}(0) \\ PEP_{1,3}(1) \\ PEP_{1,3}(2) \\ PEP_{1,3}(3) \\ E4P_{1,4}(0) \\ E4P_{1,4}(1) \\ E4P_{1,4}(2) \\ E4P_{1,4}(3) \\ E4P_{1,4}(4) \\ PEP_{2,3}(0) \\ PEP_{2,3}(1) \\ PEP_{2,3}(2) \end{bmatrix}, \sum_{i=0}^{3} PEP_{1,3}(i) = 1, \sum_{i=0}^{4} E4P_{1,4}(i) = 1, \sum_{i=0}^{2} PEP_{2,3}(i) = 1 \tag{5.21}$$

The methods search for solutions that minimize the difference between the calculated and the measured MDVs of the target amino acid, through the application of Equation 5.7. The problem is constrained in such a way that, for each precursor fragment, the sum of the generated mass isotopomers is equal to 1. Therefore, the optimization problem can be formulated as:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{s} \left(\frac{MDV_{c,aa}(i) - MDV_{m,aa}(i)}{MDV_{m,aa}(i)} \right)^2 \\ \text{subject to:} \quad & MDV_{c,aa}(k) = \sum_{\text{all combinations}} \left(\prod_{i=1}^{m} S^i(l_i) \right), k = 0, \ldots s \\ & \text{where } \sum_{i=1}^{m} l_i = k \\ & \sum_{j=1}^{s_i} S^i(j) = 1 \ , \ \ \forall \, S^i \in \mathcal{B} \end{aligned}$$

where $MDV_{c,aa}$ and $MDV_{m,aa}$ are the calculated and measured MDVs of the target amino acid, respectively, and s is the length of the vectors . The calculated vector is obtained by solving Equation 5.7 for the set of precursors' fragments ($\mathcal{B}$) of the amino acid and s_i is the length of each precursor S^i. The first constraint is the application of Equation 5.7 to the amino acid fragment.

5.5 Calculating ratio expressions

Taking the expressions stated in above sections, it is possible to combine them and build the equations that reflect how the MDV of a product is influenced from the MDVs of its precursors in the general case. Each fragment

in $\mathcal{F}$ is selected as a candidate product for a ratio expression. Therefore, when a product is selected, the process starts and searches on $\mathcal{N}$ for the fragments in $\mathcal{F}$ that can contribute to the labeling pattern of the product.

In Figure 5.5 A, there is a small excerpt of a network that is useful to illustrate how the procedure works. In this example, the fragment C_2-C_5 of metabolite Z is selected as a product and the fragments $A_{2,5}$, $B_{2,3}$, $C_{1,2}$, $D_{1,4}$ $\in \mathcal{F}$.

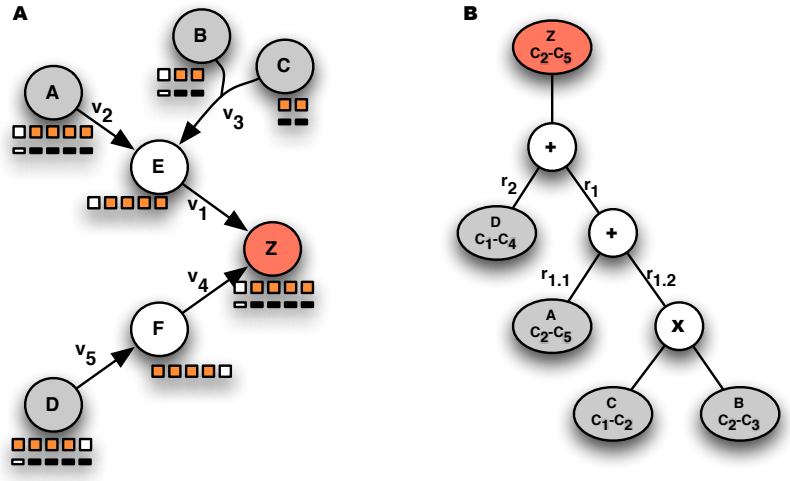


Figure 5.5: Small excerpt of a network (A) and the tree-shaped representation (B) of the expression that enables to calculate the MDV of the product fragment $Z_{2,5}$. In (A), the graph enables to trace the labeling contribution for the fragment $Z_{2,5}$, which is obtained from the fragment sources $D_{1,4}$, $A_{2,5}$, $C_{1,2}$ and $B_{2,3}$. The orange squares stand for the contiguous fragments that are visited during the search process, and the sequence of white / black rectangles identifies the measured fragments of the metabolites. In (B), the expression to calculate the MDV of the root node is depicted, from the MDV of the fragment sources on the leafs of the tree. The ratios are represented in the edges under the "+" nodes.

Starting from the product Z, the search discovers a fork that leads to the ratios of two pathways ρ_1 and ρ_2, from the transitions on v_1 and v_4, respectively. Afterwards, when searching for the carbons that originate the fragment C_2-C_5 of E, another fork is found guiding the search to the paths ρ_{11}, from v_2, and ρ_{12}, from v_3. The set $\mathcal{F}_{11}$ contains only the fragment $A_{2,5}$

, while the path ρ_{12} leads to the four carbon fragment obtained from the combination of the elements $B_{2,3}$ and $C_{1,2}$.

Finally, the branch corresponding to ρ_2, drives the search to the fragment $D_{1,4}$. From this procedure, a tree that organizes the structure of the operations that must be performed to obtain the product fragment is built (Figure 5.5 B), which, when used to solve the MDV of $Z_{1,4}$, will be translated in the expression:

$$Z_{2,5} = \tau_1.\left(\tau_{1.1}.A_{2-5} + \tau_{1.2}.(B_{2,3} \times C_{1,2})\right) + \tau_2.D_{1,4} \tag{5.22}$$

Please note that in this expression and the following, the MDV of a metabolite fragment Z is simply given by the representation of the fragment in both sides of the expression, e.g. the MDV of $Z_{2,5}$ is simply represented by $Z_{2,5}$. Also, the symbols defined in previous expressions will not be explained further to improve readability.

5.5.1 Recurrent sub-expressions

As it has been shown during the procedure to search for the expression that calculates the MDV of the product fragment in the previous example, the sources are not always found in the nodes that are immediately precedents to the product. Indeed, almost all the expressions are built upon pathways where $\mid \mathcal{T}_p \mid \gg 1$. In each iteration, every fragment that is a reactant in each transition that originates the current node of the search, i.e. an intermediate fragment, will be recursively selected as a destination node in succeeding iterations. This process can be seen as a search for the sub-expressions of the intermediates that form the global expression to the original product fragment.

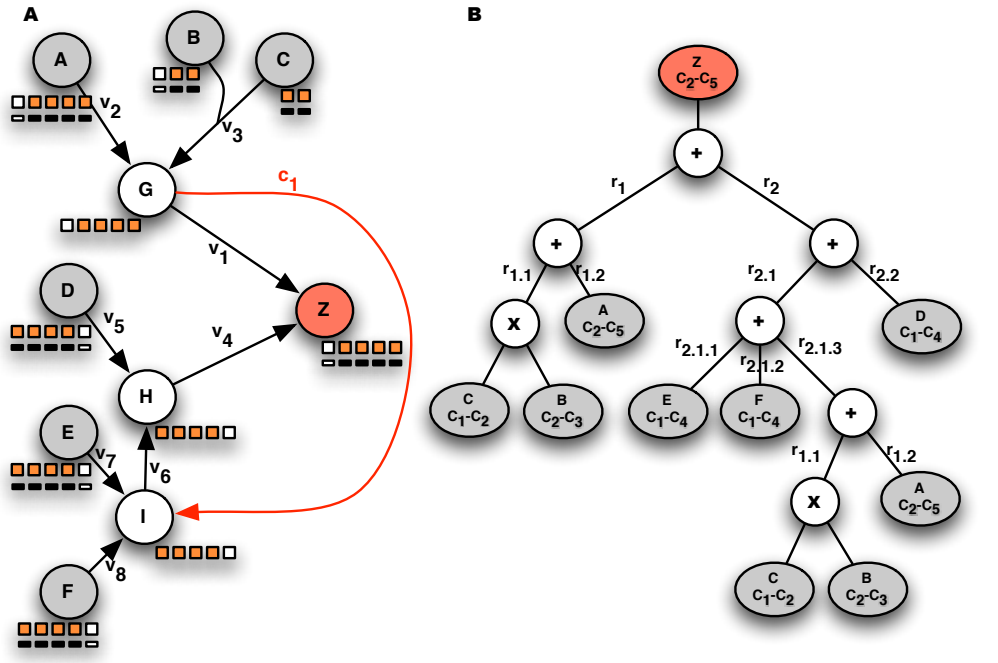


Figure 5.6: Pseudo network originating recurrent ratios sub-expressions. In highly connected graphs, the occurrence of hub nodes is common, which are vertices of many edges. In (A) the carbons C_2-C_5 are transferred to different metabolites through two different reactions (v_1 and c_1). The red arrow indicates a reaction that originates a repeated search for the same fragment of G. In (B) the fragment $Z_{2,5}$ is calculated from the operations represented on the tree, where the source fragments are the grey nodes (the leafs of the tree). Z is obtained from the fragment sources $D_{1,4}$, $A_{2,5}$, $C_{1,2}$ and $B_{2,3}$.

Metabolic networks are complex and highly connected, so the carbon atoms transitions between given metabolites are also complex. The same carbon skeleton can contribute for different metabolites of the network, and this carbon skeleton can even contribute to the same fragment from alternative pathways, as illustrated in Figure 5.6 A. Therefore, when searching for the expression that enables to obtain a fragment from its sources, it might happen to find metabolites,for which the expressions for its fragments had been previously derived. In such cases, if the expression has been already found, it is not necessary to perform the procedure all over again. Thus, a register that keeps the history of the searches is kept and updated whenever

a new fragment is found and its expression is built. If a fragment has been already visited in some previous iteration of the current search, or in the search for a different product of the input fragments set, the expression of the fragment is retrieved from the register and attached to the expression of the selected product. In Figure 5.6 B, it is possible to notice that the sub-expression, corresponding to the determination of the labeling of fragment $G_{2,5}$ is repeated in the tree.

Moreover, as it is shown, the identifiers of the ratios of the sub-expression are the same in the different branches of the tree, since the source fragments and the product of the sub-expression are the same.

Hence, the expression to determine the MDV of $Z_{2,5}$ can be written as:

$$\begin{cases} Z_{2,5} = \tau_1.G_{2,5} + \tau_2.\left(\tau_{2.2}.D_{1,4} + \tau_{2.1}.\left(\tau_{2.1.1}.E_{1,4} + \tau_{2.1.2}.F_{1,4} + \tau_{2.1.3}.G_{2,5}\right)\right) \\ G_{2,5} = \left(\tau_{1.1}.(B_{2,3} \times C_{1,2}) + \tau_{1.2}.A_{2,5}\right) \end{cases} \tag{5.23}$$

5.5.2 Undetermined ratios

The expressions obtained by the procedure are totally dependent on $\mathcal{F}$. Therefore, there is no guarantee that the sources of a given branch will be found in the network. When searching for a branch, if there are no measurements that enable to calculate the vectors that originate all the metabolites that are reached, it will not be possible to find all the sources of the labeling contributions from that branch.

To guarantee that the process terminates, a stopping criteria is defined, which can be seen as a radius (R_x) of the search procedure, allowing a maximum number of nodes to be visited in each path originating the product

fragment (Figure 5.7). When the limit is reached, it is assumed that it is not possible to find sources belonging to $\mathcal{F}$ on the path. In such scenario, the corresponding part of the expression is not conclusive, and it cannot be used to calculate the MDV of the selected product.

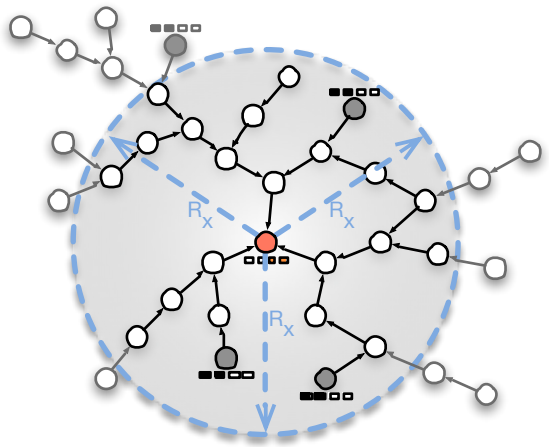


Figure 5.7: Radius of the search procedure. The radius defines the maximum number of visited fragments for each path originating the product fragment. If the radius is reached, then the search process returns an unknown fragment for the current visiting path.

However, it should not be assumed that there is no influence of that pathway on the labeling pattern of the product. Hence, one can either dismiss the expression, if the degree of uncertainty is too high (e.g. too many unknown sources), or establish the expression to estimate bounds for the ratios, translating that the labeling pattern of the product is partially influenced from the pathways in which it is possible to find sources belonging to $\mathcal{F}$.

In Figure 5.8 A, this case is illustrated, where the path that originates the product from reaction v_5 drives the search to an unknown source with respect to the fragment C_1-C_4 of metabolite H, where the process never reaches a fragment of $\mathcal{F}$. Therefore, the contribution of the pathway is compromised by this missing vector and the MDV of Z can not be accu-

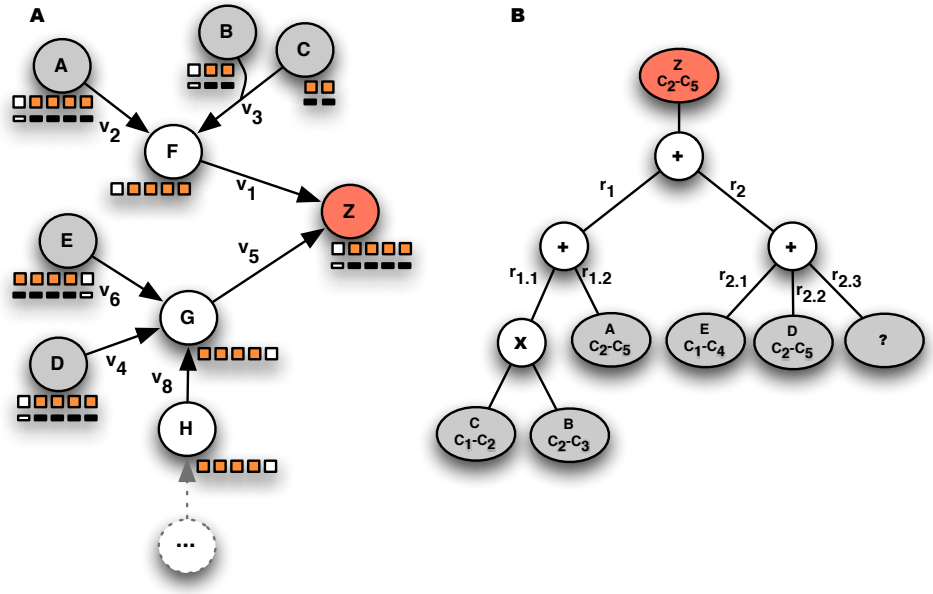


Figure 5.8: Illustration of a pseudo network with undetermined ratios. Since the set of fragments with associated MDVs changes depending on the measured data, it might happen that the measurements are not enough to determine all the sources from the different pathways that originate a certain product. In (A), the node with "$\cdots$" illustrates a branch of the network on which it is not possible to find measured sources that originate $Z_{2,5}$. In (B), the leaf with the question mark refers to the branch with no sources found.

rately calculated. However, if the branch originating from the flux v_8 is not accounted for in the expression, it can be written as

$$Z_{2,5} = \tau_1.\left(\tau_{1.1}.(B_{2,3} \times C_{1,2}) + \tau_{1.2}.A_{2,5}\right) + \tau_2.\left(\tau^*_{2.1}.E_{1,4} + \tau^*_{2.2}.D_{2,5}\right) \quad (5.24)$$

where $\tau^*_{2.1}$ and $\tau^*_{2.2}$ are the upper bounds of the ratios to $Z_{2,5}$ originating from v_6 and v_4, respectively.

If ρ_ϕ stands for a pathway where $\mathcal{F}_\phi \not\subseteq \mathcal{F}$, and τ_ϕ relates to the ratio from ρ_ϕ in the expression tree, all the ratios referring to the neighbour nodes of ϕ are assumed to define bounds to the labeling contribution of the fragment sources of the corresponding paths. Although the values for all

the other branches of the expressions are also calculated, they will not be used to build flux ratio constraints.

Depending on the value of the MDV that would translate ϕ, the values for the ratios immediately above its level could either increase or decrease. Consequently, other ratios on the expressions' tree would also be changed. The neighbour ratios of τ_ϕ, are considered as upper bounds, since their values can decrease, or be maintained, if no labeling contribution arises from ρ_ϕ. Although the ratios above the upper bound ratios could increase their values, if there was a high contribution of the missing path, the upper bound ratios would also decrease according to this high contribution of the other path. Therefore, the product between the upper bound ratios with all the ratios above them, until the root node of the tree, would always be less than the product of their estimated values.

When $\mathcal{F}$ lacks in measured fragments, the procedure to build the expressions can return more than one unknown node. In such scenario, if the number of unknowns is too high, the estimation procedure will not be able to find reliable values for the ratio variables. Moreover, even if the number of undetermined branches is not too high, the relative position between the unknown nodes affects how the expressions could be solved.

In the cases where the unknown nodes belong to separate branches of the trees (Figure 5.9), if values would be given for the missing MDVs, the ratios could be adjusted differently according to the values, so the estimation for all the ratio variables of the expressions would not be consistent. Thus, when more than one ratio is unknown under a certain node of the tree, the sub-tree under the node is removed, and the node is replaced by an unknown node. If the reduced tree does not contain more than one ratio not relating to unknown nodes, the expression cannot be solved, since it is no longer a

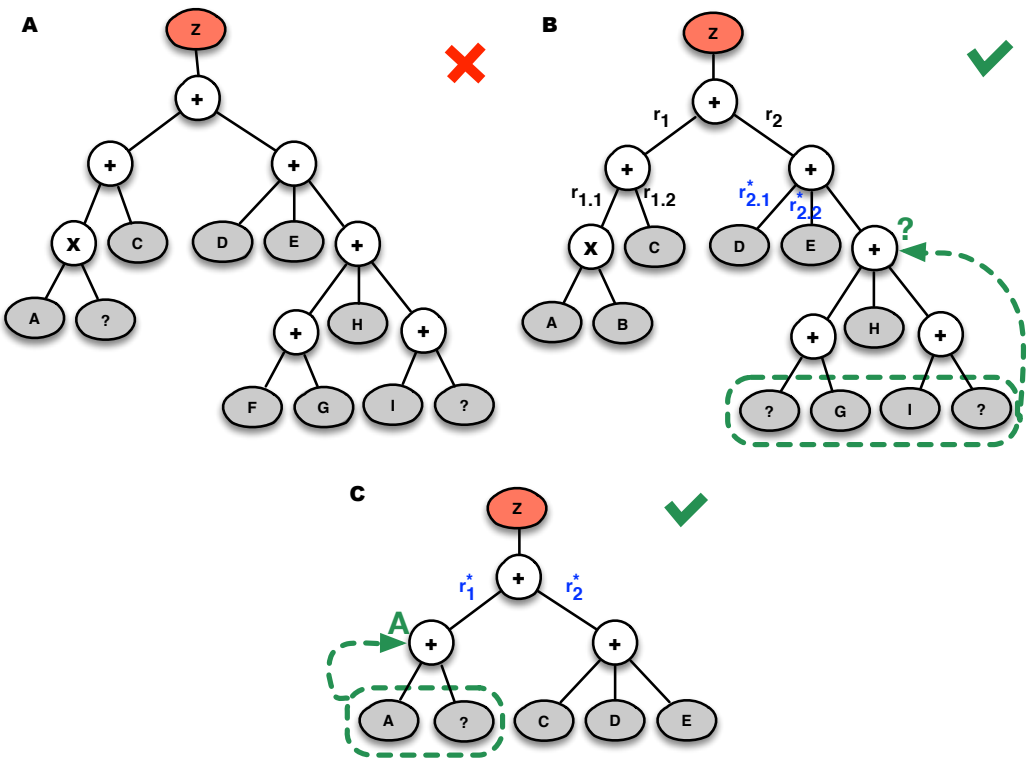


Figure 5.9: Substitution of expression sub-trees by unknown nodes. When more than one path does not contain all its sources measured, the node immediately above the sub-tree containing the corresponding unknown nodes is removed and replaced by a new unknown node. In (A) the replacement originates an expression tree without flux ratios, so it will not be allowed to be solved; (B) the sub-tree with the unknown nodes is replaced, and the expression can be solved to obtain the upper bounds for the ratios $\tau^*_{2.1}$ and $\tau^*_{2.2}$; (C) when a sub-tree with an unknown node contains only one more node, its parent node is replaced by this measured node and the ratios τ^*_1 and τ^*_2 will refer to a lower and upper bounds, respectively.

flux ratios expression. Otherwise, it can be solved and the ratios for the branches around the unknown branch will be assumed as upper bounds.

On the other hand, when a parent node (Ω) of a sub-tree with an unknown node contains only one other child node $A \in \mathcal{F}$, than the sub-tree is removed and $\Omega$is replaced by A. In such cases, the ratio from A in the new tree will be defined as a lower bound since the contribution of the unknown path could increase the contribution of the sub-tree of Ω, while the other ratios in the same level of the tree will be converted to upper bounds.

5.5.3 Carbon mapping cycles

Occasionally, when searching for a fragment of a carbon skeleton, it might happen to come across with a cycle in the involved carbon transitions. This kind of cycles happens, when a specific fragment of a certain metabolite is being searched for, and no fragment that belongs to $\mathcal{F}$ has been found, before the same fragment is reached on a further iteration. In such case, it is clear that it will not be possible to determine the expression (or a portion of the total expression) due to missing fragment sources.

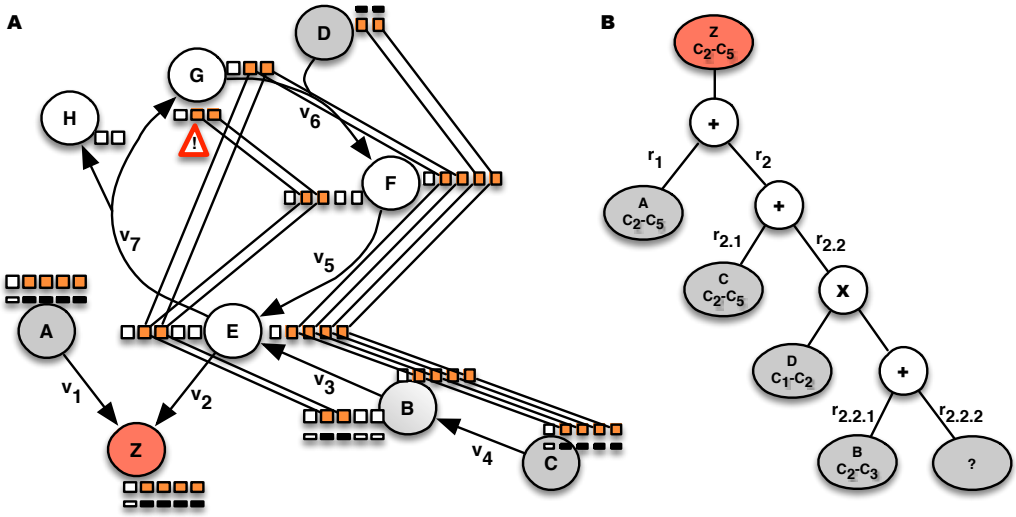


Figure 5.10: Illustration of a network with carbon mapping cycles. When searching for a specific fragment, it might happen that a fragment is reached twice without finding the sources that originated it. In (A) a cycle is illustrated where the fragment $G_{2,5}$ is visited twice, without being possible to determine its sub-expression. The thin lines connect the carbon atoms that are being transferred between the fragments being searched in the corresponding reaction. in (B) the question mark node indicates the unknown sub-expression to $G_{2,5}$.

However, a cycle is not identified if the process reaches a metabolite that has been previously visited. The stopping criterion is defined in terms of repeated fragments, since in different iterations upon the same metabolite, distinct fragments of this metabolite might be considered, due to the

partitioning of the fragments throughout the different carbon fragments in the network.

Figure 5.10 A exemplifies a search for an expression to calculate the MDV of the fragment C_2-C_5 of metabolite Z. To obtain the part of the expression that originates from the pathways that start on reaction v_5, the search is driven to the fragment C_2-C_5 of F which is obtained from the fragment C_2-C_3 of G and the fragment C_1-C_2 of D. $D_{1,2} \in \mathcal{F}$, but on the other branch, the search proceeds to the fragment of G. As it is illustrated in the transitions map, the metabolite E is visited twice, however for a different fragment (C_2-C_3) of its carbon skeleton. When searching for this fragment on the pathway originating from v_3, the search stops on metabolite B, as its fragment C_2-C_3 belongs to $\mathcal{F}$ (it did not happen for the fragment $B_{2,5}$, where the search finished only in the source $C_{2,5}$). Nevertheless, the search through the pathway from v_5 proceeds, visiting, once again, the metabolites F and G. As it is shown, the procedure does not stop either on E or F, because the fragments in each iteration are different from the visited ones. However, the fragment $G_{2,3}$ is used again to form a different fragment of F and, consequently, there is the occurrence of a cycle with respect to this fragment. It is not possible to determine, from the fragments in $\mathcal{F}$, its sources. The expression tree (Figure 5.10 B) that is obtained to this case is written as follows:

$$Z_{2,5} = \tau_1.A_{2,5} + \tau_2.(\tau^*_{2.1}.C_{2,5} + \tau^*_{2.2}.(D_{1,2} \times B_{2,3})) \tag{5.25}$$

where $\tau^*_{2.1}$ represents the upper bound for the ratio $\tau_{2.1}$ and $\tau^*_{2.2}$ the lower bound of $\tau_{2.2}$.

5.5.4 Reversible reactions

When the model contains information about the reversibility of the reactions with the carbon atom transitions, the ones that are reversible are decomposed on two irreversible reactions in each direction, i.e., each reaction $v_i \notin \mathcal{I}$ is duplicated and the carbons transitions are converted to its opposite direction. The following example, where the first reaction is the original reversible reaction, and the second ($v_{1_{inv}}$) is its conversion:

$$\begin{aligned} v_1 &: A_{abcd} + B_{efg} = C_{dcb} + D_{efga} \\ v_{1_{inv}} &: C_{abc} + D_{defg} = A_{gcba} + B_{def} \end{aligned} \tag{5.26}$$

The lowercase characters indicate the indexing of the carbon atoms that belong to the carbon skeleton of the precursors, and are transferred to the products through the corresponding reaction.

However, a reaction cannot contribute simultaneously, on both directions, to the carbon labeling of the metabolites that are involved in its pathways. Accordingly, with no insights on which direction a certain reversible reaction is taking place for a physiological context, if every possibility, in terms of alternative directions of the reversible reactions, is taken into account, many different solutions could be returned from the procedure. When a set of reversible reactions is present in the pathways that originates the selected product, the problem would be combinatorial, and a large amount of possible solutions could be found, depending on the number of reversible reactions that were considered during the configuration of the expression.

For instance, in Figure 5.11 A, $v_3 \notin \mathcal{I}$ and the carbon transitions can

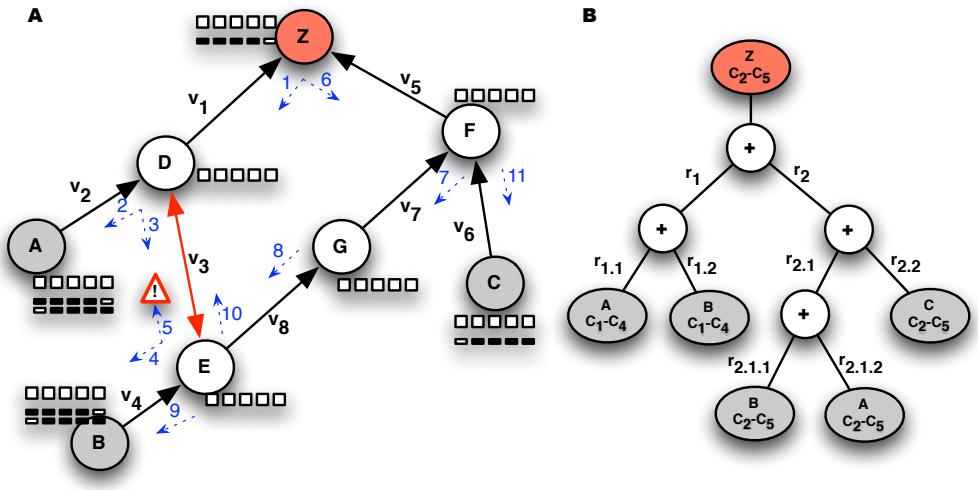


Figure 5.11: Reversible reactions might originate several expressions for different assumptions. In (A), blue arrows indicate the direction and the order taken in the process, with no previous assumption on the direction of the reversible reaction (the red arrow). The colored carbons were omitted to simplify the diagram, since the carbon fragments addressed to some metabolites vary during the iterations of the process. In (B) the general expression is illustrated without disregarding any direction of v_3.

occur on either direction. Thus, if the procedure progresses in the direction of the blue arrows, and follows its order, initially v_3 will be visited to search for the sources that originate $Z_{1,4}$. Afterwards, the process will proceed to find the sources of $E_{1,4}$ and the reverse way of v_3 will be select as a reaction that originates the fragment. However, since the reaction v_3 has been already selected on the direction from E to D to achieve this point, it does not make sense to select its reverse direction to proceed with the search for that branch. Therefore, in similar situations, the reverse direction of a reaction that was previously selected to be active on the opposite direction, for the same branch of the search, is not considered.

Nevertheless, when the search is performed to find the sources to $Z_{1,4}$, across the pathway originating from v_5, the reaction v_3 will be reached again, but on a different direction of the kept before, since at this point the search is made for the sources of the fragment $E_{2,5}$. If the correct direction

of v_3 were known to be from E to D, then the search should stop when visiting the reaction to find the sources of E, since no labeling contribution would happen from D to E. But, if the correct direction was the other way around, the expression should be totally different. Moreover, if an assumption is made that the first visiting direction is the one chosen to be fixed, when the order on the selection of the branches in which the products are derived is inverted, the obtained expression could also be different. It is important to notice that both directions cannot be considered, only during the search of the same branch of the expression, and the exclusivity of the directions of a network can be taken only in a local context.

Therefore, instead of generating the full set of possibilities of expressions that can be obtained for a given product, when reversible reactions contribute to the carbon skeleton of the product, no filter or stop criteria is defined in terms of the reversibilities of the reactions, i.e., the two possible directions of the reactions are interpreted as two different reactions as shown in Figure 5.11 B. In order to fix the expression and solve it, the proper direction of the reactions must be selected, either from the experimental measurements or from knowledge on the physiological conditions of the problem. Subsequently, the excerpts of the expression that are derived from the directions that should not be addressed have to be removed from the expression, by selecting the corresponding ratios to be equal to zero, meaning that there is no labeling contribution though the corresponding pathways.

5.6 Determination of metabolic flux ratios

After obtaining the expressions that enable to calculate the MDVs for the fragments in $\mathcal{F}$, a set of unknowns is retrieved. These unknowns consist

of the flux ratios that are detected when different pathways are converging to a certain metabolite. To compute these unknowns, two scenarios can occur. The simplest case is when the expression to calculate a product fragment is linear. In the other case, the expressions are not linear due to the occurrence of products between different flux ratios.

5.6.1 Solving linear ratios expressions

The former relates to cases where, when searching for the nodes that originate the product fragment, the method determines that the labeling contribution for these nodes originate from a single source or is a combination of the labeling of more than one source. Therefore, the tree representation for this type of expressions will have only one level, with no need to multiply ratios to calculate the MDV of the product. In such cases, it is possible to calculate the flux ratios by rearranging equation 5.8. Since $\tau_m = 1 - \tau_1 - \tau_2 - \ldots - \tau_{m-1}$, the following system for the first $m-1$ flux ratios can be written as:

$$\begin{bmatrix} MDV_1 - MDV_m \\ MDV_2 - MDV_m \\ \cdots \\ MDV_{m-1} - MDV_m \end{bmatrix} \begin{bmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_{m-1} \end{bmatrix} = \begin{bmatrix} MDV_Z - MDV_m \end{bmatrix} \tag{5.27}$$

where $m = |\mathcal{P}|$, the left-hand side is an $n \times m-1$ matrix A covering the subtraction of the MDVs of the first $m-1$ sources by the source m, τ is a vector of the first $m-1$ ratios, and the right-hand side $n \times 1$ vector (b) refers to the subtraction of the MDV of the product and the source n.

To obtain the values for the ratios in τ, QR decomposition is used, which corresponds to the least squares solution of the system of the simultaneous linear equations. The decomposition finds the vector τ that minimizes the two norm of $QR\tau - b$, where Q is an n x $(m-1)$ orthogonal matrix and R an $(m-1)$ x $(m-1)$ upper triangular matrix, so that $A = QR$.

5.6.2 Solving non linear ratios expressions

On the other hand, if the search for the sources of a product originates inner ratios, due to the occurrence of different pathways leading to an intermediate node that does not belong to $\mathcal{F}$, the expression to calculate the MDV of the product will be non linear. Here, to solve the problem that tries to find the vector τ that minimizes the difference between the calculated MDV of the product and its measured MDV, a classical gradient-based method will not adopted. Instead, a global optimization meta-heuristic, such as Differential Evolution (DE), EA or Simulated Annealing (SA), will be configured to solve the optimization problem formulated below, by iteratively trying to improve a candidate solution with regard to the objective function and constraints of the problem, defined as follows:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{s_i} \left(\frac{MDV_{c,i}(j) - MDV_{m,i}(j)}{MDV_{m,i}(j)} \right)^2 \\ \text{subject to:} \quad & \textit{all expressions to each } MDV_{c,i} \\ & \sum_{j=1}^{s_i} MDV_{c,i}(j) = 1 \ , \ \forall \, MDV_i \in \mathcal{D} \\ & \sum_{x=1}^{s_l} \tau_{l,x} = 1 \ , \ \forall \, l \in \mathcal{R} \end{aligned}$$

where $MDV_{c,i}(j)$ and $MDV_{m,i}(j)$ are the calculated and measured fractional abundances of the m_{+j} mass of the fragment i, respectively, $\mathcal{D} \subset \mathcal{F}$ the products for which the meta-heuristic is being solved, s_i is the length of MDV_i and s_l is the number of elements in the subset l in $\mathcal{R} = \{\{\tau_{1,1}, \cdots, \tau_{1,s_1}\}, \cdots, \{\tau_{l,1}, \cdots, \tau_{l,s_l}\}\}$, that encompasses all the subsets of ratios of pathways originating the same intermediate fragment. The values of $m_{_{c,i}+j}$ are calculated from the obtained expressions to the corresponding products, after the assignment of the flux ratios.

In Figure 5.12, we illustrate how a candidate solution is encoded and used to solve the problem. A solution is a vector that encodes real values corresponding to the ratios in the expressions. As it is shown, although the expressions refer to different product fragments, since the expressions share some flux ratios, all their ratios are encoded in the same solution. Hence, the size of a solution is not strictly equal to the number of ratios that are present in the expressions. In fact, when some ratios are shared, all the expressions that use these ratios will contribute to its computation, making its estimation more accurate. On the other hand, expressions that do not have intersections among their sets of ratios are not encoded and solved together with each other. Thus, for each independent expression, or group of expressions, the problem above is formulated and solved independently, meaning that each solution used for the calculation of the ratios is local to the corresponding expression, and the ratios are not globally calculated in a single solution.

In each iteration of the method, a candidate solution is evaluated by assigning the values encoded in the solution to the corresponding unknowns in the expressions. Afterwards, the expressions are solved and the calculated MDVs of the products are compared with the measured ones. Since the

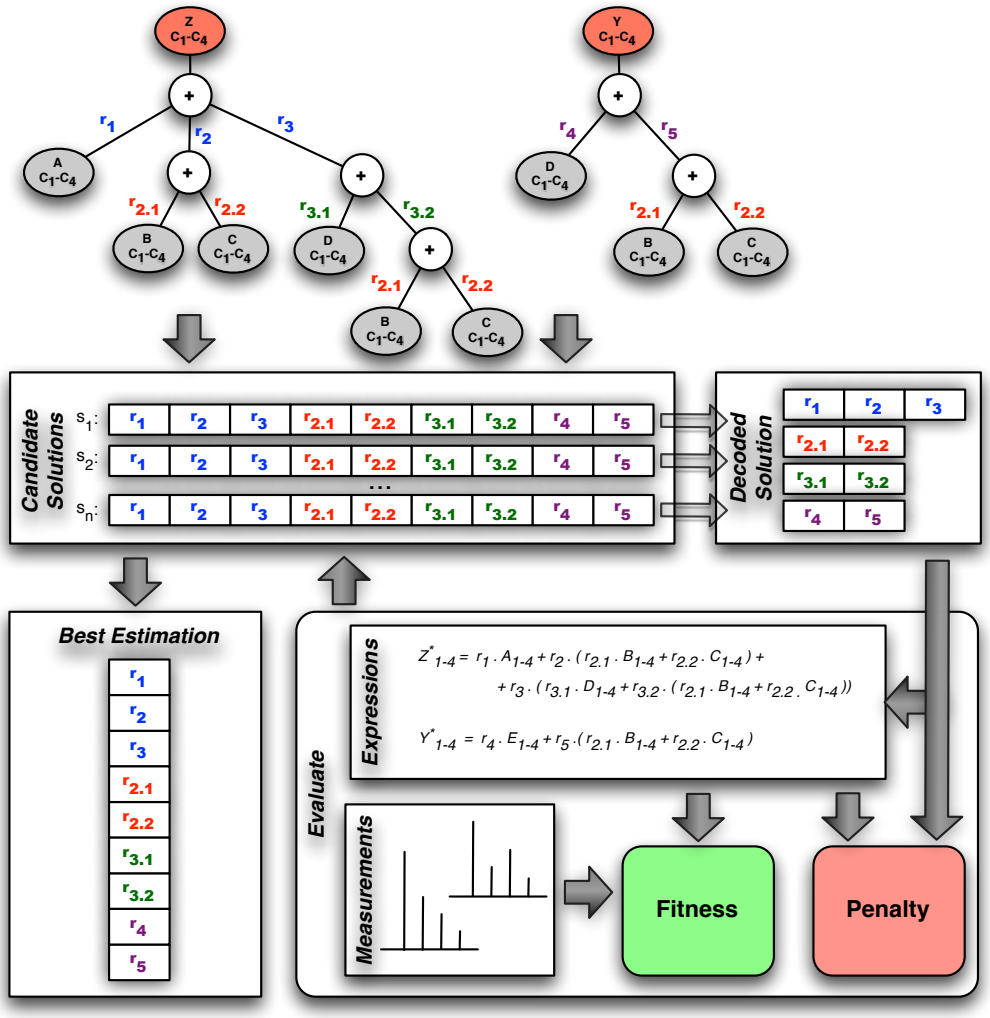


Figure 5.12: Illustration of a meta-heuristic method to estimate metabolic flux ratios. When the expression to calculate the MDV of the product fragments is nonlinear, this method is used to estimate the ratios in the expressions. The ratios of the expressions (that share ratios with the same identifier) are encoded in each candidate solution. In each generation of the iterative method, new candidate solutions are generated and evaluated by comparing the calculated and measured MDVs (e.g. of $Z_{1,4}$ and $Y_{1,4}$) and computing penalties. The best estimate is obtained from the solution that originates the best relation between the defined fitness and penalty functions.

objective function is to minimize this difference, a fitness value to each candidate solution will be proportional to the squared errors of this estimate. To guarantee that the method respects these constraints of the problem, a penalty function is configured to punish solutions that originate a calculated MDV, in which the sum of its elements is not equal to one, and also to penalize solutions that contain groups of ratios (ratios in the same level, corresponding to different pathways leading to the same intermediate node), where the sum is different from one. Therefore, a metric is defined to weigh the evaluation of the candidate solutions with respect to its fitness and penalty values.

5.6.3 Estimating mass isotopomer distribution vectors for unknown paths

Another strategy, when it comes to solve expressions that contain unknown branches, is to try to estimate the MDVs of a branch from an undetermined fragment source. Thus, instead of computing ratio bounds with respect to the missing compound fragment measurements, the method tries to estimate the MDVs of the missing paths that best approximate the calculated MDV of the product fragment to the measured one.

As in other methods mentioned in this document, meta-heuristics for global optimization are formulated to generate solutions for both the ratios and the mass abundances (Figure 5.13). In the figure, the branch identified by the fragment $\phi_{1,4}$ in the expressions' trees corresponds to a path for which no fragment sources had been found. Therefore, there is no measured MDV reflecting the influence of that path in the labeling pattern of the product fragments $Z_{1,4}$ and $Y_{1,4}$. Since the unknown fragment influences the ^{13}C pattern of two distinct product fragments, the respective expressions

must be solved simultaneously, since they share common variables. The fractional abundances of $\phi_{1,4}$ belong to these shared variables.

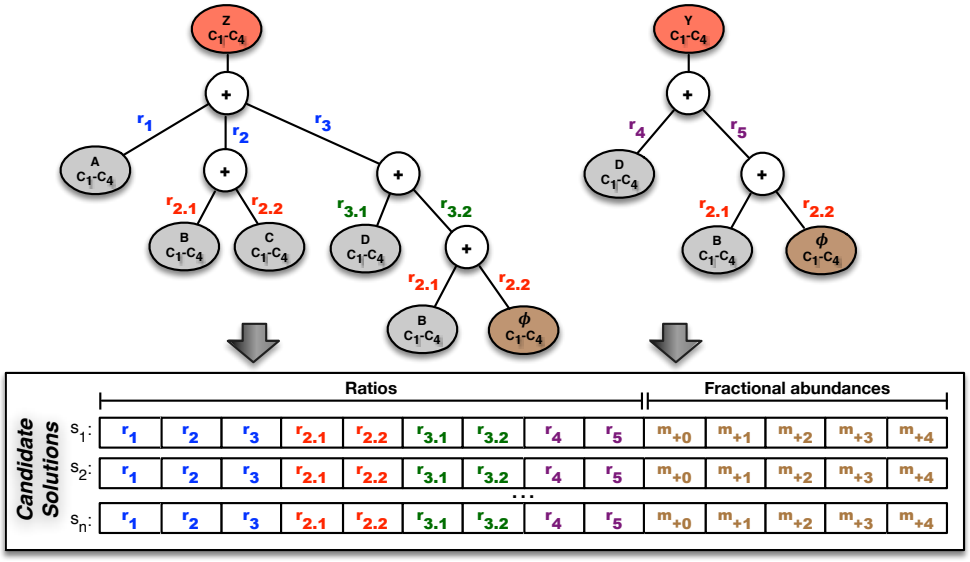


Figure 5.13: Illustration of a meta-heuristic method to estimate both flux ratios and unknown fragment MDV. Each candidate solution encodes values for the ratio variables in the expressions, and also the mass fractional abundance of each isotopomer of the unknown fragment.

The problem to estimate both ratio values and the MDV of unknown compounds' fragments is formulated as follows:

$$\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{s_i} \left(\frac{MDV_{c,i}(j) - MDV_{m,i}(j)}{MDV_{m,i}(j)} \right)^2 \\
\text{subject to:} \quad & \textit{all expressions to each } MDV_{c,i} \\
& \sum_{j=1}^{s_i} MDV_{c,i}(j) = 1 \;\; , \;\; \forall\, MDV_i \in \mathcal{D} \\
& \sum_{j=1}^{s_i} \phi_{c,i}(j) = 1 \;\; , \;\; \forall\, \phi_i \in \Phi \\
& |fl(\phi_{c,i}) - fl(substrate)| \leq \theta_{fl} \;\; , \;\; \forall\, \phi_i \in \Phi \\
& \sum_{x=1}^{s_l} \tau_{l,x} = 1 \;\; , \;\; \forall\, l \in \mathcal{R}
\end{aligned}$$

where ϕ_i is the unknown fragment i, $\Phi \not\subset \mathcal{F}$ is the set of unknown fragments in the expressions, $fl(\phi_{c,i})$ the fractional labeling of the calculated MDV of the unknown fragment i and θ_{fl} is a configured fractional labeling distance threshold.

To decode each generated solution, the values obtained for the ratios are retrieved from the individuals and associated to the corresponding variables in the expressions, and, for each unknown fragment, the fractional abundances of the mass of each isotopomer is associated to the proper vector position. After that, the expressions are solved and the calculated product vectors are compared to their measured ones.

The constraints of this problem, extend the ones referred in the formulation in Section 5.6.2, but since the solutions of the current problem also encode the fractional abundances of non measured MDVs, additional constraints are made over the generated abundances. The sum of the fractional abundances of the unknown vectors also has to be equal to one, and the fractional labeling of the generated unknown vectors has to be close to the substrate labeling content. So, if 20% [U-^{13}C] glucose is used as substrate, then the fractional labeling of ϕ_i must be around 20%.

However, the higher is the number of measured source fragments used in the solving expressions, or the number of expressions being solved, the better is the estimation. Only if the used measurements are sufficient to accurately compute the unknown MDV, is it possible to proceed with the estimation. Otherwise, different runs could generate different valid solutions according to the algorithm configuration. This inconsistency can occur, since the ratio variables and the fractional abundances in the estimated MDVs might be adjusted to compute MDVs of the products that are close to the measured ones.

Symbols used in ^{13}C-MFA formulations

Table 5.1: Symbols used in the equations and formulations of the methods of ^{13}C-MFA methods.

Symbol	Description
m_{+i}	fractional abundance of the mass of the i-th isotopomer of the fragment
$X_{i,j}$	fragment of compound X that starts in its i-th and ends in the j-th carbon atoms
$X(i)$	value of the i-th position of the MDV of the fragment X
$\mathcal{F}$	set of measured fragments
$a(isotopologue_{s,j})$	abundance of the isotopologue j of element s
I_i	naturally occurring isotope i
$f(I_i)$	frequency of isotope I_i within the isotopologue
$x(I_i)$	natural abundance of the isotope I_i
A_x^{corr}	corrected MDV of the fragment A_x
$C_{A,x}CONHSiS$	correction matrix for the fragment A_x
$C_{A,x}s$	correction matrix for the element s in the fragment A_x
$b_{unlabeled}$	fraction of unlabeled biomass
$MDV_m^{unlabeled}$	mass distribution vector of an unlabeled fragment witn m carbon atoms
$A_x^{exclusive}$	exclusive mass distribution vector of the fragment A_x
$fl(A_x)$	fractional labeling of the fragment A_x
$\mathcal{N}$	carbon atom transitions graph
$\mathcal{M}$	graph nodes associated to the metabolites
$\mathcal{T}$	set of edges representing carbon transitions between the reactants and products
$\mathcal{I}$	set of carbon transitions that are irreversible
ρ_p	path p to a product fragment
$\mathcal{T}_p$	subset of transitions in $\mathcal{T}$ that connects measured fragments to the product

Continued on next page

Table 5.1 – continued from previous page

Symbol	**Description**
$\mathcal{F}_p$	subset of fragments in $\mathcal{F}$ that are the sources of the transitions ($\mathcal{T}_p$) originating the product
$\mathcal{P}$	the set of paths originating the carbon backbone of a metabolite
τ_i	fraction of the pathway ρ_i that contributes to the labeling pattern of the in the carbons of the product fragment
R_x	radius of the search procedure
σ_{co2}	labeling content of CO_2
$\mathcal{S}^{co2}$	set of fragment sources that are combined with CO_2
$\mathcal{S}^{\sim co2}$	set of fragment sources that are not combined with CO_2
S_n	MDV of the last considered source fragment
$MDV_{c,aa}$	calculated MDV of the target amino acid generated by the meta-heuristic method
$MDV_{m,aa}$	measured MDV of the target amino acid
ϕ	unknown branch, i.e., $\mathcal{F}_\phi \not\subseteq \mathcal{F}$
τ_ϕ	fraction of a pathway that can not be calculated
Ω	parent node of a sub-tree in an ratios expression tree
$MDV_{m,i}(j)$	measured fractional abundance of the m_{+j} mass of the fragment i
$MDV_{c,i}(j)$	calculated fractional abundance of the m_{+j} mass of the fragment i
$\mathcal{D}$	the products for which the optimization problem is being solved to estimate when estimating their converging flux ratios
$\mathcal{R}$	set with all subsets of ratios corresponding to pathways originating the same intermediate fragment
$\phi_{i,j}$	unknown fragment that starts in its i-th and ends in the j-th carbon atoms

Continued on next page

Table 5.1 – continued from previous page

Symbol	Description
Φ	set of unknown fragments
$fl(\phi_{c,i})$	fractional labeling of the calculated MDV of the unknown fragment i
θ_{fl}	fractional labeling distance threshold

BIBLIOGRAPHY

[Fischer and Sauer, 2003] Fischer, E. and Sauer, U. (2003). Metabolic flux profiling of Escherichia coli mutants in central carbon metabolism using GC-MS. *European Journal of Biochemistry*, 270(5):880–891.

[Hellerstein and Neese, 1999] Hellerstein, M. K. and Neese, R. a. (1999). Mass isotopomer distribution analysis at eight years: theoretical, analytic, and experimental considerations. *The American journal of physiology*, 276(6 Pt 1):E1146–70.

[Latendresse et al., 2012] Latendresse, M., Malerich, J. P., Travers, M., and Karp, P. D. (2012). Accurate Atom-Mapping Computation for Biochemical Reactions. *Journal of chemical information and modeling*, 52(11):2970–2982.

[McAnulty et al., 2012] McAnulty, M. J., Yen, J. Y., Freedman, B. G., and Senger, R. S. (2012). Genome-scale modeling using flux ratio constraints to enable metabolic engineering of clostridial metabolism in silico. *BMC systems biology*, 6(1):42.

[Ravikirthi et al., 2011] Ravikirthi, P., Suthers, P. F., and Maranas, C. D. (2011). Construction of an E. Coli genome-scale atom mapping model for MFA calculations. *Biotechnology and bioengineering*, 108(6):1372–1382.

[Smart et al., 2010] Smart, K. F., Aggio, R. B. M., Van Houtte, J. R., and Villas-Bôas, S. G. (2010). Analytical platform for metabolome analysis of microbial cells using methyl chloroformate derivatization followed by gas chromatography-mass spectrometry. *Nature protocols*, 5(10):1709–29.

[Szyperski, 1995] Szyperski, T. (1995). Biosynthetically directed fractional 13C-labeling of proteinogenic amino acids. An efficient analytical tool to investigate intermediary metabolism.

[van Winden et al., 2002] van Winden, W. a., Wittmann, C., Heinzle, E., and Heijnen, J. J. (2002). Correcting mass isotopomer distributions for naturally occurring isotopes. *Biotechnology and bioengineering*, 80(4):477–9.

[Zamboni et al., 2005] Zamboni, N., Fischer, E., and Sauer, U. (2005). FiatFlux–a software for metabolic flux analysis from 13C-glucose experiments. *BMC bioinformatics*, 6(1):209.

CHAPTER 6

MINERAL - METABOLIC NETWORK RATIO ANALYSIS

6.1 Introduction

There are several tools to perform tasks related to stationary ^{13}C-Metabolic Flux Analysis (^{13}C-MFA) such as *COBRA Toolbox v2.0* [Schellenberger et al., 2011], *Metran* [Yoo et al., 2008], *13CFLUX2* [Weitzel et al., 2013], *C13* [Garcia-Albornoz et al., 2014], *FIA* [Srour et al., 2011], *OpenFLUX2* [Shupletsov et al., 2014], *NMR2FLUX* [Sriram et al., 2004], *influx_s* [Sokol et al., 2012] and *FiatFlux* [Zamboni et al., 2005], which were reviewed in chapter 2.

In the context of local approaches, only *FiatFlux* is available, which performs the so-called metabolic flux ratio (METAFoR) analysis [Sauer et al., 1999, Szyperski, 1995]. However, *FiatFlux* is not flexible to the users needs, in terms of metabolic networks, carbon atom tracing, and compound fragmentation. Therefore, it cannot be easily used for organisms with specific metabolic routing, that have differences when comparing to well-known models such as *Escherichia coli*, *Saccharomyces cerevisiae* or *Bacillus subtilis*.

In this scenario, *MiNeRAl* has been developed and is presented in this chapter, to provide a generic and flexible workbench to perform stationary ^{13}C-MFA following the local analysis paradigm. The aim was to develop a software not fixed to specific types of inputs, where every calculation is determined and performed according to the user inputs, without any

predefined assumptions. The main algorithms and methods used in this process have been described in the previous chapter.

6.2 Software functionalities

In the following sections, the main features of *MiNeRAl* are described, while in the following details on the implementation are provided. It is important to mention, before starting, that the user interaction follows the same principles of the *CBFA* plug-in, and therefore uses datatypes, operations and views as the main concepts. The explanation of these concepts is provided in Chapter 4.

6.2.1 Metabolic flux ratio analysis workflow

Although the utilization of *MiNeRAl* does not need necessarily to be always performed following a fixed set of steps, there is a common workflow that can be performed, starting from the user inputs up to the determination of metabolic flux ratio constraints. Independently of the tasks that will be performed, a project containing an atom transitions model and the fragmentation of the amino acids to be analyzed is created before any action is executed. Thus, the first step of the pipeline is to create a project, as shown in Figure 6.1.

After creating the project, the next step is to load experimental data and associate the spectra to the proper amino acids defined in the project. The isotope abundances will be used to generate Mass isotopomer Distribution Vector (MDV)s to each amino acid fragment, which will be further corrected for their natural isotope abundance. However, the calculated MDVs should only be used if their fractional labeling is close to the input substrate

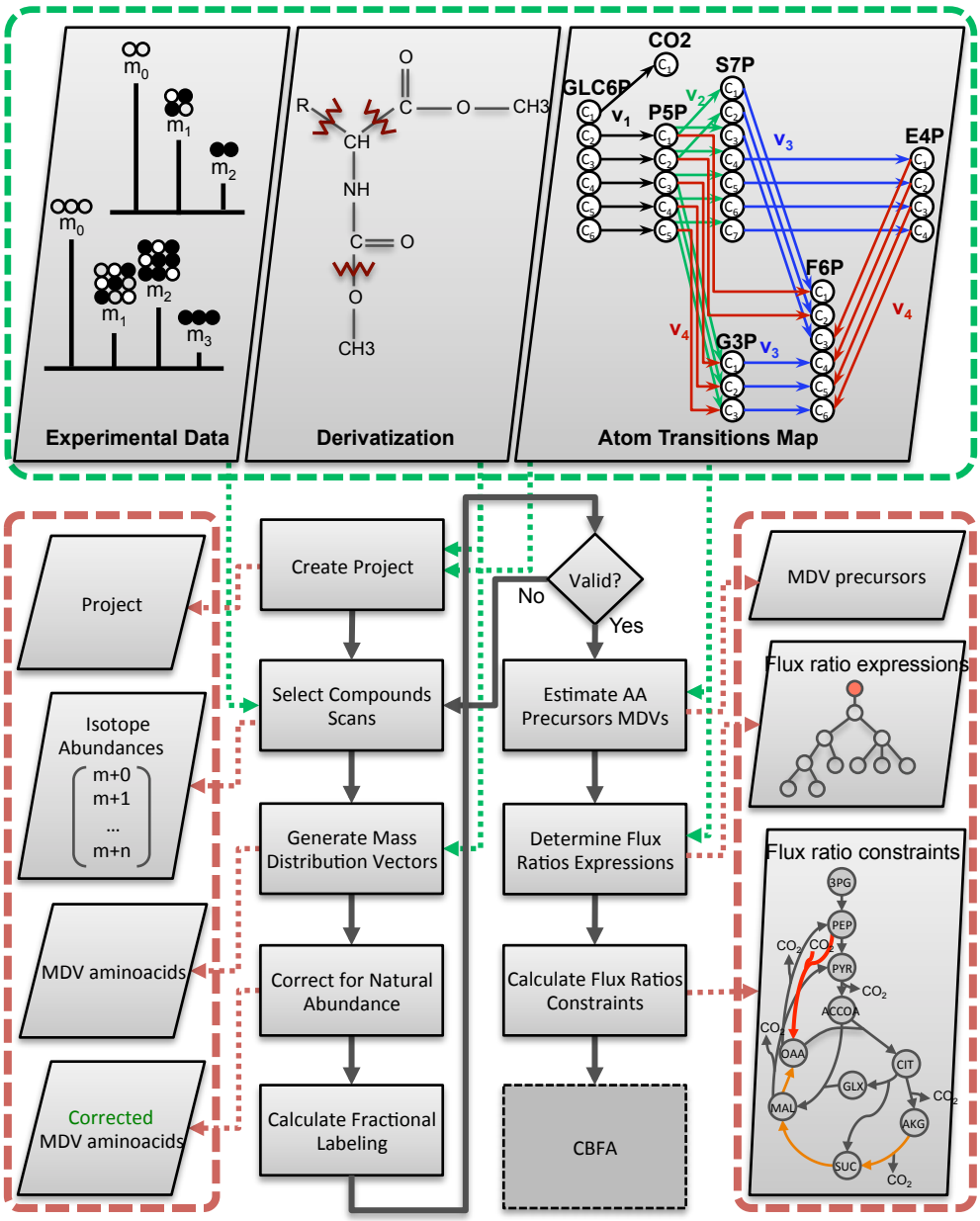


Figure 6.1: Workflow with the main tasks in MiNeRAl. The pipeline illustrates how the main operations are combined starting from the main input datatypes, until the flux ratio constraints are obtained. The last step is to link the constraints with CBFA to perform constraint-based methods. Dashed red and green boxes encompass the outputs and the inputs of the workflow, respectively. Dashed red lines represent the production of the outputs, while the green lines represent the utilization of the software inputs. Rectangles represent the main operations performed during the workflow.

labeling (e.g. if a 20% [U-^{13}C] glucose is used, the fractional labeling of the fragments should be about 20% as well). Therefore, a validation over the fractional labeling of the amino acids' fragments is performed, and the validated MDVs will be used in further analysis.

Although experimental measurements can be inputted for both amino acids and other intermediary metabolites in the central metabolism, amino acids offer richer labeling information since they are much more abundant in cell extracts and proteins. Nevertheless, the measurements from the amino acids can be used to estimate the MDVs of their precursors.

The set of MDVs stored in the datatypes can then be used to determine the metabolic flux ratios related to different paths in the atom transitions map. These ratios are calculated through expressions that are automatically built over the MDVs of the fragment sources from the associated paths. Finally, the ratios can be used to build flux ratios constraints that can be exported and used in constraint-based methods, such as the ones implemented in *CBFA*.

6.2.2 Organization of the project components

All the outputs generated during an analysis workflow are stored in datatypes within the initially created project. In Figure 6.2, the project datatypes organization is shown, where the top two datatypes, the carbon atoms transitions and the compounds fragmentation, will be common to every operation performed over the project. Since different samples can be analysed in the same project, for instance samples related to different experimental conditions, all the remaining datatypes will be stored nested in the corresponding sample. Each project contains three datatypes immediately attached to its associated root node in the workbench clipboard.

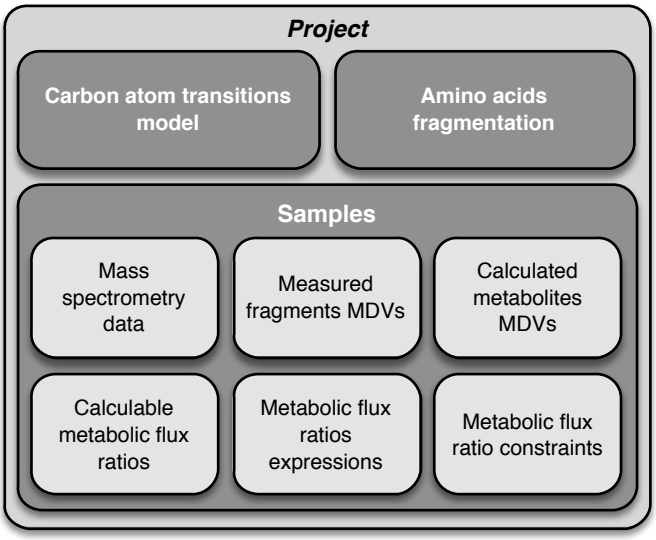


Figure 6.2: MiNeRAl project elements organization. The project contains three main datatypes: the carbon atom transitions, the compounds fragmentation and the experimental samples. All the generated datatypes during the pipelines performed in MiNeRAl are stored nested to the corresponding sample.

For each loaded sample, the selected compounds spectra, their MDVs and associated procedures will be kept and available for visualization and exportation. A separate metabolic flux ratio analysis is done for the data corresponding to each sample, since the set of fragments that are measured might change, and their labeling informations are distinct.

Some advantages arise in allowing to keep several samples nested to the same project, namely, since the carbon transitions model and the compounds fragmentation are the same, when comparing different conditions for the same case study, their configuration is made only once. Moreover, the analysis for the different samples can be performed simultaneously without the need of repeating all the loading and configuring steps for each sample.

6.2.3 Main operations

In *MiNeRAl*, there are different operations to create, and use, the datatypes related to a certain project. The available operations can be grouped according to the type of data they take or generate, as shown in Figure 6.3.

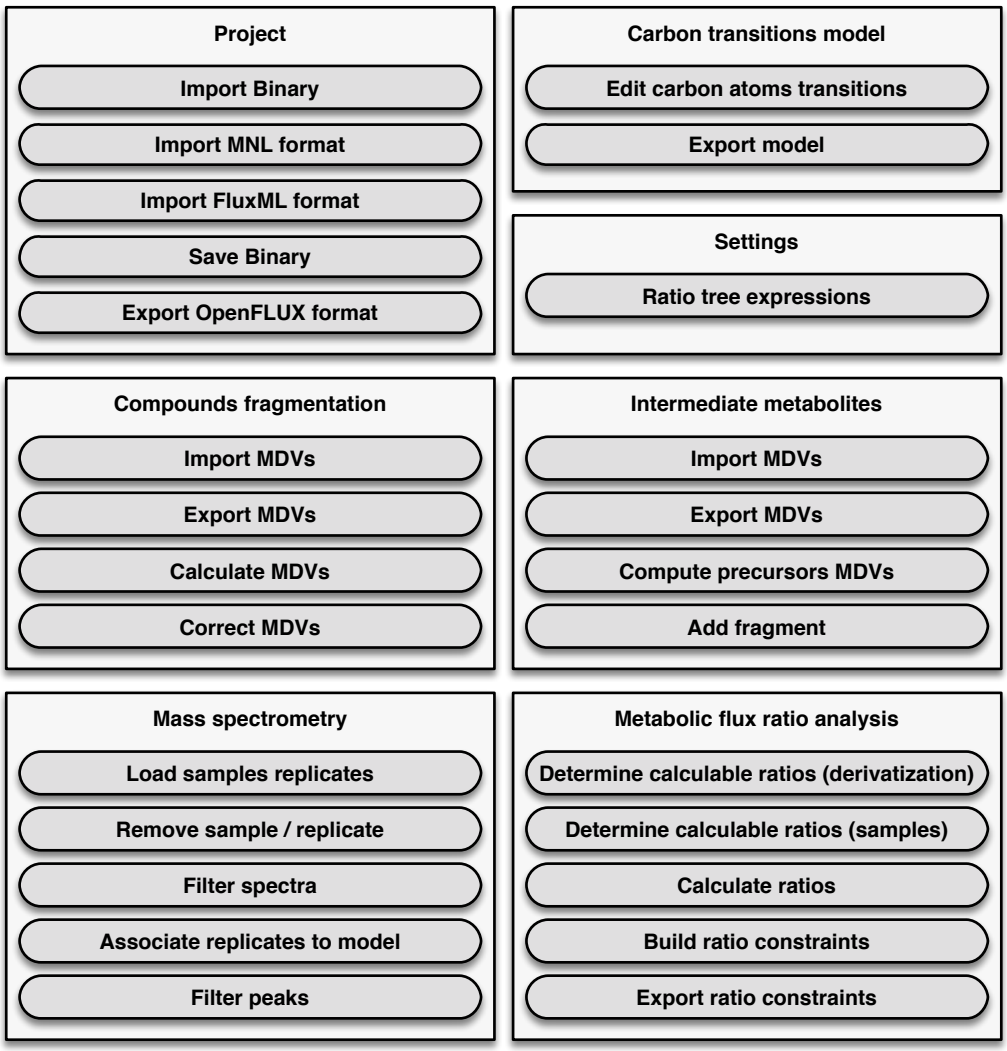


Figure 6.3: Main MiNeRAl operations. The central operations to execute the main features are separated in different groups. All the referred operations are made available in the menu bar of the workbench, or by right clicking in the datatypes that are used as input to the operations.

Below, follows a brief description about the functionality of each oper-

ation, grouped by the main datatypes:

Projects:

- **Import binary:** loads a previously binary saved *MiNeRAl* project and the previous state of the project will be restored;
- **Import MNL format:** imports the project settings written in a text file, so some steps related to the configuration of the project components, such as, the carbon transitions model and the compounds fragmentation, can be skipped in the graphical interface. Moreover, the MDVs for the fragments of the amino acids and intermediary metabolites can be provided in this kind of file, so all the steps associated to the experimental data can also be avoided;
- **Import FluxML format:** import a project in the FluxML format, which are documents for metabolic and isotope network specification, so data originating from files formatted by the *13CFLUX2* software can be loaded [Weitzel et al., 2013]. However, not all data generated in these files are compatible with *MiNeRAl*, since their approach is different from the one implemented here. The carbon atoms transitions and the isotope abundance can be imported;
- **Save binary:** save all the project components to a binary file, so the project can be posteriorly loaded in the same exact state;
- **Export OpenFLUX format:** export the project information to the format accepted by *OpenFLUX* [Shupletsov et al., 2014]. As it happens in the case of FluxML format, not all the project components are exported. The carbons transitions and the calculated MDVs are

converted to the proper format and stored with a format readable by the software.

Carbon atoms transitions:

- **Edit carbon atoms transitions:** this operation allows to change a loaded carbon transitions model. It is possible to change reactions reversibility, alter the carbons mapping, remove reactions and add new ones;
- **Export model:** If a model is edited, it might be useful to export the changes, so this operation allows to export a project's model in csv format.

Settings:

- **Ratio tree expressions:** allows to change the visualization settings of the tree expressions that are made available in the views for the datatypes containing the flux ratio expressions.

Compounds fragmentation:

- **Import MDVs:** loads previously calculated MDVs for compound derivatives;
- **Export MDVs:** export the MDVs calculated for the fragments originating from the derivatization;
- **Calculate MDVs:** calculate the MDVs for the compounds in the fragmentation using the loaded experimental data;
- **Correct MDVs:** correct the MDVs for their natural isotope abundance.

Intermediate metabolites:

- **Import MDVs:** imports MDVs for intermediary compounds that have been previously estimated / measured;
- **Export MDVs:** exports the MDVs of the estimated fragment precursors to a csv formatted file;
- **Compute precursors MDVs:** searches for the fragment precursors that originate the measured compound derivatives, and calculates their MDVs;
- **Add fragment:** allows to manually add a new compound fragment MDV to the existing set, without the need of computing it.

Mass spectrometry:

- **Load samples replicates:** import a set of samples, where each sample can contain several replicates. All the replicates' files can be organized in different sample clusters, to associate to which sample the replicates belong to;
- **Remove sample / replicate:** delete a replicate from a sample, or a full sample, and the corresponding replicates, from the project;
- **Filter spectra:** calls a method to filter the compounds occurring in all replicates (for each sample) and filters the most confident spectrum for each compound;
- **Associate replicates to model:** after loading and organizing the samples' replicates, associate the compounds in the different replicates to the model compounds;

- **Filter peaks:** filter the scan for each compound according to a selected approach, that could be the highest base peak, the center scan spectra or the one for which the MDVs would have the best fractional labeling.

Metabolic flux ratio analysis:

- **Determine calculable ratios (derivatization):** allows to determine the ratios that can be calculated, given the carbon transitions model and the compound derivatives from the loaded fragmentation, without the need of computing the MDVs from experimental data. The ratio expressions are generated;

- **Determine calculable ratios (samples):** allows to determine the ratios that can be calculated, given the experimental data. As in the previous operation, the ratio expressions are obtained;

- **Calculate ratios:** use the ratio expressions to actually compute the values for the metabolic flux ratios;

- **Build ratio constraints:** the calculated ratios are used to constrain metabolic fluxes; to achieve such constraints, ratio expressions around the reactions in the model are built;

- **Export ratio constraints:** save the generated metabolic flux ratio constraints, so they can be used in *CBFA* to constrain its simulation methods.

6.2.4 Carbon atom transitions model

In *MiNeRAl*, only models with carbon atom transitions are considered. These carbon transitions reactions are used to infer all the formulations in the different steps of the software pipeline. For each reaction, the carbon transitions are codified as letters, where the specified order allows to trace how the carbons are transferred from the reactants to the products.

The model is the basis of the workflows performed in the software, and each workflow is stored in the context of a project. Therefore, the model is the first main element that will define a project (Figure 6.4). By default, *MiNeRAl* allows to load the carbon transitions map for the central carbon metabolism of *Escherichia coli*. However, any other model can be loaded, for instance, as a Comma-Separated Values (CSV) file, with the format shown in Figure 6.4. After loading the model, a datatype is instantiated to encapsulate the stored transitions map. This will be the case of all the operations described next. For all the resulting datatypes, a view is made available and their content can be visualized on the right panel of the Graphical User Interface (GUI) of the application.

Since there are other components that compose a project, and all of them have the concept of metabolites in common, the metabolites originating from the different components have to be mapped to the metabolites in the model, so every component is standardized to the same identifiers. This association will be explained in section 6.2.7.

6.2.5 Compounds fragmentation

MiNeRAl is not specific for any kind of compound fragmentation. Unlike other approaches, the details about the obtained fragments from any

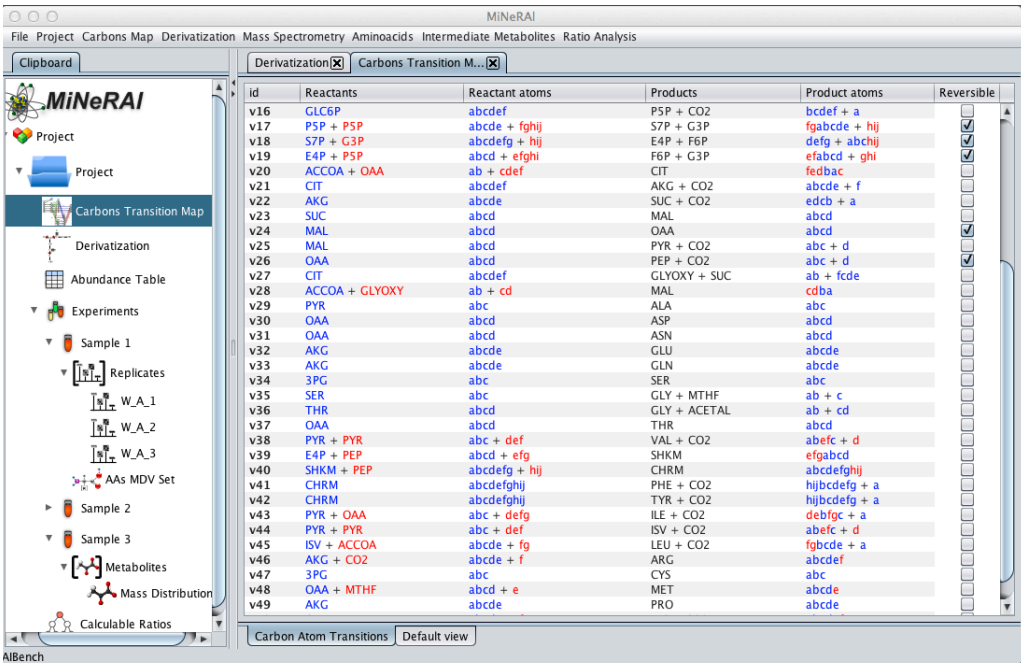


Figure 6.4: A project is defined in the context of a carbon transitions model. The colours in the atoms help to check from which reactants the carbons in the products come from.

derivatization, such as t-butyl-trimethylsilyl (TBDMS) or methyl chloroformate (MCF) [Smart et al., 2010] derivatives, can be added to a project. For each compound, its fragments are identified by the first and last atoms in its carbon skeleton, and the mass-to-charge ratio (m/z) of the lightest isotopomer of the fragments. The set of fragments configured at this stage, will be the one used in a further step when it comes to load experimental Mass Spectrometry (MS) data. In the case of the MCF fragmentation, the set of amino acids fragments that can be obtained is listed in Table 6.1.

This can be imported by the software as a CSV file with the same format as the table above. Each row corresponds to the compound identifier, followed by the compound's fragments and corresponding m/z of its lightest isotopomer. After loading the derivatization, it is always possible to add new compound fragments. In Figure 6.5, it is possible to see how this

Table 6.1: Amino acids methyl chloroformate derivatives

Amino acids	Fragments							
Alanine	C_2-C_3	102[1]						
Aspartate	C_1-C_3	160	C_2-C_4	128	C_1-C_2	146		
Glutamate	C_2-C_4	114	C_1-C_4	174	C_2-C_5	142		
Glycine	C_2	88	C_1-C_2	116				
Histidine	C_1-C_6	210	C_2-C_6	226	C_3-C_6	81		
Isoleucine	C_2-C_6	144	C_2	88	C_1-C_2	115		
Leucine	C_2-C_6	144	C_2	88	C_2-C_3	102	C_1-C_2	115
Lysine	C_2-C_6	142	C_1-C_6	244	C_2	88	C_6	88
Methionine	C_1-C_2	115	C_1-C_3	128	C_2-C_5	162		
Phenylalanine	C_1-C_9	162	C_2-C_9	178	C_3-C_9	91		
Proline	C_2-C_5	128						
Serine	C_2-C_3	100	C_1-C_2	146				
Tyrosine	C_1-C_9	236	C_2-C_9	121				
Threonine	C_1-C_2	115	C_2-C_4	100				
Valine	C_2-C_5	130	C_1-C_2	115				

[1] The integer number on the right side of each fragment corresponds to the mass-to-charge ratio of the lightest isotopomer of the fragment.

information is stored in the software.

The atom composition of the fragments depends on the used derivatization technique. Therefore, it is possible to provide the number of each type of atoms present in the fragments. This information is useful, if the analysis includes the correction of the natural isotope abundance of the amino acids' MDVs. Table 6.2 summarize the atoms composition for some of the amino acids' fragments obtained in the MCF derivatization. Although only atoms of carbon (C), oxygen (O), hydrogen (H) and nitrogen (N), are represented in the table, for other derivatizations, other atoms could also be added and taken in consideration by the software.

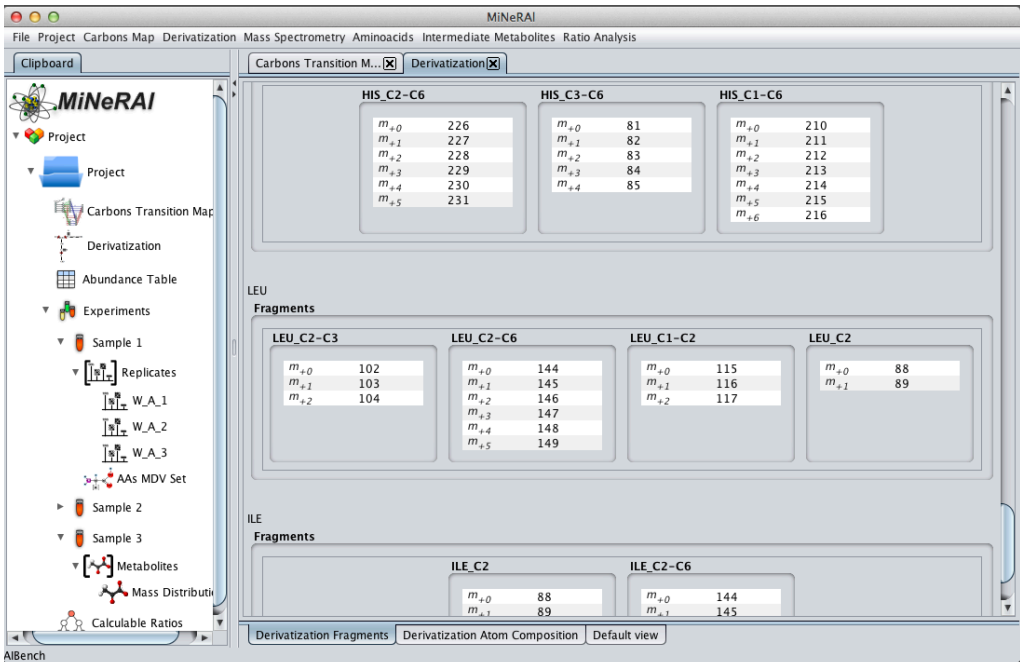


Figure 6.5: After loading a derivatization file, a datatype representing the fragmentation is created. The view in the figure is the one associated to the datatype, containing a table for each added fragment. HIS, Histidine; LEU, Leucine; ILE, Isoleucine.

Table 6.2: Methyl chloroformate derivatives fragments atoms composition

Amino acid	Fragment	Atoms			
		C	H	O	N
Alanine	C_2-C_3	2	8	2	1
Aspartate	C_2-C_4	3	10	4	1
Aspartate	C_1-C_3	2	7	3	1
Aspartate	C_1-C_2	3	8	4	1
Glutamate	C_2-C_4	2	9	2	1
Glycine	C_2	2	6	2	1
Histidine	C_1-C_6	3	11	4	2
Histidine	C_3-C_6	-	4	-	2
Isoleucine	C_2-C_6	2	14	2	1
Isoleucine	C_2	2	5	2	1
Isoleucine	C_1-C_2	1	5	3	1

Continued on next page

Table 6.2 – continued from previous page

Amino acid	Fragment	Atoms			
		C	H	O	N
Leucine	C_2-C_6	2	14	2	1
Leucine	C_2	2	5	2	1
Leucine	C_2-C_3	2	7	2	1
Leucine	C_1-C_2	1	5	3	1
Lysine	C_2-C_6	2	13	2	1
Lysine	C_1-C_6	4	17	5	2
Methionine	C_1-C_2	3	8	4	1
Phenylalanine	C_1-C_9	2	12	2	1
Phenylalanine	C_2-C_9	-	7	-	-
Proline	C_2-C_5	2	10	2	1
Serine	C_2-C_3	2	7	2	1
Serine	C_1-C_2	3	8	4	1
Threonine	C_1-C_2	1	5	3	1
Threonine	C_2-C_4	1	7	2	1
Tyrosine	C_2-C_9	-	8	1	-
Tyrosine	C_1-C_9	3	14	4	1
Valine	C_2-C_5	2	12	2	1
Valine	C_1-C_2	1	5	3	1

6.2.6 Mass spectrometry data processing

In terms of ^{13}C experimental data, *MiNeRAl* works with the ones originating from MS analytical techniques. The labeling data is used throughout the different steps of the workflow in the form of MDVs (Figure 6.1) that have been introduced previously. Here, two alternatives are available. It is possible to load MDVs that have been previously computed, or calculate the fragments MDV from inputted experimental data. There are several tools to treat MS data, from data extraction [Luedemann et al., 2008, Pluskal

et al., 2010], to data correction [Wahl et al., 2004, Huege et al., 2011]. *FiatFlux* and, more recently, *iMS2Flux* [Poskar et al., 2012] encompasses all the data processing steps.

When it comes to calculate the vectors from experimental data, it was not on the scope of this work to perform all the data extraction tasks. Instead, a module to process reports obtained from Automated Mass Spectral Deconvolution and Identification System (AMDIS), which is a freeware commonly used to deconvolute chromatograms generated by Gas Chromatography–Mass Spectrometry (GC-MS) and to identify and quantify compounds in biological samples [Stein, 1999], has been developed. This software identifies target compounds using a library, for instance, the list of amino acid fragments in a certain derivatization method. Frequently, in a chromatogram, two or more compounds elute at the same retention time, and consequently their peaks are merged, making difficult their identification and peak integration. AMDIS can resolve co-eluted peaks via deconvolution to separate the different components. After data deconvolution, it generates a report with the match factors between the deconvoluted spectrum and the target spectrum, that are above a threshold defined by the users.

Since these reports are text files, there is no need for additional libraries to access its contents. However, for each identified compound, there are several scans that are reported and also, different peaks with distinct retention times and degrees of certainty. Therefore, some algorithms to automatically select the proper compound scans have been developed.

The first step is to select the valid set of candidate spectra from all the reported ones. In Figure 6.6, the activity diagram summarizes how this procedure has been implemented. When the user selects a set of replicates

for a given sample, the first activity is to filter the spectra for which no compound has been identified. If no spectrum has been identified, the user has to verify the AMDIS reports that were imported, for instance, to verify if the proper library has been used or to check if the configurations were set according to the experimental purposes.

AMDIS exports additional details for each mass spectrum, such as, the identification of the high-intensity ion used to identity and perform the deconvolution of the spectra (reference ion), whether it corresponds to a primary or secondary peak, the relative retention time on the GC column, or the degree of certainty when attributing the scans to compounds. For instance, for MCF derivatives, the library used in AMDIS also adds an additional suffix 'MDF derivative' to the compound identification.

For the set of identified compounds, the next step is to retrieve the compound name from the headers before each spectrum. Given these terms, a verification step is performed to check if all replicates contain the compounds. Sometimes, a compound can be identified with different retention times, for such cases, the compound spectrums with retention times that occur in every replicate are kept.

For the reduced set of spectra, the most confident ones are retrieved. AMDIS adds question marks '?' following the compound names to specify the degree of certainty regarding their identification. The more '?' symbols the peak has, less certain has been the deconvolution performed by AMDIS. This low quality of matching may originate due to low concentrations , or if there is a high background signals in the region of the chromatogram.

Often, during the derivatization, more than one derivative may be produced for the same compound. A main derivative is associated to a product with higher yield, or main GC peak, where the peak has a higher inten-

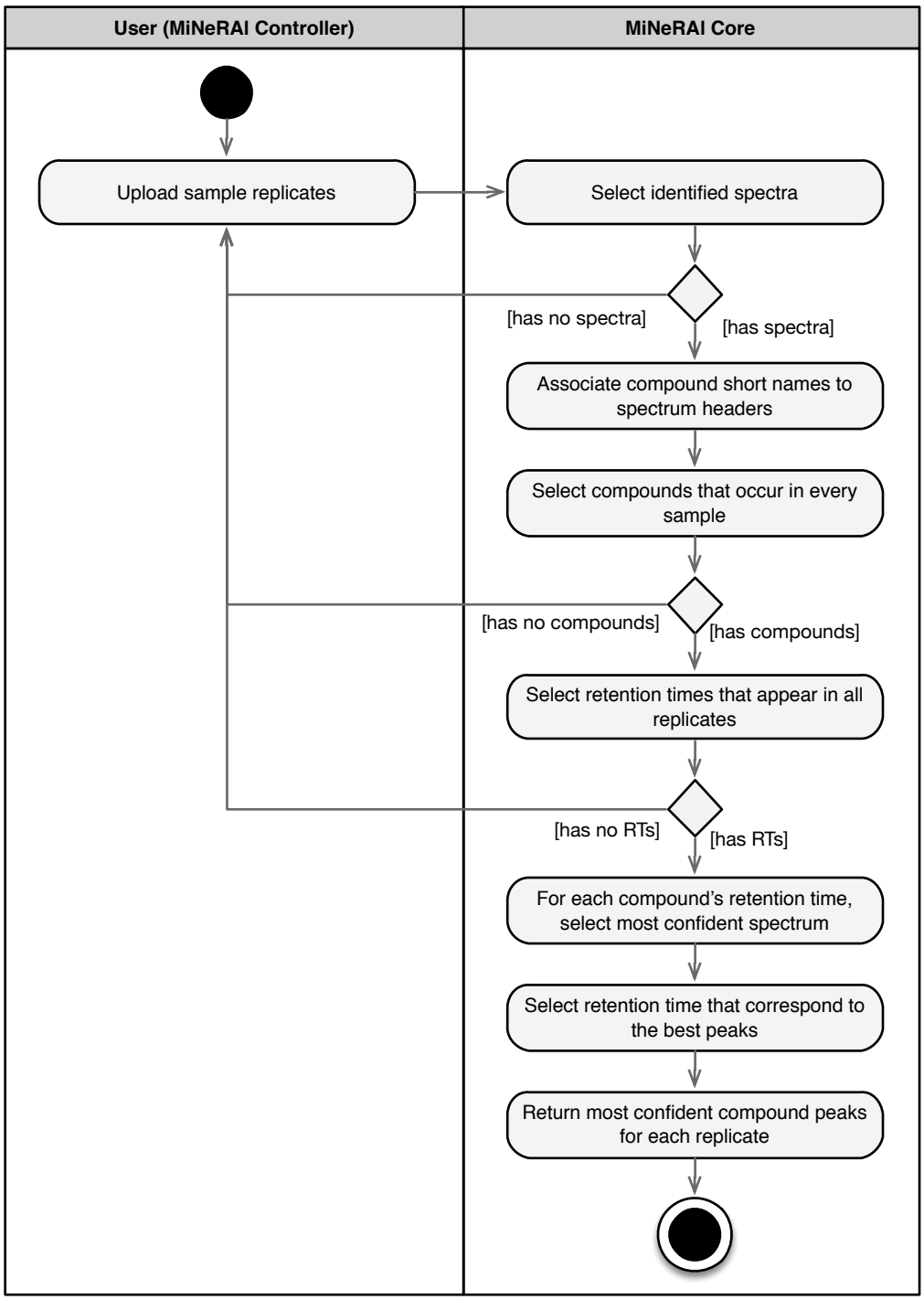


Figure 6.6: Activity diagram for filtering compounds spectra. The large black filled circle represents the start node, while the white circle with the black circle inside represents the final node. The round-cornered rectangles represent activities, which enclose actions, control flows and other elements that shape an activity.

sity (abundance) than the secondary peaks. Usually, the libraries used in AMDIS contain mass spectra information for the identification of secondary peaks and derivatization artifacts.

Secondary peaks are minor peaks that can be detected, such as the case for citric acid in MCF derivatization. These minor peaks are useful to avoid incorrect estimation of the number of unknown peaks detected in the samples. With respect to artifacts, for the case of MCF derivatives, it corresponds to peaks containing chlorine, being coproducts of the MCF reaction when it is carried out in a complex biological sample, rather than a specific metabolite itself [Smart et al., 2010]. Accordingly, the best main peaks have priority to be returned for each replicate, and only if no main peak has been identified, a secondary peal will be associated to the compound.

All the imported spectra can be graphically visualized in *MiNeRAl* as shown in Figure 6.7, both before and after the filtering process. The software allows to ignore the automatic spectra filtering, so the visualization of the identified spectra can be useful to help the user association of the compound experimental reads to the model compounds.

6.2.7 Linking metabolites from different project components

Until now, three components that defines a project (Figure 6.2) have been introduced, i.e., the carbon atoms transitions model, the compounds fragmentation and the MS experimental data. All the three components contain a common entity – compounds. However, these components originate from different sources and it is likely that the names used to identify the compounds do not match. This is a problem, since the amino acids in the model have to be the same amino acids in the experimental data, so the project

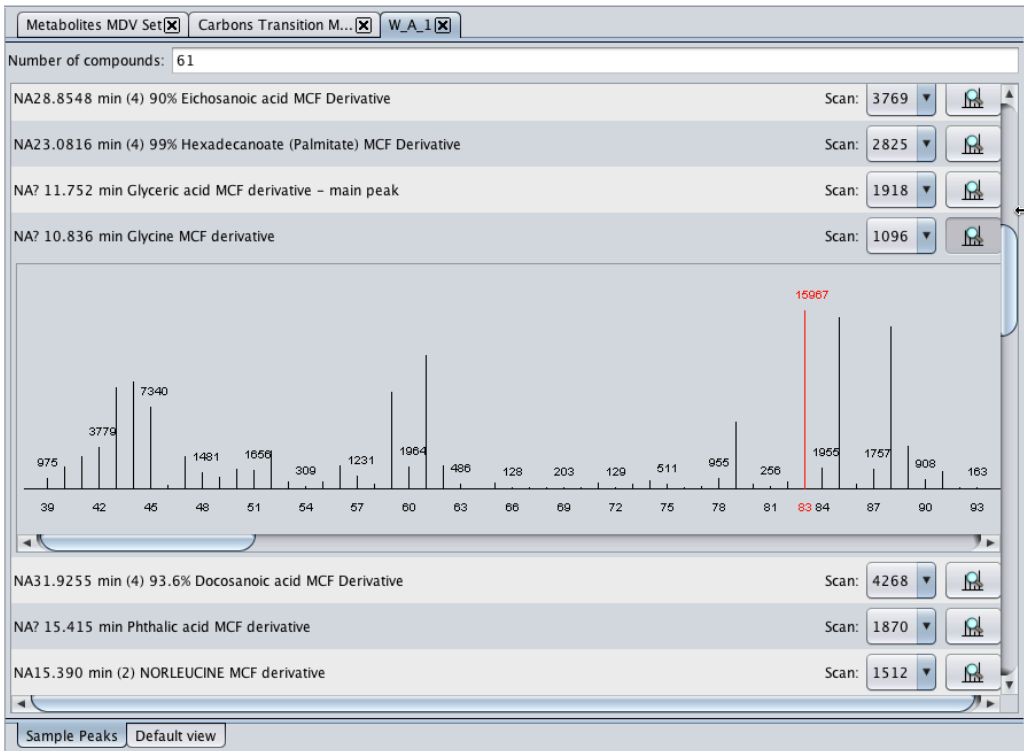


Figure 6.7: Visualization of the scan spectra obtained for a loaded sample replicate. The spectrum of each identified derivative can be visualized before and after the filtering process.

does not have two different names to the same entity.

Therefore, after configuring the transitions map, when loading either the derivatization, or the experimental data, an additional step has to be performed to associate the entities in these components to the entities in the model. To facilitate such task, a non-supervised method to generate candidate model compound names has been implemented, which is based in a measure of similarity between two strings, namely the Jaro–Winkler distance [Winkler, 1999]. Given two compound names (c_1 and c_2), the higher is the distance measure between them (Equation 6.1), the more similar they

are.

$$d_{winkler}(c_1, c_2) = \begin{cases} 0 \ , & \text{if } d_{jaro}(c_1, c_2) = 0 \\ d_{jaro}(c_1, c_2) + l(c_1, c_2) s_w \left(1 - d_{jaro}(c_1, c_2)\right) \ , & \text{otherwise} \end{cases} \tag{6.1}$$

where $l(c_1, c_2)$ is the common prefix length between c_1 and c_2 (up to a maximum of four characters), s_w is a constant representing the Winkler scaling factor (usually set to 0.1), and $d_{jaro}(c_1, c_2)$ is obtained from Equation 6.2.

$$d_{jaro}(c_1, c_2) = \begin{cases} 0 \ , & \text{if } m_c = 0 \\ \frac{1}{3}\left(\frac{m(c_1,c_2)}{|c_1|} + \frac{m(c_1,c_2)}{|c_2|} + \frac{m(c_1,c_2)-t}{m(c_1,c_2)}\right) \ , & \text{otherwise} \end{cases} \tag{6.2}$$

where $\mid c_1 \mid$ and $\mid c_2 \mid$ represents the length of c_1 and c_2, respectively, $m(c_1, c_2)$ is the number of matching characters in c_1 and c_2, i.e., if the same character in c_2 is not farther than $\frac{min(|c1|,|c2|)}{2}$ in c_1, and $t(c_1, c_2)$ refers to the number of character transpositions.

A compound name is considered as a valid candidate to another compound, if their Jaro–Winkler distance is greater than a defined similarity threshold. The default value to this threshold is 75 %.

6.2.8 Translating compounds labeling information

The basis of all calculations in *MiNeRAl* are the MDVs. The labeling pattern of amino acids can be used to calculate the labeling of its precursors. So, a key task when performing the metabolic flux ratio analysis is to properly calculate the MDVs of amino acids.

6.2.8.1 Building mass distribution vectors from AMDIS reports

When the MDVs are not provided, they have to be determined from the available experimental data. In such scenario, two project components are necessary to calculate the vectors, the project compounds fragmentation datatype, and the samples' experimental data loaded from AMDIS reports. The former is used to select the m/z of the lightest isotopomer of the derivatives, and to determine the number of carbon atoms in the fragment backbone, while the samples data are used to retrieve the intensities corresponding to each m/z.

Even after filtering the AMDIS reports, and identifying the proper compounds, there are still several scans with retention times closer to the one given in the used compounds library. To select the appropriate scan spectra three approaches have been implemented:

- highest m/z base peak;
- center scan spectra;
- best fractional labeling.

The highest m/z base peak selection, refers to an algorithm that searches for the scan that contains the base peak with the highest intensity value. After finding the scan, all the intensities for the m/z of all isotopomers of the compound fragments are retrieved from that scan. If no intensity is found for a certain isotopomer, it is assumed to have a zero value.

The center scan spectrum is provided by AMDIS, so it is straightforward to select the scan with the specified number. However, in the case of the best fractional labeling criterion, it takes a bit longer to be executed, since all the steps to build the fragments' MDVs, correcting them for the natural

abundance and calculating their fractional labeling, are necessary for all the different scans. Hence, for every fragment, the fractional labeling obtained for each scan is compared to the substrate labeling content, and the one with the lowest difference to the substrate labeling is selected.

fractional abundance of the m+i mass

If a sample contains more than one replicate, each position i in the calculated MDVs will have an average value, and the corresponding standard deviation, for the fractional abundance of the mass isotopomer i.

6.2.8.2 Correcting natural isotope abundance of mass distribution vectors

After determining the MDVs, it is still necessary correct the fragment vectors for their natural isotope abundance. Accordingly, it is necessary to calculate the abundances that are present in the correction matrices belonging to each atom specie, as it has been introduced in Section 5.2.1.1 [van Winden et al., 2002]. These calculations use the information about the atom composition of each fragment obtained in the derivatization process (Figure 6.8).

The software generates the correction matrices of each atom element present in the derivatives, based on the number of occurring elements provided by the user. The natural abundance of the naturally occurring isotopes of the main biological elements (Table 6.3) is loaded by default, and this information will be used to calculate the abundances of each isotopologue in the correction matrices (Equation 5.5).

Additionally, the labeling content of the input substrate and the fraction of unlabelled biomass are used to calculate the exclusive mass isotope distribution of the fragments. To achieve this exclusive MDV, the software

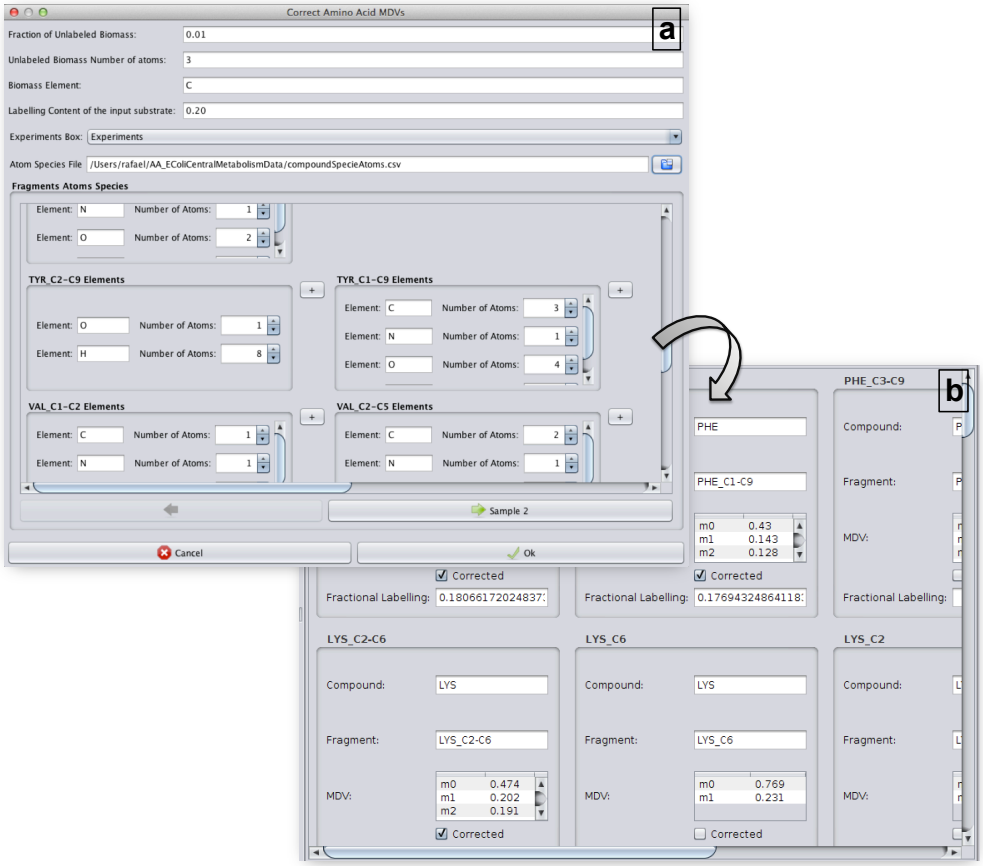


Figure 6.8: Correction of mass distribution vectors for natural isotope abundance. a) The atom composition of each fragment is configured by defining the number of each atom species that compose it. b) For the fragments where the atom composition has been provided, the corrected mass distribution vectors are updated and its fractional labeling is presented.

uses the corrected MDV and solves the equation [Fischer and Sauer, 2003]:

$$A_x^{exclusive} = \frac{A_x^{corr} - b_{unlabeled} \cdot MDV_m^{unlabeled}}{1 - b_{unlabeled}} \tag{6.3}$$

where $b_{unlabeled}$ is the fraction of unlabeled biomass and $MDV_m^{unlabeled}$ is the MDV of an unlabeled fragment with m carbon atoms calculated using Equation 5.5. Here, the two carbon isotopes are considered, and the number of carbons in the fragment stipulate the frequencies of the isotopes.

Table 6.3: Isotopic compositions of the most biologically relevant atoms

Atom symbol	Mostly occurring mass (m_{+0})	Fractional abundances			
		m_{+0}	m_{+1}	m_{+2}	m_{+3}
H	1	0.999885	0.000115		
C	12	0.9893	0.0107		
N	14	0.99632	0.00368		
O	16	0.99757	0.00038	0.00205	
Si	28	0.922297	0.046832	0.030872	
P	31	1.00000			
S	32	0.9493	0.0076	0.0429	0.0002

The data presented in this table were retrieved from [Rosman and Taylor, 1998]. H, hydrogen; C, carbon; N, nitrogen; O, oxygen; Si, silicon; P, phosphorus; S, sulfur.

To check whether the calculated MDVs can be used, their fractional labeling are calculated. Therefore, for each MDV, one can analyse its fractional labeling, and if the value is not similar to the substrate labeling content, the MDV should not be considered for further analysis. The fractional labeling of an MDV can be obtained through the equation:

$$fl(A_x) = \frac{\sum_{i=0}^{m} i \,.\, m_{+i}}{m \,.\, \sum_{i=0}^{m} m_{+i}} \tag{6.4}$$

where m is the number of carbon atoms in the fragment and m_{+i} is the fractional abundance of the mass of the i-th isotopomer of the fragment.

6.2.8.3 Computing amino acids precursors

In most organisms, the precursors of amino acids belong to the following list:

- Ribose 5-Phosphate;
- Erythrose 4-phosphate;
- 3-Phosphoglycerate;
- Phosphoenolpyruvate;
- Pyruvate;
- Oxaloacetate;
- α-Ketoglutarate.

By default, the software tries to identify compounds in the network with names that are similar to the abbreviations of the compounds in the above list, through the similarity approach previously described in this chapter. However, if the algorithm fails to identify the candidate precursors, it is possible to configure new compounds to the list of candidate precursors. This list will be the one used in the stopping criteria of the graph algorithm responsible to find the amino acids precursor fragments.

Since no ratios from different pathways are allowed when computing the MDVs of the precursors fragments, the selected set of candidate precursors has to be properly chosen, in a way where there are few steps (reactions) distant from the amino acids. In fact, if the traditional precursors are chosen, this will happen.

In the cases where the precursors MDVs need to be computed, an optimization approach is used to solve the non linear equations that are obtained for each mass isotopomer of the target amino acid. Three methods for global optimization have been implemented in *MiNeRAl*:

- Evolutionary Algorithm (EA);
- Simulated Annealing (SA);
- Differential Evolution (DE).

The methods progress by evaluating each generated vector, where after decoding its values, calculate the MDV of the target amino acid and a fitness score is calculated based on the objective function, as described in the previous chapter. In each evaluation, a penalty function is also applied to the vectors, to guide the algorithms to generate valid solutions, where the sum the generated mass isotopomers is equal to 1 for each precursor fragment.

The intermediate precursors that form the carbon skeleton fo the amino acids' fragments, obtained through MCF derivatization, have been identified by *MiNeRAl* and are listed in table Table 6.4.

Table 6.4: Intermediate precursors of methyl chloroformate derivatives fragments.

Amino acid	Fragment	Precursors
Alanine	C_2-C_3	$PYR_{2,3}$
Aspartate	C_1-C_3	$OAA_{1,3}$
	C_1-C_2	$OAA_{1,2}$
	C_2-C_4	$OAA_{2,4}$
Glutamate	C_2-C_5	$AKG_{2,5}$
	C_2-C_4	$AKG_{2,4}$
	C_1-C_4	$AKG_{1,4}$

Continued on next page

Table 6.4 – continued from previous page

Amino acid	Fragment	Precursors			
Glycine	C_1-C_2	$G3P_{1,2}$	$OAA_{1,2}$		
	C_2	OAA_2	$G3P_2$		
Histidine	C_2-C_6	$G3P_3$	$P5P_{1,4}$		
	C_3-C_6	$P5P_{1,3}$	$G3P_3$		
	C_1-C_6	$G3P_3$	$P5P_{1,5}$		
Isoleucine	C_1-C_2	$OAA_{1,2}$			
	C_2-C_6	$PYR_{2,3}$	$OAA_{2,4}$		
	C_2	OAA_2			
Leucine	C_1-C_2	$PYR_{2,3}$			
	C_2-C_6	$PYR_{2,3}$	PYR_3		
	C_2-C_3	PYR_3	PYR_2		
	C_2	PYR_3			
Lysine	C_2-C_6	$PYR_{2,3}$	$OAA_{2,4}$		
	C_1-C_6	$OAA_{1,4}$	$PYR_{2,3}$	$OAA_{2,4}$	$PYR_{1,3}$
Methionine	C_1-C_3	$OAA_{1,3}$			
	C_1-C_2	$OAA_{1,2}$			
	C_2-C_5	$OAA_{2,4}$	$G3P_3$		
Phenylalanine	C_2-C_9	$PEP_{2,3}$	$E4P_{1,4}$		
	C_1-C_9	$PEP_{2,3}$	$E4P_{1,4}$	$PEP_{1,3}$	
	C_3-C_9	PEP_3	$PEP_{2,3}$	$E4P_{1,4}$	
Proline	C_2-C_5	$AKG_{2,5}$			
Serine	C_1-C_2	$G3P_{1,2}$			
	C_2-C_3	$G3P_{2,3}$			
Threonine	C_1-C_2	$OAA_{1,2}$			
	C_2-C_4	$OAA_{2,4}$			
Tyrosine	C_2-C_9	$PEP_{2,3}$	$E4P_{1,4}$		
	C_1-C_9	$PEP_{2,3}$	$E4P_{1,4}$	$PEP_{1,3}$	
Valine	C_1-C_2	$PYR_{1,2}$			
	C_2-C_5	$PYR_{2,3}$			

6.2.9 Predict calculable metabolic flux ratios

The main purpose of *MiNeRAl* is to calculate metabolic flux ratios through different pathways in a given metabolic network, based on the inputted ^{13}C information. These flux ratios express the metabolic activity of certain reactions over another, in terms of percentage of flux that contributes to achieve a target metabolite labeling pattern. Before obtaining such ratio expressions, the values for these ratio percentages must be calculated based on probabilistic equations around the fragments' MDVs.

However, before starting all the experimental procedures, it might be convenient to inspect how the outcomes of such experiments would affect the determination of flux ratios. This is important, because the determination of the pathways' flux ratios is dependent on the measured fragments that are given to the methods. Thus, different fragments will originate different expressions that translate the flux ratios, and whether one decides to use TBDMS derivatization during a ^{13}C-tracer experiment, the metabolic ratio expressions will be different from the ratios obtained in MCF fragmentation.

The procedures to construct the expressions to calculate flux ratios are agnostic to the measured fragments in the sense that, to build the expressions, their MDVs are not mandatory. Instead, only the set of fragments that are available, for instance, all the fragments from the derivatization, are used. The graphical user interfaces in *MiNeRAl* allow to select either the measured amino acids and intermediate metabolite fragments, or a set of non measured fragments configured by the user. Given the set of fragments, the output of the operation will be a datatype representing the calculable flux ratios in the project's metabolic network. Since the expressions are dependent on the configured set of fragments, it is possible to analyse if the given fragments will actually be useful to accurately inspect the metabolic

activity on the different pathways throughout the network, even before the ^{13}C-tracer experiments have been performed.

In Figure 6.9, it is possible to see the created GUI of the operation to predict the ratio expressions of calculable ratios from the derivatization fragments inputted by the user. It is possible to add other fragments that to not belong to the derivatization, to perform a more efficient prediction of the expressions, e.g., intermediate precursors, glucose or CO_2.

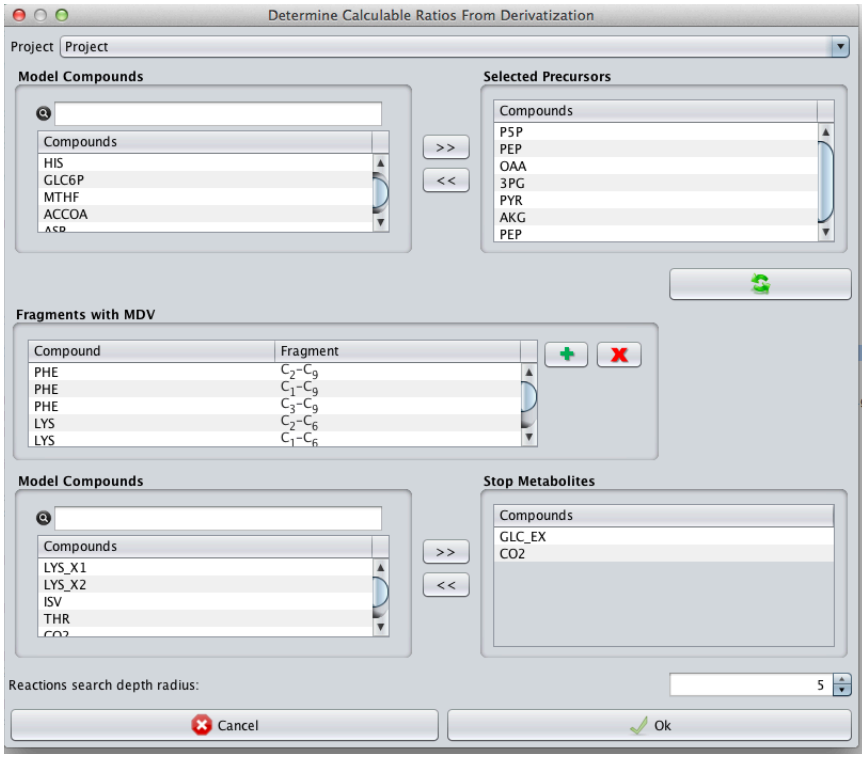


Figure 6.9: Calculable ratios operation dialog. The user can select the fragments from the derivatization for which the process will be performed. By default every fragment in the derivatization are select, as the substrate and CO_2. However, other fragments in the model can be added to the procedure.

6.2.10 Computing metabolic flux ratios

After computing the calculable flux ratios as described before, the expressions can be used to actually calculate the metabolic flux ratios along the different pathways on the atom transitions map. For the cases where the expressions have nonlinear terms, i.e., products between two or more ratio variables, as described in the previous chapter, the expressions are not solved algebraically through Equation 5.27. Instead, the values for the ratio variables are estimated through the minimization of the distance between measured MDVs and computed MDVs. As in the determination of amino acid precursors. implementations based on the three mentioned global optimization meta-heuristics have been implemented.

If a certain pathway contributing to the labeling pattern of a certain fragment is common to other analysed target fragments, the metabolic flux ratios corresponding to these pathways will be the same in all the determined expressions. Thus, expressions that share flux ratios must be computed in the same instance, so all the measurements that may influence the estimation are used.

6.3 Case studies

In this section, three case studies are provided to illustrate the use of the developed methods and computational tools. These were selected to provide typical uses of the software, rather than to provide a complete workflow with extensive data.

6.3.1 Calculable ratios for MCF derivatization in *Escherichia coli*

The aim of this study was to use *MiNeRAl* to determine the metabolic flux ratios that could be calculated when using the MCF derivatization fragments, without actually obtaining their values from experimental data. As shown in Section 6.2.8.3, when the precursors for the MCF amino acid fragments were listed, in this study the atom transitions for the central carbon metabolism of *Escherichia coli* were used.

Through the use of the software operation defined to execute such task, all the amino acids from the derivatization were selected, as did the stopping compounds, CO_2 and glucose, since these two compounds can be used as fragment labeling sources to calculate the expressions' ratio values. Moreover, the radius of the search process was configured to a maximum of 10 transitions.

In Figure 6.10, the output of this operation is shown through a visualization panel (in the view area of *MiNeRAl*), shown when the object is selected in the clipboard. In the view, all flux ratio expressions (including the ones with unreachable source fragments) are listed. For each ratio expression, it is possible to check its tree representation, by an additional sub-panel that can be activated, where the paths linking each fragment source to the product fragment are depicted in a tree form.

Some of the obtained flux ratios expressions are listed in Table 6.5. As the table shows, the ratio expression to phosphoenolpyruvate contains an unknown fragment source. This happened because when tracing the carbon atoms that originate the fragmeng C_2-C_3 of phosphoenolpyruvate, the same fragment is reached in few steps ahead, and the search process stops for that path. Through phosphoenolpyruvate carboxykinase, the fragment

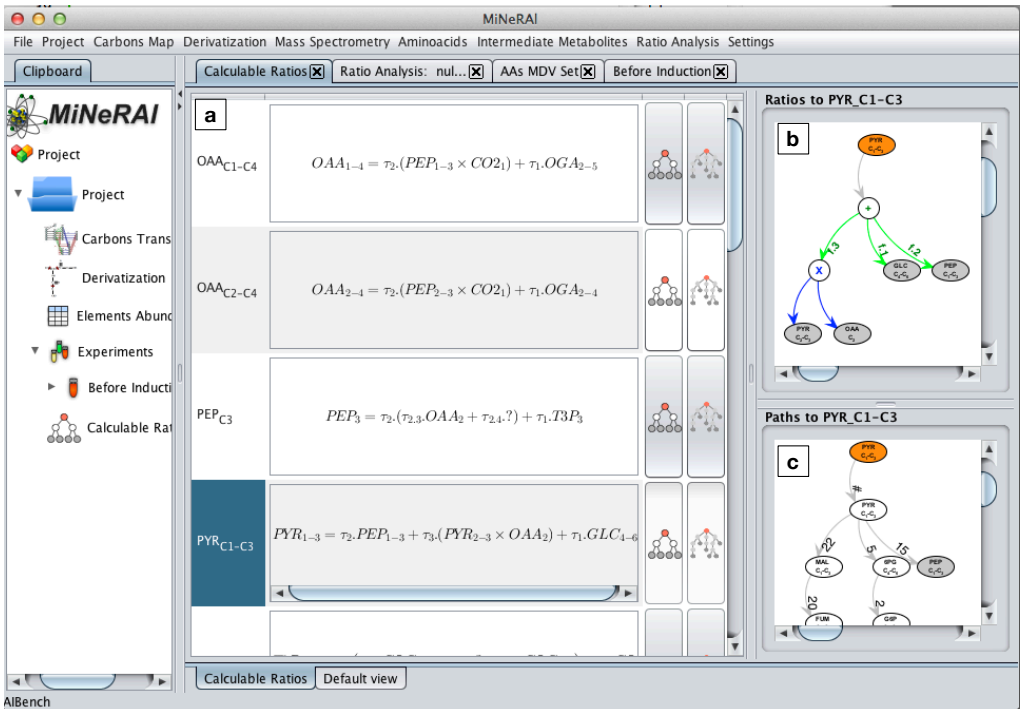


Figure 6.10: Calculable ratios datatype view. By selecting the calculable ratios datatype, its visualization panel is made available. The view lists all the calculable ratio expressions (a). For each expression, it is possible to activate a panel to check its tree representation (b) and the paths linking each source fragment to the product of the expression (c).

C_2-C_3 of oxaloacetate is reached, and two carbon transitions originate this fragment, one corresponding to the reaction malate dehydrogenase which is responsible for the combination of PYR_3 with OAA_2, and another from phosphoenolpyruvate carboxykinase. This last is the one that originates from the same fragment.

Table 6.5: Summary of some metabolic flux ratios expressions generated by MiNeRAl.

Product fragment	**Ratio expression**
$OAA_{1,4}$	$\tau_1.OGA_{2,5} + \tau_2.(PEP_{1,3} \times CO2)$
$OAA_{2,4}$	$\tau_1.OGA_{2,4} + \tau_2.(PEP_{3,3} \times CO2)$

Continued on next page

Table 6.5 – continued from previous page

Product fragment	Ratio expression
$OAA_{1,3}$	$\tau_1.PEP_{2,3} + \tau_2.(OGA_{2,3} \times PEP_{3,3} \times OAA_{2,})$
OAA_2	$\tau_1.PYR_3 + \tau_2.(\tau_{2.1}.3PG_2 + \tau_{2.2}.OAA_2)$
$PEP_{2,3}$	$\tau_1.3PG_{2,3} + \tau_2.\ (\tau_{2.1}.\ (PYR_3 \times OAA_2) + \tau_{2.2}.?)$
$PYR_{1,3}$	$\tau_1.GLC_{4,6} + \tau_2.PEP_{1,3} + \tau_3.(PYR_{2,3} \times OAA_2)$
$PYR_{2,3}$	$\tau_1.GLC_{5,6} + \tau_2.PEP_{2,3} + \tau_3.(PYR_{3,\times}OAA_2)$
$PYR_{1,2}$	$\tau_1.GLC_{4,6} + \tau_2.(\tau_{2.1}.3PG_{1,2} + \tau_{2.2}.OAA_{1,2})$ $+\tau_3.PYR_{2,3}$

6.3.2 Computing metabolic flux ratios for *Pseudomonas putida*

Another study was performed for computing metabolic flux ratios in *Pseudomonas putida* using the ^{13}C labeling data available in [Ebert et al., 2011]. The labeling substrate strategy was growing *Pseudomonas putida* in a mixture of 20% [U-^{13}C] with 80% natural glucose or 100% [1-^{13}C] glucose. The hydrolyzed proteinogenic amino acids fragmentation was obtained by TBDMS derivatization.

The MDVs of the amino acids' precursors were retrieved for two different strains: *Pseudomonas putida* KT2440, and *Pseudomonas putida* KT2440 transformed with plasmid pCOM10-*nox* to overxpress a NADH oxidase gene of *Streptococcus pneumoniae*.

The stoichiometric model of *Pseudomonas putida*'s central metabolism [Blank et al., 2008] was used, and carbon atom transitions were associated

to this model by adapting the transitions from the reactions in *Escherichia coli*. The resulting model is shown in Table 6.6.

Table 6.6: Central carbon atoms transitions model of Pseudomonas putida

Id	Reaction	Carbon transitions
1	GLC = G6P	abcdef = abcdef
2	G6P = 6PG	abcdef = abcdef
3	6PG = P5P + CO2	abcdef = bcdef + a
4	G6P = F6P	abcdef = abcdef
5	6PG = T3P + PYR	abcdef = abc + def
6.1	T3P + DHAP = F6P	abc + def = abcdef
6.2	T3P = DHAP	abc = cba
7	P5P + P5P = S7P + T3P	abcde + fghij = fgabcde + hij
8	F6P + T3P = E4P + P5P	abcdef + ghi = cdef + abghi
9	S7P + T3P = E4P + F6P	abcdefg + hij = defg + abchij
10	T3P = PGA	abc = abc
11	PGA = PEP	abc = abc
12	PEP = PYR	abc = abc
13	PYR = AcCoA + CO2	abc = bc + a
14	AcCoA + OAA = ICT	ab + cdef = fedbac
15	ICT = OGA + CO2	abcdef = abcde + f
16	OGA = FUM + CO2	abcde = edcb + a
17	FUM = MAL	abcd = abcd
18	MAL = OAA	abcd = abcd
19	MAL = PYR + CO2	abcd = abc + d
20	OAA = PEP + CO2	abcd = abc + d
21	PYR + CO2 = OAA	abc + d = abcd
22	ICT + AcCoA = MAL + FUM	abcdef + gh = abhg + fcde

Based on this model and the measured MDVs, the ratios of pyruvate originating from malate, and oxoloacetate deriving from pyruvate have been computed for both strains.

When building the metabolic ratios expression for the fragment C_1-C_2 of pyruvate, the fraction of malate originating pyruvate could arise from both fumarase reaction and from glyoxylate shunt. Since the fragment source that originates the fraction from malate obtained from glyoxylate shunt ($ICT_{1,2}$) is not reachable (no measured MDV was matched), the ratio expression to calculate the ratio of pyruvate from malate contains an unknown source:

$$PYR_{1,2} = \tau_1.(\tau_{1.1}.ACCOA_{1,3} + \tau 1.2.?) + \tau_2.PEP_{1,2} + \tau_3.GLC_{4,5} \quad (6.5)$$

Therefore, to compute the ratios in the above expression, a meta-heuristic has been used (in this case, the one based in EA). The computed bound for the ratio originating from malate ($\tau_{1.1}$) was equal to 34.4% for *Pseudomonas putida* KT2440 (Figure 6.11 a), and 27.5% for *Pseudomonas putida* KT2440 pCOM10-*nox* (Figure 6.11 c).

When building the flux ratios expression to the fragment C_1-C_4 of oxaloacetate, glyoxylate shunt is visited again, and the source fragments contributing to the labeling pattern of malate are not all reached. Therefore, another ratio expression with unknown sources is returned by the implemented method:

$$OAA_{1,4} = \tau_1.(PEP_{1,3} \times CO_2) + \tau_2.(\tau_{2.1}.OGA_{2,5} + \tau_{2.2}.?) \quad (6.6)$$

One could use a meta-heuristic to estimate the ratios as for the previous expression. However, for the computation of similar ratio expressions, where the glyoxylate shunt could contribute to the labeling of $OAA_{1,4}$ [Fischer

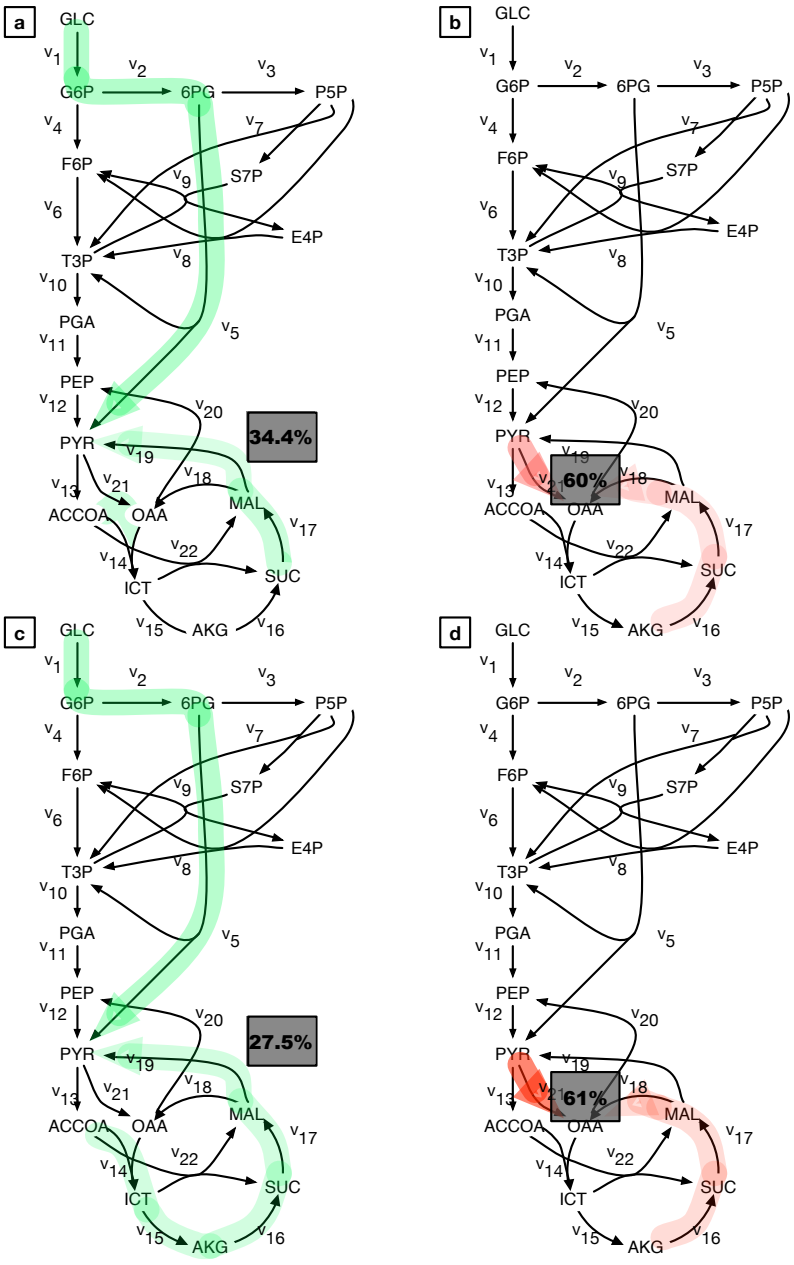


Figure 6.11: Computed metabolic flux ratios for Pseudomonas putida. The ratio analysis has been performed to calculate ratios in: Pseudomonas putida KT2440, pyruvate originating from malate (a) and oxaloacetate from pyruvate (c); Pseudomonas putida KT2440 pCOM10-nox, pyruvate originating from malate (b) and oxaloacatete from pyruvate (d)

and Sauer, 2003], this contribution is not accounted for when manually drawing the flux ratios equation. Therefore, for this example, the glyoxylate shunt was removed from the model, and the resulting expression was the linear equation that follows:

$$OAA_{1,4} = \tau_1.(PEP_{1,3} \times CO_2) + \tau_2.OGA_{2,5} \tag{6.7}$$

Since there is a contribution of CO_2 combined with pyruvate, this expression was solved by applying the equation 5.16 in the previous chapter (where the system is solved by QR decomposition). The calculated values for $\tau 1$ were 60% and 61% for *Pseudomonas putida* KT2440 (Figure 6.11 b) and *Pseudomonas putida* KT2440 pCOM10-*nox* (Figure 6.11 d), respectively, and the labeling content of CO_2 was 0.18% and 0.13% for each of the strains.

6.3.3 Metabolic flux ratio constraints

From the metabolic flux ratio expressions, *MiNeRAl* can build ratios between the reaction rates defining constraints over the network fluxes. For each computed ratio expression by using the paths linking the source fragments to the products, stored during the process to build the ratio expressions (Figure 6.10 c), it is possible to trace the relation between the reactions corresponding to the metabolic ratios.

For each path linking the source fragments, for which the ratio expression is built to the product, this tracing process checks the reactions that are visited. Moreover, the reactions not belonging to the path, but using intermediate fragments in the path, are registered so they will be subtracted in the ratio constraint.

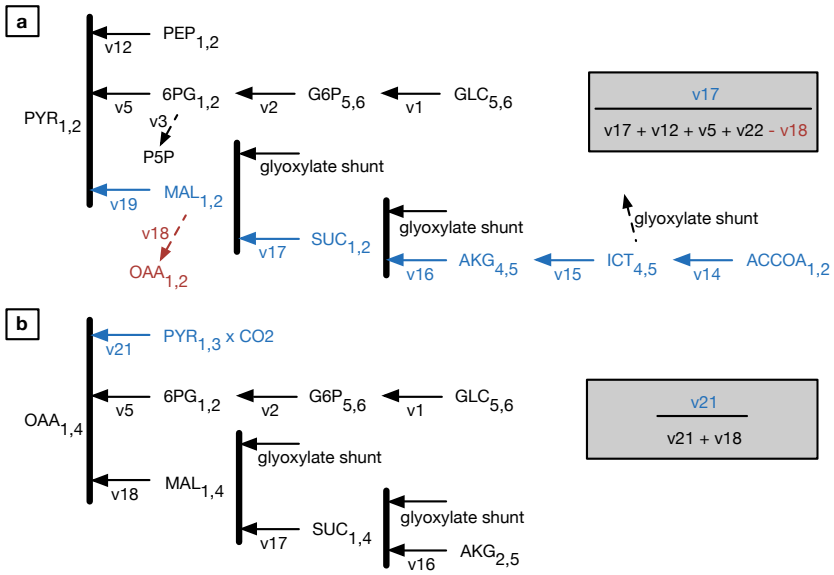


Figure 6.12: Illustration of the process to build metabolic flux ratio constraints. a) Flux ratio constraint for the fraction of pyruvate originating from malate. b) Flux ratio constraint for oxoloacetate originating from pyruvate. Blue arrows indicate the paths for which the constraint is built for (more than one path can be accounted for the same constraint), and red arrows indicate the reactions to be subtracted in the ratio constraint.

In Figure 6.12 the process for building the constraints for the two metabolic flux ratio expressions obtained for *Pseudomonas putida* is illustrated. As it is shown, for the paths linking the fragment $PEP_{1,2}$ (Figure 6.12 a), and $PYR_{1,2} \times CO_2$ (Figure 6.12 b) to their corresponding expression products, all the reactions belonging to their level in the schematic representation (in the tree representation of paths to product stored by the process of building ratio expressions) are added to the denominator of the ratio constraint. Additionally, for the ratio expression to pyruvate, the reactions that are registered to be subtracted are added to the denominator as well.

These ratios between the reactions rates are then concatenated to the flux ratio values in the corresponding ratio expressions, to form equality or inequality (bound) constraints over the model fluxes, and to be further used for the tasks of Metabolic Flux Analysis (MFA), e.g. the ones implemented

by *CBFA*.

6.4 Implementation

6.4.1 Workbench architecture

MiNeRAl is divided in two main modules as shown in Figure 6.13, both implemented in Java. The Core module is completely autonomous of the Controller module in such a way that all the functionalities in *MiNeRAl* can be used independently of the application above it. To accomplish this, the Core module was designed to have an independent Application Programming Interface (API) that allows the interaction with all its functionalities. Therefore, it is simple to call its features and develop an interface (graphical or not) to provide its functionalities to the end user.

Another aim was to develop a software that could also exploit the existing resources within *OptFlux*, in the same way *CBFA* has been built, i.e., following the *AIBench* architecture (also implemented in Java). This aspect influenced how the application layer was designed, since *AIBench* has a particular set of policies to be adopted when developing a software application. The reason to choose to build the application layer over the *AIBench* architecture is due to its plug-in based paradigm, and also because it facilitates developing a GUI based software. For instance, *MiNeRAl* can be easily coupled to *OptFlux* as a new plug-in, just as it happens with *CBFA*.

The entire application layer follows the Model-View-Controller (MVC) paradigm, where users can interact with the software through the concepts of Operations and Views over the use of Datatypes that encapsulate all the information transferred between the two modules, as explained in the

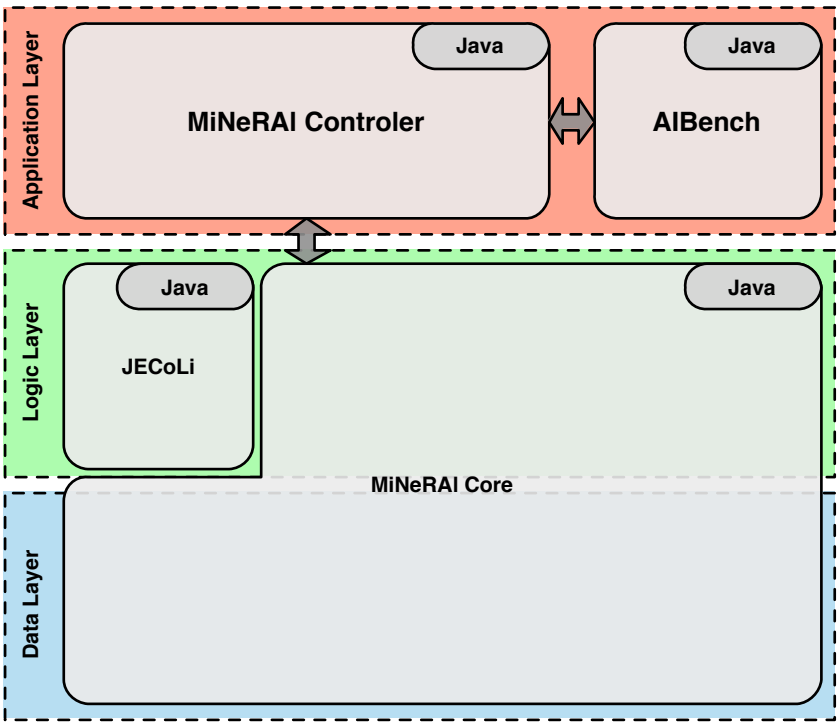


Figure 6.13: MiNeRAl layers. The application layer is the top layer, composed by the Controller and AIBench, which are connected since the Controller follows the AIBench paradigm. The implementations belonging to the data and logic layers are in the Core module, which is completely independent from the Controller module. All the layers are implemented in Java.

previous chapter.

All the main elements in the Core have a corresponding wrapper in the Controller side. Therefore, the functionalities in *MiNeRAl* are made available as operations that take as inputs the wrappers. These operations are nothing else than classes that call methods in the wrapped core classes, and no main computation is implemented, which is left to these methods. The contents of the produced datatypes can be consulted through the concept of views, which are built to show the status of the core data. Although the data should not be changed directly through the views, in some cases it might be necessary to simplify the application and use shortcuts in the view in order to make some updates. To perform such tasks, some views

have been implemented to provide components that will call operations to perform the actions over the corresponding datatype.

6.4.2 Integration with JECoLi

The core module uses the Java Evolutionary Computation Library (JECoLi) [Evangelista et al., 2009] for performing the estimation of intermediate metabolite precursors MDVs and to compute metabolic flux ratios. JECoLi is a generic multi-purpose evolutionary computation library, developed in Java, that provides implementations of meta-heuristic optimization algorithms, such as: genetic algorithms and evolutionary programming; DE; genetic programming and linear genetic programming; SA; cellular automata genetic algorithms and multi-objective optimization evolutionary algorithms.

The current version of JECoLi (version 3) encompasses several different types of solution representations, including binary, integer, real, permutations, sets and trees. For these representations, different reproduction and selection operators are available. However, if a new representation, or operator, is needed, its well defined API allows to easily develop new components.

In this work, the focus was given to the implementations for the DE, EA and SA algorithms. In terms of representation, the linear real representation was adopted for both MDV and flux ratio estimations.

6.4.3 Core module organization

The Java classes belonging to the core module are organized in different packages according to their context. The Unified Modeling Language (UML) class diagram presented in Figure 6.15 contains the organization of

the main components in terms of data structuring, including the ones for the carbon atoms transitions model, fragmentation, experimental spectra and MDVs.

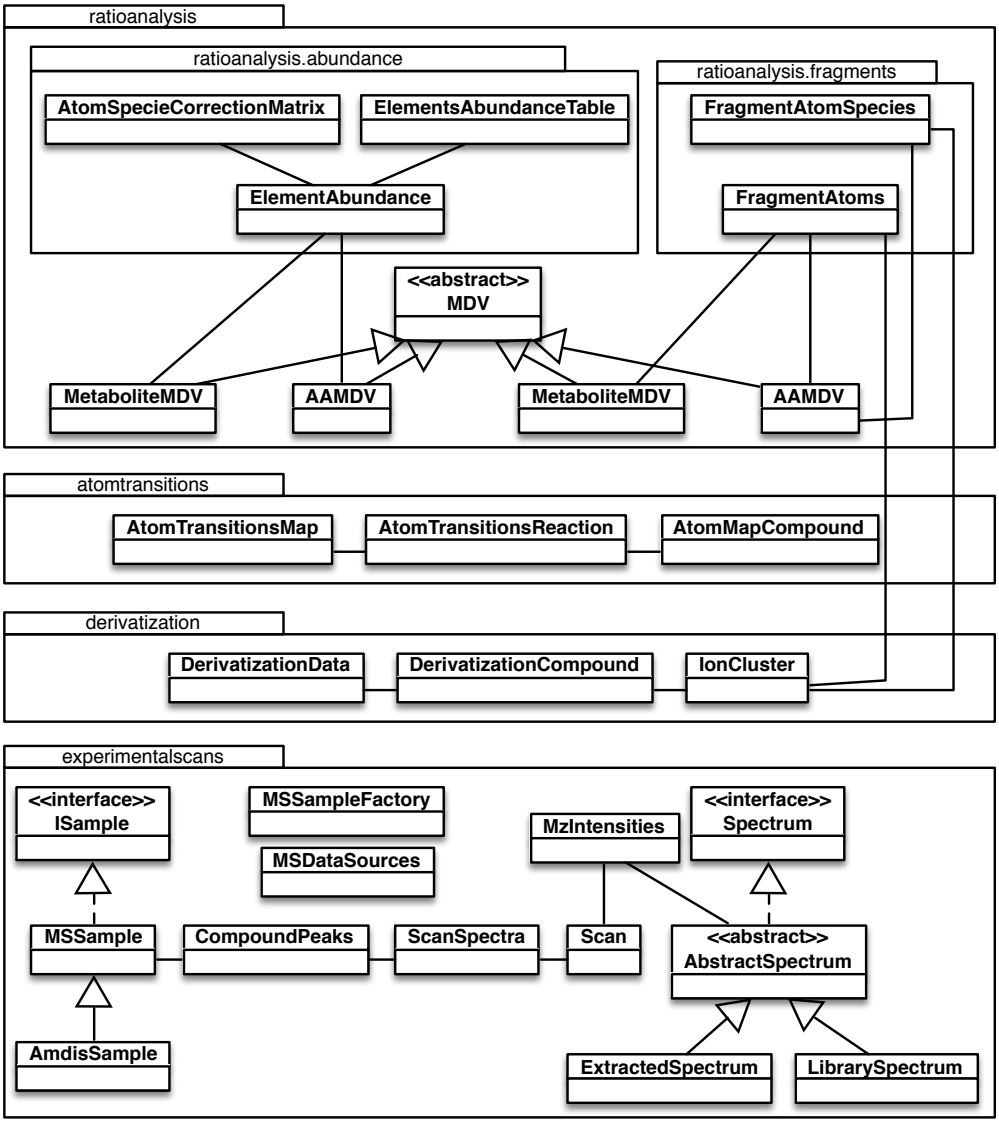


Figure 6.14: UML class diagram for the main MiNeRAl components data structuring.

The class organization for the estimation processes related to the calculations of the MDVs of the amino acids precursors, and metabolic flux ratios, is illustrated in the UML class diagram in Figure 6.15. The ab-

straction in the *Optimization* class contains the main links with the used resources in JECoLi.

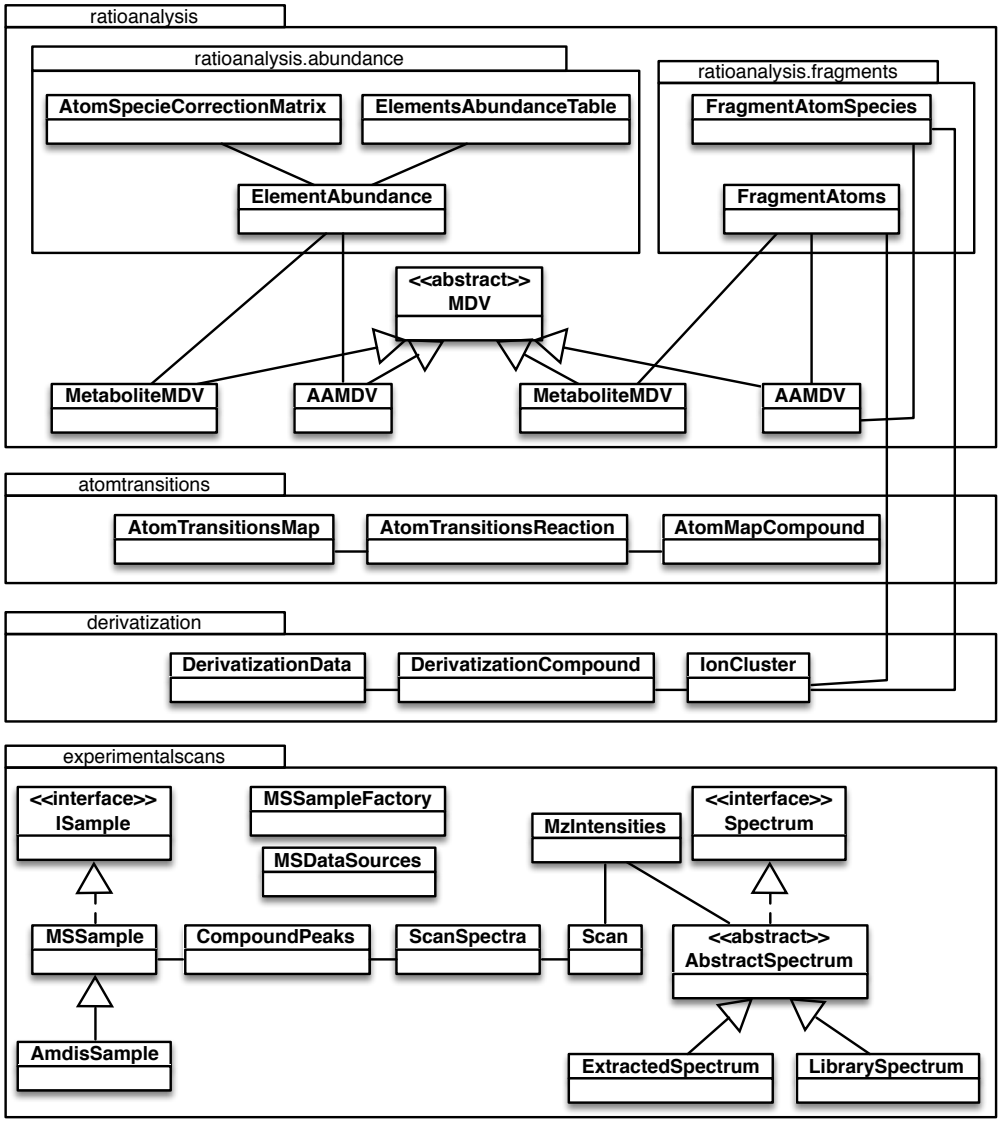


Figure 6.15: UML class diagram for the organization of the estimation processes implemented in MiNeRAl.

6.4.4 Implementing methods to compute ratios

Whenever an expression (or a set of related expressions that share metabolic flux ratio variables) is being solved, an object that encompasses all the relevant information about the problem is instantiated. This object, referred as problem instance, defines the context that is necessary to decode the generated solutions by the meta-heuristics and evaluate them. Whether the selected optimization method is the one based on EA, SA or DE, the problem instance will be the same.

All the expressions to be solved in the current problem are configured in objects from the class *RatiosToMetaboliteEquationSolver*, which are responsible to solve the expressions when values for the flux ratios are inserted. Hence, the necessary information to configure the algorithms and start solving the problem are made available and the estimation can be performed. The UML sequence diagram for the instantiation of the main entities for the estimation method based on the EA approach, and the steps to run the algorithm and retrieve its solution, is shown in Figure 6.16.

The problem instance is the instantiated object (re) from the class *RatiosEstimationConfiguration* in the diagram and its reference will be used in the following diagrams. The reference to the algorithm configuration is indicated in the representative rectangle in the diagram, after calling the method *configureAlgorithm*. The UML sequence diagram of the algorithm configuration is illustrated in Figure 6.17 and Figure 6.18.

The linear representation of the individuals will contain the same number of genes of the ratios to be solved in the expressions belonging to the problem instance, where each gene is a real value in the interval zero to one. The genes to mutate (gm) used when instantiating an object of the class *LinearGenomeRandomMutation*, as shown in the sequence diagram, is

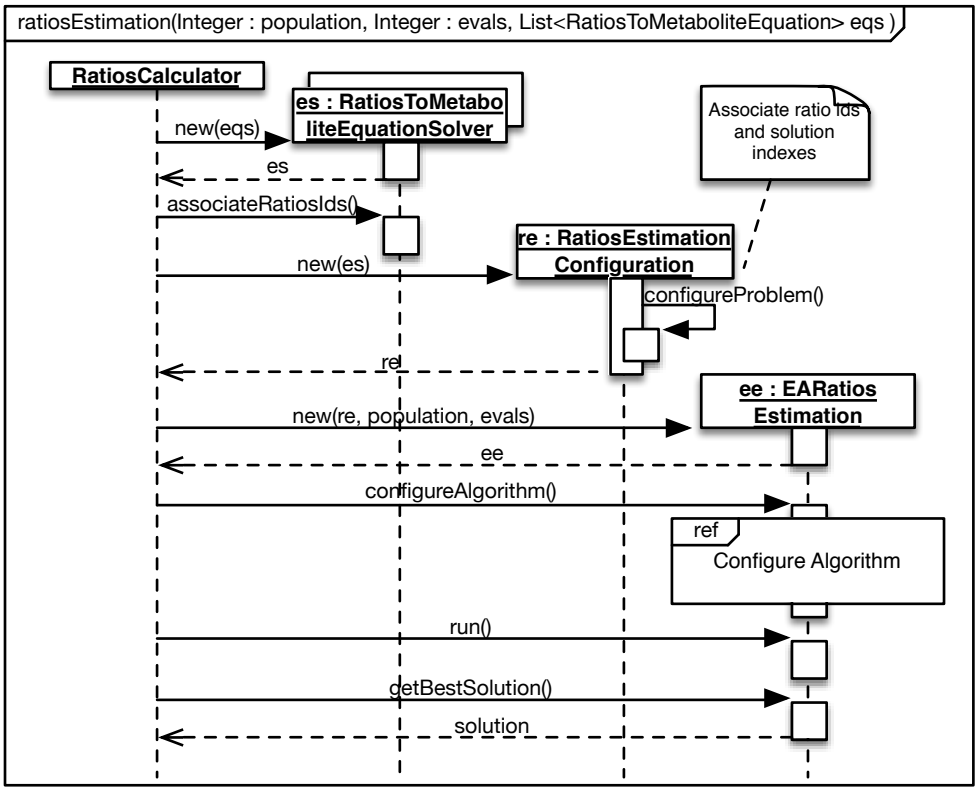


Figure 6.16: Sequence diagram of the ratios estimation method based on the EA approach. Rectangles represent Java classes, the arrows are calls to instance methods and dashed arrows indicate the results of the method executions. The parameters of the ratio estimation method are the population size, the number of function evaluations (evals) and the list of expressions to be solved (eqs).

the percentage of genes in the individuals that will be mutated for further generations.

The problem instance object is passed to the evaluation function, so the expressions can be used to evaluate the individuals. It is also used to index the genes of the individuals, so the values in those genes can be associated to the proper flux ratios variables. So, to evaluate each candidate solution, all the expressions belonging to the problem instance are solved, after decoding the individuals and linking the generated ratio values to the expressions. The difference between the calculated and measured MDVs of the target fragments defines the objective function of the problem.

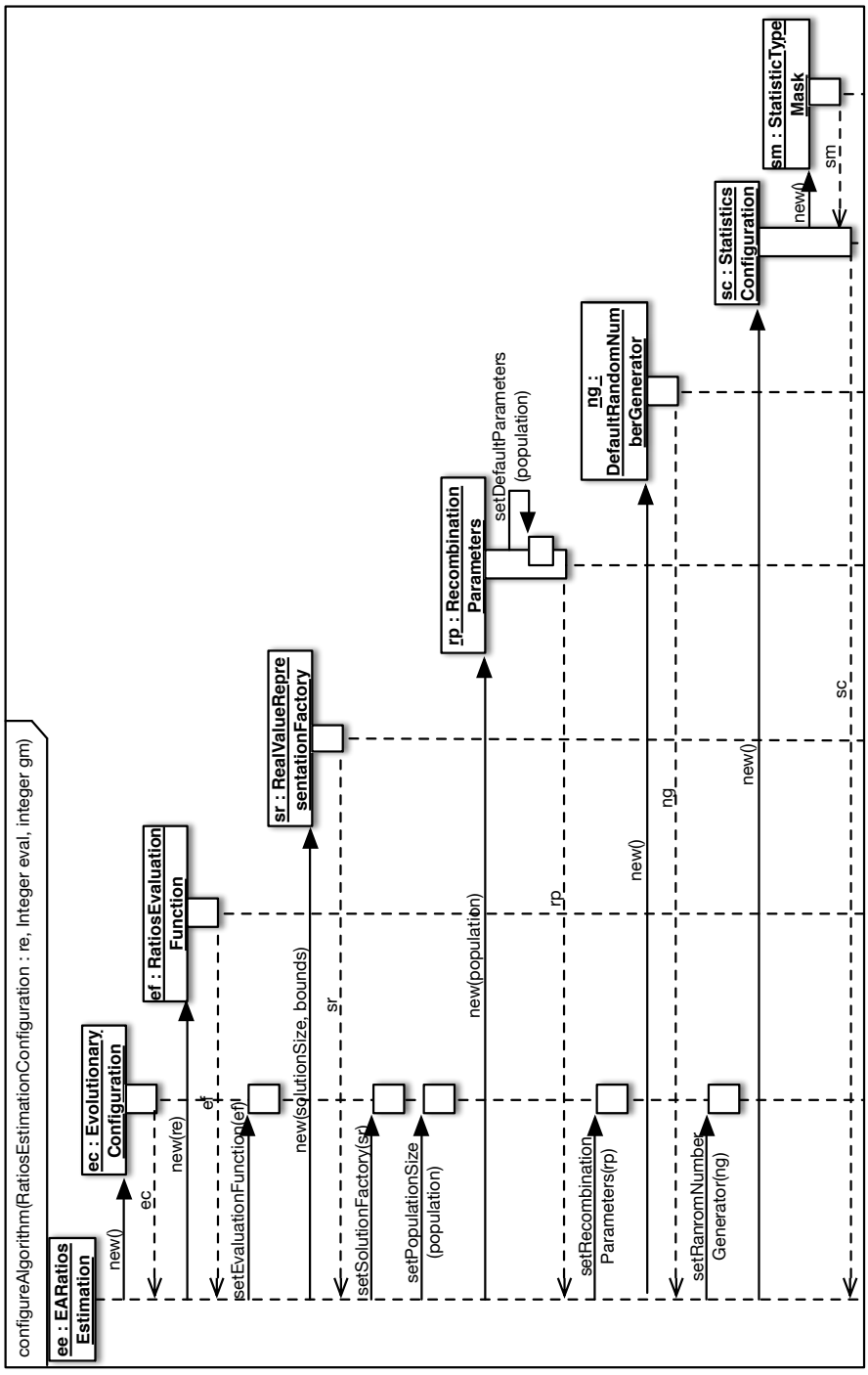


Figure 6.17: Sequence diagram of the EA configuration for the estimation of flux ratios to solve ratios to metabolite fragment expressions. The parameters for the configuration are the number of function evaluations (evals), the problem instance (re) and the percentage of genes to be mutated (gm).

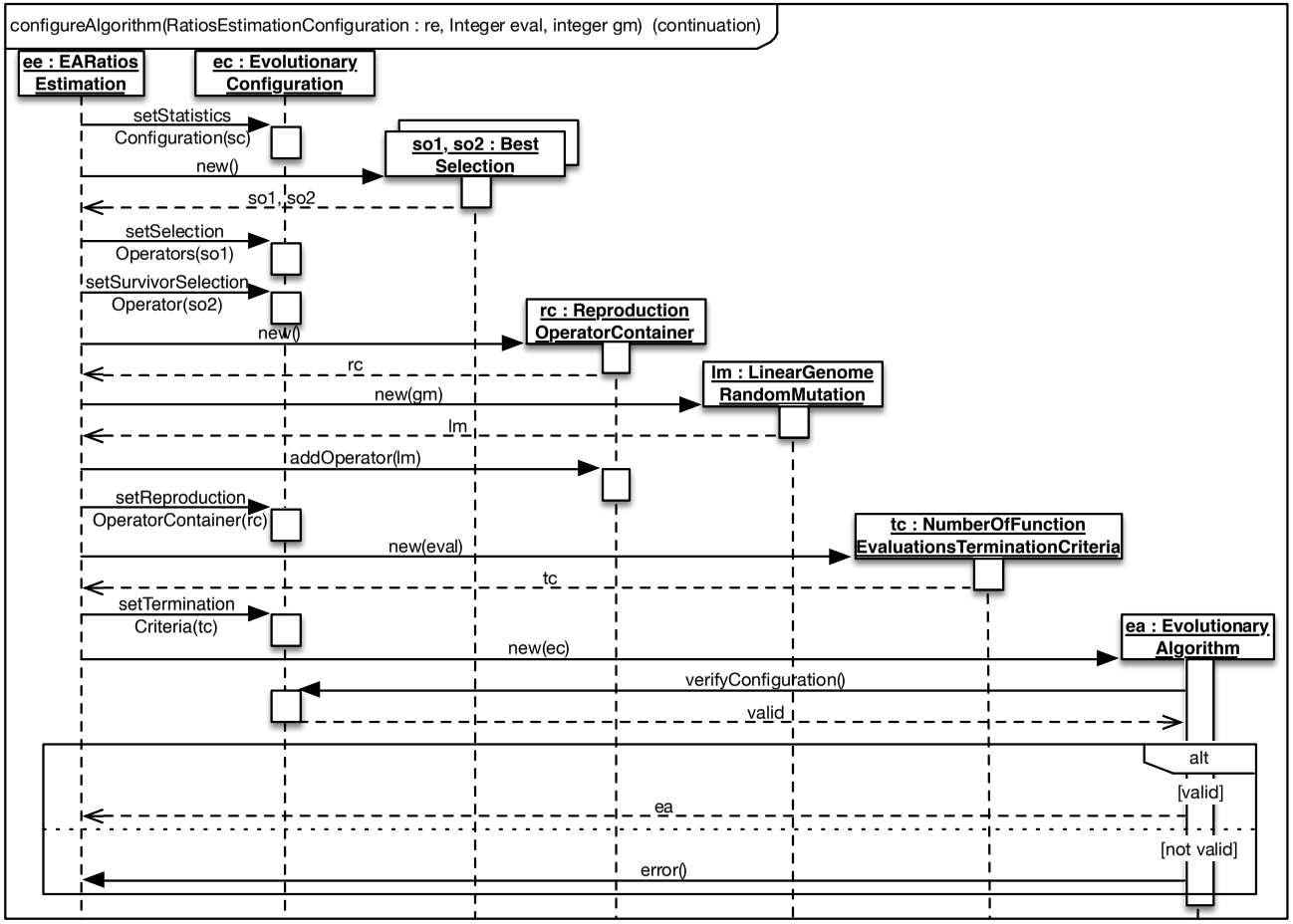


Figure 6.18: Continuation of the sequence diagram of the EA configuration for the estimation of flux ratios to solve ratios to metabolite fragment expressions.

6.4.5 Availability and requirements

In Table 6.7, the relevant technical information about *MiNeRAl* is summarized.

Table 6.7: MiNeRAl technical information.

Software name	*MiNeRAl*
Project home page	http://darwin.di.uminho.pt/mineral
Supporting files	http://darwin.di.uminho.pt/mineral
Operating system(s)	Platform independent
Programming languages	Java
Other requirements	Java JRE 1.7.x (for Mac OS users the installation of JDK 1.7 is recommended), GLPK
License	GNU-GPL, version 3

BIBLIOGRAPHY

[Blank et al., 2008] Blank, L. M., Ionidis, G., Ebert, B. E., Bühler, B., and Schmid, A. (2008). Metabolic response of pseudomonas putida during redox biocatalysis in the presence of a second octanol phase. *Febs Journal*, 275(20):5173–5190.

[Ebert et al., 2011] Ebert, B. E., Kurth, F., Grund, M., Blank, L. M., and Schmid, A. (2011). Response of pseudomonas putida kt2440 to increased nadh and atp demand. *Applied and environmental microbiology*, 77(18):6597–6605.

[Evangelista et al., 2009] Evangelista, P., Maia, P., and Rocha, M. (2009). Implementing metaheuristic optimization algorithms with jecoli. In *Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on*, pages 505–510. IEEE.

[Fischer and Sauer, 2003] Fischer, E. and Sauer, U. (2003). Metabolic flux profiling of Escherichia coli mutants in central carbon metabolism using GC-MS. *European Journal of Biochemistry*, 270(5):880–891.

[Garcia-Albornoz et al., 2014] Garcia-Albornoz, M., Thankaswamy-Kosalai, S., Nilsson, A., Väremo, L., Nookaew, I., and Nielsen, J. (2014). BioMet Toolbox 2.0: Genome-wide analysis of metabolism and omics data. *Nucleic Acids Research*, 42:1–7.

[Huege et al., 2011] Huege, J., Goetze, J., Schwarz, D., Bauwe, H., Hagemann, M., and Kopka, J. (2011). Modulation of the major paths of carbon in photorespiratory mutants of Synechocystis. *PLoS One*, 6(1):e16278.

[Luedemann et al., 2008] Luedemann, A., Strassburg, K., Erban, A., and Kopka, J. (2008). TagFinder for the quantitative analysis of gas chromatography—mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics*, 24(5):732–737.

[Pluskal et al., 2010] Pluskal, T., Castillo, S., Villar-Briones, A., and Orešič, M. (2010). MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC bioinformatics*, 11(1):395.

[Poskar et al., 2012] Poskar, C. H., Huege, J., Krach, C., Franke, M., Shacharhill, Y., and Junker, B. H. (2012). iMS2Flux – a high – throughput processing

tool for stable isotope labeled mass spectrometric data used for metabolic flux analysis. *BMC Bioinformatics*, 13(1):1.

[Rosman and Taylor, 1998] Rosman, K. J. R. and Taylor, P. (1998). Isotopic Compositions Of The Elements 1997. *Pure Appl. Chem.*, 70(1):217–236.

[Sauer et al., 1999] Sauer, U., Lasko, D. R., Fiaux, J., Hochuli, M., Glaser, R. W., Szyperski, T., Wüthrich, K., and Bailey, J. E. (1999). Metabolic flux ratio analysis of genetic and environmental modulations of Escherichia coli central carbon metabolism. *Journal of bacteriology*, 181(21):6679–88.

[Schellenberger et al., 2011] Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., and Others (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2. 0. *Nature protocols*, 6(9):1290–1307.

[Shupletsov et al., 2014] Shupletsov, M. S., Golubeva, L. I., Rubina, S. S., Podvyaznikov, D. A., Iwatani, S., and Mashko, S. V. (2014). OpenFLUX2 : 13 C-MFA modeling software package adjusted for the comprehensive analysis of single and parallel labeling experiments. pages 1–25.

[Smart et al., 2010] Smart, K. F., Aggio, R. B. M., Van Houtte, J. R., and Villas-Bôas, S. G. (2010). Analytical platform for metabolome analysis of microbial cells using methyl chloroformate derivatization followed by gas chromatography-mass spectrometry. *Nature protocols*, 5(10):1709–29.

[Sokol et al., 2012] Sokol, S., Millard, P., and Portais, J. C. (2012). Influx_s: Increasing numerical stability and precision for metabolic flux analysis in isotope labelling experiments. *Bioinformatics*, 28(5):687–693.

[Sriram et al., 2004] Sriram, G., Fulton, D. B., Iyer, V. V., Peterson, J. M., Zhou, R., Westgate, M. E., Spalding, M. H., and Shanks, J. V. (2004). Quantification of compartmented metabolic fluxes in developing soybean embryos by employing biosynthetically directed fractional 13C labeling, two-dimensional [13C, 1H] nuclear magnetic resonance, and comprehensive isotopomer balancing. *Plant Physiology*, 136(2):3043–3057.

[Srour et al., 2011] Srour, O., Young, J. D., and Eldar, Y. C. (2011). Fluxomers: A new approach for 13C metabolic flux analysis. *BMC systems biology*, 5(1):129.

[Stein, 1999] Stein, S. E. (1999). An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry*, 10(8):770–781.

[Szyperski, 1995] Szyperski, T. (1995). Biosynthetically directed fractional 13C-labeling of proteinogenic amino acids. An efficient analytical tool to investigate intermediary metabolism.

[van Winden et al., 2002] van Winden, W. a., Wittmann, C., Heinzle, E., and Heijnen, J. J. (2002). Correcting mass isotopomer distributions for naturally occurring isotopes. *Biotechnology and bioengineering*, 80(4):477–9.

[Wahl et al., 2004] Wahl, S. A., Dauner, M., and Wiechert, W. (2004). New tools for mass isotopomer data evaluation in (13)C flux analysis: mass isotope correction, data consistency checking, and precursor relationships. *Biotechnology and bioengineering*, 85(3):259–68.

[Weitzel et al., 2013] Weitzel, M., Nöh, K., Dalman, T., Niedenführ, S., Stute, B., and Wiechert, W. (2013). 13CFLUX2—high-performance software suite for 13C-metabolic flux analysis. *Bioinformatics*, 29(1):143–145.

[Winkler, 1999] Winkler, W. E. (1999). The state of record linkage and current research problems. In *Statistical Research Division, US Census Bureau*. Citeseer.

[Yoo et al., 2008] Yoo, H., Antoniewicz, M. R., Stephanopoulos, G., and Kelleher, J. K. (2008). Quantifying reductive carboxylation flux of glutamine to lipid in a brown adipocyte cell line. *Journal of Biological Chemistry*, 283(30):20621–20627.

[Zamboni et al., 2005] Zamboni, N., Fischer, E., and Sauer, U. (2005). FiatFlux–a software for metabolic flux analysis from 13C-glucose experiments. *BMC bioinformatics*, 6(1):209.

CHAPTER 7

CONCLUSIONS

7.1 Summary of the work and main contributions

The first objective of this work was to provide a tool for performing stationary Metabolic Flux Analysis (MFA), based on the use of a genome-scale model, to which a set of different measurements or knowledge (including flux measurements, knowledge on genetic modifications or on non-active reactions, and/ or metabolic flux ratios) could be added. The *Constraint-Based Flux Analysis (CBFA)* [Carreira et al., 2014] application developed in the first part of this work implements a plethora of methods to estimate flux distributions based on all this information.

Indeed, although several software tools and applications have been developed and published to perform constraint-based flux analysis, currently none provides the full set of methods to perform flux analysis in a user-friendly way, while many are commercial or based on commercial systems such as Matlab. Thus, CBFA is complementary to existing ones, since it provides a free, simple and quick way to perform flux analysis, with an extensive portfolio of methods, without all the complexity and specificity associated with experimental data processing.

CBFA allows to emulate the use of experimental data, by simply setting values (or intervals) for flux measurements or environmental conditions (e.g. media), configuring reaction/ gene knockouts, or to establish flux ratios that

may reflect the over/ under expression of certain genes or translate the flux activity of a given pathway, for instance, obtained from ^{13}C labelled substrate experiments.

The application identifies the set of applicable methods based on the constraints defined from user inputs, allowing to select the desired approach, encompassing algebraic and constraint-based simulation methods (such as Flux Balance Analysis and its variants). Anytime a set of constraints is selected, the software calculates the degrees of freedom of the configured system, and updates the admissible methods depending on whether the system is underdetermined, determined or overdetermined, as shown in Figure 7.1 (top). A method to perform robustness analysis is also implemented.

The integration of *CBFA* within the *OptFlux* framework [Rocha et al., 2010] allows the use of different model formats and the integration with complementary methods for phenotype simulation and visualization of the results. Moreover, the flux ratio constraints can be obtained from previous calculations in *MiNeRAl*, or manually defined by the user. The first option provides a straightforward way to integrate both applications in a Metabolic Engineering (ME) workflow.

In the second part of this work, a framework that allows the flexible and automatic computation of metabolic flux ratios for performing tasks of ^{13}C metabolic flux analysis has been developed [Carreira et al., 2013]. Based on a set of input fragments, that have an associated Mass isotopomer Distribution Vector (MDV) obtained from ^{13}C labeling experiments, and on a graph that represents the topology of the metabolic network and how the carbon atoms are transferred between reactants and products, our approach traces the carbon atoms through the different pathways of the network. As its main results it is able to compute MDV expressions for fragments not

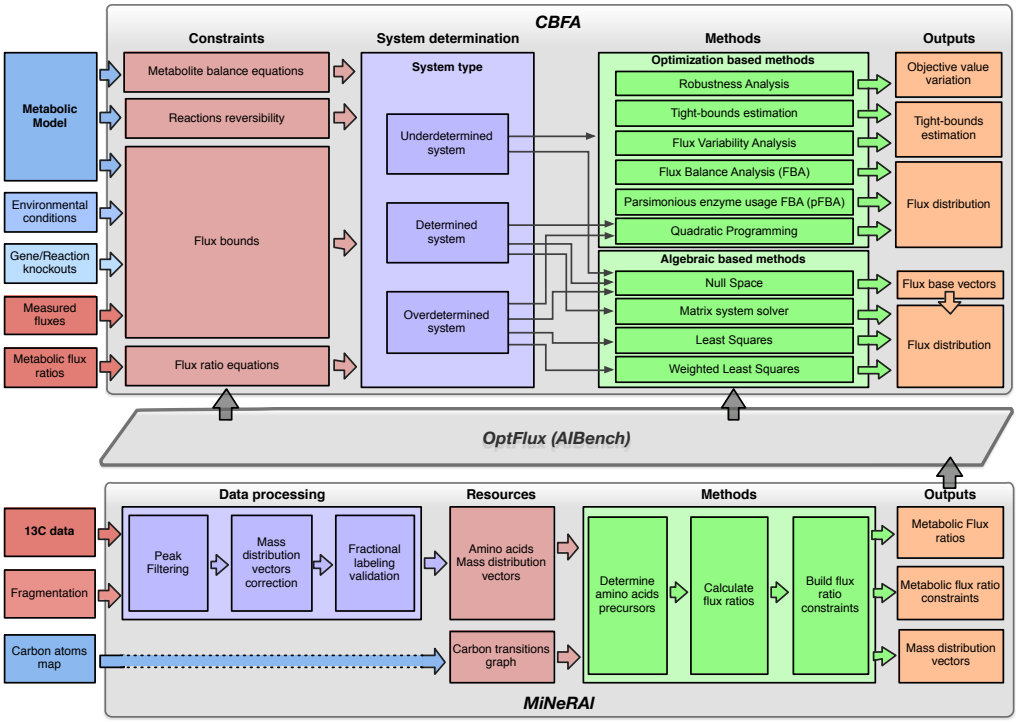


Figure 7.1: Overview of the developed software applications during this work. On the left side, the inputs are represented. The center boxes of CBFA contain its different functional blocks: the types of constraints, the determination of the system type and the supported methods. The boxes of MiNeRAl correspond to its main features, including data processing and methods for determining flux ratios. The rightmost boxes represent the outputs of the methods. The middle layer represents OptFlux, as some of its features are used and it allows the interaction between the two applications.

measured and to build expressions that enable to obtain the relevant flux ratios.

This approach does not require any assumptions on the metabolic network and reversibility of the reactions, and allows to compute ratios originating from coupled joint points of the network. Since the measured data influence the shape of the derived equations, when not enough data exist, the system tries to infer ratio bounds from the measurements. The computed ratios can be used to generate metabolic flux ratio constraints to be used within existing approaches to perform constraint-based flux analysis

methods.

Over this methodological framework, another software application, named *MetabolIc NEtwork Ratio AnaLysis (MiNeRAl)*, has been developed (Figure 7.1, bottom). This software aims at analyzing labeling experiments to infer flux constraints for stoichiometric models. Since the calculation of the flux ratios is independent of the flux distribution, *MiNeRAl* can be used independently of other flux calculation processes, and the ratios can be further exploited to reduce the degrees of freedom of systems obtained in other MFA approaches [Zamboni et al., 2005, McAnulty et al., 2012].

The main differentiating characteristics of this tool are, besides being user-friendly, the fact that it is generic for any type of metabolite fragmentation originating from GC-MS techniques and metabolic network topology. Furthermore, the software is also able to investigate what flux ratio constraints are possible to be inferred for a certain experiment beforehand.

Both applications are validated with targeted case studies, which aim to illustrate the main uses of the software for typical ME scenarios.

7.2 Future work

The main objectives of this work have been achieved. However, more can always be done, mainly in the second developed software. Some improvements and new features could be implemented, including:

- additional features for building flux ratio expressions: for the cases where few measurements are given to the software, the generated ratio expressions could give rise to some unknown branches, or large ratio expressions. To overcome this, the software could be more flexible to allow users to deactivate some carbon atoms transitions, based on

biological knowledge on the case study. In the current version, this can be done in a not so straightforward fashion, by changing the model.

- support for experimental data originating from other analytical techniques: metabolic flux ratios can also be calculated from data deriving from Nuclear Magnetic Resonance (NMR) techniques. In fact, the Metabolic Flux Ratio (METAFoR) approach has been introduced by using this type of data [Szyperski, 1995]. However, such analysis would be different from the proposed in this work, since the present methods work over mass vectors;

- other ^{13}C MFA approaches: although *MiNeRAl* is a software to calculate metabolic flux ratios following a local analysis paradigm, due to its plugin-based architecture, new plugins to perform some state of the art global methods for ^{13}C MFA could be added. These could include, for instance, a method based on the Elementary Metabolite Unit (EMU) framework [Antoniewicz et al., 2007], which could use the resources available in *MiNeRAl* in terms of carbon transitions model and calculated MDVs.

During the development of the algorithms that are the basis of *MiNeRAl*, different scenarios have been applied to test and validate the approach. However, more case studies will be applied to perform the full pipeline in a extensive manner, beginning from the ratio analysis, until the flux predictions in *CBFA*. This work is already in progress for a couple case studies.

BIBLIOGRAPHY

[Antoniewicz et al., 2007] Antoniewicz, M. R., Kelleher, J. K., and Stephanopoulos, G. (2007). Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions. *Metabolic engineering*, 9(1):68–86.

[Carreira et al., 2014] Carreira, R., Evangelista, P., Maia, P., Vilaça, P., Pont, M., Tomb, J.-F., Rocha, I., and Rocha, M. (2014). Cbfa: phenotype prediction integrating metabolic models with constraints derived from experimental data. *BMC systems biology*, 8(1):123.

[Carreira et al., 2013] Carreira, R., Rocha, M., Villas-Bôas, S. G., and Rocha, I. (2013). Algorithms to infer metabolic flux ratios from fluxomics data. In *BIBM 2013 - IEEE International Conference on Bioinformatics and Biomedicine. Shanghai, China, December 18-21*, pages 205–209. IEEE.

[McAnulty et al., 2012] McAnulty, M. J., Yen, J. Y., Freedman, B. G., and Senger, R. S. (2012). Genome-scale modeling using flux ratio constraints to enable metabolic engineering of clostridial metabolism in silico. *BMC systems biology*, 6(1):42.

[Rocha et al., 2010] Rocha, I., Maia, P., Evangelista, P., Vilaça, P., Soares, S. a., Pinto, J., Nielsen, J., Patil, K., Ferreira, E., and Rocha, M. (2010). OptFlux: an open-source software platform for in silico metabolic engineering. *BMC systems biology*, 4(1):45.

[Szyperski, 1995] Szyperski, T. (1995). Biosynthetically directed fractional 13C-labeling of proteinogenic amino acids. An efficient analytical tool to investigate intermediary metabolism.

[Zamboni et al., 2005] Zamboni, N., Fischer, E., and Sauer, U. (2005). FiatFlux–a software for metabolic flux analysis from 13C-glucose experiments. *BMC bioinformatics*, 6(1):209.

APPENDIX I : OUTPUTS OF THIS WORK

Articles

- Carreira, R., Evangelista, P., Maia, P., Vilaça, P., Pont, M., Tomb, J. F., Rocha, I., and Rocha, M. (2014). CBFA: phenotype prediction integrating metabolic models with constraints derived from experimental data. BMC systems biology, 8(1):123.

- Carreira, R., Rocha, M., Villas-Bôas, S. G., and Rocha, I. (2013). Algorithms to infer metabolic flux ratios from fluxomics data. *BIBM 2013 - IEEE International Conference on Bioinformatics and Biomedicine. Shanghai, China, December 18-21*, 205–209. IEEE.

Oral presentations

- Carneiro, S., Carreira, R., Ferreira, E.C., Rocha, I. (2012). Investigating metabolic responses in recombinant E. coli cultures using metabolomics and fluxomics data. *3rd Meeting of the Institute for Biotechnology and Bioengineering - Book of Abstracts. Lisboa, Portugal, 16-17 March*, 75-75.

- Carreira, R, Carneiro, S., Villas-Bôas, S., Rocha, I., Rocha, M. (2011). A new open-source software for metabolic flux analysis based on fluxomics data. *Metabolic pathway analysis. University of Chester, United Kingdom, September 19-23.*

- Aggio, R., Carreira, R., Carneiro, S., Liu, T., Rocha, Miguel, Rocha, I., Ferreira, E. C., Villas-Bôas, S. G. (2011). Comparing results obtained from the Pathway Activity Profiling (PAPi) algorithm with ^{13}C-based metabolic flux analysis. *7th International Conference of the Metabolomics Society. Cairns, Australia, June 27-30*, 49-49.

Poster presentations

- Carreira, R., Pont, M., Villas-Bôas, S. G., Rocha, Miguel, Rocha, I. (2014). Flexible and user friendly tools for the incorporation of fluxomics data into metabolic models. *Metabolic Engineering X. No. 15, Vancouver, Canada, June 15-19*, 10-11.

APPENDIX 2 : SUPPLEMENTARY MATERIAL

- **Additional file 1** - Feature comparison of several software tools for constraint-based methods, including a list of *CBFA* novel features under the framework of *OptFlux*;
- **Additional file 2** - *CBFA* tutorial;
- **Additional file 3** - Results for the simulations carried out in the first case study in Chapter 4: reduced model of *Saccharomyces cerevisiae*.
- **Additional file 4** - Results for the simulations carried out in the second case study in Chapter 4: GSMM of *Escherichia coli*.